

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/146320>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

**Application of Whole Genome
Sequencing and
Metagenomics for Diagnosis
of Tuberculosis**

Emma Louise Doughty

University of Warwick,
Warwick Medical School

Thesis submitted for the degree of Doctor of
Philosophy
December 2017

Table of Contents

Table of Contents	i
Acknowledgements	v
Author's Declaration	vii
Publications	viii
Abstract	ix
Abbreviations	x
Tables and figures	1
1. Introduction	3
1.1 What is tuberculosis?	3
1.2 Historical perspective	3
1.3 Clinical features, aetiology and transmission	5
1.3.1 Presentation	5
1.3.2 Aetiology	5
1.3.3 Transmission	8
1.3.4 Pathogenesis	9
1.4 Treatment and drug resistance	13
1.4.1 Acquisition and transmission of drug resistant strains	14
1.4.2 Clinical classifications of drug resistant strains	16
1.4.3 Treatment regimens	17
1.4.4 New drugs and regimens in the development pipeline	20
1.5 Diagnosis of tuberculosis	21
1.5.1 Current diagnostic methods endorsed by WHO	21
1.5.1.1 Diagnosis of infection and active disease using host responses	22
1.5.1.2 Microbiological detection and drug resistance testing	24
1.5.1.3 Epidemiological typing of the MTBC	34
1.5.1.4 New diagnostics in the development pipeline	36
1.6 Unmet needs and the global challenge	37
1.7 Objectives and outline	38
1.7.1 Thesis objectives	38
1.7.2 Thesis outline	38
2. Materials and Methods	40
2.1 Ethical approvals	40
2.2 Microbiology	40
2.2.1 Smear microscopy for all chapters	40
2.2.2 Mycobacterium tuberculosis complex culture: Chapters three and five	41
2.2.3 Microscopic-observation drug susceptibility assay: Chapter five	42
2.2.4 BACTEC 960 PZA resistance testing: Chapter five	42

2.2.5 Wayne's pyrazinamidase assay: Chapter five	42
2.2.6 Spoligotyping: Chapter three	43
2.3 Sample collection and handling	43
2.3.1 Sputum samples: Chapters three and four	43
2.3.2 Cultured samples: Chapter five	43
2.3.3 Storage and maintenance of samples: All chapters	44
2.4 DNA preparation	44
2.4.1 Chapter three DNA preparation	44
2.4.1.1 Sputum decontamination with NaOH-NALC: Chapter three methods A and B, chapters four and five	44
2.4.1.2 Sputum decongestion with NALC: Methods C-F	44
2.4.1.3 Depletion of human DNA: Methods E and F	44
2.4.1.4 Mechanical lysis and Qiagen kit extraction: Methods A and D	45
2.4.1.5 CTAB DNA extraction: Methods B, C and E	45
2.4.1.6 Nucleospin tissue kit extraction: Method F	46
2.4.2 Chapter four DNA preparation	46
2.4.2.1 Sputum decongestion with sputolysin	46
2.4.2.2 Depletion of human DNA: Method G	46
2.4.2.3 Depletion of human DNA: Method H	47
2.4.2.4 Depletion of human DNA: Method I	47
2.4.2.5 Depletion of human DNA: Method J	47
2.4.2.6 Depletion of human DNA: Method K	48
2.4.2.7 Nucleospin tissue kit extraction	48
2.4.3 Chapter five DNA preparation	48
2.4.3.1 CTAB extraction from culture	48
2.5 DNA quantification and fragment length determination: All chapters	48
2.6 DNA sequencing	49
2.6.1 Nextera XT library preparation: Chapters three and five	49
2.6.2 TruSeq Nano library preparation: Chapter four	49
2.7 Bioinformatic analyses	49
2.7.1 Read alignment: All chapters	49
2.7.2 Phylogenetic placement: Chapters three and five	50
2.7.3 Small variant calling: Chapter five	50
2.7.4 Genomic prediction of drug resistance: Chapter five	51
2.8 Statistical analyses: Chapter four	51
3. Initial Attempts at Metagenomic Diagnosis of Tuberculosis	52
3.1 Introduction	52
3.1.1 The need for new TB diagnostics	52
3.1.2 Short-comings of current diagnostic methods	53
3.1.3 Genomic sequencing in diagnosis	55
3.1.4 Rationale for this study	56

3.2 Methods	57
3.2.1 Sample collection, selection and processing	59
3.2.2 Library preparation and sequencing	60
3.2.3 Analysis of human and mycobacterial sequences	60
3.2.4 Species and lineage assignment using low-coverage SNPs	61
3.3 Results	62
3.3.1 DNA quantification and validation of controls	62
3.3.2 Selection of best extraction method	62
3.3.2 Detection of MTBC in sputum samples	63
3.3.3 Phylogenetic placement	66
3.3.4 Assessment of human DNA quantity	69
3.4 Discussion	70
4. Metagenomic Diagnosis with Depletion of Human DNA	74
4.1 Introduction	74
4.1.1 Study rationale and objectives	79
4.2 Methods	83
4.2.1 Sample collection, selection and processing	85
4.2.2 Library preparation and sequencing	86
4.2.3 Bioinformatic and statistical analysis	86
4.3 Results	87
4.4 Discussion	93
4.4.1 Performance of treatment methods in this study for removing human DNA	94
4.4.2 Statistical analysis approaches and limitations of this study design	96
4.4.3 Unusual fragmentation of DNA	97
4.4.4 Conclusions	98
5. Performance of drug susceptibility testing approaches in Peru	99
5.1 Introduction	99
5.1.1 TB in Peru	99
5.1.2 Drug susceptibility testing	100
5.1.2.1 Use of MODS in drug susceptibility testing	100
5.1.2.2 Whole genome sequencing in drug susceptibility testing	101
5.1.2 Pyrazinamide	102
5.1.2.1 Pyrazinamide in the treatment of TB	102
5.1.2.2 Pyrazinamide resistance	103
5.1.2.3 Pyrazinamide susceptibility testing	106
5.1.8 Rationale for the Study	107
5.2 Methods	108
5.2.1 Sample collection and processing	110
5.2.2 DNA extraction and genome sequencing	110
5.2.3 Genomic prediction of drug resistance	111
5.2.4 Lineage typing and identification of mixed lineages	111

5.3 Results	112
5.3.1 Lineage and drug resistance predictions	112
5.3.2 Genotypic definition of pyrazinamide resistance	128
5.3.2.1 Pyrazinamide	128
5.3.2.2 Rifampicin	135
5.3.2.3 Isoniazid	136
5.3.2.4 Other drugs	136
5.3.3 Performance of susceptibility testing and prediction methods	137
5.4 Discussion	146
5.4.1 Mutations in pyrazinamide resistance genes	147
5.4.1.1 Resistance-causing mutations	147
5.4.1.2 Other resistance-associated mutations	149
5.4.2 Performance of pyrazinamide susceptibility detection methods	151
5.4.3.1 MGIT	152
5.4.3.2 Wayne's test	152
5.4.3.3 MODS	153
5.4.3.4 Sequence-based tools	154
5.4.3 Evaluation of MODS for susceptibility prediction in rifampicin and isoniazid	155
5.4.4 Evaluation of sequence-based tools for prediction of drug susceptibility	157
5.4.4.1 All whole genome sequencing analyses	157
5.4.4.2 Individual sequence-based tools	158
5.4.4.3 Characterisation of heterogeneity using sequence-based tools	160
5.4.5 Conclusions	161
6. Discussion	163
6.1 Genomics in the diagnosis of tuberculosis	163
6.1.1 Direct sequencing of tuberculosis	164
6.1.2 Drug susceptibility testing from genomic sequence data	167
6.1.3 Heterogeneity analysis using whole genome sequencing	169
6.1.4 The future of diagnostic genomics	170
6.2 Diagnosis in developing countries and the role of genomics	171
6.3 Conclusions	172
References	173
Supplementary materials	209

Acknowledgements

This is a thesis about tuberculosis and genomics, but it reflects a PhD experience that goes far beyond that. It's provided invaluable lessons in global health, culture and travel, public speaking and presentations, writing, philosophy, strategy and planning, (surely!) every type of organisation, overcoming challenges, perseverance and so much more than I can even be aware of. It wouldn't have been possible to get to this point without the multitude of people who have supported me throughout the way and to whom I want to offer my appreciation.

First and foremost, I want to thank my supervisors. I am grateful to Mark Pallen and Martin Antonio for both giving me the opportunity to undertake this PhD then mentoring me to undertake the research and develop as a scientist. I truly appreciate Chrystala Constantinidou for later taking me on as her student and guiding me to complete this thesis. I learnt a lot about leadership from her and became much more confident in my writing. Through them all, I have learnt a lot that is both reflected in the pages of this thesis and more widely throughout my life.

My experience at Warwick Medical School has been a great journey. Gemma Kay has played an instrumental role in undertaking this PhD, providing both laboratory and moral support throughout and input of caffeine as required. It's made this experience so significantly easier to have had her at my side and I appreciate her enormously. My bioinformatic work has also benefitted from the kindness of Andrew Millard, Martin Sergeant and Zhemin Zhou who have accommodated my novice skills and helped me to improve and complete tasks that would have made no sense to me 4 years ago! I could list the entire microbiology department in my thanks for the great memories, learning experiences and friends.

It has been a phenomenal life-forming experience to live and work in The Gambia as part of this PhD; one that I have enjoyed immensely and will surely have changed the course of my life and career going forward. The kindness

of the people I met there and the spirit of community is something that I'll never forget. Special thanks go to Madikay Senghore, Catherine Okoi, Francis Mendy, Florian Gehre and Sheikh Jarju for aiding the work that I've undertaken at MRC Unit, The Gambia. More generally, I'd like to thank the TB diagnostics lab and molecular microbiology labs for so warmly accommodating and welcoming me. Of the friends that I've met in The Gambia, particular thanks go to Matt Routledge, Steph Migchelson Felicity Coulter, Claire Laxton, Gilleh Thomas, Suzanne Anderson and Nana Tawiah for their friendship, academic advice and pep talks.

The research conducted in chapter five would never have happened without the initiation of Patricia Garrido-Matta and backing from Mirko Zimic and Patricia Sheen at Universidad Peruana Cayetano Heredia, Peru. Through them, I gained a richer perspective of tuberculosis diagnostics. I'm glad of the academic discussions with them and hope to meet in person someday.

Heart-felt thanks go to my friends and family. My parents have coached me to this point, through 26 years of guidance and commitment to my education that I admire and will always be indebted for. If it weren't for them, I wouldn't have been able to even start this PhD, let alone finish it. The warm welcome from all the people that I've lived with during this time, especially as I've written my thesis, has been humbling. Notable thanks go to Andrew Taylor and the Taylor family, Felicity Coulter, Charlie Sykes, Jana Bartlett, Jake Davies, the entire Davies/Tate family and my grandparents. These people have gone above and beyond in their friendship and helped me through the toughest of times. I credit them for helping me through challenges when I was unsure of the path and will forever be grateful. Further friends, Kathy Cox, Blake Hemingway, Liam Lockyer, Tom Rich and James Senior, are also owed thanks for their kindness and support throughout this period. This PhD has not just been an academic experience but an undertaking that has permeated throughout my life making my personal relationships even more important in the quest to earn it.


My completion of this PhD is also an achievement by all of the people mentioned here; thank you all. I truly appreciate you.

Author's Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree in this or any other university.

The work presented (including data generated and data analysis) was carried out solely by myself, except, in the cases indicated below:

- The collection of clinical samples and their characterisation by phenotyping and routine diagnostic methods was carried out by staff at the MRC Unit, The Gambia and Universidad Peruana Cayetano Heredia.
- Phylogenetic placement by pPlacer in chapter three was undertaken by Martin Sergeant.

Signed..........

Date.....

Publications

Parts of this thesis have been published by the author here:

Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. *Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer*. PeerJ. 2014;2:e585. doi: 10.7717/peerj.585.

Other publications by the author of this thesis:

Senghore M, Otu J, Witney A, Gehre F, **Doughty EL**, Kay GL, Butcher P, Salako K, Kehinde A, Onyejebu N, Idigbe E. *Whole-genome sequencing illuminates the evolution and spread of multidrug-resistant tuberculosis in Southwest Nigeria*. PLoS One. 2017 Sep 19;12(9):e0184510.

Senghore M, Bayliss SC, Kwambana-Adams BA, Foster-Nyarko E, Manneh J, Dione M, Badji H, Ebruke C, **Doughty EL**, Thorpe HA, Jasinska AJ. *Whole-genome sequencing reveals transmission of Staphylococcus aureus from humans to green monkeys in The Gambia*. Applied and Environmental Microbiology. 2016 Jul 29:AEM-01496.

Abstract

Globally, tuberculosis kills more people than any other infectious disease. Control of the epidemic is impeded by poor diagnostic approaches. My original contribution to knowledge presented in this thesis is towards diagnosis of tuberculosis by (meta)genomic approaches.

In the work presented here, I established proof-of-principle that tuberculosis can be detected, identified and somewhat characterised using a shotgun metagenomics approach. I developed an approach for DNA extraction directly from sputum followed by metagenomic sequencing that allowed me to detect sequences from the *M. tuberculosis* complex in all sixteen samples with low coverage of the H37Rv reference genome. This allowed me to assign the lineage of the MTBC species in thirteen of these samples. This was the basis of the first publication to sequencing tuberculosis without prior culture.

I determined that the proportion of human reads in the resulting metagenomic data was a major limitation to characterising the MTBC organisms with greater resolution and sought, though unsuccessfully, to determine methods to remedy this. In doing so, I identified a number of considerations that need to be made when designing studies of human DNA depletion from such heterogeneous clinical samples in the future.

Addressing some of the other limitations to using genomics in the diagnosis of tuberculosis, I studied the genotype-phenotype association of first-line drug resistance found in patients in Peru and evaluated the performance diagnostic approaches used. This identified novel mutations associated with pyrazinamide resistance, flaws in the MODS method of antibiotic resistance testing and variation in resistance prediction tools.

Abbreviations

AFB	Acid-fast bacilli
AG	Aminoglycosides
AMI	Amikacin
BCG	Bacillus Calmette–Guérin
bp	Base pairs
BR	Broad Range
BSA	Bovine serum albumin
CAP	Capreomycin
CD4	Cluster of differentiation 4
CFP	Culture filtrate protein
CI	Confidence interval
CRI	Colorimetric redox assay
CTAB	Cetrimonium bromide
CXR	Chest X-ray
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
DOTS	Directly Observed Treatment short course
DR	Drug resistant
DS	Drug sensitive
DST	Drug susceptibility testing
E. coli	Escherichia coli
EB	Elution buffer
EDTA	Ethylenediaminetetraacetic acid
ESAT	Early secretory antigenic target
ETB	Ethambutol

FLD	First-line drug
FLQ	Fluoroquinolones
GAMSTEP	The Gambian Nationwide Tuberculosis Prevalence survey
HIV	Human immunodeficiency virus
HS	High sensitivity
IGRA	Interferon-gamma release assay
INDEL	Insertion/ deletion
INH	Isoniazid
IS	Insertion sequence
KAN	Kanamycin
LF-LAM	Lateral flow urine lipoarabinomannan assay
LJ	Lowenstein–Jensen
LPA	Line probe assay
LSP	Large Sequence
MAC	<i>Mycobacterium avium</i> complex
MAF	<i>Mycobacterium africanum</i>
MDR	Multidrug resistant
MgCl ₂	Magnesium chloride
MGIT	Mycobacterial growth indicator tube
MIC	Minimum inhibitory concentration
MIRU	Mycobacterial interspersed repeat-unit
MODS	Microscopic-observation drug susceptibility
MOTT	Mycobacteria other than tuberculosis
MRCG	Medical Research Council Unit, The Gambia
MTB	<i>Mycobacterium tuberculosis</i>
MTBC	<i>Mycobacterium tuberculosis</i> complex
NaCl	Sodium chloride
NALC	N-acetyl-l-cysteine

NRA	Nitrate reductase assay
NTM	Non-tuberculous mycobacteria
PANTA	Polymyxin, amphotericin, nalidixic acid, trimethoprim, azlocillin
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
POA	Pyrazinoic acid
pp	Posterior probability
PPD	Purified peptide derivative
PZA	Pyrazinamide
PZAse	Pyrazinamidase
qPCR	Quantitative polymerase chain reaction
R	Resistant
RD	Region of Difference
REML	Restricted maximum-likelihood mixed effects model
RFLP	Restriction fragment length polymorphism
RIF	Rifampicin
rma	Random-effects meta-analysis
RNase	Ribonuclease
RR	Rifampicin resistant
rRNA	Ribosomal ribonucleic acid
S	Sensitive
SLD	Second-line drug
SLID	Second-line injectable drug
SM	Streptomycin
SNP	Single nucleotide polymorphism
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
Spoligotyping	Spacer oligonucleotide typing
TB	Tuberculosis

TB-LAMP	Tuberculosis loop-mediated isothermal amplification
TDR	Totally drug resistant
TE	Tris(hydroxymethyl)aminomethane-Ethylenediaminetetraacetic acid
tmRNA	Transfer-messenger ribonucleic acid
TNF	Tumour necrosis factor
Tris-HCl	Tris(hydroxymethyl)aminomethane-hydrochloric acid
TST	Tuberculin skin test
UK	United Kingdom
USA	United States of America
VNTR	Variable-number of tandem-repeat
WGS	Whole genome sequencing
WHO	World Health Organisation
XDR	Extensively drug resistant
XXDR	Extremely drug resistant
ZN	Ziehl-Neelsen

Abbreviation	Amino Acid
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
Cys	Cysteine
Glu	Glutamic acid
Gln	Glutamine
Gly	Glycine
His	Histidine
Ile	Isoleucine
Leu	Leucine
Lys	Lysine
Met	Methionine
Phe	Phenylalanine
Pro	Proline
Ser	Serine
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine

Abbreviation	Nucleotide
A	Adenine
G	Guanine
C	Cytosine
T	Thymine

Tables and figures

Table 1.2: Advantages and disadvantages of diagnostic methods for the detection of tuberculosis.....	31
Table 1.3: Advantages and disadvantages of methods used for drug susceptibility testing in MTBC	32
Table 2.1: Ziehl-Neelsen smear results and grades.....	41
Table 3.1: Methods of DNA preparation used to detect TB from sputum by metagenomics.....	60
Table 3.2: Evaluation of methods detect TB from sputum by metagenomics.....	62
Table 3.4: Phylogenetic placement by pPlacer and spoligotyping: Method E.....	67
Table 3.5: Phylogenetic placement by pPlacer and spoligotyping: Method F	68
Table 3.6: Proportions of reads aligning to the human reference genome, hg19, in each sample	69
Table 4.1: Prior studies attempting to deplete human DNA from clinical samples or laboratory-derived proxies.....	77
Table 4.2: Methods used to deplete human DNA from sputum samples	86
Table 4.3: Adjusted proportions of reads aligning to the human reference genome per method group, as calculated through random effects meta-analysis	93
Table 5.1: All drug resistance associated mutations, as predicted by any bioinformatic analysis tool, and the major lineages predicted by MGplacer	114
Table 5.2: Mixed lineage samples and their associated lineage branches and proportions, as reported by MGplacer	124
Table 5.3: Heteroresistant mutations predicted to cause resistance by Mykrobe and PhyResSE	125
Table 5.4: Low frequency resistance alleles reported by Mykrobe predictor	126
Table 5.5: Results of drug susceptibility testing for pyrazinamide with the various phenotypic methods and genomic prediction tools	129
Table 5.7 Rate of prediction of resistance for susceptibility-associated mutations found in this study Minor alleles are not considered here.....	143
Table 5.8 Performance of drug susceptibility testing methods for Pyrazinamide relative to BATEC 960 MGIT PZA assay.....	146

Figure 1.1: The phylogenetic relationships between the MTBC and other mycobacteria, from Gutiérrez <i>et al</i> (22)	Error! Bookmark not defined.
Figure 1.2: Spectrum of host-mycobacteria interactions in tuberculosis disease	10
Figure 3.1: Methods undertaken to analyse samples in this chapter	58
Figure 4.1: Methods undertaken to analyse samples in chapter four	84
Figure 4.2: Representative Bioanalyzer graphs showing shorter fragment lengths than typically seen in genomic DNA extractions	88
Figure 4.3: Distribution of read lengths in all control subsamples	89
Figure 4.4: Percentage of reads derived from human DNA, as distributed amongst groups of subsamples	91
Figure 5.1: Workflow for Peruvian TB study	109

1. Introduction

1.1 What is tuberculosis?

Briefly, tuberculosis (TB) is an infectious disease of humans and other mammals caused by the bacterium, *Mycobacterium tuberculosis*, and its close relatives in the *M. tuberculosis* complex (MTBC). Tuberculosis most commonly affects the lungs, but can also manifest in other parts of the body. In most cases, the infection remains latent but in around 10% of cases latent infection progresses to active disease, which if left untreated, kills about half of those infected. Active pulmonary TB presents with a chronic cough with blood-containing sputum (haemoptysis), fever, night sweats and/or weight loss. Despite being a treatable disease, TB is the ninth leading cause of death worldwide, killing 1.3 million HIV-negative people and a further 374 000 HIV-positive people in 2016 (1). In the same year, 10.4 million people were infected with the disease (1).

1.2 Historical perspective

Tuberculosis is an ancient disease that has plagued humans for thousands of years and been documented throughout medial history. Hippocrates (460-370 BC) described phthisis (2), a wasting disease that was thought to be a hereditary condition at the time:

“Early in the beginning of spring, and through the summer, and towards winter, many of those who had been long gradually declining, took to bed with symptoms of phthisis; Many, and, in fact, most of them died, and of those confined to bed, I do not know of a single individual survived for any considerable time, Consumption was the most considerable of the diseases which then prevailed, and the only one which proved fatal to many persons. Most of them were affected by these diseases in the following manner; fevers accompanied with rigors, ... constant sweats, ... extremities very cold, and warmed with difficulty; bowels disordered, with bilious, scanty, unmixed, thin, pungent, and frequent dejections. The urine was thin, colourless, unconcocted, or thick, with a deficient sediment. Sputa small, dense, concocted, but brought up rarely and with

difficulty; and in those who encountered the most violent symptoms there was no concoction at all, but they continued throughout spitting crude matters."

Later, Aristotle (384-322 BC) described phthisis in pigs, concluding that the disease was most likely infectious rather than hereditary (3). By the 5th century AD the symptomology of the disease had been defined by Roman physicians and a treatment of 'fresh air, milk and sea voyages' was recommended (2).

Though the written record of TB is sparse throughout the middle ages, the disease was clearly present, given palaeopathological evidence of TB at widespread sites (4). By the 1800s, tuberculosis was epidemic throughout Europe and known as 'the white plague' for the romanticised white appearance of its victims (2,5). The mortality rate in cities across Europe and America was as high as 800-1000 deaths per 100,000 per year (6,7). René Théophile Hyacinthe Laennec, a French anatomist and the inventor of the stethoscope, described the pathogenesis of TB and the physical signs of its pulmonary manifestation in a book first published in 1819. This initiated the modern understanding of tuberculosis and introduced many of the terms used to describe it today (8,9). The understanding of TB aetiology remained poor and there was division in opinion as to whether tuberculosis was an inherited or an infectious disease. In 1865, the infectious aetiology of tuberculosis was demonstrated experimentally by a French surgeon, Jean-Antoine Villemin, who inoculated a rabbit with the liquid of a tuberculous cavity and then that it had developed tuberculous lesions three months later (10). Only a year later, William Budd published that TB "is disseminated through society by specific germs ... cast off by persons ... suffering from the disease" (11). Robert Koch developed methods for isolating pathogens in pure culture and hypothesised that an organism he had grown in the laboratory could cause TB, and went on to demonstrate this using guinea pigs (2,12,13). He gave this organism the name *Mycobacterium tuberculosis*.

By the mid-19th century, mortality rates associated with TB had begun to decline in Europe and America. This decline is still continuing (14). Although the reason for this 19th century decline is yet to be concluded upon, it is thought to be associated with improved living conditions, better nutrition and herd

immunity from long-term natural selection of a more resistant human population (15). The first anti-TB drugs were developed in the 1950s and public health measures have further reduced the disease burden in developed countries. These measures led to only modest decreases relative to the enormous decline in the 19th century (16). A high burden of TB remains in some parts of world, where the consequences of poverty - poor sanitation, poor nutrition, poor living conditions and poor access to healthcare - and a high prevalence of HIV underlies transmission and infection with the disease (2).

1.3 Clinical features, aetiology and transmission

1.3.1 Presentation

Tuberculosis can present as latent or active disease. Around a third of the global population is estimate to have latent infection, where the tubercle bacilli are present in their bodies but do not cause symptoms or disease. Only 5-15% of people infected with TB will develop active disease that causes illness and may result in transmission (1,17). Active disease most often manifests as a respiratory condition causing a cough that may produce sputum or blood, fatigue, chest pain, fever, night sweats, loss of appetite and/or weight loss and without treatment almost certainly leads to death. In 15-20% of people with active disease the infection affects extra-pulmonary locations (EPTB), most commonly the pleurae, central nervous system, lymphatic system, genitourinary system, bones and joints, each with a different clinical manifestation (18). Miliary tuberculosis is characterised by many tiny tuberculous lesions distributed throughout the body though it is classified as pulmonary TB since some lesions appear in the lungs (19). Some patients present with TB at multiple sites due to metastatic spread of the disease after the initial infection.

1.3.2 Aetiology

Tuberculosis disease is caused by organisms in the *Mycobacterium tuberculosis* complex (MTBC), a group of bacteria from the genus *Mycobacterium*, the

family Mycobacteriaceae, the order Actinomycetales and the phylum Actinobacteria. Whilst the MTBC originally included only *M. tuberculosis* sensu stricto and *M. bovis* (20), it now encompasses a number of species or 'ecotypes' each with a primary host association (21).

M. tuberculosis (*Mtb*), *M. africanum* (*Maf*) and the 'smooth tubercle bacilli', including *M. canettii*, are adapted to primarily infect humans. They are considered almost obligate human pathogens given they are not known exist in any substantial non-human reservoir and transmission of their diseases cannot be sustained in animals. Most human disease is caused by *Mtb*, but *Maf* is responsible for a significant proportion of disease burden concentrated in West Africa (22). There are two *Maf* sub-types with distinct geographic restrictions: *M. africanum* type 1 (*M. africanum* 1, lineage 5) dominates in the east West Africa and *M. africanum* type 2 (*M. africanum* 2, lineage 6) dominates in west West Africa, with overlap mainly in Ghana and Benin in south West Africa (23,24). *Maf* causes variable proportions of total tuberculosis disease throughout the region, from 31% of all TB cases in Mali and Burkina Faso, up to 66% in Benin (22). Outbreaks of *Maf* have also occurred outside the region, including an outbreak of *Maf* 1 in France with a Brazilian index case (25). The mortality rate in HIV-negative people treated for *M. africanum* is very similar to that when infected with *Mtb* at around 3% (26). *Maf* is generally considered less virulent than *Mtb*, given that

- Although *Maf* is efficiently transmitted from human-to-human through aerosols, exposure to *Maf* is less likely to lead to infection than exposure to *Mtb*. Reduced pathology in *Maf* relative to *Mtb* leads to comparatively low transmission rates for *Maf* with the likelihood of transmission correlated with proximity of contact (27).
- People infected with *Maf* tend to be less immunologically robust than people infected with *M. tuberculosis*: A survey in The Gambia showed that the species has a greater tendency to infect people who are older, malnourished or HIV-positive (22), although a study in Ghana found the species has no preference for HIV-positive people (28). The fact that people become diseased at a greater age may reflect the slow progression of the disease.

- Relative to people with *Mtb*, patients infected with *Maf* are less likely to experience symptoms within 2 years (27) and reactivated disease is more common than a recently transmitted infection (28).
- *Maf* grows more slowly than *Mtb*, taking approximately 10 weeks to culture relative to ~4 weeks for *Mtb* (29,30).

Infection with the smooth tubercle bacilli is considered rare with around sixty cases of tuberculosis caused by *M. canettii* having ever been found, predominantly in Horn of Africa or in cases with links to this region (31).

The other members of the MTBC are considered to be predominantly associated with animals. *M. bovis* widely infects cattle across the globe though it has a much broader host range including other livestock and wild animals. Other animal-associated MTBC species include *M. microti* (voles) (32,33), *M. caprae* (goats) (34), *M. pinnipedii* (seals) (35), *M. suricattae* (meerkats) (36), *M. mungi* (banded mongooses) (36), *M. orygis* (antelopes) (37), 'the dassie bacillus' (36) and 'the chimp bacillus' (38) that have not been found to cause frequent zoonotic infection. The *M. bovis* Bacillus Calmette–Guérin (BCG) vaccine is also considered part of the MTBC, having been derived from *M. bovis*. It causes disease known as BCGosis in some immunocompromised individuals.

Several other members of the genus *Mycobacterium* are collectively known as non-tuberculous mycobacteria (NTM), mycobacterium other than tuberculosis (MOTT), environmental mycobacteria or atypical mycobacteria. These do not cause tuberculosis in the strict sense, although some species may cause pulmonary disease that closely resembles it. Each species persists in its own niche environmental reservoir(s), including soil and water. The prevalence of pulmonary disease caused by NTMs is as high as 15% of suspected TB cases in Nigeria (39) whilst in the USA the prevalence was 47 people per 100,00 in 2007 (40). NTMs pose a serious public health threat of their own, but owing to significant differences to TB disease, including treatment regimens, they are not addressed as part of the global epidemic of TB.

Recognising the aetiological species causing tuberculosis is important for appropriate treatment of infection. It is also necessary for implementing public health interventions to prevent further cases of disease arising in the population.

1.3.3 Transmission

Pulmonary TB, responsible for the over-whelming burden of global tuberculous disease, is spread by aerosolised droplets containing the pathogenic bacilli. Most commonly, these droplets are dispersed from infectious people by coughing—a symptom of pulmonary TB—but they can also be spread after sneezing, talking, spitting and singing (1). People may transmit bacilli when active TB that remains untreated or partially or improperly treated. This type of transmission requires that people are in close proximity since droplets will fall to the ground within a few feet. Since the infectious dose is very low, potentially a single bacillus, infectious patients are estimated infect as few as 10-15 people with each year of active disease (41).

Some individuals are thought to spread more TB than others, especially those with smear-positive TB (42) or lung cavities, large gas-filled spaces caused by tissue necrosis which accumulate bacilli that are readily expelled on coughing (43,44). In contrast to its effect on other diseases, HIV-TB co-morbidity is associated with lower rates of TB transmission, perhaps due to the lower prevalence of cavitary TB amongst people with low cluster of differentiation (CD)-4+ counts a consequence of untreated HIV (45).

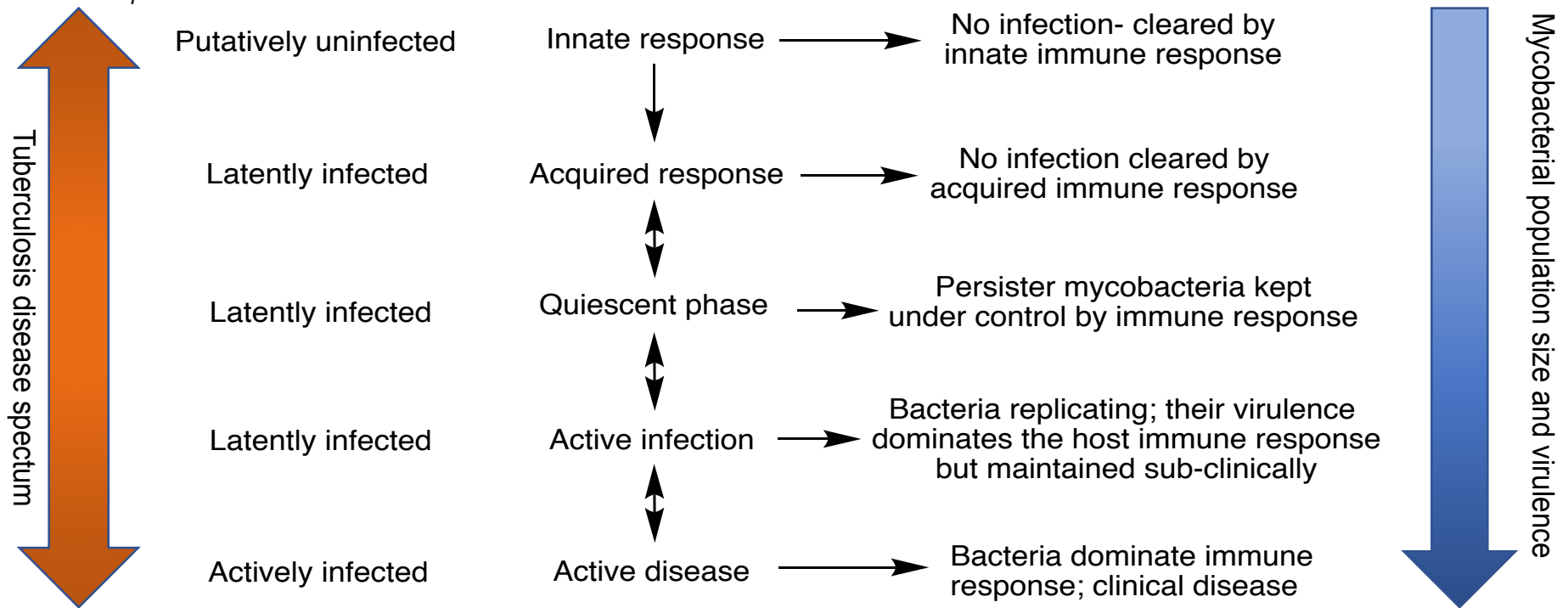
Human-to-human transmission is not typically observed with extra-pulmonary TB so the patient is not considered infectious (46). A recent case of nosocomial transmission illustrates that this may occur in rare instances; a nurse, who nine months after dressing the suppurating nodes of the first extra-pulmonary TB patient, presented with peritoneal *M. tuberculosis* of the identical strain type as this original EPTB patient (47). Zoonotic transmission from diseased animals, most commonly cattle, occurs through direct contact with the animal or by consumption of its meat or unpasteurised milk (48).

Zoonotic TB is commonly associated with MTBC species other than *M. tuberculosis* or *M. africanum* and with extra-pulmonary disease (49). Since the introduction of pasteurisation, the incidence of human *M. bovis* cases has decreased in high- and middle-income countries (48).

1.3.4 Pathogenesis

The clinical outcomes after exposure to tuberculous mycobacteria are highly variable and are dependent on the interactions of the host immune system with the microbial virulence factors. Since the molecular bases of these interactions are not central to this thesis, they are summarised here only briefly in the text and in terms of their clinical and immunological outcomes in Figure 1.2.

Figure 1.1: The spectrum of host-mycobacteria interactions in tuberculosis disease, adapted from Barry *et al* (50) Within the tuberculosis disease spectrum, clinical classifications of disease are associated with dynamic states as mycobacteria interact with the human immune response.



After mycobacteria are inhaled into a human host, they are phagocytosed into alveolar macrophages initiating the innate immune response. Despite prolonged exposure to contacts with infectious tuberculosis, some people have delayed or absent tuberculin skin test (TST) conversion, which detects the biomarkers of the adaptive immune response to mycobacteria (51). This indicates that their innate immune responses alone may be sufficient to kill mycobacteria cells prior to initiating the adaptive immune response. People achieving early clearance of the pathogens in this way, are then considered 'putatively uninfected'; this phenomenon has been observed in populations across Africa, Asia and America since the 1930s (52,53).

Should mycobacteria not be cleared from the body by the innate immune response, a further cascade of adaptive immune responses can be initiated. Clinically, these are described in a bimodal model as either latent or active disease (50). Latent disease is characterised by absence of symptoms but the detection of a mycobacterial specific immune response by either a tuberculin skin test (TST) or the interferon-gamma release assay (IGRA), discussed in section 1.5.1.1. In active disease, replication of the mycobacteria is not controlled by the host immune response and infected people are able to transmit the disease via aerosolisation mechanisms, typically coughing. By comparison, latently infected individuals are non-infectious, due to control of replication and often containment of mycobacteria within caseous granulomas (54).

This bimodal clinical model of tuberculosis disease fails to reflect the broad range of host-pathogen interactions that results from exposure to TB. This diversity of interactions creates heterogeneous microenvironments in which mycobacteria reside leading to a variety of mycobacterial phenotypes including slow replication, non-progressing 'percolating' infection or diminishing virulence in the latent phase (55). As a result of these diverse interactions, a number of immunological states may occur within the clinical latent phase, summarised in Figure 1.2. Contrary to earlier assumptions about latency, it is unclear whether mycobacteria ever actually adopt a state that might be considered 'dormant'. Latent disease includes people who were infected but no longer harbour the pathogen because the infection was cleared

after initiation of the adaptive immune response. They are clinically considered latently infected since they have the biomarkers targeted by TST or IGRA tests but this is simply an acquired response phase that may not lead to replication of bacilli (53). Should infection not have been eliminated by either the innate or adaptive immune responses, the quiescent phase commences where the mycobacterial population is at a constant size but is continually active and it may be said to occupy a 'persister' state (56). That the bacteria are still replicating during the latent phase is indicated by the accumulation of genetic mutations during this period in the absence of disease symptoms (57,58). A growing population of mycobacteria is considered active infection and part of the clinically latent phase as no symptoms of disease are present. Spontaneous deactivation, where activated disease becomes inactive, has been reported to occur in up to 50% of patients prior to the introduction of antibiotics in the 1950s (59). This state is different from latency in that chest x-ray abnormalities are present.

An important element of mycobacterial infection and pathogenesis is the potential dissemination of disease throughout the host beyond the lungs. After the initial infection of alveolar macrophages, mycobacteria replicate and survive by modulating phagosome trafficking via inhibiting fusion of the phagosome and lysosome (60). Some infected macrophages also leave the granuloma early in infection, disseminating through the lungs and blood and seeding secondary granulomas, first in the pulmonary lymph nodes, then in the liver and spleen (61,62). These secondary sites of infection may result in disseminated active disease, known as extrapulmonary TB, but most often remain sites of latent infection and have a role in activating the adaptive immune response (63,64).

Fitting with the binary interpretation, the transition from a clinically latent state to an active one is known as reactivation. Reactivation may occur after decades of latent infection, triggered by weakened immune control over the infection associated with a number of risk factors. The only well-established mechanisms for reactivation are associated with HIV and anti-tumour necrosis factor (anti-TNF) treatments. They lead to low CD-4+ counts (65) and neutralised TNF (66) respectively, that impair the adaptive immune response

after infection has been established. These, along with other risk factors, account for just a minority of reactivated infections.

The interactions of mycobacteria and the host immune response throughout pathogenesis are complex and lead to a spectrum of immunological states, clinically summarised as active and latent disease. This broad spectrum of states has equally diverse implications, including which patients may progress to active disease, respond to preventative therapy or relapse after treatment of active disease (67).

1.4 Treatment and drug resistance

Without treatment, the death rate from tuberculosis is high given that about 70% of smear-positive people and 20% smear-negative-culture-positive succumb within ten years (59). The aim when treating tuberculosis is to rid the patient of the infection whilst avoiding adverse side-effects, such that the patient survives and further transmission of the disease is avoided. Such treatment relies on antibiotic drugs that are able to combat the aetiological MTBC organism. Owing to the unusual multi-layered structure of the mycobacterial cell wall, these antibiotics are distinct from those used to treat other bacterial infections. The particular drugs and the combinations used in a treatment regimen vary depending on the manifestation of disease or the species and drug resistance profile of the aetiological organism. Neither *M. bovis* nor *M. canettii* can be treated by pyrazinamide, a drug that is ideally used in all other treatment regimens except where it is contraindicated. Failure to identify *M. bovis* has led to treatment with regimens including pyrazinamide and fatal consequences (68). Some MTBC strains have also evolved drug resistance to particular agents, as discussed further below, which further restricts antibiotic choices.

For patients with no TB symptoms, positive IGRA or TST and normal chest X-ray, latent tuberculosis can be diagnosed. This may be treated to prevent activation of the disease or the infection may be left untreated depending on the balance of benefits and harms of treatment and the length of time since the initial exposure (69). The WHO recommends one of four regimens to treat

latent TB including isoniazid monotherapy, rifampicin monotherapy or a combination of rifampicin and isoniazid (69).

1.4.1 Acquisition and transmission of drug resistant strains

Some strains of tuberculosis have acquired resistance to one or many of the antibiotic agents used to treat the disease. There are two mechanisms by which resistance develops in strains. The first is directly through genetic mutations by Darwinian evolution with inadequate antibiotic therapy providing a purifying pressure to select drug resistant strains. Inadequate treatment includes poor adherence to drugs and sub-therapeutic doses or pharmacokinetic variability owing to non-optimised dosing (70). It was exacerbated by early monotherapy treatments. The rate of genetic mutation is high in tuberculosis and mutations even develop during latency that cause drug resistance (57,70–74). Multiple drug resistance may have developed incrementally through this mechanism, given the heritable nature of genetic mutations, with one drug resistance mutation after another being introduced to the strain, presumably as a result of continued or repeated episodes of inadequate treatment (75). The second mechanism is via epigenetic changes in gene expression or protein modification that lead to phenotypically drug tolerant, non-growing persister bacilli (76–78). The presence of persisters requires longer durations of treatment and is associated with increased risk of relapse after successful treatment (79,80). A phenotypic change in this way may simultaneously facilitate tolerance to multiple drugs and may eventually lead to the development of genetic drug resistances for one or more drugs (70,81). Resistance can evolve dynamically in a single organism by both mechanisms (82).

The mutations underlying genetic resistance have not all been identified and it is increasingly becoming clear that the relationship between a resistance-associated mutations and drug resistance phenotype is complex. Thus far, there are in excess of 100 genetic loci that have been associated with resistance (83–87). Clinical resistance to one given drug may be caused by one or a

number of mutations and may involve many genes (88). Though the genetic basis of resistance is relatively well understood for rifampicin, isoniazid, further investigations to identify the genetic basis of resistance to other drugs need to be conducted (1). From what we know so far, some resistance-associated mutations directly cause the resistance, others compensate for the loss of fitness or virulence associated with the resistance and some may be passenger mutations (89,90). Notably, confusion has occurred as some mutations in genes associated with drug resistance have been associated with particular lineages but not had any impact on resistance phenotypes (86). As such, when resistance-associated mutations are identified, their role in the resistance phenotype must be evaluated through studies of functional or structural impacts or by allelic exchange (91). Genetic mutations associated with drug resistance are more likely to be acquired into particular genomic locations, such as genes associated with cell wall permeability phenotypes, transcriptional regulation and DNA repair (92), and in particular genetic backgrounds, such as the Beijing lineage which has a disproportionate abundance of MDR/XDR strains (93–95).

Any TB strain, including those with drug resistance mutations, can be transmitted from person-to-person spreading the disease. Transmission of drug resistant strains has been thought to be reduced relative to drug sensitive strains (96) and the abundance of drug resistant strains has been attributed to acquired resistance. A number of studies using high-resolution methods to analyse transmission have shown that the majority of drug resistance in the populations that they have studied can be attributed to transmitted resistance, challenging this prior notion (75,83,97–102). Strains with drug resistance may be particularly prone to transmission given prolonged delays at diagnosis, longer duration of treatment and increased likelihood of treatment failure. Though drug resistant strains can be identified, the global burden from strains that have acquired resistance within the patient versus those where the drug resistant strains have been transmitted to the patient is unclear. It is important to understand the derivation of this resistance for public health purposes as distinct interventions are required depending on whether resistant strains are being acquired or transmitted in the given population (75,103).

1.4.2 Clinical classifications of drug resistant strains

Considerations when choosing the drugs and treatment regimen for a patient with active tuberculosis include the known or expected susceptibility profile of the infecting organism. This drug susceptibility profile should ideally be determined using diagnostic approaches as described in section 1.5.1.2. Drug susceptibility testing (DST) is not available in all countries and it is only considered reliable and reproducible for rifampicin, isoniazid, fluoroquinolones and the injectable agents (104–106). When drug susceptibility testing is unavailable, resistance may be predicted based on recent surveillance data and the patient's clinical history including whether they have previously been treated for TB or have had significant contact with a patient with known drug resistant TB or are failing to respond to treatment (107).

The drug resistance profiles of tuberculosis strains are classified into groups for the purposes of patient management and surveillance. TB is classified by the WHO (108) as

- Drug sensitive (DS) when it is not resistant to rifampicin (RIF)
- Rifampicin resistant (RR) when resistant to rifampicin, but not isoniazid (INH), and any other antibiotic agents
- Multidrug resistant (MDR) when resistant to at least rifampicin and isoniazid.
- Extensively drug resistant (XDR) when resistant to rifampicin, isoniazid, a fluoroquinolone (FLQ) and an injectable agent, (amikacin (AMI), kanamycin (KAN) or capreomycin (CAP))

Other classifications are used, including

- Pre-extensively drug resistant (Pre-XDR) when resistant to rifampicin, isoniazid, a fluoroquinolone or an injectable agent (109).
- 'Extremely drug resistant (XXDR) TB' and 'Totally drug resistant (TDR) TB', both proposed in response to cases resistant to all first- and second-line drugs tested (97,110–112).

There is on-going debate about how to classify strains that are considered more resistant than XDR-TB (113,114). The WHO has been reluctant to name these given issues with the concept of 'total' resistance for a number of

reasons: it is subject to drugs and DST available in-country; DST for second-line drugs has issues with reliability and reproducibility; there is no consensus on methods or critical concentrations of DST and the term may fuel stigma around the disease (115–117). Additionally, the relationship between DST results and clinical responses has not been established for all drugs. Acknowledging the inadequacies of DST and national drug availability, patients who cannot be treated have been considered to have ‘programmatically incurable tuberculosis’ (114,118–121). The drug resistance classifications used both inform clinical choices about patient treatment and are useful for epidemiology, public health interventions and research.

1.4.3 Treatment regimens

The first-line drug (FLD) regimen for adults consists of two phases: The first ‘intensive’ stage uses isoniazid (H), rifampicin (R), pyrazinamide (Z) and ethambutol (E) daily for two months; the second ‘continuation’ stage then follows with only isoniazid and rifampicin three times per week for a further four months (104,122). Ethambutol may be continued for the full six months in settings with high levels of INH resistance and where INH testing is not implemented (104,122).

There are now two second-line drug (SLD) regimens, a shorter and longer one, used for people with drug resistant tuberculosis, including RIF monoresistance. For people with pulmonary RR- or MDR-TB, less than one month of previous exposure to SLDs, no resistance to FLQ or injectable agents and without known or strongly suspected resistance to any of the drugs to be used, the shorter regimen of 9-12 months may be followed (107). DST against FLQ and injectable agents is recommended prior to commencing this treatment, in addition to DST for RIF and INH (107). This involves a four- to six-month intensive phase using gatifloxacin or moxifloxacin, kanamycin, prothionamide, clofazimine, pyrazinamide, ethambutol and high-dose isoniazid, until conversion to sputum smear-negative (see section 1.5.1.2). This is followed by a five-month continuation phase of gatifloxacin or moxifloxacin, clofazimine, pyrazinamide and ethambutol. No drug may be substituted or excluded from this regimen (107).

For people with strains resistant to rifampicin (RR-, MDR-, Pre-XDR and XDR-TB) who are not eligible for the shorter regimen whether HIV positive or negative, the longer regimen is used (107). This has an eight-month intensive phase using at least five drugs to which the strain is susceptible, ideally including the group D1 agents, pyrazinamide plus high-dose isoniazid and/or ethambutol for patients with RR- or MDR-TB without high-MIC isoniazid resistance, and four further agents (Table 1.2), of which one should be from group A, one from group B and two from group C. If owing to further drug resistance or drug availability this may not be possible, then one agent from group D2 and others from group D3 should be added to total five drugs (107). Pyrazinamide is added routinely unless contraindicated by evidence of resistance by reliable DST or expectations or risks of significant toxicity, in which case a further agent from group C or D may be substituted (107). After the initial treatment phase, a less intense regimen is continued for a total duration of 18-24 months, typically around 20-months, in patients who have not been previously treated for tuberculosis but this may be guided by monitoring the patient's response to therapy (123). For patients who have previously been treated for TB, the duration of treatment is guided by the patient's response to treatment but optimal likelihood of treatment success was found after 27.6 and 30.5 months of treatment (124). For patients with pre-XDR- or XDR-TB treatment options are severely limited so inclusion of group D2 and D3 agents is almost always necessary and often drugs are used at the highest doses that patients can tolerate (107). For these strain classifications, use of more drugs is also associated with increased likelihood of treatment success (107,118). With increasingly drug resistant strains and access to some drugs being severely limited in many countries, some patients are considered programmatically incurable then discharged into the community until their death (121). Using the national or subnational drug resistance surveillance data, a standardised SLD regimen may be developed for use in specific countries (104).

Table 1.1: Drugs recommended in the longer treatment regimen for drug resistant tuberculosis *At least five drugs to which the strain is susceptible should be used in the longer treatment regimen, ideally including the group D1 agents, one drug from group A, one from group B and two from group C. Drugs in group D2*

and D3 provide further treatment options to overcome limitations imposed by resistance to preferred drugs.

Group		Drugs
Group A: Fluoroquinolones		Levofloxacin Moxifloxacin Gatifloxacin
Group B: Second-line injectable agents		Amikacin Capreomycin Kanamycin Streptomycin
Group C: Other core second-line agents		Ethionamide / prothionamide Cycloserine / terizidone Linezolid Clofazimine
Group D: Add on agents	D1	Pyrazinamide Ethambutol High-dose isoniazid
	D2	Bedaquiline Delamanid
	D3	p-aminosalicylic acid Imipenem–cilastatin* Meropenem* Amoxicillin-clavulanate* Thioacetazone [§]

*Carbapenems to be used with clavulanate, formulated as amoxicillin-clavulanate; †Drugs in group A and C presented in order of preferred use; ‡Thioacetazone must not be used in HIV positive individuals

The choice of regimen may also be influenced by factors including the patient's needs and likelihood of treatment success (104). The FLD and shorter SLD regimens are standardised and drugs cannot be substituted (107). They were designed to maximise the rate of treatment success, minimise severe adverse side effects and reduce the likelihood of acquisition of drug resistance in the strains, however they cannot be tailored to the specific contraindications of the patient. The longer SLD regimen is adaptable in the drugs used and the duration of treatment such that only agents expected to

be effective are used and the balance of benefits of treatment to harms of adverse side effects can be obtained. This can also accommodate various patient groups including children, pregnant women and those infected with extra-pulmonary tuberculosis (107). Additionally, the regimen may be designed with consideration of difficulties with drug availability, financial costs, suitable preparations and dosages for children and quality assurance of the drug (107). Global treatment outcomes are reported as 85% success for DS-TB using FLD but only 54% for RR/MDR-TB and 30% for XDR-TB, both of which are treated with SLD therapy (however, these statistics are based on studies prior to the introduction of the shorter SLD regimen and regrouping of drugs in the longer SLD regimen) (1). For patients meeting the inclusion criteria for this new shorter SLD regimen, the treatment success rate of 90% (patients cured and those who completed treatment, not those who relapsed, died or whose treatment failed) was significantly improved relative to the success rate of longer regimens at 78% (15). Additionally, these shorter regimens using less toxic drugs were associated with higher rates of treatment adherence and a lower financial cost per patient at US\$1000 for the shorter SLD regimen versus US\$2000-5000 for the longer SLD regimen (1).

Treatment of RR/MDR-TB may include partial lung resection in addition to SLD therapy for HIV negative patients. This is associated with higher rates of treatment success than chemotherapy alone, though this may be because of patient selection bias in that only patients who were fit enough for surgery could undergo this treatment. It is not associated with decreased risk of death however, though this could be due to patient selection bias again should the patients selected for surgery have been more unwell than those who received drugs alone. Without knowledge of the patient prior to surgery, the relative performance of this treatment versus drug therapy alone cannot be evaluated (107).

1.4.4 New drugs and regimens in the development pipeline

Drug resistance really threatens our ability to effectively treat tuberculosis. As strains evolve further resistance, they become more difficult to effectively

characterise and treat with effective antibiotics. There is evidence that programmatically incurable strains are being transmitted throughout society (125) and the emergence of strains that may be resistant to all existing antibiotics would make curing these patients medically impossible. There is an arms race between tuberculosis evolving new drug resistance and clinicians which can only be won by development of new drugs, optimising drug dosages and defining improved treatment regimens. There are a number of new drugs in the development pipeline from the discovery stage up to stage 3 clinical trials (126), including those in new chemical classes, and these provide hope for additions to the treatment arsenal for the foreseeable future. The recent inclusion of bedaquiline and delamanid for treatment of DR-TB marks a success of these endeavours (107). The attrition rate from the pipeline is high so continuing discovery of new compounds is essential to sustain this success. Similarly, new treatment regimens are being trialed with the aim to increase treatment efficiency and the likelihood of successful treatment outcomes, reduce adverse side effects and emergence of additional resistance and decrease treatment time with a number of new regimens in the pipeline (127). The shorter SLD regimen for treatment of DR-TB was a product of such trials, being release in May 2016 (107).

1.5 Diagnosis of tuberculosis

1.5.1 Current diagnostic methods endorsed by WHO

Diagnosis of tuberculosis is paramount both to patients being cured and the epidemic of disease being managed. Depending on the way that patients present to the healthcare setting, they may be assessed using immunological assays, chest X-rays, patient history and/or variable laboratory analyses. These might seek to identify any infection with MTBC, to differentiate latent disease from the active form, to identify active tuberculosis from other clinical infections or to characterise the aetiological mycobacteria. For active disease, the ideal diagnostic outcomes would be to detect tuberculosis, identify the aetiological species, test the drug susceptibility of the strain(s) and epidemiologically type the sublineage (128). These steps are not all routine globally, given limited resources and the obligation to prioritise healthcare

efforts (129). Detection of MTBC tends to be prioritised followed by drug susceptibility testing (DST). The World Health Organisation (WHO) endorses diagnostic tools for use under prescribed circumstances, after evaluating their accuracy and impact on both patients' clinical outcomes and public health (105). Though national tuberculosis programmes may use diagnostic methods that have not been endorsed, the focus of this review is on only the superior methods that have gained WHO recommendation.

1.5.1.1 Diagnosis of infection and active disease using host responses

Immunological assays are used to identify people who are clinically considered infected with *M. tuberculosis*, regardless of whether the disease might be latent or active or where in the body infection may be. There are two assays that are used routinely depending on the income level of the country (69). The tuberculin skin test (TST) involves intradermal injection of tuberculin purified peptide derivative (PPD) into the forearm of the patient. Should a patient possess acquired immunity to mycobacteria, a larger hard, red raised induration will form at the injection site by 48-72 hours after administration. A positive test result depends on size of the induration relative to risk-stratified cut-offs determined locally and pre-test risk of infection (130,131). A limitation with TST is that PPD antigens are not specific to MTBC that cause tuberculosis. Exposure to environmental mycobacteria or prior Bacillus Calmette–Guérin (BCG) vaccination will cause false positive reactions which are particularly problematic in countries close to the equator where people are more likely to be exposed to non-tuberculous mycobacteria (132,133). This non-specific reaction is not a problem with the interferon-gamma release assays (IGRAs); these group of *in vitro* tests each measure the response to early secretory antigenic target ESAT-6 and culture filtrate protein CFP-10 that are produced specifically in MTBC organisms other than *Mycobacterium bovis* BCG (132). For this reason, IGRAs are more specific than TST as an indicator of tuberculosis infection, however they are only recommended by the WHO in higher income settings owing to their higher costs and other issues may mean that TST is a better choice (134,135). Neither TST nor IGRAs can differentiate people who are currently infected with

tuberculosis from those who were previously infected but have acquired immunity (50).

Chest x-ray (CXR) and patient history may be used to indicate active pulmonary tuberculosis. Patient history includes consideration of signs, symptoms (discussed in section 1.3.1), prior TB contact and social risk factors. Symptoms, and not necessarily all TB-associated symptoms, are only present in actively diseased patients and are not present in people who have latent TB (54). These symptoms are also common to other diseases and further investigation is necessary to confirm that a person does have tuberculosis, even after immunological tests have indicated that they have been infected with tuberculosis. Chest x-ray allows visualisation of the lung structure by passing electromagnetic x-ray radiation through the body, where different tissues types allow the radiation pass through to varying extents onto an x-ray sensitive film or electronic detector (136). CXR can highlight show abnormalities in the lungs and has a high sensitivity for diagnosing TB. It can only be used to rule out pulmonary TB amongst differential diagnoses in a symptomatic person, however, as features on the radiograph are common to other diseases and it has a low specificity (137).

The lateral flow urine lipoarabinamannan (LF-LAM) assay may be used for the diagnosis of TB in HIV positive patients who have low CD4 cell counts or are seriously ill (138). Urine is applied to the test strip, which detects the LAM antigen, released from mycobacterial cell walls during metabolic activity or dying cells. Whilst it has good specificity (96%), its sensitivity is very poor (23%) so other methods of detection, such as GeneXpert, smear microscopy or culture, should be used in addition to LF-LAM (138). Given such poor sensitivity, it is not suitable for TB screening. The sensitivity is much lower in patients who do not have HIV co-infection. It was recommended by the WHO given its simplicity, minimal biosafety requirements and low financial cost, in addition to implementation in many high-burden HIV settings prior to endorsement.

WHO-endorsed diagnostics based on host responses or pathology are limited to only detecting tuberculosis infection or pulmonary abnormalities. When

active disease is suspected, all of the methods must be followed up with laboratory analyses.

1.5.1.2 Microbiological detection and drug resistance testing

Laboratory detection methods should be used to confirm active tuberculosis infection then ideally followed up with drug resistance testing. Until the last decade, diagnosis of tuberculosis had relied on the microscopy and culture approaches pioneered in the 1880s (13,139). Development of new diagnostics was revived after the first outbreak of XDR-TB, in South Africa in 2006 (140). This led to the WHO endorsement of the further methods described here. Some are used to detect active tuberculosis and others for drug susceptibility testing, both for initial diagnosis of patients monitoring the infection throughout the patient's treatment. DST includes both phenotypic and molecular approaches that generally perform well for RIF and INH, but poorly for other drugs. Molecular DST approaches fail to distinguish susceptibility between individual FLQ and the group of second-line injectable aminoglycoside drugs, though this is possible with phenotypic approaches using the same drugs as provided in local treatment regimens (117).

The WHO describes laboratory testing as being conducted within a hierarchical network of laboratories at the (105):

- Peripheral-level to service sub-districts and communities
- Intermediate-level for regional and district service
- Central-level (reference laboratories)

These are guidelines, however, and the diagnostic methods used and intended outcomes (e.g. screening, case-finding, treatment follow up, surveillance) may not be the same throughout laboratories at a similar level in different countries or regions. The diagnostic tools available in a setting laboratory are subject to their ability to meet the needs of the epidemiological landscape of the region they serve, alongside limitations to financial support, trained laboratory staff, laboratory infrastructure and biosafety measures. In turn, this informs diagnostic algorithms that vary greatly between countries (105).

Smear microscopy is used to detect tuberculosis directly from sputum. The WHO recommends that two consecutive smears are assessed in the same day for TB to be diagnosed by this method (141). Sputum samples are prepared either with the Auramine-O or Ziehl-Neelsen (ZN) staining methods where the thick lipid-containing cell walls of mycobacteria preserve stains after decolourisation with acid, earning mycobacteria the name 'acid-fast bacilli' (AFB) (105). The sputum preparations are visualised with a light-emitting diode fluorescence or Brightfield microscope, respectively, and TB is diagnosed after counting at least one AFB across 100 microscopic fields, in one or both smear preparations (141,142). Smear grades can be designated by the number of AFB seen and used as a proxy for bacterial load both to monitor the progress of treatment and as a guide to infectiousness (143,144). Conversion from smear positive to negative is used to determine treatment success. The Auramine-O procedure is considered more sensitive than ZN approaches and equally as specific (145) though the sensitivity of smear microscopy is generally low, requiring a minimum of 5000 AFB per millilitre of sputum to produce a positive result (105). It is expected that smear microscopy is available in every tuberculosis diagnostic laboratory and recommended that smear is performed as an initial diagnostic, at the end of the intensive phase of treatment and at treatment completion (105).

Culture of mycobacteria can be undertaken on solid media or in liquid media, with the latter achieved manually or through automation. Sputum must be decongested and decontaminated prior to inoculation of the media to prevent over-growth of contaminants, though this may still occur. Cultures must be incubated at 36-37 °C for several weeks, requiring incubators to do this, and liquid cultures should be agitated continuously. Mycobacterial growth may be detected visually and confirmed with rapid immunochromatographic assays to confirm and identify mycobacteria or through automated detection of metabolites (146). Commercial liquid culture systems with automated detection of metabolites, such as the BD BACTEC™ MGIT™ automated mycobacterial detection system (Becton Dickinson) (BACTEC), revolutionised culture of MTBC by reducing time to culture positivity and providing increased culture yield relative to propagation on solid media (105,147). It came with the disadvantages of slightly decreased sensitivity, an

increased risk of contamination and higher biosafety requirements than solid media culture, also with increased financial expense. Culture-based diagnostics are available throughout most countries, regardless of income level, but not necessarily in every laboratory.

Drug susceptibility testing can be conducted by incorporating drugs into the media used for culturing tuberculosis in a number of ways:

- In the proportion method, a standardised inoculum of mycobacteria is grown on both drug-free and drug-containing solid media and the resistant proportion is determined by the relative number of colonies on each plate. When more than 1% of colonies are resistant (for RIF and INH), therapy is likely to fail and the patient should be treated with a drug resistant therapy regimen (148). Automated methods used modified versions of this approach (149–151).
- The absolute concentration incorporates dilutions of the drug into solid or liquid media and adds a standardised inoculum of culture. The minimum inhibitory concentration is determined as the lowest concentration of the drug with which culture grows (152).
- The resistance ratio method is as the absolute concentration method except it divides the MIC value of the clinical isolate by the MIC value for a susceptible strain. When the ratio is 2 or less the clinical isolate is considered susceptible and when the ratio is eight or more, the sample is considered resistant. Only inoculum size needs to be standardised in this approach (150,151).

These methods represent the gold standard for DST, though they may fail to detect some mutations causing rifampicin resistance (153). Though many of the methods used are able to provide minimum inhibitory concentrations (MIC) for a particular strain against a given drug, resistance is defined using critical concentrations of the drug used by the WHO (154). These methods are slow, taking 9-12 weeks to produce a result using solid media or 3-5 weeks using liquid media (105,146).

The Xpert MTB/RIF assay which automates a closed-cartridge nucleic acid amplification test using the multi-pathogen GeneXpert diagnostic platform was endorsed for use by the WHO in 2010 and recommended as an alternative

to smear microscopy in 2013 (155,156). It simultaneously detects the presence of MTBC species and predicts resistance against rifampicin directly from sputum or some extra-pulmonary clinical samples in less than 2 hours. Detecting rifampicin resistance is particularly important as it determines whether first- or second-line drugs should be used for the treatment of disease and indicates that further DST should be undertaken. An update of this product, the GeneXpert Ultra, was released this year (2017) improving the sensitivity of detection by 5%, particularly in children and HIV-infected people, though decreasing specificity by 3.2% due to increased detection of non-viable mycobacteria (157). This assay improves on the sensitivity of smear microscopy, making it an attractive alternative as a first-line detection tool when the additional expense, consumable reagents and local infrastructure are not prohibitive.

Another smear alternative, the tuberculosis loop-mediated isothermal amplification (TB-LAMP) or Loopamp™ *Mycobacterium tuberculosis* complex (MTBC) detection kit, Eiken Chemical Company, was endorsed by the WHO in 2016 for detection of pulmonary TB (no DST) in patients with signs and symptoms suggestive of TB (158). It was recommended that this assay should not replace other rapid molecular tests that also detect rifampicin resistance, like the GeneXpert and despite lower financial costs of the assay, it is unclear whether it offers operational advantages over the GeneXpert assay.

Line probe assays (LPA), such as the GenoType MTBDRplus assay (Hain Lifescience, Nehren, Germany), can be used as an initial test to determine genetic resistance to RIF, INH, FLQ, ETB and second-line injectable drugs (SLIDs) directly from smear-positive sputum or cultured MTBC isolates (106,159,160). Testing of these SLDs with LPAs can be conducted on smear-negative sputum (106). LPAs involve DNA extraction directly from sputum, polymerase chain reaction (PCR) of the relevant resistance-determining region with biotinylated primers then application of the PCR product to test-strip embedded with oligonucleotide probes. Hybridisation of the amplicon and probes will occur in the absence of resistance-associated mutations promoting colorimetric development on the test-strip indicative of drug sensitivity. They were endorsed by the WHO in 2008 for RIF and INH testing

then in 2016 for FLQ, ETB and SLID testing (106). LPAs provide a relatively rapid assessment of resistance to the most important drugs for determining which patients are suitable for shorter SLD regimen or which drugs may be suitable in longer SLD regimen. This method of DST may be completed in 24-48 hours directly from sputum, though may not be suitable for use in peripheral-level laboratories given moderate needs for staff training, consumables, laboratory infrastructure and biosafety measures (105). Where available, LPAs should be used in preference to phenotypic DST.

The WHO has recommended non-commercial DST approaches under highly specified circumstances for the detection of RIF and INH resistance (161):

- The microscopic observation drug susceptibility (MODS) assay involves culture of decontaminated sputum or prior cultured isolates in liquid media with and without drugs. The cultures are visualised with an inverted microscope and resistance to the drug is determined by the characteristic cords of mycobacterial growth in both cultures with and without the drug (161). Susceptibility to the drug is indicated by the absence of growth in its presence but not in the drug-free culture. Growth may be detected in 2-21 days from inoculation with sputum or in 3-4 weeks after inoculation with culture and requires significant training and moderate laboratory infrastructure, comparable with commercial liquid-culture DST (105).
- The nitrate reductase assay (NRA) exploits the biochemistry of mycobacteria since they reduce nitrate to nitrite during culture. Culture can be undertaken in the presence of the drug and growth of the organism, indicative of resistance to the drug, can be detected by colorimetric change that occurs by addition of Griess reagent (162). The NRA can be performed directly from sputum in 6-9 days or from culture in 7-11 weeks with minimal laboratory infrastructure and moderate needs for consumables, staff training and biosafety precautions, similar to solid culture of TB (105).
- The colorimetric redox indicator (CRI) assay uses liquid cultures of mycobacteria and exposes them to the drug of interest. Colorimetric indicator added to the culture causes colour change that is proportional to the number of viable mycobacteria (105). This method is slow relative

to molecular tests or automated liquid-culture DST and requires a high-level of biosafety precautions, but also less expensive and requires minimal infrastructure.

These methods are suitable for use in intermediate-level and central laboratories as an interim option until the capacity for GeneXpert, LPAs or automated culture-based DST can be established (161). There are concerns about reproducibility of the MODS assay, NRA and CRI method alongside elevated issues with biosafety risks and staff training, relative to the performance of methods preferred by the WHO.

There are a number of other short-comings amongst diagnostic tools that lead to substantial inadequacies in diagnosis throughout the world. Both the advantages and disadvantages of individual detection and DST method are summarised in Table 1.2 and 1.3, respectively. The main concerns are associated with accuracy and rapidity of diagnosis and the feasibility to implement these tools in all settings across the world. Collectively, these approaches also fail to provide clinically useful information about

- susceptibility to all drugs used within the treatment regimens in a short time frame. Should drug resistance go unidentified, regimens including poorly suited drugs may be used and lead to treatment failure.
- mixed infection. Infection with multiple strains of TB may impede adequate treatment if these have mixed clinical susceptibility, should the resistance not be detected (163). This may lead to fixation of resistance within the infecting TB population and to transmission of this resistant strain (164–166).
- the species of infecting MTBC. This is important as both *M. canettii* and *M. bovis* are intrinsically resistant to pyrazinamide (167). Unidentified infection with NTMs, may lead to use of typical drug regimens that are insufficient for their treatment (68,168). Immunocompromised individuals may also have become infected with the *M. bovis* Bacillus Calmette- Guérin (BCG) vaccine strain, an infection called BCGosis, which may offer the first indicator of their immune status.
- epidemiological type of the strains that may inform local public health interventions (discussed further in section 1.5.3.1)

Even the most advanced diagnostic technologies implemented together are inadequate for providing all information required for prompt treatment of all patients.

The diagnostic methods used and their roles in the diagnostic algorithm of a tuberculosis laboratory network vary between countries and sub-regions depending on the epidemiological demands of the area and resources available. In highly resourced networks, the most highly performing methods are used, including GeneXpert, rapid DST and gold standard DST by culture in the presence of further drugs (beyond RIF, INH, FLQ and SLIDs (154)), along with culture for subsequent epidemiological characterisation. In low resource countries, laboratories are limited to using the financially cheapest and least demanding diagnostic methods, which may simply amount to only smear microscopy. The capacity of most laboratories falls somewhere between these two ends of the spectrum.

Table 1.2: Advantages and disadvantages of diagnostic methods for the detection of tuberculosis

Diagnostic method	Advantages	Disadvantages
Smear microscopy	<ul style="list-style-type: none"> • Inexpensive • Simple • Rapid • Suitable for peripheral-level labs and higher • Low biosafety precautions required • Suitable for monitoring infection 	<ul style="list-style-type: none"> • Relatively insensitive, especially in children, HIV-positive patients and for extra-pulmonary TB. • Training and quality assurance for laboratory staff in necessary • Provides no information beyond that AFB are present
Culture	<ul style="list-style-type: none"> • Increased sensitivity vs microscopy • Provides cultured material for DST or epidemiological strain typing, either as single or multiple colonies • Suitable for monitoring infection 	<ul style="list-style-type: none"> • Slow, especially for solid culture • Complex • Expensive • Higher safety precautions required • Requires specific laboratory equipment • Contaminants may over-grow MTBC culture • Suitable for intermediate-level and central reference laboratories
GeneXpert	<ul style="list-style-type: none"> • Rapid • Sensitivity: 88%* • Specificity: 99%* • Detects TB in smear-negative samples • Improved sensitivity for people with HIV vs smear microscopy • Low biosafety precautions required • Minimal staff training required • May be used in peripheral level laboratories 	<ul style="list-style-type: none"> • Expensive • Requires stable electricity supply • Room temperature around machine cannot exceed 30 °C • Cartridges must be stored below 28 °C and used before expiry • A computer must accompany the device • GeneXpert system requires annual calibration • Not suitable for monitoring infection

Diagnostic method	Advantages	Disadvantages
	<ul style="list-style-type: none"> • High-throughput when GeneXpert selected has many cartridge slots 	
TB-LAMP	<ul style="list-style-type: none"> • Rapid • Sensitivity vs smear microscopy: 7.1 to 13.2% but poor in HIV co-infected patients • Cheaper than GeneXpert • Low biosafety precautions required • No sophisticated equipment required • Suitable for peripheral-level labs and higher • Low infrastructure requirements • High-throughput 	<ul style="list-style-type: none"> • Specificity vs smear microscopy: -1.8 to -3.2% • Laboratory technicians require further training

* Improved sensitivity and specificity with GeneXpert Ultra, released 2017; information in table obtained primarily from WHO policy reports (105,158,161)

Table 1.3: Advantages and disadvantages of methods used for drug susceptibility testing in MTBC

Diagnostic method	Advantages	Disadvantages
Culture and DST	<ul style="list-style-type: none"> • Gold-standard diagnosis of drug resistant TB though fails to detect some high MIC rifampicin mutations 	<ul style="list-style-type: none"> • Slow • Suitable for use only in central laboratories • Accuracy of detection varies according to drug • Requires high-level of staff training • Extensive infrastructure required • High-level biosafety precautions necessary
GeneXpert	<ul style="list-style-type: none"> • Rapid • Sensitivity: 95%* • Specificity: 98%* 	<ul style="list-style-type: none"> • Expensive • Tests for RIF resistance only • Requires stable electricity supply

Diagnostic method	Advantages	Disadvantages
	<ul style="list-style-type: none"> • Low biosafety precautions required • Minimal training required • May be used in low level laboratories 	<ul style="list-style-type: none"> • Room temperature around machine cannot exceed 30 °C • Cartridges must be stored below 28 °C and used before expiry • A computer must accompany the device • GeneXpert system requires annual calibration • Assay must be repeated when RIF resistance is predicted in patients considered at low risk of drug resistance • Discordance with phenotypic DST may occur, requiring DNA sequencing, if available, in a reference laboratory
Line-probe assays	<ul style="list-style-type: none"> • Rapid • High-throughput • High sensitivity and specificity (variable for each drug) 	<ul style="list-style-type: none"> • Conventional culture and DST still required • Recommended for smear-positive sputum and <i>Mtb</i> samples only (for FLDs) • Suitable for use at central/regional laboratories only • Sensitivity for INH resistance lower than for culture methods; INH sensitive samples should be tested by culture-based DST, particularly in high INH-resistance regions • Lower accuracy for determining MDR-TB than other methods • Culture still required for DST of other drugs; there are further disadvantages associated with culture. • 3 rooms with unilateral workflow required
MODS	<ul style="list-style-type: none"> • High sensitivity (98%) • High specificity (RIF: 99%; INH: 91%) 	<ul style="list-style-type: none"> • Slow, though quicker than conventional culture directly from culture • Requires highly-skilled staff

Diagnostic method	Advantages	Disadvantages
	<ul style="list-style-type: none"> • Cheaper than commercial culture-based DST • Non-proprietary reagents 	<ul style="list-style-type: none"> • Requires inverted light microscope, used uniquely for this TB diagnostic method • Moderate biosafety measures are required • Must be used at higher levels of laboratory network
NRA	<ul style="list-style-type: none"> • High sensitivity (97%) • High specificity (RIF: 100%; INH: 99%) • Cheaper than commercial culture-based DST • Non-proprietary reagents • Requires minimal infrastructure or equipment 	<ul style="list-style-type: none"> • Slow, though quicker than conventional culture directly from culture • Requires moderate staff skills • Requires inverted light microscope, used uniquely for this TB diagnostic method • Moderate biosafety measures are required • Must be used at higher levels of laboratory network
CRI	<ul style="list-style-type: none"> • High sensitivity (RIF: 98%; INH: 97%) • High specificity (RIF: 99%; INH: 98%) • Cheaper than commercial culture-based DST • Non-proprietary reagents • Requires minimal infrastructure or equipment 	<ul style="list-style-type: none"> • Testing must be carried out on prior culture • Slow • Requires highly-skilled staff • High-level biosafety precautions are required • Must be used at higher levels of laboratory network

* Improved sensitivity and specificity with GeneXpert Ultra, released 2017; information in table obtained primarily from WHO policy reports (105,158,161)

1.5.1.3 Epidemiological typing of the MTBC

Epidemiological typing is a useful part of the diagnostic paradigm that is often overlooked (128). Various molecular epidemiology techniques have been developed to discriminate between clinical strains of MTBC. These techniques are based on the genetic variations between strains of mycobacteria, both amongst the MTBC species and NTMs. While the MTBC

was thought to be genetically monomorphic until the 1990's, recent techniques have instead highlighted their genomic diversity. Over time, the genetic resolution with which we can analyse these species has improved and it continues to expand our abilities to analyse transmission networks, to identify contamination and discriminate between relapse and re-infection (169–171),

During the last few decades, various molecular techniques of DNA fingerprinting have been used to discriminate between clinical strains of MTBC. A subset of these are used more commonly in public health microbiology, including

- IS6110 restriction fragment length polymorphism (*RFLP*) typing, developed in 1991 (172,173). It differentiates strains by the copy number and location of the insertion sequence (IS) 6110, a mobile genetic element found exclusively in the MTBC. The method is practically onerous, difficult to standardise for comparison between laboratories and not suited for strain classification (174). It has been beneficial to the study of transmission and to identify outbreaks (173,175–181).
- Spacer oligonucleotide typing (spoligotyping), developed in 1997 (182). It is based on the presence or absence of (usually) 43 unique regions intercalated between direct repeats in the Clustered Regularly Interspaced Short Palindromic Repeats region (CRISPRs) of the MTBC genome. Relative to other classification methods, it is able to discriminate fewer lineages and these usually have low statistical support (174). Owing to its ease of use, cost-effectiveness, reproducibility, it has been widely adopted for use in laboratories (183) to provide groupings of genotype families, though it has limited use for understanding the phylogeny or transmission of strains.
- Mycobacterial interspersed repeat-unit-variable-number of tandem-repeat (MIRU-VNTR) typing (184,185), developed in 2006. This classifies MTBC strains by the number of tandem repeats at VNTR loci across the genome. Carrying out this strain typing is labour intensive but it is considered the gold standard approach epidemiological typing method for tuberculosis and its discriminatory ability increases with the rising

number of loci investigated (183). The classifications are globally comparable.

More recently, whole genome sequencing has been used to compare TB strains using two main genomic features. *Region of Difference* (RDs) or *Large Sequence Polymorphism* (LSP) (186) are deletions of genomic regions and unidirectional markers of evolution since horizontal gene transfer (HGT) is rare in the MTBC (187). As such, they were used to define six primary lineages within the MTBC which are each geographically associated with particular global locations and a seventh was later added (187,188). *Single nucleotide polymorphisms* (SNPs) are substitutions of individual nucleotides in one strain relative to another that have surpassed a critical frequency threshold in the bacterial population. Below this threshold, they are considered 'mutations' and the term 'single nucleotide variant' is used to define both SNPs and mutations (189). Owing to low SNP frequency and absence of HGT or genetic recombination amongst the MTBC, SNPs are reliable phylogenetic markers that help to resolve sub-lineages of MTBC and distinguish individual strains (190). Comas *et al* produced the first global phylogeny of the human-associated MTBC derived from whole genome variants in 2010 (191), providing the basis for phylogenetic studies going forward. The numbers of SNPs that differentiate two strains, called the 'SNP difference', can be used to define outbreak populations and indicate direct transmission between individuals. Whole genome based epidemiological analysis is not available globally owing to the unaffordable costs of sequencing for much of the world, but it provides the ultimate genetic resolution with which to compare TB strains (192).

1.5.1.4 New diagnostics in the development pipeline

In 2016, there was a huge gap between the estimated incidence of 10.4 million tuberculosis cases and the 6.3 million new cases being reported (1). Rather than the target that every TB patient should receive a rifampicin susceptibility test, this was only the case for 41% of patients (1). These inadequacies prevent the timely diagnosis and treatment of patients from this curable disease.

Having evaluated the diagnostic tools available, there is evidently a lacking technical capacity for diagnosis of tuberculosis. New and improved tools are needed to circumvent the short-comings of those currently available. For this reason, target product profiles have been constructed to focus developers on the needs of the end users (193). Alongside targets for improved biomarker and triage tests, there is the aim to provide an upgraded test to replace smear microscopy for the detection of TB at point-of-care and to offer a rapid DST assay for detection of first-line drug resistance in peripheral-level laboratories. Fitting with this there are a number of rapid molecular tools scheduled for WHO evaluation in the next few years including sequencing technologies (discussed further in section 3.1.3), expansions of the GeneXpert system and other molecular drug resistance assays, some of which include resistance testing to confirm XDR-TB (1). This is just the tip of a comprehensive diagnosis pipeline that has a multitude of new products aiming to meet the target product profiles (193,194).

1.6 Unmet needs and the global challenge

Tuberculosis is now the leading cause of death from infectious disease (1). Despite being a treatable disease, it is one of the top ten causes of death worldwide and killed 1.7 million people in 2016 (1). Alongside this unacceptable burden of mortality, 10.4 million people are newly infected each year (1). For those individuals, life typically includes significant social, economic and psychological trauma alongside the severe symptoms of the illness.

Tuberculosis was explicitly mentioned in the Sustainable Development Goals set out by the United Nations for the years 2015-2030 by calling for an “end to the epidemic”. In line with this, the Global Plan to END TB was developed with the goals to reduce mortality by 90% and the incidence rate by 80% relative to those in 2015 (195). To meet this, the annual decline in incidence rate must fall to 10% by 2025, according to models. To put the enormity of this challenge in context, a 10% decline has only ever been achieved in countries with well-performing universal health care and high levels of socioeconomic development. Steps to achieve these goals focus on early and comprehensive

TB diagnosis, actively finding cases amongst the people at highest-risk of disease, treatment of all disease including drug-resistant tuberculosis and latent disease in people at high-risk of progression to active disease, and implementation of an effective anti-TB vaccination. This is reliant on both socio-political support and significant research for implementation of new diagnostics, drug and vaccines, underpinned by an improved understanding of pathogenesis and epidemiology (195).

1.7 Objectives and outline

1.7.1 Thesis objectives

Combatting the TB epidemic requires a new diagnostic approach that is timely, cost-effective, informative, sensitive, specific and feasible in resource-deprived locations. Though the work presented in this thesis, I have taken the first steps in exploring shotgun (meta)genomics for this purpose. The overarching aims were to:

1. Develop a clinically-feasible shotgun metagenomics method that could potentially diagnose infection with any pathogen from a clinical sputum sample, using tuberculosis diagnosis to illustrate this in principle (chapters three and four).
2. Investigate the genotype-phenotype relationship in drug resistant MTBC in Peru (chapter five)
 - a. Predict drug resistance profile using whole genome sequencing, including identification of novel mutations associated with pyrazinamide resistance.
 - b. Evaluate the performance of drug resistance detection approaches.

1.7.2 Thesis outline

The work presented in chapters three and four was carried out in collaboration with colleagues in The Gambia. This West African country ranks 173 of 188 countries in the 2016 human development index, with a life expectancy of just 60.5 years and a mean of just 3.3 years of schooling (196). I

spent approximately a year in The Gambia developing methods for extraction of DNA directly from sputum such that MTBC organisms could be detected after shotgun metagenomic sequencing. The approaches that I took were feasible in this resource-poor setting within the facilities of the MRC Unit, The Gambia. Chapter three describes my initial attempts to identify a method of extracting DNA from sputum with which I could detect MTBC by metagenomic sequencing. In chapter four, I then attempted to optimise the method by addressing the biggest challenge to characterising the MTBC organisms, the overwhelming proportion of human DNA in the extract.

A distinct project, presented in chapter five, addressed other challenges that need to be overcome before genomic and metagenomic approaches could fulfil their full clinical potential. This study was conducted in collaboration with researchers in Peru, using methods of phenotypic drug resistance testing and DNA extraction that are feasible in this resource-poor country. As yet, our understanding of the genetic basis of resistance is limited for some of the antibiotics used in anti-TB therapy, particularly pyrazinamide. In this study, we used numerous phenotypic methods for determining pyrazinamide resistance, as well as the MODS method for determining rifampicin and isoniazid resistance, and compared these outcomes with predictions made by bioinformatic software after whole genome sequencing of the cultured organisms. I then looked for new mutations in resistance-associated genes that had not been identified by the bioinformatic tools. This allowed me to evaluate both novel genetic patterns associated with resistance and the methods used for determining this resistance.

In both projects, I was able to conduct cutting-edge genomic research in resource-limited developing countries, with exchanges of knowledge that benefitted both me and my foreign collaborators. My understanding of the limitations of tuberculosis diagnosis and healthcare in these countries has been expanded and I was able to train collaborators in DNA sequencing techniques and share my knowledge of genomics to colleagues in The Gambia.

2. Materials and Methods

2.1 Ethical approvals

The metagenomic projects described in chapters three and four were undertaken at the Medical Research Council Unit, The Gambia (MRCG) using sputum samples obtained in two on-going projects:

- the Enhanced Case Finding project (197)
- The Gambian Nationwide Tuberculosis Prevalence survey (GAMSTEP) (198)

Through these parent projects, informed consent was obtained and six months of Directly Observed Treatment short course (DOTS) was provided free of charge by the National TB control Program for all participants. The sputum metagenomics investigations were approved by the MRCG's Scientific Co-ordinating Committee under reference SCC 1232.

Ethical approval for work on Peruvian samples, described in chapter five, was obtained from the institutional review board of Universidad Peruana Cayetano Heredia. Consent was not obtained from participants because the data were analysed anonymously after checking for duplicated patients. Institutional approval for the study was obtained from the Peruvian Ministry of Health.

2.2 Microbiology

All microbiology methods were undertaken by collaborators at MRCG for chapters three and four or Universidad Peruana Cayetano Heredia for chapter five.

2.2.1 Smear microscopy for all chapters

Aliquots of decontaminated sputum (see section 2.4.1.1) were stained with Auramine-O and evaluated by smear microscopy, according to standard procedures. In samples where acid-fast bacilli (AFB) were detected, Ziehl-

Neelsen (ZN) staining was used to confirm and quantify the number of tuberculous bacilli present. ZN smear grades were given as +, ++ or +++ according to the number of ZN-stained AFB observed as specified in Table 2.1.

Table 2.1: Ziehl-Neelsen smear results and grades

Sputum is graded by the density of acid-fast bacilli (AFB) observed with microscopy in the specified number of fields for that smear grade, as detailed below.

AFB observed	No. fields examined	Result (grade)
10 AFB per oil immersion field	20	Positive (+ + +)
1-10 AFB per oil immersion field	50	Positive (+ +)
10-99 AFB in 100 oil immersion fields	100	Positive (+)
1-9 AFB in 100 oil immersion fields	100	Scanty
No AFB in 100 oil immersion fields	100	No AFB seen

2.2.2 Mycobacterium tuberculosis complex culture: Chapters three and five

At the MRCCG, all diagnoses of TB were confirmed by culture of the *M. tuberculosis* complex. Sputa were decontaminated by the sodium hydroxide and N-acetyl-L-cysteine (NaOH/NALC) method as described by Kent *et al* (199). Decontaminated sputa were inoculated into the BACTEC™ MGIT™ 960 System (Becton Dickinson) according to the manufacturer's protocol. Subsequently, liquid culture was transferred to Lowenstein–Jensen (LJ) slopes for solid media culture. In Peru, cultures from samples were obtained during the microscopic-observation drug susceptibility (MODS) assay described in section 2.2.3. Liquid cultures from the control tests, that did not use drugs, were sub-cultured onto solid LJ medium for further study.

2.2.3 Microscopic-observation drug susceptibility assay: Chapter five

MODS testing was conducted on separate sputum aliquots for rifampicin, isoniazid and pyrazinamide as described by Caviedes (200) with the following modifications:

- Sputum-medium mixture preparation: 2 mL of each decontaminated sputum aliquot was diluted in 15 mL of Middlebrook 7H9 broth containing a final concentration of 20 $\mu\text{l/ml}$ PANTA Antibiotic Supplement Stock and 10% Oleic Albumin Dextrose Catalase.
- Antibiotic testing: 900 μL of each sputum-medium solution was inoculated into five wells of a 24-well plate each with 100 μL drug solution, to result in a final concentration of either 0.4 $\mu\text{g/ml}$
- , 1.0 $\mu\text{g/ml}$ RIF, 400 $\mu\text{g/ml}$ PZA, 800 $\mu\text{g/ml}$ PZA or 100 μL 7H9 broth to act as a negative control.
- Determining results: After five days, wells were examined every one or two days for the presence of mycobacteria under an inverted light microscope at x40 magnification. An aliquot was considered sensitive to the given drug after nine days of incubation without observed growth in its presence. Observation of microbial growth in the no-drug-control and in the presence of an antibiotic was determined as resistance to it.

2.2.4 BACTEC 960 PZA resistance testing: Chapter five

The BACTEC™ MGIT™ 960 PZA assay (Becton Dickinson, Sparks, MD) was performed on isolates from solid culture according to the manufacturer's instructions using a critical concentration of 100 $\mu\text{g/ml}$ pyrazinamide.

2.2.5 Wayne's pyrazinamidase assay: Chapter five

The Wayne's pyrazinamidase assay that measures activity of the *pcnA* gene as a proxy for determining pyrazinamide resistance, was conducted on cultured isolates in chapter five as described by Wayne (201). In total, two isolates from each sputum expectorate were subjected to this assay in separate tubes, the first was examined after four days, then again four hours later if

negative, and the second examined after seven days should the first have been negative both times. If a pink colour was present in either of these assays, the result was positive, indicating pyrazinamidase activity associated with PZA sensitivity.

2.2.6 Spoligotyping: Chapter three

Spoligotyping, used in chapter three, was undertaken from cultured isolates, which were derived from a single colony on solid medium, using standard protocol (182,202).

2.3 Sample collection and handling

2.3.1 Sputum samples: Chapters three and four

Sputum samples were collected from participants in the ECF project at MRCG between February 2012 and February 2014 and from participants in the GAMSTEP survey between March and August 2015. Patient-identifying information was removed before samples were processed, but basic biographical details were retained. Participants included males and females, ranging in age from 16 to 73 years. HIV status was unknown. Samples were collected before or after TB treatment had been started. Patients expectorated sputum into a sterile cup that was immediately sent to the laboratory and aliquotted. As part of the standard diagnostic procedures, sputum aliquots from each patient expectorate were subjected to microbiological analysis. Samples selected for further analysis had an even distribution in terms of colour and viscosity between the analysis groups.

2.3.2 Cultured samples: Chapter five

Patients from the DIRESA-Callao region and Hospital Dos de Mayo, Lima, Peru provided sputum expectorates in the periods July-December 2016 and October- December 2016, respectively as part of an on-going project, the MP project. The 73 selected samples were smear-positive and had previously been characterised by rifampicin, isoniazid and pyrazinamide phenotypic

sensitivity testing with complete metadata and DNA suitable for Nextera XT library preparation.

2.3.3 Storage and maintenance of samples: All chapters

At the MRCG, sputum aliquots were refrigerated until use. After culture on LJ slopes, multiple colonies were collected with a sterile loop and inoculated into storage medium for storage at -80 °C. In Peru, all DNA extracts were maintained in EB or TE buffer at -20 °C.

2.4 DNA preparation

2.4.1 Chapter three DNA preparation

2.4.1.1 Sputum decontamination with NaOH-NALC: Chapter three methods A and B, chapters four and five

Sputa were decontaminated using the method described in section 2.2.2.

2.4.1.2 Sputum decongestion with NALC: Methods C-F

Sputa of various positive smear grades were decongested using a modification of the decontamination protocol in 2.2.2, replacing the NaOH with an equal volume of sterile water. After the final centrifugation, pellets were retained.

2.4.1.3 Depletion of human DNA: Methods E and F

A protocol modified from Lim and colleagues (203) was initially used to attempt to deplete human DNA from the decongested sputum. Pellets resulting from sputum decongestion were resuspended in 10 mL sterile deionised water, incubated at room temperature for 15 min then centrifuged at 3330 x g for 20 min. The supernatants were discarded and this resuspension and centrifugation was repeated. The pellets were resuspended with minimal water, 25 µL DNase (2.73 Kunitz/ mL) and 100 µL RDD provided in the

RNase-Free DNase Set (Qiagen, Hilgen, Germany) then incubated at room temperature for 2 hours with repeated inversion of the tubes. The suspensions were centrifuged at 3800 x g for 15 min and the pellets were resuspended in 10 mL TE buffer, then this was repeated before a final centrifugation at 3800 x g for 15 min to collect the pellets for DNA extraction.

2.4.1.4 Mechanical lysis and Qiagen kit extraction: Methods A and D

The DNA from decontaminated samples or samples decongested by NALC was extracted using methods modified from Aldous et al (2005). The pellets were resuspended in 0.5 mL TE and added to 2 mL microcentrifuge tubes containing 0.1 mm zirconia beads (Scientific laboratory supplies, Nottingham, UK). The tubes were centrifuged for 5 min at 1800 x g then incubated at 99 °C for 10 min. After a brief centrifugation to remove the beads, 200 µL of the supernatants were incubated with 50 µL TE buffer containing 0.08 mg/mL lysozyme and 150 U/mL mutanolysin at 37 °C for one hour. The solutions were purified using the Qiagen DNA mini kit (Qiagen, Hilgen, Germany) according to the tissue protocol but using EB buffer to elute DNA.

2.4.1.5 CTAB DNA extraction: Methods B, C and E

Sputum pellets resulting from decontamination, human-depletion or decongestion by NALC (chapter three) or cultured material (chapter five) were resuspended in 0.5 mL TE buffer, incubated at 80 °C for 20 min to kill live bacilli, then supplemented with 50 µL 10 mg/mL lysozyme and incubated on a rocker at 37 °C overnight. After incubation, 100 µL 20% Sodium diethyl sulphate and 10 µL 10 mg/mL proteinase K were added. The solutions were mixed gently and incubated at 60 °C for 40 min. Next, 100 µL of 5 M NaCl and 100 µL of cetyltrimethylammonium bromide (CTAB) were added, the solutions were vortexed until they appeared 'milky' and then incubated at 65 °C for 10 min. This was followed by the addition of 750 µL chloroform: isoamyl alcohol (24:1) and the solutions were vortexed for >10 seconds. The aqueous phases were separated by centrifugation at 11000 rpm for 8 min then gently mixed with 5 µL glycoblue (Ambion, Applied

Biosystems, Foster City, CA, USA) and 550 μL $-20\text{ }^{\circ}\text{C}$ isopropanol. The tubes were incubated at $-20\text{ }^{\circ}\text{C}$ overnight and centrifuged for 15 min at 15000 rpm. The pellets were resuspended in 1 mL $-20\text{ }^{\circ}\text{C}$ 70% ethanol, the tubes were inverted to wash out the isopropanol then centrifuged at 11000 rpm for 5 min. The supernatants were discarded and the pellets were allowed to air dry before resuspending them in 20 μL EB buffer (Qiagen). The DNA extracts were kept at $4\text{ }^{\circ}\text{C}$ overnight before storage at $-20\text{ }^{\circ}\text{C}$.

2.4.1.6 Nucleospin tissue kit extraction: Method F

The human-depleted sputum pellets were resuspended in 0.5 mL TE buffer then heated to $80\text{ }^{\circ}\text{C}$ for 20 min. The DNA was extracted using the Nucleospin® Tissue kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's support protocol for hard-to-lyse bacteria including overnight incubation with proteinase K.

2.4.2 Chapter four DNA preparation

2.4.2.1 Sputum decongestion with sputolysin

Sputum aliquots were added to 970 $\mu\text{g}/\text{mL}$ sputolysin (Sigma Aldrich, St Louis, MO, USA), using an equal volume of sputum to sputolysin, and vortexed vigorously for 30 seconds. The solutions were incubated at room temperature for 15 min then centrifuged at $3220 \times g$ for 20 min. The pellets were resuspended in 1 mL phosphate-buffered saline (PBS), which was then divided equally between two tubes to provide equivalent treatment and control samples from each sputum aliquot. These suspensions were then centrifuged again at $13000 \times g$ for 5 min to produce pellets to be used for human DNA removal.

2.4.2.2 Depletion of human DNA: Method G

The sputum pellets were resuspended in 50 mL PBS, mixed thoroughly by vortexing then centrifuged at $300 \times g$ for 5 min to pellet human cells. The

supernatant solutions were transferred to new tubes and centrifuged at 3220 x g for 20 min to collect remaining cells. The supernatants were discarded.

2.4.2.3 Depletion of human DNA: Method H

The sputum pellets were resuspended in 10 mL sterile deionised water. After incubation at room temperature for 15 min, the suspensions were centrifuged at 3220 x g for 20 min and the supernatants were discarded. The resuspension and centrifugation were repeated. The resulting pellets were resuspended in 0.5 mL water, mixed thoroughly with 0.9 mL AMPure XP (Beckman Coulter, Indianapolis, IN, USA) to bind extracellular DNA and placed on a magnetic rack for 10 minutes to separate the bound DNA from remaining bacilli. These supernatants were centrifuged at 13000 x g for 5 min and the resulting supernatants were discarded.

2.4.2.4 Depletion of human DNA: Method I

Sputum pellets were resuspended in 10 mL sterile deionised water and incubated at room temperature for 15 min. These solutions were centrifuged at 3220 x g for 20 min, the supernatants were discarded then the resuspension, incubation and centrifugation were repeated. The supernatants were discarded, the pellets were resuspended in 100 U DNase in 1.5 mL buffer as supplied by Biolabs (Ipswich, MA, USA) and these suspensions were incubated at 37 °C with gentle rocking for two hours. After centrifugation at 13000 x g for 5 min, the supernatants were discarded.

2.4.2.5 Depletion of human DNA: Method J

The pellets were resuspended in 7.2 mL 1 mM Tris-HCl pH 8 and rocked at room temperature for 90 min. Tris-HCl and MgCl₂ were added to make 200 mM and 10 mM respectively and a total volume of 8 mL. The solutions were rocked with 250 U benzonase (Novagen, EMD Millipore, Darmstadt, Germany) at 37 °C for 2 hours. The benzonase digestion was quenched by vortexing with 88 µL 0.5 M EDTA and 880 µL 1.5 M NaCl. The suspensions were centrifuged at 3220 x g for 20 min and the supernatants were discarded.

2.4.2.6 Depletion of human DNA: Method K

Filters were prepared by washing with 3 mL PBS with 3% BSA. The sputum pellets were resuspended in 10 mL PBS and sequentially passed through Whatman filters with 80 μM , 20 μM then 10 μM pore sizes using a syringe. 10 mL PBS was passed through the filters to collect any remaining bacterial cells. The filtrate was centrifuged at 3220 x g for 20 min and the supernatant was discarded.

2.4.2.7 Nucleospin tissue kit extraction

Extraction undertaken as for chapter three, described in section 2.4.1.6.

2.4.3 Chapter five DNA preparation

2.4.3.1 CTAB extraction from culture

The CTAB extraction was performed by our Peruvian collaborators in the way described in section 2.4.1.5 starting from multiple colonies of culture.

2.5 DNA quantification and fragment length determination: All chapters

DNA was quantified using the Qubit 2.0 fluorimeter (Invitrogen Ltd., Paisley, United Kingdom) and either the HS (high sensitivity) or BR (broad range) Qubit® dsDNA Assay Kit according to manufacturer's protocol. The quantitation range for the two kits are 10 pg/ μL to 100 ng/ μL and 100 pg/ μL to 1000 ng/ μL , respectively. In chapters three and four, negative control samples were subjected to HS quantification to determine the absence of detectable DNA. When appropriate, DNA fragments were analysed using the Bioanalyzer high sensitivity DNA assay (Agilent technologies, Santa Clara, CA, USA).

2.6 DNA sequencing

2.6.1 Nextera XT library preparation: Chapters three and five

For Nextera XT library preparation used in chapters three and five, the DNA concentrations of samples were adjusted to 0.2 ng/ μ L and processed into sequencing libraries using the Illumina Nextera XT library preparation kits according to the manufacturer's instructions (Illumina UK, Little Chesterford, United Kingdom). The selection between a v2 and v3 Nextera XT kit was made based on estimates of the total DNA quantity to be sequenced in one run of the MiSeq. For TruSeq library preparation used in chapter four, DNA extracts were diluted to have a 60 μ L total volume with up to 200 ng DNA.

2.6.2 TruSeq Nano library preparation: Chapter four

The TruSeq Nano DNA HT library preparation kit was used with modifications to the manufacturer's instructions (Illumina UK, Little Chesterford, United Kingdom) to prepare sequencing libraries. All steps to fragment DNA were omitted. The protocol to select library size was amended to use 200 μ L AMPure XP (Beckman Coulter, Indianapolis, IN, USA) to clean up the DNA in place of the SPB reagent and the removal of small fragments was omitted. For the PCR enrichment, 15 PCR cycles were used rather than 8. Libraries were sequenced using the Illumina MiSeq and reagents facilitating either 2 x 251 bp paired-end reads (chapter three and four) or 2 x 301 bp (chapter five) paired-end reads to generate FASTQ files.

2.7 Bioinformatic analyses

2.7.1 Read alignment: All chapters

Sequence reads were aligned to reference genomes in a number of ways, depending on the application:

- ‘Relaxed’ alignment of sequence reads to the *Mycobacterium tuberculosis* H37Rv (GenBank accession number AL123456) was undertaken using the --very-sensitive-local pre-set option using Bowtie 2 version 2.1.0 (204).
- ‘Stringent’ alignment against the H37Rv reference was undertaken using Bowtie 2 with the options --ignore-quals --mp 10,10 --score-min L,0,0.725 --local --ma 1, as adapted from recent studies on ancient DNA (205,206).
- In the initial metagenomics study, chapter three, alignment against the hg19 human reference genome (GenBank Assembly ID GCA_000001405.1) used Bowtie 2 with --fast-local pre-set option.
- In the follow-up metagenomics study, chapter four, alignment against hg19 was performed using Bowtie 2 with the --sensitive-local pre-set parameters.

Summary statistics about alignments were obtained using Qualimap (207).

2.7.2 Phylogenetic placement: Chapters three and five

For chapter three, H37Rv-aligned reads were placed into a phylogenetic tree previously built by Dr. Martin Sergeant and Dr. Zhemin Zhou from published TB strains with designated lineages as described by Doughty et al (208). Dr. Sergeant used pPlacer software (209) to define single lineages of MTBC present in samples with very few reads. In chapter five, I used MGplacer (210) was used to define samples with mixed lineages in a similar manner.

2.7.3 Small variant calling: Chapter five

Small nucleotide polymorphisms (SNPs), insertions and deletions (INDELs) were determined from the Peruvian whole genome sequences after relaxed alignment of reads to H37Rv. Previously defined repetitive regions (191) were excluded from further analysis with a custom script. Variants with a minimum coverage of 5 bases, quality score above 20, minimum variant frequency over 0.1 and a p-value greater than 0.001 were determined using Varscan2 (211). SNPs that fell near to INDELs were excluded and the remaining SNPs for each sample were compiled into a list.

2.7.4 Genomic prediction of drug resistance: Chapter five

Prediction of drug resistance phenotypes in the Peruvian whole genome sequences was undertaken by a number of approaches. Initially, software tools, Mykrobe predictor, KVarQ and PhyResSE, were used to identify known resistance-causing mutations from databases of known genotype-phenotype associations (212,213). These web-tools identified resistance-causing mutations to corroborate most resistance phenotypes, but there was discordance in the phenotypes and genotypes of some samples relating to pyrazinamide and rifampicin susceptibility. For all samples, the lists of small variants previously generated were searched for mutations in resistance-associated genes and upstream of their predicted operon start sites in the following genomic regions relative to the reference genome NC_018143:

- Pyrazinamide
 - *pcnA*: 2289607-2289707; 2288689-2289249
 - *rpsA*: 1833449- 1834991
 - *panD* 4044041- 4044460; 4046297-4046397
 - *hadC*: 731830-731930; 732828-733328
- Rifampicin
 - *rpoB*: 759659- 763328
 - *embB*: 4235452-4235552; 4246692- 4249988
 - Rv2629: 2955677-2956901

Mutations present in the resistance-associated regions of phenotypically resistant strains, but not in any phenotypically sensitive strains were considered potentially-resistance-causing.

2.8 Statistical analyses: Chapter four

The statistical analyses in chapter five were conducted in collaboration with Dr. Peter Kimani. The percentage of reads per sample that aligned to the human reference genome were calculated to summarise the data. Using the R package, metaphor (214), a restricted maximum-likelihood mixed effects model (REML) was fitted to the data using random-effects meta-analysis (rma).

3. Initial Attempts at Metagenomic Diagnosis of Tuberculosis

The work presented in chapter three has been published here:

Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. *Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer*. PeerJ. 2014 Sep 23;2: e585. doi: 10.7717/peerj.585.

3.1 Introduction

3.1.1 The need for new TB diagnostics

Currently available tools fail to diagnose and characterise all active TB cases in a timely manner. Some patients are failed at every stage of the diagnostic process, from diagnostic facilities not being available to them to poor performances of those laboratory diagnostic technologies available. The WHO estimates that 10.4 million people developed new active tuberculosis disease in 2015, however, only 6.1 million cases were notified (108). As a consequence of those diagnostic failures, it is assumed those missing were unable to obtain treatment. Even of the bacteriologically confirmed cases, only 30% were able to obtain drug susceptibility testing (DST) against even rifampicin meaning potential drug resistant infections went both undiagnosed and untreated (108). From every stage of diagnostic failure, people have been left ill, as potential vectors to spread disease and undermine TB control efforts.

A core component of the End TB Strategy is early detection of TB with universal DST (215). Mathematical modelling employed by the End TB strategy predicts that improved implementation of current diagnostics could lead to a 10% annual decline in global incidence by 2025 (215). The ultimate

goal, to end the TB epidemic by 2035, requires accelerated reduction in decline rate that cannot be achieved with the tools presently endorsed by the WHO (215). New technologies that address the gaps in active TB diagnosis are needed both for the market in resource-limited countries and in resource-rich settings as challenges exist in each setting. The priorities including development of a higher sensitivity rapid test to replace smear-microscopy for detection of MTBC species and a rapid drug susceptibility test for all first-line drugs that can be used at peripheral levels of the healthcare system (193,216,217). DST must include capacity for testing antibiotics that are likely to be part of new drug regimens in the near future (218).

3.1.2 Short-comings of current diagnostic methods

The diagnostic ideal is for detection, species identification, full susceptibility testing and epidemiology of tuberculosis disease in a timely manner with high levels of sensitivity and specificity (128). Diagnostic algorithms (discussed in greater detail in section 1.5) currently rely heavily on smear microscopy and culture in all settings, leaving challenges with detection sensitivity and rapidity. Settings with greater resources may have access to automated culture systems and molecular testing that can reduce the rapidity of detection and DST to a matter of hours to weeks, depending on the system being used and the drug in question. In a few of the most resource-rich countries, whole genome sequencing is being used increasingly to support drug susceptibility testing and epidemiological requirements (192,219–226).

Species identification is often overlooked by commonly used diagnostic approaches when epidemiological analysis is not immediately undertaken. This is important because species identification is needed to inform successful patient drug regimens. Within the *M. tuberculosis* complex, the *M. bovis* and *M. canettii* lineages are intrinsically resistant to pyrazinamide (68,168) whilst mycobacteria outside the complex often do not respond to conventional TB treatment (68,168). Species identification is also important epidemiologically, for example highlighting the likelihood of transmission from animals and the need for appropriate public health interventions (227). Strain evaluation beyond the species level using traditional epidemiological typing methods

are onerous and lack the resolution needed to reconstruct transmission chains, with the exception of WGS of cultured samples (192,228).

Current diagnosis methods, often including WGS, fail to recognise mixed infection because conventional diagnostic approaches sometimes rely on isolation and follow-up of single colonies. WGS of liquid cultures, such as the MGIT, can lead to identification of mixed infections though growth of the organisms may distort the ratio of lineages identified (229). Failure to identify mixed infection can lead to inadequate treatment of resistant strains, resulting in amplification of drug resistance and relapse (165,230–233). It also distorts the understanding of the epidemiology of disease in a region.

Drug susceptibility testing is essential to inform the most appropriate treatment choices, whether that's the standard first-line treatment regimen or identifying the appropriate combination of second-line drugs. As the GeneXpert can identify rifampicin resistance, it can determine whether a patient should be treated with a first or second-line drug regimen. It cannot determine the full drug resistance profile for an organism so rifampicin resistant organisms should undergo further DST to determine an appropriate drug combination. This testing may be conducted using rapid line probe assays, slower phenotypic susceptibility testing of cultured organisms or through whole genome sequencing. That second-line drug testing is usually only conducted after first-line drug resistance is identified meaning causing long delays before the correct drug regimen can be commenced (234).

Even in resource-rich countries, it typically takes one two months to fulfil the entire diagnostic paradigm because of the continued dependency on culturing tuberculosis which is timely and onerous (235). Combinations of the available approaches can diagnose most patients, given enough time and resources, however there are numerous difficulties and many patients are not adequately diagnosed. Moreover, in resource-limited settings, many of these technologies are not suitable owing to unavailability of electricity, skilled-workers, laboratory facilities or finances. This leaves gaps in the diagnostic paradigm and these settings are more likely to rely on older technologies that take longer and perform with poorer sensitivities and specificities. New

diagnostic methods are needed to address the current short-comings in species identification, detection and characterisation of mixed infections, high-resolution epidemiological typing, timely susceptibility testing for all drugs and accessibility to improved diagnostics in all settings.

3.1.3 Genomic sequencing in diagnosis

Whole genome sequencing is beginning to be used for epidemiological analysis of MTBC species, but not for the full diagnostic paradigm due to the reliance on traditionally slow culture of the organism prior to DNA extraction and sequencing. The ideal sequence-based diagnostic test for TB (193,216,236) would:

- work on sputum, with minimal safe sample-preparation steps.
- detect *M. tuberculosis* and provide a drug resistance profile, with high sensitivity and specificity relative to culture.
- provide results within 20 minutes.
- cost just a few dollars per sample.

Before commencing this study in September 2013, a number of steps had been made, paving the way for further investigation of metagenomic DNA sequencing in TB diagnosis:

- Whole genome sequencing had been adopted into some TB reference laboratories for sensitivity testing and epidemiology (192,225,237).
- Steps to reduce culture time prior to WGS had been made. In a study by Köser *et al* (238), sufficient culture was obtained by growth in the MGIT system for 3 days to extract DNA directly from the MGIT then sequence it. From the WGS, a mixture of two *M. tuberculosis* strains was identified and resistance was predicted to nine drugs, making the sample extremely drug resistant. The genetic prediction of the DST profile correlated with phenotypic results and WGS was able to identify and characterise the mixed infection that MIRU-VNTR had not. The time to culture positivity, 3 days, was shorter than for most samples and even combined with Illumina sequencing, the time from sample receipt to results was significantly improved relative to typical methods.

- Metagenomic analysis of mummified lung tissue had previously identified and characterised *M. tuberculosis* after DNA extraction directly from the tissue and sequencing using an entire run of the Illumina MiSeq (205). That the DNA was obtained from mummified tissue was expected to pose different challenges than from contemporary samples, given that for the ancient samples the resulting DNA read lengths would be shorter and human DNA would have been degraded.
- A metagenomics approach had been employed to the investigation of other pathogens including
 - Many viruses (239)
 - The aetiological agent causing diarrheal disease, *Campylobacter jejuni*. It was detected by unbiased metagenomic sequencing of DNA extracted directly from a patient's faecal sample indicating the potential to diagnose any pathogen using this approach (240).
 - The *Escherichia coli* O104:H4 outbreak strain of 2011. Retrospective analysis of the faecal samples was used to reconstruct and characterise the outbreak strain (241).

3.1.4 Rationale for this study

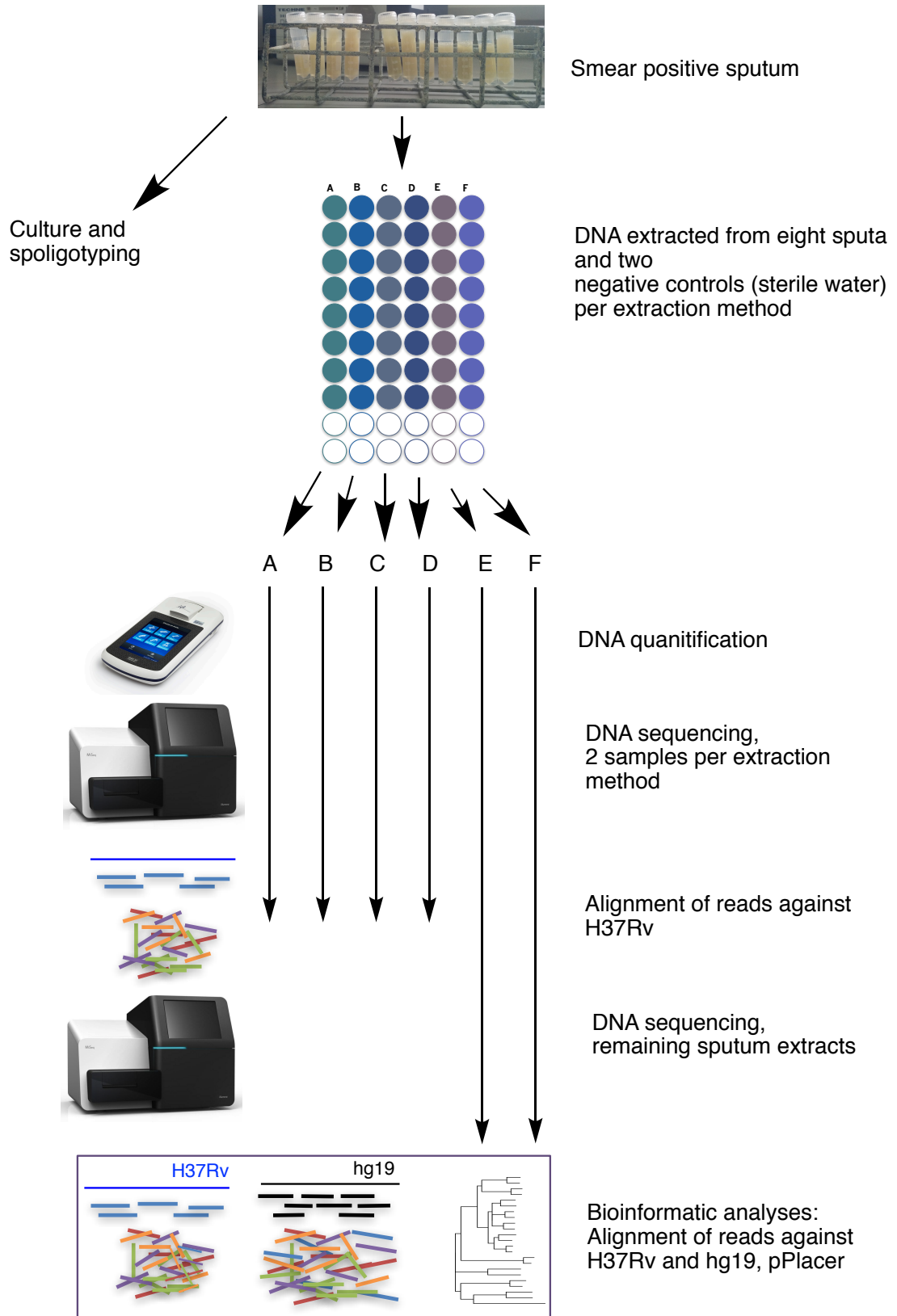
At the time of starting this study in September 2013, shotgun metagenomics showed theoretical potential as a diagnostic method for tuberculosis but this had not been demonstrated experimentally (128). It had been speculated that this approach might fulfil in a single laboratory step with fully automated analysis all steps in the diagnostic paradigm: detection, identification, DST and epidemiology. Simultaneously, the cost of DNA sequencing had been falling and the speed of sequencing had been rising taking this potential approach closer to being financially feasible for use in clinical practice (242). As such, this study aimed to experimentally explore the feasibility for shotgun metagenomics to detect and characterise MTBC strains. DNA extraction methods were employed that might progress both towards fulfilling the ideals of a new TB diagnostic and providing a universal diagnostic method for any pathogen in a clinical sample. Smear-positive sputum samples from patients from The Gambia in West Africa were used owing to their ready-

availability and the DNA extractions were carried out at the MRC Unit, The Gambia.

3.2 Methods

Details of general materials and methods can be found in chapter two. The laboratory and bioinformatics workflow for this study is shown in Figure 3.1.

Figure 3.1: Methods undertaken to analyse samples in chapter three



3.2.1 Sample collection, selection and processing

For this project, 48 sputum were selected for metagenomic analysis from specimens collected between February 2012 and May 2014 under the auspices of the Enhanced Case Finding project at the MRC Unit, The Gambia and processed at their TB diagnostic laboratory(197). Prior to selection for metagenomic investigation, an aliquot of each sample had been subjected to microbiological analyses by MRC diagnostic staff: Smear microscopy, culture on LJ slopes and in the BACTEC MGIT 960 Mycobacterial Detection System, and spoligotyping of cultured isolates. Spoligotype patterns were converted to SITVIT clades using the knowledge-based Bayesian network (KBBN) method via the TB Lineage online tool (243). The samples used were all smear-positive and culture-positive with unknown HIV status of the donor patients. The selected samples were split into six groups, each for use with a distinct sample processing method. For each method, eight sputum samples and two negative controls consisting of sterile water were used. The sample names were prefixed with A-F according to the extraction method used and numbered 1-8 to denote their derivation from unique sputum expectorates. Samples numbered 9 and 10 were negative controls.

DNA extraction was performed in the TB laboratory at the MRC Unit, The Gambia. Sputum aliquots were processed by one of six methods (Table 3.1). These steps addressed the anticipated obstacles to handling sputa and detecting TB, principally sputum viscosity, human DNA content and TB cell lysis.

Table 3.1: Methods of DNA preparation used to detect TB from sputum by metagenomics

Six methods of DNA preparation, A-F, were applied to sputum in an attempt to detect M. tuberculosis by downstream metagenomic analysis. Each extraction method comprised of a method sputum decongestions, human DNA depletion then extraction of DNA from bacteria present.

Method	Decongestion method	Human DNA depletion method	DNA extraction method
A	Decontamination		Mechanical lysis and kit
B	Decontamination		CTAB
C	NALC	None	CTAB
D	NALC	None	Mechanical lysis and kit
E	NALC	Differential lysis and DNA degradation	CTAB
F	NALC	Differential lysis and DNA degradation	Nucleospin tissue kit

3.2.2 Library preparation and sequencing

DNA extracts were sent to the Warwick Medical School where all further laboratory and bioinformatics analyses were performed. The concentration of DNA present in each extract was determined and negative control samples were considered uncontaminated when there was no detectable DNA. From each processing method, two samples with adequate DNA were initially converted to sequencing libraries using the Nextera XT protocol and sequenced on one run of the Illumina MiSeq. The remaining samples from the two most-promising methods were later sequenced in the same way.

3.2.3 Analysis of human and mycobacterial sequences

The resulting sequence reads were assessed by alignment to the *M. tuberculosis* reference genome, H37Rv (GenBank accession numbers AL123456) using Bowtie 2 version 2.1.0 (204) with a relaxed protocol that exploited the option --very-sensitive-local. However, the ‘relaxed’ alignment

parameters were found to be too non-specific, matching reads from other bacterial species in the sample to the reference genome. An alternative 'stringent' alignment protocol was thus developed that avoided this. The stringent protocol allowed only limited mismatches (3 per 100 base pairs) and soft clipping of poor quality ends, by exploiting the options `--ignore-quals --mp 10,10 --score-min L,0,0.725 --local --ma 1`.

The number of reads aligning to hg19, the human reference genome, and total number of reads were determined using Bowtie 2 under the `--fast-local` settings and Qualimap.

3.2.4 Species and lineage assignment using low-coverage SNPs

Epidemiological analysis of the strains was achieved by phylogenetic placement. A phylogenetic tree was built from representative genomes from each MTBC species and major lineage, using the lineage designations reported by PolyTB (244). These genome sequences were obtained from entries to the short-read archive, ERP000276 and ERP000124 (245), and mapped against *M. tuberculosis* H37Rv with Bowtie 2 under its default settings (204). The SNPs were called using VarScan2 (211) excluding those that fell within previously published repetitive regions (191). A phylogenetic tree was constructed from these SNPs with RAxML version 7 (246) using its default parameters and the GTR-gamma model.

Reads from the samples in this study were mapped against the reference strain, *M. tuberculosis* H37Rv, again using the default settings in Bowtie 2 (204). SNPs were identified without quality filtering from the majority base called at each mutational position and when no base was present at a position, a gap was used. At this point, the pPlacer program (209) was used to assign each SNP sequence to a species and lineage of the phylogenetic tree.

3.3 Results

3.3.1 DNA quantification and validation of controls

Most extracts contained enough DNA for Nextera XT library preparation, with the exception of those obtained by method B (Table 3.2) For the negative control samples to be considered uncontaminated, they must have contained no detectable DNA when analysed with the Qubit HS kit, which is sensitive to 10 pg/ μ L DNA. DNA was detected in one negative control sample of method C and both negative control samples of method D.

Table 3.2: Evaluation of methods detect TB from sputum by metagenomics

The extracts obtained by the six preparation methods used, A-F, were initially evaluated by Qubit fluorimetry to determine whether adequate DNA, 0.2 ng/ μ L, was present in each DNA extract and whether any DNA could be detected at all in negative controls. The outcomes of the initial evaluations are summarised for each extraction method and were used to inform analysis going forwards.

Method	Adequate DNA in each sample	Detectable DNA in negative controls
A	8/8	0/2
B	2/8	0/2
C	8/8	1/2
D	8/8	2/2
E	8/8	0/2
F	8/8	0/2

3.3.2 Selection of best extraction method

Two samples from each extraction method were sequenced, this included a negative control sample from methods C and D as these contained detectable DNA and the remaining DNA extracts sequenced had been obtained from sputum. The sequence reads were aligned under the relaxed settings to the *Mycobacterium tuberculosis* H37Rv reference; Method A resulted in very few

sequences that aligned to the H37Rv genome. The negative control samples from methods C and D yielded reads aligning to H37Rv whilst the DNA extracted from sputa by these methods did not yield a significantly greater proportion of reads aligning to H37Rv. As such it was not considered beneficial to investigate these extraction methods further. The largest proportion of reads aligning to H37Rv were from samples processed by method B, however it had already been established that only two of eight samples contained enough DNA for sequencing and the number of aligned reads was not significantly greater than methods E and F. Methods E and F gave samples with the largest proportions of reads derived from TB without being compromised by sample contamination and consistently provided enough DNA for sequencing.

3.3.2 Detection of MTBC in sputum samples

Metagenomic sequences were obtained from all smear- and culture-positive sputum samples processed by methods E and F. The number of sequence reads in each sputum-derived metagenome ranged from 989,442 to 2,818,238.

Sequence reads from all DNA extracts resulting from methods E and F were aligned under the 'relaxed' parameters against the H37Rv reference genome (Table 3.4). These parameters are typically used for alignment of reads derived from whole genome sequencing of cultured isolates but were found not to be suitable for reads generated from metagenomic DNA. Two values were calculated to assess the quality of read alignment against H37Rv. Average genomic coverage describes the average number each nucleotide of the whole H37Rv reference could be covered by a read and assumes random distribution across the genome. Average read depth on the other hand is a calculation of the average number each nucleotide is actually covered by a read. Therefore, if the average coverage observed was less than one against H37Rv the expectation for a uniformly sequenced genome would be that the read depth would also be one, although this could also approach two if pair end sequencing was used and the DNA fragments were short. Comparing the two values helps with evaluating how evenly the reads cover the genome. When coverage plots were generated using Qualimap, it became apparent

that the average genome coverage observed was less than one for all but one sample, whereas the read depth was as high as 32.8x. This suggested that the matched sequence reads were unevenly distributed across the reference genome. The high average read depth and clustering of reads across the genome under 'relaxed' alignment conditions arose where conserved sequences in both MTBC and other species matched the H37Rv reference.

In order to identify the reads truly derived from MTBC organisms, alignment parameters must match only reads derived from MTBC to the H37Rv reference genome. In previous metagenomic analyses of mummified material containing TB (205), 'stringent' alignment parameters had been developed that only matched reads that were very closely related to the reference genome. Similar parameters were used that allowed reads to align with soft-clipping ends of poor quality reads and only three mismatched positions per 100 bases. This resulted in average genome coverage ranging from 0.002 to 0.693 (Table 3.4) and read depths that never exceeded 2x, in line with expectation for when only MTBC species are aligning.

Table 3.3: Detection of *M. tuberculosis* complex sequences in sputum samples

Reads from each sample were aligned to the H37Rv M. tuberculosis reference genome under relaxed and stringent mapping conditions. For each sample, the number of reads aligning, genomic coverage and average read depth across the genome were determined and it could be observed that these metrics were altered different when comparing those obtained under relaxed and stringent mapping conditions. An average read depth between one and two was expected, given the average genomic coverage around one, and accepted as an indicator or specific alignment of the metagenomic reads to the reference.

Sample	Under relaxed mapping conditions			Under stringent mapping conditions		
	Bases aligning	Coverage	Average read depth	Bases aligning	Coverage	Average read depth
E1	34750	0.008	1.3	15698	0.004	1.3
E2	3614561	0.819	2.9	1058815	0.240	1.7
E3	3864651	0.876	2.0	1824712	0.414	1.5
E4	44004	0.010	1.9	11515	0.003	1.4
E5	3640048	0.825	13.6	174053	0.039	1.4
E6	63818	0.014	1.9	21342	0.005	1.2
E7	123748	0.028	3.2	21177	0.005	1.3
E8	2916936	0.661	32.8	26385	0.006	1.6
F1	410228	0.093	2.2	141906	0.032	1.3
F2	5685901	1.289	2.3	3057187	0.693	1.9
F3	99643	0.023	1.3	54413	0.012	1.2
F4	40019	0.009	1.9	10840	0.002	1.3
F5	732623	0.166	2.5	238451	0.054	1.3
F6	94023	0.021	2.3	34704	0.008	1.7
F7	1366309	0.310	11.4	50873	0.012	1.5
F8	1725816	0.391	7.7	109514	0.025	1.3

3.3.3 Phylogenetic placement

Strain typing is an important yet often neglected part of the full diagnostic paradigm (128). The methods typically used to characterise TB strains, such as spoligotyping and MIRU-VNTR, rely on analysis of specific repeat regions that cannot be reconstructed by short read shotgun sequencing. Likewise, lineage assessment from whole genome data relies on high-confidence SNPs and RDs from across the full genome that is not possible with the partial genomic data available here. pPlacer (209), offers an alternative way of characterising TB using the few short reads available from metagenomic sequencing with maximum likelihood and posterior probability phylogenetic placement into a reference phylogenetic tree. Of the 16 samples analysed, 13 could be phylogenetically placed at the strain, lineage and sub-lineage level with the posterior probability >0.97 in all but sample E1 which had a posterior probability of 0.5 with a very low coverage of the genome. (Tables 3.5 and 3.6). The placement for all samples corresponded with the spoligotypes of the cultured isolates, except for those in sample E2. Interestingly, sample E2, where the sample was designed as the LAM clade by pPlacer but the T clade by spoligotyping, was from the same patient as sample F1 where both methods were concordant in their LAM clade result. The majority of strains were found to be part of the globally dominant *M. tuberculosis* species and 4 were found to be *M. africanum* clade 2, concurrent with knowledge of the TB population structure through West Africa.

3.3.4 Assessment of human DNA quantity

The amount of human DNA as a proportion of total DNA in samples was anticipated to be a major limitation to the number and proportion of reads for detection and characterisation of MTBC. The proportion of reads aligning to hg19 the human reference genome, relative to the total reads per sample was calculated. The proportion of reads in each metagenome mapping against the human genome ranged from 20% to 99% (Table 3.7), with the remaining reads associated with other organisms.

Table 3.6: Proportions of reads aligning to the human reference genome, hg19, in each sample

Seq. ref. no.	% reads aligning to hg19
E1	98.16
E2	23.92
E3	19.28
E4	95.89
E5	4.03
E6	95.87
E7	49.65
E8	18.71
F1	73.71
F2	78.46
F3	99.3
F4	97.22
F5	74.17
F6	97.47
F7	50.59
F8	20.29

3.4 Discussion

This study provides the first proof-of-principle that TB can be detected and characterised by metagenomic analysis of DNA extracted directly from sputum. Furthermore, this has been achieved with simple, low-cost, non-commercial approaches to DNA extraction. The most successful method included sputum viscosity reduction and a first attempt at human DNA depletion without target-specific capture, amplification or microbiome depletion.

Some of the hypothesised advantages of metagenomic TB diagnosis have been confirmed by this study (235). Here, the best approaches to sample processing facilitated detection of TB in all sputum extracts and epidemiological results in 13 of 16 of these. Coverage of the genomes was low, especially compared to WGS of cultured isolates, limiting the extent to which they could be characterised. In this study, the samples could not be characterised in terms of drug resistance but the species could be identified in most cases and epidemiological typing by pPlacer matched that by spoligotyping of the cultured samples in all but one case. The epidemiological resolution provided by spoligotyping was greater than by pPlacer but only when the spoligotype patterns were analysed by SITVIT. Standard interpretations of spoligotype patterns provided by MRCG had lower epidemiological resolution. Should both spoligotyping and pPlacer been accurate, the discordance between the results for sample E2 with those from sample F1 may have been caused by mixed infection. This highlights that the potential for mixed infection and heterogeneous strains in sputum cannot be appreciated either via this metagenomic approach with low coverage of the MTBC species' genome or by methods using a single colony culture from sputum. It was not possible to determine which samples contained mixed infections from this level of genomic coverage. In a clinical laboratory, employment of this single-laboratory method would replace many onerous approaches for detection and epidemiological typing of TB (235) circumventing the bottleneck of culture in timely epidemiological analysis.

The samples sequenced from method A, decontamination and extraction by mechanical lysis and the Qiagen kit, yielded very few reads aligning to H37Rv, even with the relaxed alignment approach. The abundance of the reads must have been obtained from either human DNA or the DNA of other organisms in the microbiome. This may be owing to the failure of the mechanical lysis and simple Qiagen kit extraction to lyse the mycobacterial cells whilst it would have been adequate to lyse other bacteria and human cells.

Method B, decontamination followed by a CTAB extraction, resulted in a very low concentration of DNA in the extract. It is expected that this is due to depletion of much of the microbiome that would have yielded DNA during the decontamination step and perhaps poor extraction efficiency by the CTAB method. It highlights a potential problem with a direct sequencing approach that depletes non-target DNA, that after this kind of extraction approach, the concentration of the resulting DNA may be too low for sequencing. An additional amplification step might be required.

The methods, C and D, both using decongestion then either extraction by the CTAB or mechanical lysis and Qiagen kit method, resulted in DNA in the negative control samples. Two samples obtained by each of these methods were sequenced to determine if a large and significant proportion of the reads aligned to H37Rv, warranting repeating these extraction methods. This showed that there was very little difference in the quantity of DNA in the samples obtained from sputum and the contaminated negative controls. There was also not a huge proportion of reads aligning to H37Rv from the extracts obtained from sputum. It was decided to not investigate these samples further in this study and that repeating the extractions was unlikely to aid the identification of a better extraction method for the goals of this study. The extractions in these methods were likely contaminated by not using sterile tubes for the final elution of DNA, due to a laboratory evacuation whilst the samples were being processed. After evacuation, the only elution tubes available were not sterilised and could not be sterilised in the necessary timeframe. This highlights the issue with equipment contamination that is

pervasive throughout many molecular methods and is particularly problematic in metagenomic investigations (247).

Methods E and F, using decongestion with NALC, attempted differential lysis of human cells then removal of human DNA followed by DNA extraction by either the CTAB or Nucleospin tissue kit addressed the goals in this study most successfully. It is unclear how much impact the attempts to remove human DNA made on the proportion of human to bacterial sequence reads but it was very apparent that these human reads made up an overwhelming proportion of the total reads sequenced from each sample. There was little difference in the concentration of DNA or proportion of reads aligning to H37Rv between the methods using the Nucleospin Tissue kit and CTAB extraction steps. These approaches were the best given that they consistently resulted in DNA concentrations above 0.2 ng/ μ L and provided reads aligning to the H37Rv genome.

The diagnostic metagenomics approach illustrated here shows several obvious shortcomings and areas of unclear clinical performance remain:

- Though the DNA extraction method is inexpensive, the full diagnostic method including the sequencing is costly, at around £50 per sample.
- It requires a sophisticated laboratory and bioinformatics set-up and the laboratory workflow is too laborious relative to minimum requirements specified for an improved laboratory diagnosis test (193,216). Nonetheless, this is early in the development of this diagnostic approach and feasible improvements in this method could address these issues.
- The depth-of-coverage obtained is insufficient to confidently call all relevant SNPs required for prediction of drug resistance profiles, for full single-nucleotide resolution epidemiology and inferring transmission networks. This is limited by the small proportion of the total reads that align to the H37Rv genome; the vast majority of reads in each sample align to the human reference genome.
- It is insensitive: all sputum samples used here are smear-positive, but some are negative on metagenomics, so metagenomics has a poorer sensitivity than microscopy.

- It is of uncertain specificity. Whilst average read depth and evenness of coverage suggest that stringent parameters aligned reads derived only from MTBC, specificity remains unknown without its application to TB-negative samples.
- Information about HIV-seropositivity of patients was not available so performance of this method in HIV co-infected individuals could not be assessed. Given that smear microscopy performs poorly in HIV-positive people and the elevated incidence, severity, mortality and transmission associated with co-infection, this is an important patient cohort to address (248–250).

Despite these drawbacks, a metagenomic approach appears promising for the future of TB diagnosis and more widely in microbiology. The approach developed here could be used as a one-size-fits-all approach for any sample containing any pathogen that may facilitate the re-unification of TB diagnosis with the rest of microbiology. Prior success of metagenomic diagnosis of other respiratory pathogens, especially viruses, is encouraging and suggests that identification of co-infection between two species of pathogen may be possible (251–255). A key challenge is that the number of reads derived from the TB bacillus were overwhelmed by human reads or those derived from other microbes. Depletion of human reads from samples before sequencing could potentially increase the number of reads from *M. tuberculosis* obtained from each sample—this challenge is addressed in the subsequent chapter.

4. Metagenomic Diagnosis with Depletion of Human DNA

4.1 Introduction

In the previous chapter (chapter three), it was established that shotgun metagenomics can detect and epidemiologically type tuberculosis directly from sputum. The diagnostic information obtained using this method was limited by the absolute number of sequence reads associated with the pathogen that could be obtained from the sample. In most samples, human-derived sequence reads overwhelmed those from the MTBC. Consequently, the coverage of the tuberculosis-associated genomes obtained was too low to reliably call SNPs for antibiotic resistance prediction or for single nucleotide resolution epidemiology as can be obtained via WGS for transmission analysis.

Previous metagenomic and PCR-based studies have explored approaches to minimise the challenge presented by high ratios of target to non-target DNA. The 'target' might be a specific organism or a group of organisms, like all bacteria or all microorganisms, whilst the 'non-target' might be host or human cells or microorganisms that are not of interest. There are three main approaches to improving the detection of such target organisms within metagenomics samples:

- Increasing the depth of sequencing such that more reads are associated with every organism represented in the metagenomic DNA extract.
- Enrichment of the target organism(s) or their DNA, by culture; target-specific capture; or random or target-specific amplification.
- Increasing the relative amount of target DNA by specific depletion of non-target cells and/or DNA during DNA extraction.

Employing the first approach would mean increased cost per sample, particularly since the depth of sequencing may need to increase significantly. This approach has been taken to identify tuberculosis from historical samples,

purely for research purposes (205,206) but it would be financially infeasible in a clinical diagnostic laboratory, especially for TB as it is most prevalent in resource-poor settings. It does not fulfil the aim of improving an extraction method to facilitate metagenomic diagnosis.

The second approach risks compromising the open-endedness of diagnostic metagenomics by targeting only expected organism(s), prohibiting metagenomics as a universal diagnostic method. When this study commenced, two publications had characterised tuberculosis in this way:

1. Bouwman *et al* (256) used hybridization capture of 260 specific polymorphic regions of the *M. tuberculosis* genome to enrich target ancient DNA extracted from the rib of a 19th century skeleton. From this approach, 664,500 of the total 726,848 Sequencing by Oligonucleotide Ligation and Detection (SOLiD) reads (91.4%) aligned to the *M. tuberculosis* reference genome in the targeted regions which enabled characterisation of 218 of 247 SNPs, eight of nine RDs, TbD1, the IS1081 and IS6110 insertion sequences, and the *mtp40* locus. The strain was classified within lineage two and no attempt to analyse drug susceptibility was reported. The main disadvantage of this approach was the extent of missing genomic data, both from the regions that were not captured by the baits and from the uneven coverage across these sequenced regions. It is also unclear how this would perform on species other than *M. tuberculosis*, if cross contamination of amplicons between samples might be a concern and how expensive this approach might be.
2. In another study by Köser *et al* (238), one sample of decontaminated smear-negative sputum was cultured in the MGIT automated system. Once it flagged positive, the DNA was extracted and sequenced. Though 72% of the sequence reads failed to align to the *M. tuberculosis* reference genome, H37Rv, the remaining reads were sufficient to provide 52-fold coverage of the genome. This facilitated DST, lineage classification and identification of mixed infection characterising the strain as XDR-TB and indicating a more appropriate treatment regimen. Relative to the ideal, this culture-first approach was limited by the additional time and cost associated with culture. It also employed a large number of sequencing reads, totalling 1.98Gb, adding to the cost of the method, though many

of the reads sequenced were not the target. When the same MGIT culture was inoculated onto an LJ slope and further cultured, improved results were obtained with only 1% of the reads failing to align to H37Rv, however this added yet more time and cost to the diagnostic process and jeopardized the ability to detect mixed infection.

Approaches aiming to enrich target DNA in an extract have failed to provide the results that are required for a clinical metagenomic diagnostic method.

The third approach, to deplete non-target cells or DNA, seemed most likely to fulfil the goals of a universal diagnostic method using metagenomics. Though this approach has not been employed in studies that have identified tuberculosis disease, it has been investigated a number of times in other studies where removal of human DNA has been necessary for detection of bacteria (Table 4.1). These studies have sought to reduce human DNA, thereby enriching the bacterial proportion of DNA obtained, from various clinical sample types, for various purposes, with various target organisms and reporting the efficiency of the human DNA depletion in various incomparable ways. It is not possible to determine which approaches have been most successful, nor which might have the greatest impact on the composition of DNA extracted from sputum. The methods all involve some loss of bacterial DNA with the human DNA which may impede detection of bacteria, would restrict biodiversity analyses and potentially compromise the universal potential for diagnostic metagenomics. Commercial methods, including the NEBNext DNA library preparation kits (New England BioLabs, Hitchin, UK (257)) and MolYsis kits for human DNA removal (Molzylm GmbH & Co. KG, Bremen, Germany), are available and have showed significant reductions in human DNA content with minimal impact on microbial DNA. Whilst effective, these commercial methods are prohibitively expensive for routine diagnostic purposes. For the purposes of metagenomic diagnosis of infections, including tuberculosis, from sputum it was unclear which approach might be most effective.

Table 4.1: Prior studies attempting to deplete human DNA from clinical samples or laboratory-derived proxies

Method	Sample type	Outcome	Assessment method	Reference
Selective lysis using buffer or water, digestion of differentially released human DNA with benzonase	Dental plaque	Statistically significant depletion of human DNA relative to no treatment control, little impact on bacterial DNA	qPCR	(258)
Differential centrifugation by pulsing to 5000 x g to remove human cells	Dental plaque	Statistically significant depletion of human DNA relative to no treatment control, large bacterial DNA loss	qPCR	(258)
Differential centrifugation by pulsing to 5000 x g to remove human cells after treatment with detergent (Triton X-100 or Tween 20)	Dental plaque	Statistically significant depletion of human DNA relative to no treatment control, bacterial DNA loss	qPCR	(258)
Differential centrifugation by pulsing to 5000 x g to remove human cells after treatment with trypsin	Dental plaque	Statistically significant depletion of human DNA, relative to no treatment control, significant reduction in bacterial DNA; clinically insignificant impact on human-bacterial DNA ratio	qPCR	(258)
Filtration through 6 μm Whatmann paper	Mouth wash	89% reduction in human DNA; proportionate enrichment of bacterial DNA	qPCR	(259)
Enzymatic digestion of skin samples, sequential filtration through nylon filters of 80 μm , 20 μm and 11 μm ; flow cytometry to remove mitochondria	Skin	0.02% reads assigned to human genome; 95% reads assigned to bacteria;	Shotgun metagenomics	(260)

Method	Sample type	Outcome	Assessment method	Reference
Differential centrifugation at 2000 x g for 30s	Urine	Sufficient DNA for sequencing in 23/35 samples; limited human DNA contamination (associated with sample type)	Shotgun metagenomics	(261)
Treatment with β-mercaptoethanol, osmotic lysis and treatment with DNase, prior to bacterial DNA extraction	Sputum	13-97% human DNA, able to characterise microbiome from remaining reads	Shotgun metagenomics	(203)
Selective lysis of human blood cells using ox bile and degradation of human DNA by micrococcal nuclease, prior to bacterial DNA extraction	Blood spiked with <i>Salmonella</i> Typhi	Significant reduction in human DNA and enrichment of <i>Salmonella</i> Typhi DNA leading to 1000-fold enhancement of <i>Salmonella</i> detection by qPCR	qPCR	(262)
Selection of human and bacterial DNA with non-methylated GATC sequences by binding with DpnII restriction endonuclease under conditions where DNA is not cut and can be removed using magnetic beads	Human DNA spiked with <i>E. coli</i>	Enriched human DNA with little binding to <i>E. coli</i> DNA	qPCR	(263)

4.1.1 Study rationale and objectives

In the previous chapter, the goal of characterisation and detection of MTBC genomes by metagenomics was limited by the small number of reads associated with these pathogens. Simultaneously, many of these samples contained a high proportion of reads that aligned to the human reference genome, minimising information about mycobacterial infections and the rest of the sputum microbiome. After assessing the potential approaches to enriching mycobacterial reads in the metagenomic sequences, it was considered that depleting human DNA offered the greatest hope to achieve this whilst maintaining the universal nature of metagenomics to potentially detect any other pathogen present in the sputum. The work presented in this chapter aimed to take initial steps to explore methods for depletion of human DNA and human-derived reads from metagenomic sequences, generating preliminary data for on-going method development towards a universal sputum diagnostic approach.

A number of methods aiming to reduce the proportion of human DNA in sputum extracts were investigated in this study. These methods were chosen based upon the availability of reagents and equipment in The Gambia, where this work was conducted, the cost of extracting DNA from each sample given that it should be clinically feasible and prior research found in the literature. The methods using osmotic lysis were of particular interest since the only metagenomics study with sputum had used this approach successfully. Five methods were devised with the following considerations of the method steps:

- Centrifugation to remove human cells at $300 \times g$ for 5 mins: A method in the literature had used centrifugation to $5000 \times g$ and successfully reduced the amount of human DNA in the final extract but had also removed much bacterial DNA (258); another study had centrifuged urine at $2000 \times g$ resulting in little human DNA contamination but this was considered more likely to be due to the sample type, urine, where human cells are less likely to be found (261). Being conscious to not pellet and discard bacteria, like in this first study, minimal centrifugation was used in an attempt to pellet only human cells. The

centrifugation speed and duration was chosen to reflect conditions typically used by cell biologists (personal communication).

- Osmotic lysis of human cells in water then degradation of extracellular DNA using DNase: This method was repeated from chapter three, using a higher concentration of DNase, such that the method could be evaluated relative to no-treatment controls and because it was felt the insufficient DNase may have limited the human DNA degradation in the previous work.
- Osmotic lysis of human cells in Tris-HCl then depletion of extracellular DNA with benzonase: Looking at ways to improve the osmotic lysis method, an alternative lysis solution and DNA degradation enzyme was investigated based on the statistically significant depletion of human DNA from dental plaque metagenomes using these methods (258).
- Osmotic lysis of human cells in water then removal of extracellular DNA using AMPure XP: Again, looking at ways to improve osmotic lysis, use of AMPure XP was investigated for its alternative mechanism of removing extracellular DNA.
- Filtration was undertaken given the 89% reduction of human DNA reported from mouthwash (259) when 6 μm Whatman paper was used. Since the sputum samples were still quite viscous after decongestion, 80 μm and 20 μm Whatman paper was used to remove large structures in the sample and avoid clogging the final filter which was intended to allow only small structures, ideally only bacteria, to pass. The final filtration was through 10 μm Whatman paper since this would be large enough to inhibit human cells from entering the filtrate and 6 μm Whatman paper could not be sourced.

Steps other than those aiming to deplete human DNA were standardised across the study for processing a sample from sputum to metagenomic sequence. These steps were amended from the previous study described in chapter three and included replacement of the decongestant agent sputolysin (dithiothreitol) with the aim to improve decongestion and sample handling. Sputolysin was chosen for its known capacity as a decongestant, availability

at the MRCCG and given that it was not known to cause human cell lysis. The Nucleospin Tissue Kit DNA extraction method was chosen from the previous analyses given its rapidity and practicality relative to the CTAB method without discernible difference in performance.

The study design for evaluating the methods was chosen after considering the number of samples available; heterogeneity in human DNA proportions and physical characteristics of sputum; and the aim of this study to generate preliminary data for further method development. Whilst the Gambia, where samples were collected for this study, has a higher burden of TB than the UK, it was challenging to obtaining excess sputum for research purposes. Only 40 samples were available in an appropriate time frame, even including both smear positive and negative sputa. As reported in chapter three, previous work had indicated that there was a wide range and distribution of human DNA proportions in sputum, from 4 to 99% and a median of 74% human DNA in the samples extracted after osmotic lysis and extracellular DNA degradation. There was no evidence about the human DNA proportions in untreated extracts, as methods C and D had been compromised, but it was assumed to be equally variable. With only 40 samples, these were not enough to conduct this study with adequate samples in each group to which a treatment (or no treatment, as a control) would be applied to overcome the between sample heterogeneity. The samples had to be split to provide treatment and control subsamples from each clinical sample. Ideally, there would have been large volumes of each sputum sample so each could be split between all five groups. This would have maximised the number of samples per group and minimised the heterogeneity of samples between the groups. With small volumes of excess sputum available, typically less than 1.5 mL, this was not feasible so the study was instead designed with a pairwise control for every sample. To implement this, the samples available were split between the five groups with eight samples per group and each sample was split in half to provide a treatment and control subsample for extraction, analysis and comparison. This study design, with few samples per group, had the potential to result in low statistical power to evaluate and compare treatments on the groups samples, especially should the effect size of the treatment have been small like reported in some studies found in the literature. Given that effect

size could not be indicated through the compromised results in chapter three and enormous variation, even between similar methods, presented in the literature, it remained quite unknown. It was determined that investigating all five treatment methods proposed was most congruent with the aim to generate preliminary data indicating further avenues of development of a human depletion approach that could be investigated with a larger sample size in the future. Moreover, should a method have provided clinically significant depletion to all samples, this would have been immediately clear without the need for statistical evaluation. The sputum expectorates had also been highly variable in terms of viscosity, homogeneity and colour. In attempt to circumvent some of the issues associated with this sample heterogeneity, distribution of samples between groups was considered, attempting to create groups of samples with roughly equal characteristics.

The best way to evaluate the effectiveness of human DNA depletion was considered carefully. Initially use of a quantitative PCR approach using primers targeted at a universal molecular barcode (e.g. the 16S rRNA gene) was considered. However, the use of universal 16S primers to assess the quantity of all bacteria present was likely to also lead to amplification of laboratory contaminants due to the high sensitivity of 16S qPCR. Such an approach would also not assess the impact of depletion on sequence data, only on DNA proportions. Sequencing followed by mapping of reads to the human reference genome was therefore used as a quantitative method to replicate the clinical diagnostic approach as closely as possible. Technical replicates from sputum samples and sequencing libraries were not used given that only a small volume of sputum was available from each patient and because this genomic sequencing approach results in such a large number of reads per sample.

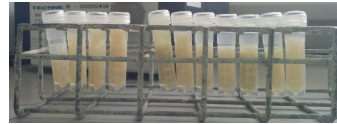
The number of reads aligning to the human reference genome, as a percentage of all sequenced reads, was used as the metric for evaluating each treatment and control subsample. Under the advice of Dr. Perter Kimani, the comparison of human read percentage between the treatment and control sample subgroups was undertaken by a random-effects meta-analysis (rma) that was used to fit a restricted maximum-likelihood mixed effects model

(REML) to the data. This statistical approach was chosen to represent the study design; it evaluated a small number of samples as part of a much larger population of data by transforming the normally distributed continuous data of the dependent variable of interest so that random effects caused by parameters other than the one of interest, i.e. heterogeneity between samples, had no effect. Using a linear regression, proportions of human DNA were then predicted for an entire population of sputum, accounting for its full range of variability. The difference between treatment and control subsample groups was used to evaluate and compare the methods.

4.2 Methods

Detailed materials and methods can be found in chapter two. The laboratory and bioinformatics workflow for this study is shown in Figure 4.1.

Figure 4.1: Methods undertaken to analyse samples in chapter four



Sputum from suspected TB patients, smear positive and negative



Eight sputum samples plus two negative controls per treatment method, split for one half to be treated and the other to be the paired control



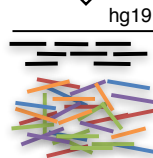
DNA quantification



DNA fragment length analysis



DNA sequencing



Alignment against hg19



Statistical analysis

4.2.1 Sample collection, selection and processing

In total, 40 sputum samples were collected from different patients enrolled in the ECF or GAMSTEP projects at the MRC Unit, The Gambia between March and July 2015. Upon collection samples were subjected to smear microscopy that defined each as 'No AFB seen' or as positive with ZN grade +, ++ or +++. Samples were visually characterised in terms of colour and viscosity.

Depletion steps and DNA extractions were performed at the MRCG. Samples were divided between five groups, each for treatment with a different human DNA depletion method, with approximately equal division of the smear grade, viscosity and colour characteristics of the samples in each group. Each group also contained two negative control samples consisting of sterile water, which were processed in the same way as the sputum samples. All samples were decongested with sputolysin to reduce the viscosity of samples.

After decongestion, each sputum sample and negative control was homogenised by gentle vortexing and divided in half to create two subsamples from each sputum aliquot. One subsample was treated with a method that attempted to reduce the proportion of human DNA, before extraction of the remaining DNA. For the other 'paired-control' subsample, DNA extraction was performed without any human-DNA depletion. Five treatments to reduce human DNA were attempted (Table 4.2). The 40 sputum samples given a letter G-K corresponding to the depletion method and were then numbered sequentially, 1-8. For each sample, the treated subsample was denoted with 't' and the paired-control with 'c'.

DNA was extracted from the solutions resulting from the treated and the paired-control subsamples with the Nucleospin Tissue Kit, used previously.

Table 4.2: Methods used to deplete human DNA from sputum samples

Five methods of human DNA depletion, G -K, were applied to sputum in an attempt to alleviate the limitation in detecting M. tuberculosis or other pathogens present in metagenomic reads derived directly from sputum.

Method	Description
G	Centrifugation to remove human cells
H	Osmotic lysis of human cells in water then degradation of extracellular DNA using DNase
I	Osmotic lysis of human cells in Tris-HCl then depletion of extracellular DNA with benzonase
J	Osmotic lysis of human cells in water then removal of extracellular DNA using AMPure XP
K	Filtration

4.2.2 Library preparation and sequencing

DNA extracts were shipped to Warwick Medical School for library preparation and sequencing. The concentration and quality of DNA in each extract was assessed using the Qubit fluorimeter and Bioanalyzer high sensitivity assay, respectively. Library preparation was performed with the TruSeq Nano kit using a modified protocol that allowed the fragmentation and size selection steps to be omitted. All 80 libraries, prepared from the treated samples and their paired-controls, were sequenced together on a single MiSeq run.

4.2.3 Bioinformatic and statistical analysis

The sequenced reads were aligned against the human reference genome, hg19, using the Bowtie 2 pre-set parameters `--sensitive-local`. Given that the TruSeq library preparation method did not fragment the DNA nor select the read lengths, read lengths were variable. The 'local' parameter for alignment against the human reference was chosen as this includes soft-clipping of mismatched ends to somewhat compensate for the variable read lengths. The total number of reads and number of human reads per subsample were quantified using Qualimap, then the number of non-human reads was deduced.

Statistical analyses were applied to determine and compare the extent of human read depletion by each of the treatment methods.

- The percentage of human reads in each method's pooled treated and paired-control subsamples was calculated and used to deduce the difference in percentage human reads.
- A random-effects meta-analysis (rma) was used to fit a restricted maximum-likelihood mixed effects model (REML) to the data. Confidence intervals and p-values were calculated to describe the statistical accuracy of the estimated reduction in human reads by a given treatment. This was implemented in the R package, *metaphor* (214), in collaboration with Dr. Peter Kimani.

4.3 Results

DNA extracts were successfully obtained from all sputum subsamples using all methods, though some contained very low concentrations of DNA. No DNA was detected in the negative control samples using Qubit fluorimeter with the high sensitivity kit. Analysis of DNA fragment length using the Bioanalyzer indicated that the DNA was highly fragmented so in much shorter lengths than expected (Figure 4.2). The fragment lengths were too short to support use of the Nextera XT library preparation kit so the TruSeq library preparation kit was used. Metagenomic sequences were successfully obtained from every subsample, despite DNA concentrations obtained from some samples being lower than recommended for conventional TruSeq library preparation. A mean average of 456,625 reads was achieved for each subsample with a range from 33,310 to 1,032,324 reads. The lengths of the reads within each subsample were variable and the read lengths were unequally distributed between subsamples and between the groups of subsamples (Figure 4.3). There was no apparent association between the length of the reads and any of the treatment methods when the read length distribution was visualised.

Figure 4.2: Graphs, as produced by the Bioanalyzer software, showing the frequency of DNA fragment lengths from two representative DNA extracts
The numbers associated with each peak, indicated by the bar across the base of the peak, show average fragment size in base pairs from integration of the peaks. Though a Bioanalyzer graph from a sample without hyper-fragmented DNA is not available since it is not typically analysed, these fragments are much shorter than typically expected from a genomic DNA preparation.

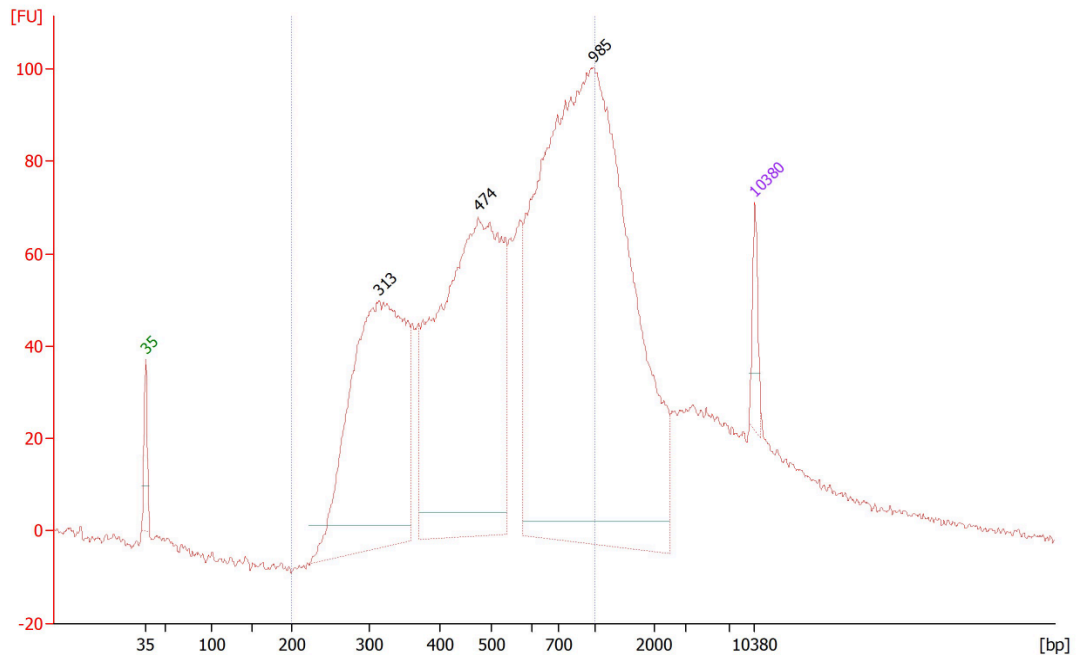
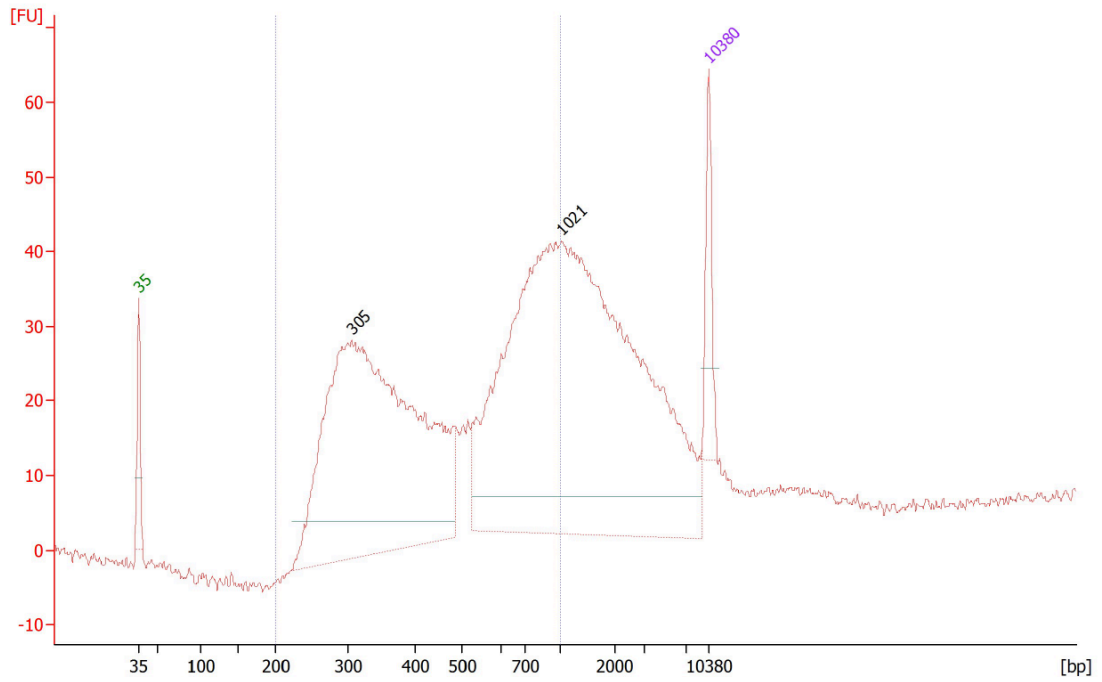
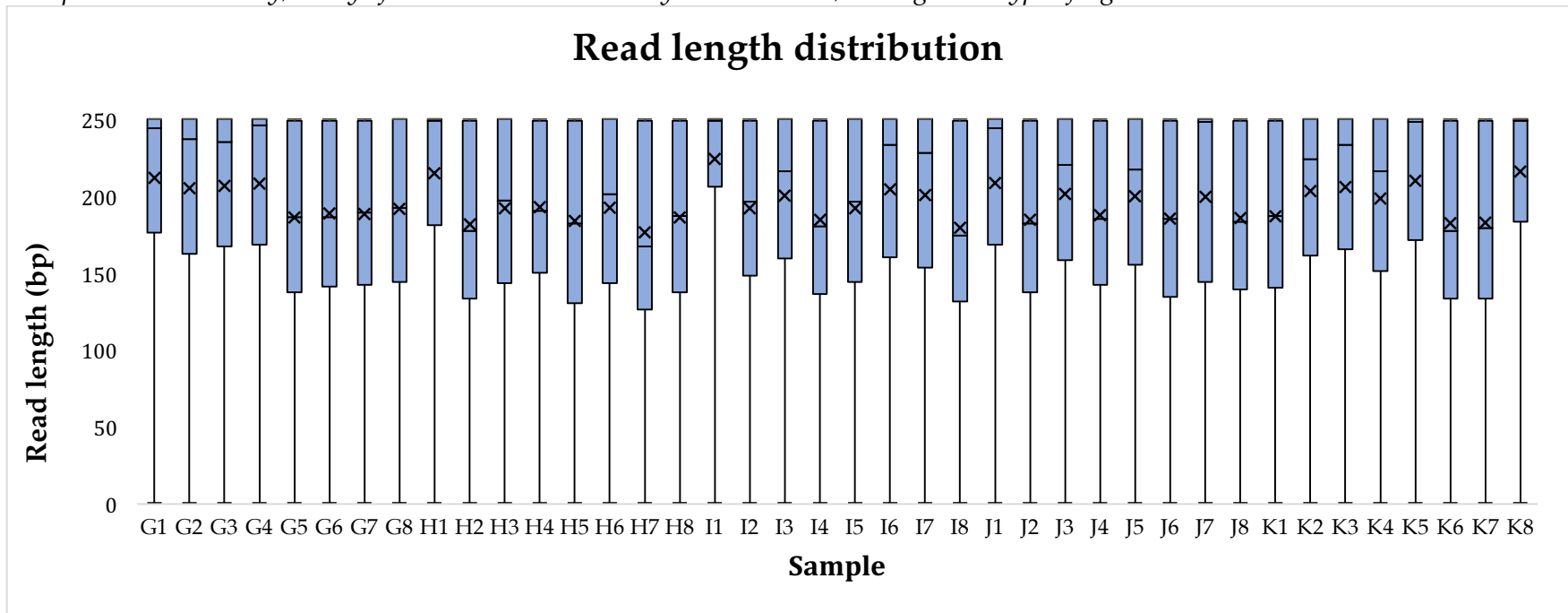


Figure 4.3: Distribution of read lengths in all control subsamples

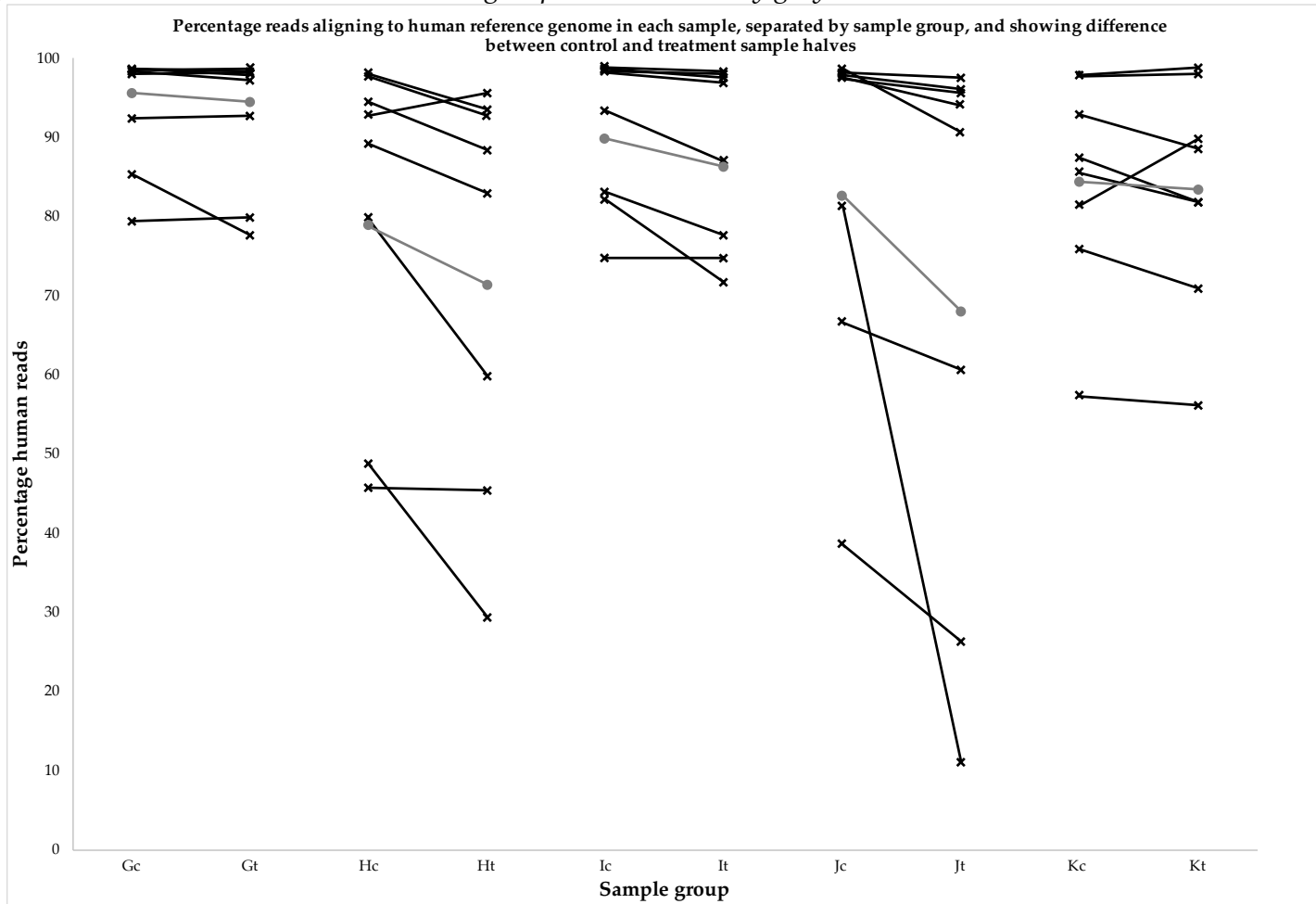
Box and whisker plots show distribution of read lengths in each of the control samples. Whiskers represent the maximum and minimum read lengths. The top and bottom of the box represent the upper and lower quartile, respectively; the intermediate bar represents the median and the mean read length is marked by a cross. The maximum read length and upper quartile are very close for most samples. In typical sequence reads generated from NextEra XT libraries and sequenced with V2 sequencing kit, read lengths would all be 250 base pairs. In this study, many of the reads were evidently much shorter, having been hyper-fragmented.



The aim of this study was to develop a method of depleting human DNA from sputum such that, after DNA extraction and sequencing, the resulting reads data would contain the lowest possible proportion reads derived from human DNA. In order to evaluate the treatment methods, each was assigned a group of eight sputum samples and two water samples that had been halved; the treatment method was applied to one subsample (treated subsample) but not the other (control subsample) then the DNA was extracted from both halves of each sample in the same way. The sequence data was analysed to identify the number of human and non-human reads in each subsample and statistical approaches were used to determine the impact that each treatment method had had on the number of human reads in the treated samples relative to their controls.

In order to evaluate a method, the impact of the treatment on the whole group of samples to which it was applied, needed to be determined. Plotting the percentage of human derived reads in each of the treatment and control samples (Figure 4.4) illustrated the significant variation in the proportion of human reads between samples, even amongst the controls which had all been prepared in using identical methods across the study. Furthermore, the initial frequency of human reads in the control subsample seemed to influence the frequency of human reads in its paired treated subsample.

Figure 4.4: Percentage reads aligning to human reference genome in each sample, grouped by treatment method
Black lines between cross markers link the percentage of human reads in the control and treated half of each sample. Grey dot markers represent the mean average human reads in a control or treatment group and are linked by grey lines.



REML estimation of adjusted proportions of human reads resulting from a given treatment method and the impact of the treatment method relative to no treatment in the paired-controls was investigated with a random-effects meta-analysis (Table 4.4). This approach accounted for the variability in the numbers of human reads between samples in a group. Comparison of the difference in human reads between treated and control groups for each method indicated that method J had the greatest impact on any sample, regardless of the characteristics of the individual samples that it was applied to. This was followed by methods H, I, K then G, in order of decreasing depletion of human reads. However, these summary effects for human read depletion by the treatment were not supported by confidence intervals or p-values for any of the treatment methods. Large p-values, much above the typical significance level of $p < 0.05$, showed that it was likely chance alone was responsible for the differences in human reads observed. The confidence intervals were large and spanned zero, indicating that the impact of the treatment on a sample may have resulted in either increase or decrease of the proportion of human reads relative to not treating it. Conclusions about the relative impact of the treatment methods on any given sputum sample, those other than used in this study, cannot be made from this analysis.

The REML analysis also provided descriptive statistics of heterogeneity that were able to separate the variation caused by the treatment method from underlying variability in the human proportion of reads between samples in each group. These indicated that the residual heterogeneity or variability that was unaccounted for by the difference in human reads between treated and control subsamples was 100% for all treatment groups.

Table 4.3: Adjusted proportions of reads aligning to the human reference genome per method group, as calculated through random effects meta-analysis

The impact of the treatments on a sputum sample is determined by the 'difference in adjusted human percentage' and given with associated 95% confidence intervals and p-values.

Method	Adjusted human proportion of total reads (%)		Difference in adjusted human % (C-T)	Difference in human 95% CI		Difference p-value
	Controls	Treated		Lower	Upper	
G	93.7	92.6	1.1	-6.9	9	0.79
H	80.8	73.5	7.3	-15.8	30.4	0.53
I	96.2	87.8	4.8	-4.6	14.2	0.32
J	84.5	71.5	13.0	-15.5	41.6	0.37
K	84.5	83.2	1.3	-12.3	14.8	0.85

4.4 Discussion

The aim of the work presented in this chapter was to generate preliminary data used to develop a method to reduce human DNA in sputum extracts. None of the methods, as presented here, were able to provide clinically significant depletion of human DNA. Depletion of human DNA in most samples was so minimal that statistical evaluation of the relative performance of each method with the REML analysis was not supported by p-values or confidence intervals. These indicated that most of the variation in frequency of human reads between the treated subsamples and controls could be attributed to the residual heterogeneity between samples rather than the treatment method itself. The few samples used here along with the small treatment effect was not enough to provide statistical power. The study was compromised by hyper fragmentation of DNA in the extracts that may have compromised the results of the study.

4.4.1 Performance of treatment methods in this study for removing human DNA

In this study, all five methods achieved little depletion of human DNA and, regardless of the treatment method used, a lot of human DNA was left in most treated subsamples. Such minimal depletion was neither enough to be clinically impactful for a diagnostic approach nor statistically significant. REML analysis indicated the average extent of depletion in each sample group, therefore the relative performance of the methods, though these results were not statistically supported.

The results of the statistical analyses provide cautionary indicators to inform further method development. Methods G and K, centrifugation to remove human cells and filtration, had very little impact on the removal of human DNA. Methods H, I and J, all of which employed osmotic lysis, had a greater, though still minimal, impact on the depletion of human reads. The two methods, H and J, using sterile water for the lysis stage may have had slightly better impact on the resulting human read frequency than method I, using Tris-HCl for lysis. This is in contrast with the results of applying similar approaches to selectively remove human DNA from dental plaque, where Tris-HCl had a markedly bigger impact than water for causing human cells lysis and removing human DNA (258). The proportion of human reads was reduced slightly more when AMPure XP beads were used in an attempt to remove extracellular DNA after differential osmotic lysis than when DNase was used.

There was a distinction between the treatment methods that had slightly greater impact (methods H, I and J) and those that had very little or no impact (methods G and K). The more impactful methods exploited cell lysis and removal of extracellular DNA whilst the less impactful methods relied on human cells being in-tact and would have incorporated extracellular DNA in the final extract for sequencing. Treatment with sputolysin may have been associated with human cells but it is not thought to lyse mycobacterial cells, given its use prior to mycobacterial culture. Similarly, though sputum samples were processed as quickly as possible after collection from patients,

this delay may also have resulted in cell lysis. It would have been useful to use a microscope to visualise the human cells at various stages in the sample processing to determine if any steps caused cell lysis, though this was not possible. Despite some success in removing the extracellular DNA by the methods using lytic approaches, many human reads remained in the subsamples that had been processed in this way. This may have been due to failure to lyse the human cells, failure to remove or degrade human DNA by use of AMPure beads, DNase or benzonase at their concentrations employed, extracellular mycobacterial DNA being degraded or a combination of these.

The filtration and centrifugation methods failed to remove human cells or captured both human and bacterial cells in equal proportion. This may have resulted from methods failing to separate the human cells from bacterial cells, or lysis of these cells before or during the processing step to remove the human cells. Viscosity of the samples, having not reduced to a water-like liquid, may have prohibited these mechanical methods. Cell lysis may have occurred through centrifugation with excess force, through shearing of the cells as they were pushed through the filter, or during processing prior to attempting human cell removal as described above. Providing similar outcomes, a prior study attempting to selectively remove human DNA from dental plaque found methods of lysing the human cells produced favorable results to centrifugation (258).

Though the aim of this study was to develop a method to reduce the proportion of human reads in the metagenomic sequences of sputum, the approaches used here were only partially successful: some reduced human reads but none could achieve this to an extent that would be useful for a diagnostic application. As presented here, these methods do not improve on the prior attempts, described in section 4.1, to remove non-target DNA to enrich target reads amongst the metagenomic sequences. The methods that attempted to lyse human cells before removing extracellular DNA were better able to fulfil this aim than those that attempted to remove whole cells only. Going forwards, improvements on the capacity to capture or degrade extracellular human DNA would require assessment of why human DNA remained using microscopy to visualise in-tact cells at various stages in the

DNA extraction protocol. Differential lysis might be pursued preferentially to the mechanical separation techniques, either by improving the proportion of human cells that are lysed or, should this already be adequate, by improving the removal or degradation of the extracellular DNA. A combination of both mechanical removal of in-tact human cells and extracellular DNA may also prove beneficial.

4.4.2 Statistical analysis approaches and limitations of this study design

This study was primarily intended to provide guidance in future method development, investigating many approaches to human DNA depletion. REML estimation could not provide statistically significant results to conclude the impact of the treatment methods on samples. The statistics of heterogeneity associated with REML analysis indicated that there was more variation between samples than between the treated and control subsample derived from the same sputum aliquot. This was reflected in the insignificant p-values above 0.05, confidence intervals that spanned 0.00 and acceptance of the null hypothesis from the REML estimation. In the visual illustration of human DNA in each sample (Figure 4.4), it could also be seen that the proportion of human reads in the group of treated subsamples broadly corresponded to the proportion in their paired control group.

The considerations of sample selection and study design, including splitting samples between groups based on physical characteristics and using pairwise controls, were insufficient to overcome the heterogeneity amongst sputa and provide statistical power to determine human DNA depletion. As such, the results presented for each treatment are only relevant to the specific samples that they were applied to in this study and are not reflective of the impact the treatment method would have on any given sputum sample. Failure to account for variation in the clinical samples represents a flaw in the design of other similar studies that conclude changes in human DNA or read proportions (203,258). In order to achieve statistical power, future studies should consider the heterogeneity amongst sputa, as presented here for the first time. These studies would either need to use large sample sizes or achieve

greater reduction in the proportion of human reads with treatment in order to overcome the confounding effects of heterogeneity and produce statistically significant results.

4.4.3 Unusual fragmentation of DNA

The DNA resulting from the extractions in this study was found to be highly fragmented, resulting in short DNA fragments and subsequently short sequence read lengths. For this reason, the Nextera XT library preparation could not be used in this study and a modified TruSeq library preparation, usually used for ancient DNA, was used instead. The short fragments of DNA are unusual after DNA extraction. Supporting this, similar DNA extraction methods used in chapter three did not have this issue and library preparation was achieved using the Nextera XT kit.

There were delays in transporting the DNA extracts from the MRC Unit, The Gambia, where the extractions were performed, to the Warwick Medical School where the fragment length analysis and DNA sequencing was undertaken. The transportation process took approximately two weeks rather than a few days as was advised, expected and had occurred with the transport of the samples in chapter three. The samples originally left the MRCG packaged with ice packs in an insulated container, were returned to MRCG a few days later and then were sent again to the UK on the next available flight, but it is unclear how the samples were stored throughout the duration of this time. The climate of The Gambia is particularly warm. Certainly, when the samples arrived at the Warwick Medical School they were packaged in the same manner as when they originally left the MRCG, though the ice packs were then at ambient temperature. DNA extracts are typically stored at -20°C in the long-term or at 4°C for short periods (e.g. overnight) whilst they can be handled at room temperature for at least a few hours. If the DNA extracts were not stored properly during this delay period, warmer temperatures may have contributed to the hyper-fragmentation.

The impact of the DNA fragmentation on the outcomes of this study cannot be determined. It is not clear whether, or how, the shorter read lengths would

affect the likelihood of a read to align the human reference genome. Longer read lengths facilitate improved assemblies of genomes and identification of the origin of reads when it is unknown and this fragmentation would have had an impact if these bioinformatic processes were necessary. This issue, likely caused by delays in transportation, highlights one of the difficulties of conducting research between multiple countries, particularly those with challenging infrastructures. Ideally, this whole study would have been conducted at the MRCG and, though this was not possible in this research project, this may be considered for future projects.

4.4.4 Conclusions

This study was not able to identify an adequate method for depletion of human DNA to suit the needs of a metagenomic diagnostic. It indicated some avenues for future improvement of the method, how to conduct a similar investigation and the potential difficulty of undertaking genomic sequencing across distinct sites, meeting the aims of the chapter. Ideally future studies of this ilk would be conducted in iterations, analysing whether cells were in tact at various stages throughout the extraction process and responding accordingly. The impact that the unexpected hyper-fragmentation of DNA that occurred in this study is unclear and would be most easily avoided by undertaking sample analysis in the same laboratory where DNA was extracted. This could be achieved using either qPCR or metagenomic sequencing with the highly portable Oxford Nanopore MinION (264).

5. Performance of drug susceptibility testing approaches in Peru

5.1 Introduction

5.1.1 TB in Peru

Statistics on the state of tuberculosis in Peru paint a picture of a high TB burden (>40 cases per 100,00) with especially high rates of drug resistance. According to WHO estimates, in 2016 there totalled 37,000 (117/ 100,000) new cases of TB in Peru of which 3500 cases were MDR/RR-TB (9.5%) (265). Peru is in the WHO's list of high MDR-TB burden countries for 2016-2020 given its incidence rate, identifying it as a location where significant progress is required to reach the END TB targets (266). Of a population of 32 million in 2015, around 9 million people live in the capital Lima and TB disproportionately afflicts people in this city. Around 60% of Peru's TB cases are concentrated in the capital and the adjacent region, Callao, including 80% of MDR-TB infections and 90% of XDR-TB (267,268).

As in much of the world, gaps exist in the diagnosis and successful treatment of TB in Peru. Of 31,079 TB cases notified in 2016, 82% of pulmonary TB cases were bacteriologically confirmed and nearly all of these, 25,144 (81%), were tested for rifampicin resistance (269). However, this lead to estimates of 2300 MDR/RR-TB cases of which only 1457 (63%) were laboratory confirmed. In 2016, 1645 patients, including those who had been diagnosed before 2016 or whose cases had not been laboratory confirmed, were started on treatment (269). From these confirmed MDR/RR-TB cases in 2016, 1104 (76%) cases were tested for second-line drug resistance, 88 cases of lab confirmed XDR-TB were notified and all were started on treatment (269). From the new TB cases registered in 2015, 87% of total TB cases were treated successfully, but the latest statistics, from 2014, show that treatment was a success for only 34% of

MDR/RR-TB cases and 65% of XDR-TB patients (269). It appears that in Peru, rates of first-line drug resistance testing and treatment success is particularly poor, despite the numbers of patients with laboratory confirmed TB being similar to the number being treated.

5.1.2 Drug susceptibility testing

Drug susceptibility testing is an essential element of the tuberculosis diagnostic paradigm and is increasingly important as resistance threatens to undermine global TB control (see section 1.5.1.2 of this thesis). The 'gold-standard' method for characterising drug resistance is via a series of MTBC cultures, one against each drug of interest. This takes weeks to months to produce results causing delays to treatment. The Gene Xpert and line probe assays have also been endorsed by the WHO, providing rapid molecular approaches to detect mutations known to cause resistance to a drug (106,155,156). These depend on the resistance-causing mutations being closely located on the genome for target by the DST approach, limiting their sensitivity (270). Their specificity is limited since they cannot differentiate determine that an amino acid substitution is non-synonymous and may not impact the resistance phenotype (271). The Gene Xpert provides prediction of RIF susceptibility whilst the LPA can predict susceptibility to RIF, INH, FLQ, ETB and SLIDs (106,159,160). The WHO-endorsed methods of DST are also often difficult to access across the world given to their prohibitive reliance on infrastructure and financial expense. As such, in many settings DST is performed only for select patient demographics and in others DST is not performed at all. Which DST methods are used depends on the resources and disease epidemiology of the setting in question.

5.1.2.1 Use of MODS in drug susceptibility testing

The first-choice DST methods, such as the culture-based approaches with the BACTEC MGIT system, are prohibitively expensive for use in developing countries yet the need to characterise drug resistance remains. The microscopic observation drug susceptibility (MODS) assay was developed in Peru as a lower cost DST method that can diagnose TB and drug resistance

reliably and efficiently in less than 2 weeks (272). The median time to culture is comparable for MODS and MGIT and quicker than growth on LJ slopes albeit with slightly lower sensitivity and specificity (273–277).

MODS has been employed throughout the world and used for the diagnosis of rifampicin and isoniazid resistance (276) though is recommended by the WHO for use only in highly specified settings (161). For rifampicin, the sensitivity of the assay is 98.0% (95% CI 94.5–99.3) and the specificity is 99.4% (95% CI 95.7–99.9) (278). For assays using 0.1 µg/mL isoniazid, the sensitivity and specificity are 97.7% (94.4–99.1) and 95.8% (88.1–98.6), respectively (278). When using 0.4 µg/mL isoniazid, sensitivity falls to 90.0% (84.5–93.7) but specificity increases to 98.6% (96.9–99.4) (278). Recent investigations using pyrazinamide in the MODS assay have been disadvantaged by the lack of a gold standard for pyrazinamide-susceptibility testing and small number of samples but have shown discordance with results obtained using both LJ slopes and the MGIT system (279–281). Lower sensitivity and specificity combined with insufficient concordance of the MODS assay with other DST methods have prevented use of the assay with the other first-line drugs, ethambutol and streptomycin (273,274). Second-line DST using this approach has yielded sensitivity from 88.1% to 100% and specificity from 92.3% to 100% in the small number of studies that have investigated it (282,283).

5.1.2.2 Whole genome sequencing in drug susceptibility testing

Whole genome sequencing has been widely used to study the epidemiology of tuberculosis (284) and is an emerging technology in TB reference laboratories of high-resource countries (192,225,237). Owing to the dependency on culture, investment in laboratory infrastructure, staff training and expense (as discussed in chapters three and four), genomics is not presently appropriate for the detection of tuberculosis. Within the landscape of well-established and widely-used diagnostic tools, it instead provides the opportunity for characterisation of any mutation that may cause resistance to any drug after an MTBC organism has been detected by culture. This may be most appropriately implemented once rifampicin resistance has been established. Additional advantages of WGS in the diagnostic paradigm

include identification of the lineage of aetiological strains, comparative use of the data for ultimate-resolution epidemiological approaches and characterisation of mixed infections when sequencing is not undertaken from an isolated colony.

Beyond obtaining adequate coverage of the genome (which isn't as issue for WGS from cultured samples), the sensitivity and specificity of genomic DST is dependent on the bioinformatic tools used to analyse sequence data. In the MTBC, this means detection of mutations based genome reconstruction and SNV-calling algorithms then interpretation of the clinical consequences of these mutations based on our prior understanding of the genotype-phenotype relationships. A number of publicly available bioinformatic tools exist to fulfil this purpose, each reliant on its own algorithms and databases of genotype-phenotype associations though these databases are not thought to be publicly available. (212,213,285). When previously uncharacterised mutations are discovered in genomic regions associated with resistance, these should be characterised phenotypically and the outcomes of this added to ever-growing databases of resistance-associated mutations.

5.1.2 Pyrazinamide

5.1.2.1 Pyrazinamide in the treatment of TB

Pyrazinamide (PZA) is an important anti-TB drug. It is now considered critical to TB treatment regimens for its ability to kill non-replicating persister bacilli, those with low metabolic activity (286). Inclusion of PZA in the treatment standard regimen enabled the reduction of treatment time from 9-12 months to 6 months (287). Pyrazinamide is included in all new anti-TB regimens currently undergoing stage II or stage III clinical trials (288) and is likely to be important for future drug regimens, due to synergistic interactions with new TB drugs, including new drugs, bedaquiline and PA-824 (289–295).

Efficacy of pyrazinamide was initially demonstrated in mice, before being shown to work against MTBC organisms in humans. It exhibits no *in vitro* activity against actively growing *M. tuberculosis* under normal culture

conditions (296). Its antibacterial activity was later found to be dependent on a low extracellular pH of 5-6, as in the phagosome of macrophages, and preferentially towards non-replicating, low metabolic activity persisters (286,294,296). It has long been a held assumption that PZA activity requires this low extracellular pH, however has recently been challenged by pyrazinamide susceptibility of *M. tuberculosis* cultured on neutral and alkaline media under other diverse stressors (297–299).

Our understanding of the mechanism of action of PZA remains incomplete. Pyrazinamide is a prodrug that enters the TB bacilli through passive diffusion. The pyrazinamidase enzyme, a nicotinamidase that is encoded by the *pncA* gene, converts it to pyrazinoic acid (POA) in the cytoplasm (300). POA is pumped out of the cell and then, under acidic extracellular conditions, is protonated to HPOA which can be reabsorbed. Inside the cell, HPOA releases its proton, acidifying the cytoplasm and becoming POA once again to continue this cycle. The cell becomes increasingly acidic and POA accumulates in the cell disrupting the balanced energy requirements of the non-replicating persister bacilli (291). The consequences are lethal and include inhibition of various enzymes, destruction of the proton motive force and effecting membrane transport. POA is also thought to bind to the ribosomal S1 protein encoded by *rpsA* (301), although this has recently been contested (302). It is proposed that POA competes with tmRNA to inhibit trans-translation, a process that manages stalled ribosomes during translation in non-replicating bacilli (301). In turn synthesis of proteins translated from damaged mRNA is prevented and stress survival, associated with diverse conditions, is aided (298,303–305).

5.1.2.2 Pyrazinamide resistance

Our understanding of pyrazinamide resistance, including its prevalence, remains inadequate despite its essential role in TB treatment (306). In a meta-analysis by Whitfield *et al.* (307), it was estimated that

- Globally 1.4 million pyrazinamide-resistant TB cases occur annually.
- The worldwide prevalence of pyrazinamide resistance is 60.5% amongst MDR cases and 16.5% in patients with any form of TB.

- The Americas region has highest proportion of pyrazinamide resistance of any WHO-region, with 21.9% of all cases being PZA resistant.

These estimates are limited by a paucity of global data about pyrazinamide susceptibility that arises as testing against the drug is not undertaken routinely (308). Both *M. bovis* and *M. canettii* are known to be intrinsically resistant to the drug, though species identification is also not undertaken routinely (167).

In a meta-analysis of mutations underlying pyrazinamide resistance, undertaken in 2015, around the commencement of the study presented here, Ramirez-Busby *et al* found that 83% of all PZA-resistant isolates had a mutation in *pncA* or its promoter region (306). Such mutations reduce PZAse activity leading to absent or reduced conversion of PZA to POA. Assessing 2760 resistant isolates, they identified 641 unique mutations in the region containing the *pncA* gene and its promoter (306). Similarly, the meta-analysis by Whitfield *et al.* (307) found 608 unique SNPs at 397 positions in the 561bp *pncA* gene. Of the mutations described in the Ramirez-Busby *et al* analysis, 79% caused an amino acid substitution, 16% were a nucleotide insertion or deletion and other isolates had more than one mutation (306). Isolates from one study they described were reported with complete deletion of the *pncA* gene. Resistance-associated mutations are dispersed throughout the gene's coding sequence and promoter region (309). The effects of *pncA* mutations on PZA susceptibility are highly inconsistent:

- Mutations are associated with diverse MICs.
- Mutations causing non-synonymous amino-acid substitutions may be associated with PZA sensitivity (310,311). Amongst 3329 sensitive samples assessed, Ramirez-Busby *et al* found 96 mutations in the *pncA* gene (306).
- Mutations causing synonymous amino acid substitution have been associated with phenotypic resistance as well as sensitivity (312).
- Mutations have been identified that are common to both phenotypically resistant and sensitive samples, as the only mutations found in the *pncA* gene (306).

No individual mutation, or even subset of mutations, has a strong predictive value for pyrazinamide resistance.

In the Ramirez-Busby *et al* analysis, 18% of resistant samples had a wild type *pncA* (306). These are instead thought to have mutations in other genes or the susceptibility phenotype may be considered 'unexplained' (291). Studies of pyrazinamide resistant samples without *pncA* mutations have led to identification of mutations in the *rpsA*, *panD* and *hadC* genes,(313,314,309,315–320). Other studies have concluded that these mutations are not associated with resistance so their role in susceptibility remains unclear (320,317).

- The *rpsA* gene encodes the 30S ribosomal protein S1. Over-expression of *rpsA* is associated with *in vitro* resistance to pyrazinamide and mutations in *rpsA* alter POA binding to the ribosomal S1 protein and allowing trans-translation to occur uninhibited (301). Most, though not all, variation between PZA sensitive and resistant samples occurs in the C-terminal domain of the RpsA protein, a potential target of POA (301,309). Mutations in *rpsA* have been associated with phenotypic resistance and sensitivity, including via synonymous amino acid substitution (309,315,318). Many samples have been described with mutations in neither *rpsA* or *pncA* (316,317). Further characterisation of the role of these mutations, including the associated MICs, is needed to verify the utility of *rpsA* in diagnosing PZA resistance (309,316–318).
- The gene, *panD*, encodes the enzyme aspartate-alpha-decarboxylase involved in beta-alanine biosynthesis. Mutations in this gene have been identified in *in vitro* generated PZA resistant mutants; *M. canettii* which is naturally resistant and in naturally occurring *M. tuberculosis* without mutations in either *pncA* or *rpsA* (313,314). Resistance-associated mutations cluster in the C-terminal domain of PanD; in one study, 24/30 mutations in PanD resulted in the same change, Met117Ile in the C-terminal domain (314). Whilst PanD is unlikely to be the main target of PZA or POA, activity of the aspartate-alpha-decarboxylase enzyme is inhibited by POA (314,320). It has been suggested that PZA resistance may be caused by binding of POA to this C-terminal domain, inhibiting production of pantothenate and co-enzyme A from beta-alanine and interfering with the cell's metabolic functions (321).
- Mutations have been identified in the *hadC* gene amongst pyrazinamide resistant strains without *pncA* and *rpsA* mutations. The gene encodes the

beta-hydroxyacyl-acyl carrier protein dehydratase involved in mycolic acid elongation during growth of the cell envelope (313). Owing to the role of *hadC* in actively growing bacilli, rather than persister cells, it is thought to be unlikely that it has a role in PZA resistance (321).

5.1.2.3 Pyrazinamide susceptibility testing

Until recently, phenotypic approaches have formed the basis of pyrazinamide susceptibility testing. The WHO recommends culture-based susceptibility testing for pyrazinamide but has endorsed no methods for doing so; tests are rarely performed in the diagnostic workflow (308). Phenotypic tests present unreliable results, particularly when the clinical inoculum is too large and physiological stressors such as an acidic environment are not present (322–325). Nonetheless, several are available, including:

- The BACTEC™ MGIT™ PZA test (Becton Dickinson, Sparks, MD), described in section 1.5.1.2, based on growth of MTBC in an automated system with a critical PZA concentration of 100 µg/mL.
- The Wayne's test (201), a colorimetric test that determines whether the PZAse enzyme is active. It requires subjective interpretation of the colorimetric changes and is limited by unacceptable rates of false-resistance (294,322,326).
- The Microscopic-observation drug susceptibility (MODS) assay (200), described in section 5.1.2.1, determines microscopic growth in the presence of 800 µg/ml PZA.

Genotypic approaches to PZA resistance testing have been proposed, though none have yet been commercialised. It has been reported that identification of mutations in *pncA* and its promoter can determine PZA resistance in MDR-TB isolates with high positive-predictive values and can rule out PZA resistance in non-MDR-TB with high negative-predictive values (322,327). A review by Ramirez-Busby *et al* estimated that this would lead to a sensitivity of 83% and specificity of 90% (306). Caveats arise from the unclear relationship between PZA susceptibility genotypes and phenotypes and the 17% of PZA resistant samples that would go undetected. A simple gene-based assay, such as the GeneXpert, is not going to be feasible for improved

prediction of PZA resistance since even the known mutations are distributed across the entire large *pncA* gene. As at least three other genes have a role in the clinical resistance, these too may need to be amplified and screened or sequenced to achieve the desirable sensitivity and specificity for the assay. However, addition of *rpsA*, *panD* and *hadC* to a genotypic diagnostic is thought to lead to insignificant improvement in sensitivity. Instead, a comprehensive molecular test, such as whole genome sequencing as discussed in section 5.1.2.2. would need to target the whole *pncA* gene sequence and screen for 600+ variants (327–331).

5.1.8 Rationale for the Study

Given the threat presented by the global burden of drug resistant TB, efficient detection and treatment of strains with resistance to any drug is central to worldwide elimination efforts. Methods currently used for characterisation of drug susceptibility have a number of short-comings, including delays in detection, prohibitive expense and poor accuracy of the test. Improving on these methods is essential to on-going success in patient treatment and meeting the WHO goals for control of TB.

Collaboration with researchers in Peru was established in summer 2015. A sample set was available that had previously been phenotypically characterised by MODS for rifampicin, isoniazid and pyrazinamide resistance, and by the BACTEC 960 PZA MGIT and Wayne's assays. Using whole-genome sequencing at the University of Warwick, it was decided to characterise the genetic basis of resistance in these strains and how this related to the performance of the phenotypic diagnostic assays used. The aims and rationale of undertaking this research were to:

- Determine the mutations associated with pyrazinamide resistance in this dataset, adding further evidence to conversations about the role of *pncA* and the involvement genes, *rpsA*, *panD* and *hadC*, whose role in PZA susceptibility remains unclear.
- Investigate the relationship between phenotypic pyrazinamide susceptibility test outcomes and the genetic basis of resistance. Poor reproducibility of PZA diagnostics has undermined its routine use in the

clinic and understanding the genetic mechanisms of susceptibility that might contribute to this, particularly for PZA MODS for which no previous studies were found to address this, is limited.

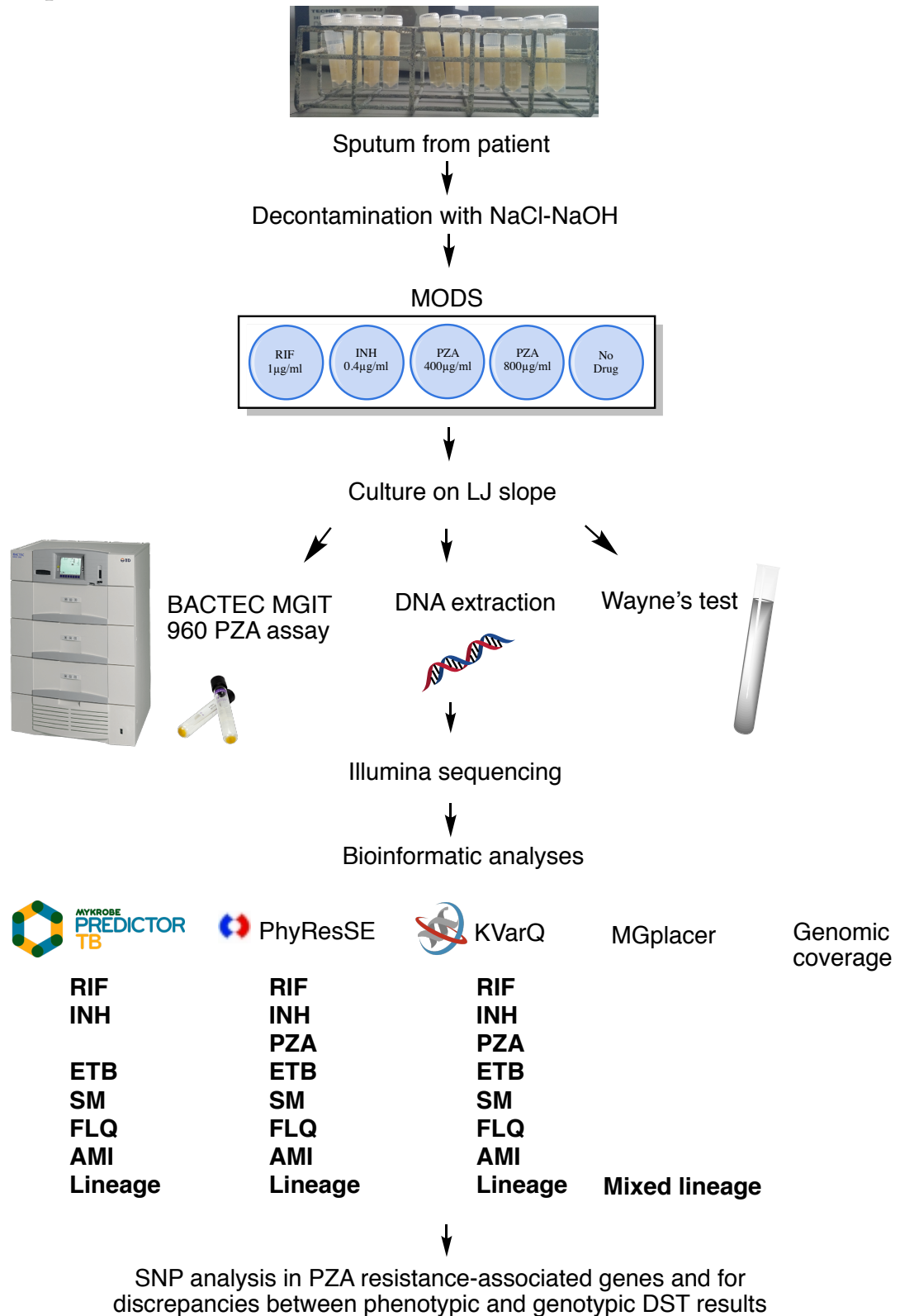
- Characterise the genetic basis of susceptibility as determined by the MODS assay for rifampicin and isoniazid resistance, adding to the understanding of the shortcomings in accuracy of these methods.
- Compare and evaluate the performance of whole genome sequencing and individual bioinformatic tools for prediction of drug susceptibility from whole-genome sequence data. All available high-throughput widely available bioinformatic tools were evaluated. It was expected that, being based on different algorithms for identifying mutations and using different databases that relate genotype to phenotype, that their susceptibility predictions would also differ. By relating this to the phenotypic results, this would allow evaluation of performance that could feedback to improvements in these tools.

The MODS assay is conducted after inoculation of culture media directly with sputum and the outcomes reflect the strains present in potentially heterogeneous sputum. In order to be directly comparable, this heterogeneity was maintained in the other phenotypic DST methods and in the whole-genome sequences. The influence of heterogeneity on performance of these diagnostic was therefore also evaluated.

5.2 Methods

Full details of materials and methods can be found in chapter two. The laboratory and bioinformatics workflow for this study is shown in Figure 5.1

Figure 5.1: Workflow of methods undertaken to analyse samples in chapter five



5.2.1 Sample collection and processing

Samples were collected, cultured and subjected to phenotypic analyses by my collaborators in Peru. Sixty-nine sputum samples were obtained, each from a different patient, in the DIRESA-Callao region and Hospital Dos de Mayo, Lima, Peru in the periods July-December 2015 and October-December 2015, respectively. Samples were processed at Universidad Peruana Cayetano Heredia. As part of an on-going project, the sputa had been characterised as smear-positive by microscopy, decontaminated and primary cultures were obtained on solid 7H10 medium. Cultured material was stored at -80 °C. The samples selected for this study were selected from the project independent of drug resistance profile (Table 5.1).

Drug susceptibility testing for the first-line antibiotics rifampicin, isoniazid and pyrazinamide was performed by staff in Peru using the following phenotypic methods:

- Microscopic-observation drug susceptibility (MODS) assay critical drug concentrations: 0.4 µg/ml INH, 1.0 µg/ml RIF, 400 µg/ml PZA, 800 µg/ml PZA
- BACTEC™ MGIT™ 960 PZA assay (MGIT), inoculated with cultured isolates and a critical drug concentration of 100 µg/mL PZA.
- Wayne's pyrazinamidase assay of cultured sample

Phenotypic testing was unavailable for first-line drugs, ethambutol and streptomycin and for all second-line

5.2.2 DNA extraction and genome sequencing

DNA extractions were performed at Universidad Peruana Cayetano Heredia from multiple-colony sweeps from solid media, after re-culture from multiple-colony frozen stocks. DNA was extracted using the CTAB method and the protocol as described previously. DNA extracts were sent to the University of Warwick where all further laboratory and bioinformatic analyses by the author. DNA extracts were quantified and converted to

sequencing libraries using the Illumina Nextera XT method. All 69 samples were sequenced on one run of the Illumina MiSeq using 2 x 301 bp paired-end reads to produce FASTQ files.

5.2.3 Genomic prediction of drug resistance

The quantity of sequence reads and the coverage of the TB reference genome was determined after alignment against H37Rv using the Botwie2 pre-set parameter `--very-sensitive-local` then visualised with Qualimap. Clinically relevant drug resistance was predicted from genetic variants after comparison to known resistance-associated mutations catalogued in databases, using the recently developed software tools Mykrobe predictor (212), PhyResSE (213), and KVarQ (285). All three tools were used as the results varied owing to differences in variant calling protocols and in the catalogues of drug resistance mutations used. Approaches that provided the number of reads from each allele enabled identification of hetero-resistance at positions with a depth of coverage > 10 and the minor allele as >10% of the total reads.

For phenotypically resistant samples where no known resistance-associated mutations could be identified, genetic variants were sought in the relevant resistance-associated genes. This approach was also applied to the identification of mutations in the *pncA*, *rpsA*, *panD* and *hadC* gene regions associated with pyrazinamide resistance. These genetic variants between the clinical isolates from Peru and the H37Rv TB reference genome were identified with VarScan2, excluding:

- variants in known repetitive regions (191)
- SNPs that fell near INDELS
- positions with coverage less than 5 bases, quality score below 20, minimum variant frequency below 0.1 or a p-value less than 0.001.

A maximum-likelihood phylogenetic tree was computed by PhyResSE.

5.2.4 Lineage typing and identification of mixed lineages

The lineages of strains present were identified using MGplacer and verified by comparison to the epidemiological types provided by PhyResSE and

Mykrobe. MGplacer was also used to screen for mixed lineages and identify the lineages present. Samples were reported to have mixed lineages when MGplacer reported that over 5% mixed reads were present and this was confirmed by visual inspection of reads with Artemis.

5.3 Results

Sputum from patients in the Lima and Callo regions of Peru was characterised to understand the genetic basis of drug resistance in these samples, particularly against pyrazinamide, and to evaluate the performance of DST approaches. The 69 smear-positive samples selected for this study had been analysed by our Peruvian collaborators by MGIT, MODS and Wayne's methods of phenotypic pyrazinamide susceptibility testing and MODS for phenotypic rifampicin and isoniazid testing. Analyses were conducted from primary cultures of sputum, either using multiple colonies from an LJ slope or broth culture, to capture the diversity of *M. tuberculosis* within the sample. Genomic coverage from each sample ranged from 25 to 91-fold coverage of the genome. Sequences were used to predict genotypic drug resistance against rifampicin, isoniazid, pyrazinamide, ethambutol, streptomycin, fluoroquinolones and the aminoglycoside agents using the Mykrobe predictor, PhyResSE and KVarQ software tools. When resistance-associated mutations could not be identified in a phenotypically resistant sample, SNP calling was undertaken in the relevant resistance-associated genes. For all samples, SNP-calling was undertaken in pyrazinamide resistance-associated genes. Heterogeneity amongst the strains cultured from each sputum sample was characterised by detecting mixed coverage by PhyResSE, defining mixed lineages by MGplacer and identifying heterogeneous alleles at resistance-associated genomic positions; this informed the understanding of the DST method performances. All samples were identified as containing *M. tuberculosis* sensu stricto strains.

5.3.1 Lineage and drug resistance predictions

Mykrobe, PhyResSE and KVarQ were first used to identify resistance-associated mutations in all samples. *M. tuberculosis* lineages were assigned

using these same tools and MGplacer (Table 5.1). These mutation profiles were the used to evaluate the performance of the tools for testing and predicting drug resistance.

Table 5.1: All drug resistance associated mutations, as predicted by any bioinformatic analysis tool, and the major lineages predicted by MGplacer

Clinical definitions have been predicted based on this full drug resistance profile. A coloured background to the cell indicates that resistance is associated with the mutation; a white background to the cell indicates drug sensitivity; a different colour is used for each drug for ease of visual inspection only. For clinical definitions, a dark grey background is used for strains that may be considered XDR-TB given the genomic predictions and the shade of grey gets progressively lighter for pre-XDR-TB and MDR-TB, then the background is white for DS-TB.

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP050	2.2.1	Leu452Pro in RpoB	Ser315Thr in KatG	Pro62Ser in PncA Arg212Arg in RpsA	Gly406Ser in EmbB	Lys43Arg in RpsL	Asp94Gly in GyrA	A1401G in gene <i>rrs</i>	XDR
MP282	2.2.1	Ser450Leu in RpoB	Ser315Thr in KatG	Ile6Ser in PncA Arg212Arg in RpsA	Met306Ile in EmbB	Lys43Arg in RpsL	Ala90Val in GyrA	A1401G in gene <i>rrs</i>	XDR
MP275	2.2.1	Ser450Leu in RpoB	Ser315Thr in KatG	Arg212Arg in RpsA	Met306Ile in EmbB	Lys43Arg in RpsL			MDR
MP308	2.2.1	Ser450Leu in RpoB	Ser315Thr in KatG	Arg212Arg in RpsA	Met306Ile in EmbB	Lys43Arg in RpsL			MDR
MP176	2.2.1			Arg212Arg in RpsA		Lys43Arg in RpsL			DS

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP035	2.2.1			Arg212Arg in RpsA					DS
MP124	2.2.1			Arg212Arg in RpsA					DS
MP306	2.2.1			Arg212Arg in RpsA					DS
MP241	4.1.1	Ser450Leu in RpoB	Ser315Thr in KatG	Gln10Pro in PncA			Ala90Val in gyrA		Pre-XDR
MP169	4.1.1		Ser315Thr in KatG						DS
MP186	4.1.1								DS
MP360	4.1.1.3	Ser450Leu in RpoB	Ser315Thr in KatG		Gly406Ala in EmbB				MDR
MP063	4.1.2.1	Asp435Val in RpoB	Ser315Thr in KatG intergenic (Rv1483- 15nt)						MDR

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP042	4.1.2.1	His445Leu in RpoB	Ser315Thr in KatG		Gly406Ser in EmbB				MDR
MP026	4.1.2.1		Ser315Thr in KatG						DS
MP268	4.1.2.1		intergenic (Rv1483- 15nt)						DS
MP126	4.1.2.1						Asp94Gly in GyrA		DS
MP051	4.1.2.1								DS
MP082	4.1.2.1								DS
MP111	4.1.2.1								DS
MP199	4.1.2.1								DS
MP200	4.1.2.1								DS
MP210	4.1.2.1								DS
MP307	4.1.2.1								DS

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP315	4.1.2.1								DS
MP190	4.3.2	Ser450Leu in RpoB	Ser315Thr in KatG			C491T in gene <i>rrs</i>			MDR
MP028	4.3.2		Ser315Thr in KatG			C491T in gene <i>rrs</i>			DS
MP203	4.3.2					C491T in gene <i>rrs</i>			DS
MP248	4.3.2					C491T in gene <i>rrs</i>			DS
MP027	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Lys48Thr in PncA	Met306Val in EmbB		Asp94Gly in GyrA	A1401G in gene <i>rrs</i>	XDR
MP202	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Lys48Thr in PncA	Met306Val in EmbB		Asp94Gly in GyrA	A1401G in gene <i>rrs</i>	XDR
MP276	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Lys48Thr in PncA	Met306Val in EmbB		Asp94Gly in GyrA	A1401G in gene <i>rrs</i>	XDR
MP320	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Lys48Thr in PncA	Met306Val in EmbB		Asp94Gly in GyrA	A1401G in gene <i>rrs</i>	XDR

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP078	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA	Tyr319Ser in EmbB		Ser91Pro in GyrA	A1401G in gene <i>rrs</i>	XDR
MP177	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA	Tyr319Ser in EmbB		Ala90Val in GyrA		Pre-XDR
MP096	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA	Tyr319Ser in EmbB				MDR
MP123	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA	Tyr319Ser in EmbB				MDR
MP170	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA	Tyr319Ser in EmbB				MDR
MP247	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA	Tyr319Ser in EmbB				MDR
MP249	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA	Tyr319Ser in EmbB				MDR
MP287	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA	Tyr319Ser in EmbB				MDR
MP341	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA					MDR

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP342	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA					MDR
MP364	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG	Gln10Arg in PncA					MDR
MP182	4.3.3	Asp435Val in RpoB	Ser315Thr in KatG						MDR
MP013	4.3.3	Asp435Tyr in RpoB	Ser315Thr in KatG						MDR
MP054	4.3.3	Ile491Phe in RpoB	Ser315Thr in KatG						MDR
MP031	4.3.3	Ser450Leu in RpoB	intergenic (Rv1483- 15nt)	His51Arg in PncA Ala381Val in RpsA					MDR
MP294	4.3.3	Ser450Leu in RpoB	intergenic (Rv1483- 15nt)	His51Arg in PncA Ala381Val in RpsA					MDR

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP313	4.3.3	Ser450Leu in RpoB	intergenic (Rv1483- 15nt)	His51Arg in PncA Ala381Val in RpsA					MDR
MP367	4.3.3	Ser450Leu in RpoB	intergenic (Rv1483- 15nt)	His51Arg in PncA Ala381Val in RpsA					MDR
MP032	4.3.3		intergenic (Rv1483- 15nt)	His51Arg in PncA Ala381Val in RpsA					DS
MP040	4.3.3		intergenic (Rv1483- 15nt)	His51Arg in PncA Ala381Val in RpsA					DS
MP175	4.3.3		intergenic (Rv1483- 15nt)	His51Arg in PncA Ala381Val in RpsA					DS

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP366	4.3.3	Ser450Leu in RpoB	Ser315Thr in KatG	His71Arg in PncA					MDR
MP369	4.3.3	Ser450Leu in RpoB	Ser315Thr in KatG	His71Arg in PncA					MDR
MP041	4.3.3		Ser315Thr in KatG						DS
MP005	4.3.3		Ser315Thr in KatG						DS
MP073	4.3.3		Ser315Thr in KatG						DS
MP118	4.3.3		Ser315Thr in KatG						DS
MP069	4.3.3								DS
MP092	4.3.3								DS

Sample	Lineage	Rifampicin	Isoniazid	Pyrazinamide	Ethambutol	Streptomycin	Fluoroquinolones	Amikacin/ Capreomycin/ Kanamycin	Clinical definition
MP167	4.3.3								DS
MP075	4.3.4.2	Ser450Leu in RpoB Ile491Val in rpoC	Ser315Thr in KatG	His57Leu in PncA					MDR
MP120	4.3.4.2	Ser450Leu in RpoB Val483Gly in rpoC	Ser315Thr in KatG						MDR
MP059	4.4.1.1								DS
MP284	4.7								DS
MP019	4.8	Leu430Pro in RpoB	intergenic (Rv1483-15nt)						MDR
MP038	4.9					A514C in gene <i>rrs</i>			DS

Heterogeneity within samples became apparent on bioinformatic analysis. The software tool, KVarQ, reported 'mixed coverage' in samples MP031, MP050, MP054, MP069, MP075, MP118, MP182, MP186, MP241, MP275, MP307, MP313, MP364 and MP367. On inspection of SNPs called in these samples, heterogeneous alleles could be observed at numerous positions across the genome, confirming that there were heterogeneous strains within these samples. Understanding this heterogeneity was important to assessing its impact on drug susceptibility typing, a primary aim of this study.

MGplacer, a tool that places metagenomic sequence reads onto a phylogenetic tree, was used to identify the lineages of strains within the samples. MGplacer identified samples with multiple lineages present and the major strain in each sample concurred with those identified by other software tools. The outputs from MGplacer were filtered such that only samples with two or more distinct lineages, excluding the ancestral lineage (lineage 7), could be considered mixed (Table 5.2). Read alignments were visually inspected at the lineage defining SNP positions to validate the determination of mixed infection.

In some samples, heterogeneous alleles were observed at sites associated with drug resistance by the Mykrobe and PhyResSE tools. A number of samples were identified with heteroresistance to one or more drugs (Table 5.3), with variable frequencies of resistant reads depending on the SNV-calling algorithm used by the tool. All samples showing heteroresistance were phenotypically resistant to the relevant drug, by MODS for rifampicin and isoniazid or BACTEC MGIT for pyrazinamide, with the exception of resistance to isoniazid by the *inhA* promoter mutation in sample MP313. For second-line drugs, phenotypic susceptibility results were not available. Both mixed lineages and heteroresistant alleles could be identified in some samples, MP118, MP186 and MP313.

Table 5.2: Mixed lineage samples and their associated lineage branches and proportions, as reported by MGplacer

Sample	Major lineage					Minor lineage				
	Proportion (%)	CI 95%	Branch beginning (lineage)	Branch end (lineage)	Location in branch	Proportion (%)	CI 95%	Branch beginning (lineage)	Branch end (lineage)	Location in branch
MP118	79.3	77.6-80.8	4.3	4.3.3	0.98	20.7	19.2-22.4	4.1	4.1.1	0.84
MP313	84.8	83.4-86.1	4.3	4.3.3	0.99	15.2	13.9-16.6	4.1.2	4.1.2.1	0.98
MP275	92.7	92.0-93.4	2.2.1	2.2.1.2	0.03	7.3	6.6-8.0	4.3	4.3.3	0.8
MP210	94.8	94.0-95.6	4.1.2	4.1.2.1	0.92	5.2	4.4-6.0	4.3	4.3.2	0.9
MP182	98.5	98.2-98.8	4.3	4.3.3	1	1.5	1.2-1.8	2.2.1	2.2.1.2	0.9
MP186	98.9	98.7-99.2	4.1.1	4.1.1.3	1	1.1	0.8-1.3	2.2	2.2.1	0.76
MP315	99.0	98.7-99.3	4.1.2	4.1.2.1	1	1.0	0.7-1.3	4.3	4.3.3	0.71
MP096	99.3	98.7-99.6	4.3	4.3.3	1	0.7	0.4-1.3	4.3.4	4.3.4.2	0.83
MP307	99.5	99.3-99.7	4.1.2	4.1.2.1	1	0.5	0.3-0.7	4.3	4.3.3	0.8
MP169	99.6	99.4-99.7	4.1.1	4.1.1.1	0	0.4	0.3-0.6	4.3	4.3.3	0.58
MP284	99.6	99.5-99.7	4	4.7	1	0.4	0.3-0.5	2.2	2.2.1	0.85

Table 5.3: Heteroresistant mutations predicted to cause resistance by Mykrobe and PhyResSE

Sample	Mutation	Mykrobe read ratio (R/S)	PhyResSE read ratio (R/S)
MP031	Ser450Leu in RpoB	5/12	11/32
MP118	Ser315Thr in KatG	18/7	25/16
MP313	<i>inhA</i> promoter mutation -15(C>T)	27/8	34/11
MP313	Ser450X in RpoB	22/2	29/2
MP313	His51Arg in PcnA	n/a	26/4
MP186	A1401X in <i>rrs</i>	3/25	n/a
MP241	Ala90X in GyrA	20/4	28/6
MP367	His51Arg in PncA	n/a	13/2
MP050	Lys43Arg in RpsL	32/6	48/8

Mykrobe also predicted low frequency drug resistance alleles or ‘minor alleles’ where a low proportion reads were associated with a resistance mutation and the sample was not otherwise predicted to be clinically resistant (Table 5.4). Some of these minor alleles were found as the only mutation for the given drug whilst others were found in addition to a second mutation that was predicted to cause resistance to the same drug. There were 21 samples with only a minor allele mutation where no further mutation that was predicted to cause resistance in the same drug. All but one of these were sensitive by the phenotypic methods available and Mykrobe did not predict that these samples were resistant.

Table 5.4: Low frequency resistance alleles reported by Mykrobe predictor
Colours represent susceptibility status indicated by the drug susceptibility test method; red indicates resistance-association whilst green indicates an antibiotic sensitive- association. Fractions indicate the number of reads in which the resistance-associated allele (numerator) and sensitive alleles (denominator) were reported by genomic prediction tools.

Sample	Drug	Phenotypic DST	Minor allele mutation	Major allele mutation
MP320	RIF	R	Gly42X in RpoB: 1/25	Asp435X in RpoB: 25/0
MP248	RIF	S	Asp335X in RpoB: 1/32	—
MP341	RIF	R	Phe425X in RpoB: 2/14	Asp435X in RpoB: 16/0
MP050	RIF	S	Phe425X in RpoB: 2/22	Leu452X in RpoB: 30/0
MP360	RIF	R	Phe425X in RpoB: 2/31	Ser450X in RpoB: 28/0
MP082	RIF	S	Phe425X in RpoB: 1/30	—
MP118	RIF	S	Gly426X in RpoB :1/17	—
MP059	RIF	S	Gly426X in RpoB: 1/24	—
MP248	RIF	S	Gly426X in RpoB: 1/33	—
MP282	RIF	R	Gly426X in RpoB: 1/26	Ser450X in RpoB: 24/0
MP182	RIF	R	Thr427X in RpoB: 1/33	Asp435X in RpoB: 29/2

Sample	Drug	Phenotypic DST	Minor allele mutation	Major allele mutation
MP186	RIF	S	Ser428X in RpoB: 1/23	—
MP118	RIF	S	Ser431X in RpoB: 1/18	—
MP118	RIF	S	Ser431X in RpoB: 1/18	—
MP268	RIF	S	Gln432X in RpoB: 1/33	—
MP075	RIF	R	Phe433X in RpoB: 1/14	Ser450X in RpoB: 18/0
MP186	RIF	S	Met434X in rpoB: 1/27	—
MP275	RIF	R	Asp435X in RpoB: 1/32	Ser450X in RpoB: 31/0
MP073	RIF	S	Asn438X in RpoB: 2/47	—
MP026	RIF	S	Pro439X in RpoB: 2/39	—
MP200	RIF	S	Gly442X in RpoB: 1/23	—
MP031	RIF	R	Gly442X in RpoB: 2/20	Ser450X in RpoB: 5/12
MP028	RIF	S	Gly442X in RpoB: 2/66	—
MP005	RIF	R	His445X in RpoB: 2/57	—
MP176	RIF	S	Arg447X in RpoB: 2/46	—
MP186	RIF	S	Ser428X in RpoB: 1/23	—
MP092	RIF	S	Ser450X in RpoB: 1/23	—
MP176	RIF	S	Ser450X in RpoB: 2/41	—
MP248	INH	S	<i>inhA</i> promoter mutation - 16nt (A>X): 1/20	—
MP176	INH	S	<i>inhA</i> promoter mutation - 8nt (T>X): 1/33	—
MP190	ETB	n/a	Met306X in EmbB: 1/32	—
MP038	ETB	n/a	Met306X in EmbB: 2/22	—
MP073	ETB	n/a	Met306X in EmbB: 2/46	—

Sample	Drug	Phenotypic DST	Minor allele mutation	Major allele mutation
MP313	FLQ	n/a	Asp89X in GyrA: 1/32	—
MP202	FLQ	n/a	Pro86X in GyrA: 1/23	Asp94X in GyrA: 26/0
MP005	FLQ	n/a	Pro86X in GyrA: 2/58	—
MP170	SM	n/a	C517X in gene <i>rrs</i> : 1/32	—

5.3.2 Genotypic definition of pyrazinamide resistance

Before the performance of susceptibility tests could be evaluated, the resistance status associated with each mutation needed to be defined. Given that no gold-standard DST is available for pyrazinamide, the available evidence about susceptibility was evaluated to draw conclusions. Neither MODS nor genotypic analysis methods could be considered as gold standard for rifampicin and isoniazid susceptibility testing, so it was not possible to conclude the susceptibility status of samples to these drugs, especially those with discordance across the DST methods.

5.3.2.1 Pyrazinamide

Phenotypic pyrazinamide susceptibility testing was undertaken using the MGIT, MODS and Wayne's assays. Predictions of pyrazinamide resistance were made using PhyResSE and KVarQ and all mutations in the *pncA*, *rpsA*, *panD* and *hadC* gene regions associated with pyrazinamide resistance were identified by calling SNPs (Table 5.5). It is well established that not every mutation in the genes associated with pyrazinamide resistance necessarily causes resistance (310,311) so the results of all DST methods were taken into consideration when making predictions about which mutations would cause clinical resistance.

Table 5.5: Results of drug susceptibility testing for pyrazinamide with the various phenotypic methods and genomic prediction tools

Red font indicates that the result is associated with resistance; green font indicates association with sensitivity; black font indicated that no susceptibility prediction was made. Fractions indicate the number of reads in which the resistance-associated allele (numerator) and sensitive alleles (denominator) were reported by genomic prediction tools. The number of resistant and sensitive reads are not reported by KvarQ. Mykrobe does not report pyrazinamide resistance.

Sample	MGIT	Wayne	MODS	PhyResSE	KVarQ	SNP calling
MP005	S	S	S	—	—	—
MP013	S	S	S	—	—	—
MP041	S	S	S	—	—	—
MP042	S	S	S	—	—	—
MP051	S	S	S	—	—	—
MP054	S	S	S	—	—	—
MP059	S	S	S	—	—	—
MP063	S	S	S	—	—	—
MP069	S	S	S	—	—	—
MP073	S	S	S	—	—	—
MP082	S	S	S	—	—	—
MP118	S	S	S	—	—	—
MP120	S	S	S	—	—	—
MP126	S	S	S	—	—	—

Sample	MGIT	Wayne	MODS	PhyResSE	KVarQ	SNP calling
MP167	S	S	S	—	—	—
MP169	S	S	S	—	—	—
MP182	S	S	S	—	—	—
MP186	S	S	S	—	—	—
MP190	S	S	S	—	—	—
MP199	S	S	S	—	—	—
MP200	S	S	S	—	—	—
MP203	S	S	S	—	—	—
MP210	S	S	S	—	—	—
MP284	S	S	S	—	—	—
MP307	S	S	S	—	—	—
MP315	S	S	S	—	—	—
MP360	S	S	S	—	—	—
MP320	S	S	S	—	Lys48Thr in PncA	Lys48Thr in PncA: 49/0
MP276	S	S	S	—	Lys48Thr in PncA	Lys48Thr in PncA: 55/0
MP275	S	S	S	—	—	Arg212Arg in RpsA: 40/2

Sample	MGIT	Wayne	MODS	PhyResSE	KVarQ	SNP calling
MP308	S	S	S	—	—	Arg212Arg in RpsA:47/0
MP124	S	S	S	—	—	Arg212Arg in RpsA: 60/1
MP306	S	S	S	—	—	Arg212Arg in RpsA: 71/0
MP092	S	R	S	—	—	—
MP019	S	S	R	—	—	—
MP026	S	S	R	—	—	—
MP028	S	S	R	—	—	—
MP038	S	S	R	—	—	—
MP111	S	S	R	—	—	—
MP248	S	S	R	—	—	—
MP268	S	S	R	—	—	—
MP202	S	S	R	—	Lys48Thr in PncA	Lys48Thr in PncA: 63/0
MP027	S	S	R	—	Lys48Thr in PncA	Lys48Thr in PncA: 69/0
MP050	S	S	R	—	Pro62Ser in PncA	Pro62Ser in PncA: 52/0 Arg212Arg in RpsA: 66/0
MP176	S	S	R	—	—	Arg212Arg in RpsA: 52/0

Sample	MGIT	Wayne	MODS	PhyResSE	KVarQ	SNP calling
MP035	S	S	R	—	—	Arg212Arg in RpsA: 58/0
MP367	R	S	R	His51Arg in PncA: 13/2	His51Arg in PncA	His51Arg in PncA: 13/2 Ala381Val in RpsA: 28/3
MP313	R	S	R	His51Arg in PncA: 28/4	His51Arg in PncA	His51Arg in PncA: 26/4 Ala381Val in RpsA: 29/8
MP175	R	S	R	His51Arg in PncA: 83/0	His51Arg in PncA	His51Arg in PncA: 76/0 Ala381Val in RpsA: 54/2
MP369	R	S	R	His71Arg in PncA: 29/0	His71Arg in PncA	His71Arg in PncA: 28/0
MP078	R	S	R	Gln10Arg in PncA: 24/0	Gln10Arg in PncA	Gln10Arg in PncA: 22/0
MP249	R	S	R	Gln10Arg in PncA: 45/0	Gln10Arg in PncA	Gln10Arg in PncA: 42/0
MP364	R	S	R	Gln10Arg in PncA: 49/0	Gln10Arg in PncA	Gln10Arg in PncA: 46/0
MP177	R	S	R	Gln10Arg in PncA: 69/0	Gln10Arg in PncA	Gln10Arg in PncA: 66/0
MP342	R	S	R	Gln10Arg in PncA: 27/0	—	Gln10Arg in PncA: 25/0
MP123	R	R	S	Gln10Arg in PncA: 40/0	Gln10Arg in PncA	Gln10Arg in PncA: 36/0
MP170	R	R	S	Gln10Arg in PncA: 48/0	Gln10Arg in PncA	Gln10Arg in PncA: 43/0

Sample	MGIT	Wayne	MODS	PhyResSE	KVarQ	SNP calling
MP247	R	R	R	Gln10Arg in PncA: 38/0	—	Gln10Arg in PncA: 34/0
MP241	R	R	R	Gln10Pro in PncA: 50/0	Gln10Pro in PncA	Gln10Pro in PncA: 47/0
MP031	R	R	R	His51Arg in PncA: 30/0	His51Arg in PncA	His51Arg in PncA: 29/0 Ala381Val in RpsA: 34/0
MP032	R	R	R	His51Arg in PncA: 67/1	His51Arg in PncA	His51Arg in PncA: 60/0 Ala381Val in RpsA: 62/0
MP040	R	R	R	His51Arg in PncA: 35/0	His51Arg in PncA	His51Arg in PncA: 33/0 Ala381Val in RpsA: 30/0
MP294	R	R	R	His51Arg in PncA: 39/0	His51Arg in PncA	His51Arg in PncA: 34/0 Ala381Val in RpsA: 32/0
MP075	R	R	R	His57Leu in PncA: 22/0	His57Leu in PncA	His57Leu in PncA: 20/0
MP366	R	R	R	His71Arg in PncA: 55/0	His71Arg in PncA	His71Arg in PncA: 54/0
MP096	R	R	R	Gln10Arg in PncA: 28/0	Gln10Arg in PncA	Gln10Arg in PncA: 27/0
MP341	R	R	R	Gln10Arg in PncA: 36/0	Gln10Arg in PncA	Gln10Arg in PncA: 45/0
MP287	R	R	R	Gln10Arg in PncA: 82/1	Gln10Arg in PncA	Gln10Arg in PncA: 79/0

Sample	MGIT	Wayne	MODS	PhyResSE	KVarQ	SNP calling
MP28 2	R	R	R	Ile6Ser in PncA: 20/1	Ile6Ser in PncA	Ile6Ser in PncA: 19/0 Arg212Arg in RpsA: 43/0

Some samples were considered to certainly be resistant. For 12 samples, the phenotyping tests concordantly determined that samples were pyrazinamide resistant. The genomic software tools collectively predicted resistance in all of these samples and these were associated with the Ile6Ser, His51Arg, His57Leu, His71Arg, Gln10Pro or Gln10Arg mutations in PncA. These mutations were also identified in other samples, all of which were resistant by the MGIT assay and a second phenotypic method. Samples that were MGIT sensitive did not have these mutations and were resistant by only one method or concordantly sensitive. The relationship between this set of mutations and MGIT resistance was clearly very strong, being supported by a secondary DST method. The mutations had previously been identified to cause resistance such that they were included in the databases of the WGS tools so, along with MGIT resistance, they were considered associated with pyrazinamide resistance in this study.

Two further mutations were reported by KVarQ in association with pyrazinamide resistance but had mixed associations with phenotypic resistance:

- The mutation, Lys48Thr in PncA, was found in two samples that were resistant by MODS only and a further two samples that were concordantly pyrazinamide sensitive.
- The Pro62Ser mutation in PncA was found in sample MP050 that was phenotypically resistant by the MODS method only. A second mutation, Arg212Arg, in resistance-associated gene, RpsA, was also found in this sample.

Both mutations, Lys48Thr and Pro62Ser in PncA, were reported by the KVarQ software in association with pyrazinamide resistance, and could be identified by PhyResSE but without association with resistance. Given that these

mutations were associated with phenotypic sensitivity by the MGIT and Wayne's assays and with resistance by only the KVarQ prediction tool, they were considered to be associated with phenotypic sensitivity in this study.

The Arg212Arg mutation in RpsA was found by SNV-calling in four concordantly phenotypically sensitive samples and as the only mutation in two samples resistant by MODS only. It was found in sample MP282 with the secondary mutation, Ile6Ser in PncA which has already been established to have a strong association with phenotypic resistance, including by the MGIT assay, and in sample MP050, discussed above. All instances of this mutation were found in samples containing the 2.2.1 lineage of *M. tuberculosis*. Another mutation, Ala381Val in RpsA, was identified by all genotyping methods but not associated with resistance by the tools. It was observed to accompany every instance of the known-resistance mutation, His51Arg in *pncA* which occurred in phenotypically resistant samples, but was absent from any other sample. Given that neither Arg212Arg nor Ala381Val in RpsA were predicted to cause resistance through any sequence-based tool and that they were never associated with MGIT resistance in the absence of a secondary pyrazinamide-associated mutation, these mutations were not considered to underlie resistance in this study.

The samples without mutations in pyrazinamide-associated genes were considered to be truly sensitive to the drug given that they were never found to be associated with MGIT resistance and only ever found to be associated with either MODS or Wayne's test resistance, never both.

5.3.2.2 Rifampicin

For rifampicin, susceptibility was determined with MODS, Mykrobe, PhyResSE and KVarQ. The mutations, Asp435Val, Asp435Tyr, Leu430Pro, Ser450Leu, His445Leu, Leu452Pro and Ile491Phe were found in RpoB and associated with MODS resistance with a few exceptions (Table 5.7). Despite MODS indicating resistance, no mutation was identified by the software in the sample, MP005. For this sample, SNP calling was conducted in the genes, *rpoB*, *embB* and Rv2629, and their putative promoter regions but no mutation

could be identified. Compensatory mutations, Ile491Val and Val483Gly in RpoC, were additionally reported by KVarQ in samples MP075 and MP120, respectively. Many 'minor alleles' were identified by Mykrobe that were not considered resistant by the tools due to the low frequency resistant alleles at the mutation position. These minor alleles were found both as the only mutation in a sample and in conjunction with other low frequency mutations or mutations that were predicted to cause resistance.

5.3.2.3 Isoniazid

For isoniazid, assessment by MODS, Mykrobe, PhyResSE and KVarQ was used to determine susceptibility. Only two genetic mutations were responsible for isoniazid resistance in these isolates, Ser315Thr in KatG and a mutation 15 bases upstream of the *inhA* gene. One sample, MP063, was identified with both of these mutations. Two minor alleles were reported by Mykrobe but not associated with resistance with the tool.

5.3.2.4 Other drugs

Resistance to ethambutol, streptomycin, fluoroquinolones and aminoglycosides was not determined phenotypically. Evidence of resistance to these drugs was only be provided by the sequence-based predictions and included the mutations Tyr319Ser, Met306Ile, Gly406Ser and Gly406Ala in EmbB; C491T, A514C and A1401G in gene *rrs*; Lys43Arg in RpsL; Asp94Gly, Ser91Pro and Ala90Val in GyrA. It is unclear whether a mutation in genes associated with resistance to these other drugs might confer true clinical or phenotypic resistance from these data presented here alone. All that can be determined is whether a tool predicted resistance or not. Only one resistance-causing mutation was ever predicted per sample. Minor alleles were also identified amongst mutations associated with resistance to these drugs.

5.3.3 Performance of susceptibility testing and prediction methods

The rate of detection of resistance predicted by each DST method varied for each of the mutations identified (Table 5.6). Some discrepancies between the susceptibility results from each DST method were associated with samples with mixed lineages, heteroresistance or “mixed coverage” reported by PhyResSE (Table 5.7). The overall sensitivity and specificity of each DST method for PZA resistance were evaluated relative to the conclusions of the MGIT assay, given its strong association with susceptibility predicted in the mutations (Table 5.8). Neither MODS nor genotypic analysis methods could be considered as gold standard for rifampicin and isoniazid susceptibility testing, so it was not possible to determine a susceptibility status of samples in association with these drugs. Comparisons of the results of each method for each mutation could be assessed nonetheless.

Table 5.6 Rate of prediction of resistance for susceptibility-associated mutations found in this study Minor alleles are not considered here.

Mutation	Drug	Susceptibility conclusion	Total samples	Detection rate of mutation (%)					
				MGIT	Wayne'stest	MODS	Mykrobe	PhyResSE	KVarQ
His51Arg in PncA Ala381Val in RpsA	PZA	R	7	100	57.1	100	n/a	100	100
Gln10Arg in PncA	PZA	R	11	100	54.5	81.8	n/a	100	81.8
His71Arg in PncA	PZA	R	2	100	50	100	n/a	100	100
Gln10Pro in PncA	PZA	R	1	100	100	100	n/a	100	100
Ile6Ser in PncA Arg212Arg in RpsA	PZA	R	1	100	100	100	n/a	100	100
His57Leu in PncA	PZA	R	1	100	100	100	n/a	100	100
Arg212Arg in RpsA	PZA	S	6	0	0	33.3	n/a	0	0
Lys48Thr in PncA	PZA	S	4	0	0	50	n/a	0	100

Mutation	Drug	Susceptibility conclusion	Total samples	Detection rate of mutation (%)					
				MGIT	Wayne'stest	MODS	Mykrobe	PhyResSE	KVarQ
Pro62Ser in PncA Arg212Arg in RpsA	PZA	S	1	0	0	100	n/a	0	100
No mutation in pyrazinamide-associated genes	PZA	S	35	0	2.9	20	n/a	0	0
Asp435Val in RpoB	RIF	n/a	17	n/a	n/a	88.2	100	100	100
Asp435Tyr in RpoB	RIF	n/a	1	n/a	n/a	0	100	100	100
Leu430Pro in RpoB	RIF	n/a	1	n/a	n/a	0	100	100	100
Ser450Leu in RpoB	RIF	n/a	14	n/a	n/a	100	100	100	92.8
His445Leu in RpoB	RIF	n/a	1	n/a	n/a	100	100	100	100
Leu452Pro in RpoB	RIF	n/a	1	n/a	n/a	0	100	100	100

Mutation	Drug	Susceptibility conclusion	Total samples	Detection rate of mutation (%)					
				MGIT	Wayne'stest	MODS	Mykrobe	PhyResSE	KVarQ
Ile491Phe in RpoB	RIF	n/a	1	n/a	n/a	100	0	100	0
No mutation in rifampicin-associated genes	RIF	n/a	33	n/a	n/a	3.0	0	0	0
Ser315Thr in KatG	INH	n/a	37	n/a	n/a	100	100	100	97.3
<i>inhA</i> promoter mutation -15 C>T)	INH	n/a	9	n/a	n/a	11.1	100	100	100
Ser315Thr in KatG and <i>inhA</i> promoter mutation -15 (C>T)	INH	n/a	1	n/a	n/a	100	0	100	100
No mutation in isoniazid-associated genes	INH	n/a	22	n/a	n/a	0	0	0	0
Tyr319Ser in EmbB	ETB	n/a	11	n/a	n/a	n/a	0	100	0

Mutation	Drug	Susceptibility conclusion	Total samples	Detection rate of mutation (%)					
				MGIT	Wayne'stest	MODS	Mykrobe	PhyResSE	KVarQ
Met306Val in EmbB	ETB	n/a	4	n/a	n/a	n/a	100	100	100
Met306Ile in EmbB	ETB	n/a	3	n/a	n/a	n/a	100	100	100
Gly406Ser in EmbB	ETB	n/a	2	n/a	n/a	n/a	0	100	0
Gly406Ala in EmbB	ETB	n/a	1	n/a	n/a	n/a	0	100	100
C491T in gene <i>rrs</i>	SM	n/a	4	n/a	n/a	n/a	0	0	100
A514C in gene <i>rrs</i>	SM	n/a	1	n/a	n/a	n/a	100	100	100
Lys43Arg in RpsL	SM	n/a	5	n/a	n/a	n/a	100	100	100
Asp94Gly in GyrA	FLQ	n/a	6	n/a	n/a	n/a	100	100	100

Mutation	Drug	Susceptibility conclusion	Total samples	Detection rate of mutation (%)					
				MGIT	Wayne'stest	MODS	Mykrobe	PhyResSE	KVarQ
Ser91Pro in GyrA	FLQ	n/a	1	n/a	n/a	n/a	100	100	100
Ala90Val in GyrA	FLQ	n/a	3	n/a	n/a	n/a	100	100	66.7
A1401G in gene <i>rrs</i>	AG	n/a	8	n/a	n/a	n/a	100	87.5	87.5

Table 5.7 Rate of prediction of resistance for susceptibility-associated mutations found in this study Minor alleles are not considered here.

Mutation	Drug	Susceptibility conclusion	Discrepancy		
			Method	Sample(s)	Observations
His51Arg in PncA Ala381Val in RpsA	PZA	R	Wayne's test (S)	MP313	Mixed lineage, PZA heteroresistance, "mixed coverage" by KVarQ
				MP367	PZA heteroresistance; "mixed coverage" by KVarQ
				MP175	
Gln10Arg in PncA	PZA	R	Wayne's(S)	MP078, MP177, MP342, MP249	
				MP364	"mixed coverage" by KVarQ
			MODS (S)	MP123, MP170	
			KVarQ (S)	MP247, MP342	
His71Arg in PncA	PZA	R	Wayne's(S)	MP369	
Gln10Pro in PncA	PZA	R	n/a	n/a	
Ile6Ser in PncA Arg212Arg in RpsA	PZA	R	n/a	n/a	
His57Leu in PncA	PZA	R	n/a	n/a	
Arg212Arg in RpsA	PZA	S	MODS (R)	MP035, MP176	
	PZA	S	MODS, KVarQ (R)	MP027, MP202	

Mutation	Drug	Susceptibility conclusion	Discrepancy		
			Method	Sample(s)	Observations
Lys48Thr in PncA			KVarQ (R)	MP276, MP320	
Pro62Ser in PncA Arg212Arg in RpsA	PZA	S	MODS, KVarQ (R)	MP027	
				MP050	“mixed coverage” by KVarQ
No mutation in pyrazinamide-associated genes	PZA	S	Wayne’s(R)	MP92	
			MODS (R)	MP111, MP248, MP268	
Asp435Val in RpoB	RIF	n/a	MODS (S)	MP027	
Asp435Tyr in RpoB	RIF	n/a	MODS (S)	MP013	
Leu430Pro in RpoB	RIF	n/a	MODS (S)	MP019	
Ser450Leu in RpoB	RIF	n/a	KVarQ (S)	MP031	RIF heteroresistant; “mixed coverage” by KVarQ
His445Leu in RpoB	RIF	n/a	n/a	n/a	n/a
Leu452Pro in RpoB	RIF	n/a	MODS (S)	MP050	
Ile491Phe in RpoB	RIF	n/a	Mykrobe, KVarQ (S)	MP054	“mixed coverage” by KVarQ
No mutation in rifampicin-associated genes	RIF	n/a	MODS (R)	MP005	
Ser315Thr in KatG	INH	n/a	KVarQ (S)	MP027	

Mutation	Drug	Susceptibility conclusion	Discrepancy		
			Method	Sample(s)	Observations
<i>inhA</i> promoter mutation -15 C>T)	INH	n/a	MODS (S)	MP019, MP032, MP040, MP175, MP268, MP294	
				MP031	"mixed coverage" by KVarQ
				MP313	Mixed lineage, INH heteroresistant, "mixed coverage" by KVarQ
Ser315Thr in KatG and <i>inhA</i> promoter mutation -15 (C>T)	INH	n/a	Mykrobe (S for <i>inhA</i> promoter mutation; R for Ser315Thr in KatG)	MP063	
No mutation in isoniazid-associated genes	INH	n/a	n/a	n/a	n/a
Tyr319Ser in EmbB	ETB	n/a	PhyResSE (R)	MP078	
Met306Val in EmbB	ETB	n/a	n/a	n/a	n/a
Met306Ile in EmbB	ETB	n/a	n/a	n/a	n/a
Gly406Ser in EmbB	ETB	n/a	PhyResSE (R)	MP042, MP50	
Gly406Ala in EmbB	ETB	n/a	Mykrobe (S)	MP360	

Mutation	Drug	Susceptibility conclusion	Discrepancy		
			Method	Sample(s)	Observations
C491T in gene <i>rrs</i>	SM	n/a	KVarQ (R)	MP028, MP190, MP203, MP248	
A514C in gene <i>rrs</i>	SM	n/a	n/a	n/a	n/a
Lys43Arg in RpsL	SM	n/a	n/a	n/a	n/a
Asp94Gly in GyrA	FLQ	n/a	n/a	n/a	n/a
Ser91Pro in GyrA	FLQ	n/a	n/a	n/a	n/a
Ala90Val in GyrA	FLQ	n/a	KVarQ (S)	MP241	FLQ heteroresistant, "mixed coverage" by KVarQ
A1401G in gene <i>rrs</i>	AG	n/a	Mykrobe (R)	MP186	Mixed lineages, AG heteroresistant, "mixed coverage" by KVarQ

Table 5.8 Performance of drug susceptibility testing methods for Pyrazinamide relative to BATEC 960 MGIT PZA assay

Testing method	Sensitivity relative to BACTEC (%)	Specificity relative to BACTEC (%)
MODS	91.3	73.9
Wayne's test	60.9	97.8
PhyResSE	100	100
KVarQ	91.3	89.1

5.4 Discussion

This study looks at three topics using this single sample set: The mutations that underlie pyrazinamide resistance and performance of methods for

determining resistance to the drug; performance of the MODS method, and performance of sequence-based drug susceptibility prediction tools. In doing so, it has assessed strain heterogeneity in samples in terms of mixed lineages and heteroresistance in DST.

5.4.1 Mutations in pyrazinamide resistance genes

Owing to the unclear relationship amongst pyrazinamide phenotypes and genotypes, consideration of which samples were resistant to the drug was necessary before evaluating the testing and prediction methods. In doing this, the study presented here has defined the role of mutations in pyrazinamide resistance-genes by their susceptibility to the drug. The study recognised the strong association between MGIT resistant samples and resistance-causing mutations that were included in the drug resistance databases of sequence-based prediction tools. Novel allelic substitutions associated with MGIT-determined pyrazinamide resistance were also identified and the previous observation, that some mutations in genes associated with pyrazinamide resistance are not causative of resistance to the drug.

5.4.1.1 Resistance-causing mutations

In this study, 24 of 69 samples (34.8%) were considered PZA resistant. Resistance was caused by the mutations, His51Arg, Gln10Arg, His71Arg, Gln10Pro, Ile6Ser and, His57Leu in PncA. Mutations have been previously observed at all of the amino acid positions however, the substitutions of His57Leu and Ile6Ser in PncA have not been previously observed. These mutations were identified by KVarQ and PhyResSE, as well as by SNV-calling, indicating the ability of these tools to remain 'open-minded' about allele substitution. Some of these resistance-causing mutations were observed with additional mutations in pyrazinamide resistance genes.

Reference to the mutation, His57Leu in PncA could not be found in the literature but many other allele substitutions at this position have been identified. In *M. tuberculosis sensu stricto*, a number of amino acid changes occur resulting in glutamine, proline, asparagine, aspartate or tyrosine

residues, all with a 0.985 probability of being PZA resistant (327). Codon 57 is essential in binding of the Fe²⁺ metal ion; site-directed mutagenesis of *M. tuberculosis* to introduce alanine in place of wild-type histidine led to suppression of metal binding and reduced enzymatic activity from 81.9 U/mg protein in the wild-type to just 0.5 U/mg protein in the mutant (332). In the intrinsically pyrazinamide resistant species, *M. bovis*, an aspartate residue takes the equivalent place to His57, and both the His57Ala and *M. bovis* PZAse have comparable enzyme activities and metal ion binding properties (332). Clearly substitution at this codon with a number of amino acids, with various side chain properties, has a high likelihood of causing pyrazinamide resistance. Histidine has a positively charged side chain, whilst, like alanine and proline, leucine is a non-polar amino acid. Similar to other substitutions described, leucine would likely disrupt the metal binding properties of the enzyme, rendering it resistant to pyrazinamide. The His57Leu mutation was concluded to be associated with PZA resistance and was concordantly predicted to cause drug resistance by all DST methods.

The mutation, Ile6Ser in PncA has not been well described in the literature, though other allele substitutions at this amino acid position are better characterised (333). Codon six is included in the hydrophobic core of the PZAse protein. Amino acid substitution at this position is not an absolute indicator of resistance as substitutions of the non-polar wild-type residue, isoleucine to another non-polar residue, leucine, have been found in 120 PZA sensitive samples and only 8 PZA resistant samples (306). However, amino acid substitution to the polar residue, threonine, has been associated with a 0.98 probability of causing resistance, a 0.015 probability of sensitivity (327) and an MIC of 100-300 ug/mL (306). As serine, found here, is also a polar residue, it may also be likely to disrupt the hydrophobic interior of the protein which is concurrent with the absent PZAse activity as indicated by the Wayne's assay of this sample. The mutation was concluded to cause resistance and all DST methods were concordant on this susceptibility.

The His51Arg mutation in *pncA* was only ever seen here in combination with the Ala381Val mutation in RpsA. This latter mutation was only identified by SNP calling in the pyrazinamide resistance-associated genes and, though the

mutation was identified by PhyResSE, the software tools did not report this mutation as causing resistance. The samples with this combination of mutations all fell in the same sublineage, 4.3.3, and share common resistance profiles for all drugs. The mutations may be observed together due to the derivation of the strains from a common ancestor. The Ala381Val mutation in RpsA hasn't been described in the literature and it remains unclear whether this represents a second resistance-causing mutation, a helper mutation or is irrelevant to pyrazinamide resistance.

The Arg212Arg mutation in RpsA was observed in addition to the Pro62Ser and Ile6Ser mutations in PncA but it was also found as the only mutation in a pyrazinamide associated gene in other samples. It is discussed in greater depth in section 5.4.1.2.

5.4.1.2 Other resistance-associated mutations

Beyond samples with a resistance-causing mutation, a further eleven samples in this study had other mutations in genes associated with pyrazinamide resistance, *pncA* and *rpsA*. The role of these mutations was evaluated with view of the predictions made by the sequence-based tools, the phenotypic susceptibility results and information provided in the literature.

In the study presented here, four samples had a Lys48Thr mutation in PncA. These were phenotypically sensitive by the BACTEC MIGIT PZA assay and Wayne's test but two of the four samples showed resistance by the MODS assay. This mutation has been observed in previous studies of Peruvian tuberculosis isolates and infrequently elsewhere throughout the world (328,334,335). Conclusions vary between studies as to the degree of susceptibility or resistance-associated with this mutation. The Wayne's assay reported susceptible results with this mutation, however, strains with this mutation had markedly reduced PZAse activity at 10.45 mmol POA /min/mg PZAse relative to 38.4 mmol POA /min/mg PZAse in the wild-type (329,336). In PZA-broth, samples with the Lys48Thr mutation were able to grow to 20% of the wild-type growth level, rendering the strains resistant by this metric (336). The MIC has previously been determined as less than or

equal to 50 $\mu\text{g}/\text{mL}$, at the upper end of the 8-64 $\mu\text{g}/\text{mL}$ epidemiological wild-type cut off (329,337). Whilst this is below the 100 $\mu\text{g}/\text{mL}$ critical concentration of PZA used in the MGIT assay, explaining the sensitive MGIT result, this may not be below the clinical exposure levels in patients. Work by Gumbo *et al* (338) showed that in patients of 48 kg body mass receiving 2 g pyrazinamide per day, the standard dosage, the breakpoint MIC was 37.5 $\mu\text{g}/\text{mL}$ at pH 5.8 or 50 $\mu\text{g}/\text{mL}$ at pH 6.0; but that the breakpoint fell to as low as 25 $\mu\text{g}/\text{mL}$ (pH 6.0) with a 2 g dose as body mass increased. This suggests that a standard drug regimen, especially for a heavier patient, would not be sufficient to treat infection with TB strains that have a Lys48Thr mutation in the *pncA* gene, though increased dosage may be beneficial. In instances such as these, the strain could be said to have intermediate sensitivity to pyrazinamide, in accordance with the SIR classification (Sensitive/Intermediate/Resistant) of susceptibility (339). Whilst MGIT resistance is thought to indicate true resistance, it may fail to detect low-level resistance given the high critical concentration of PZA that it uses (329). MIC data for the other mutations found in MGIT sensitive samples are not available meaning that the impact of these mutations on clinical susceptibility may not be predicted.

One isolate carries a Pro62Ser mutation in PncA and appears to be sensitive by the MGIT and Wayne's assay and is resistant only by MODS. Some isolates with a Pro62Arg substitution have previously been defined as phenotypically sensitive, whilst other samples with this mutation have been determined to be resistant (327). Other substitutions, Pro62His and Pro62Leu in PncA, are exclusively associated with resistance and found to have MICs of >800 $\mu\text{g}/\text{mL}$ and 400 $\mu\text{g}/\text{mL}$, respectively. Both show an absence of PZAse activity in the Wayne's assay (336,340). Clearly some mutations at this position can cause high levels of PZA resistance. However, without full characterisation of the PZA phenotype in terms of MIC, it remains uncertain whether this novel mutation is associated with clinical resistance.

The synonymous Arg212Arg mutation in RpsA was observed exclusively in samples of the 2.2.1 sublineage and in all of the samples in this sublineage. It was not reported to cause resistance by any of the sequence-based resistance

prediction tools though the mutation was identified by PhyResSE. Instead PhyResSE indicated that the mutation defined a lineage (341). This mutation was observed by SNP-calling in both MGIT sensitive and resistant samples. In the former, it was the only mutation in the pyrazinamide resistance-associated genes or in combination with the Pro62Ser mutation in PncA. In the resistant sample, it was found combination with the Ile6Ser mutation in PncA which was thought to confer the observed resistance, as discussed above. The presence of this synonymous mutation clearly does not determine pyrazinamide susceptibility.

The role of some mutations to cause resistance to pyrazinamide is clear but for other mutations, their role in pyrazinamide resistance is less well defined. Further investigation into these mutations, including correlation with MIC data and clinical outcomes with a variety of lineage backgrounds, is necessary to determine their role in the SIR classification of resistance and how to treat strains with these mutations in the clinic (339). The mutations that were considered to underlie PZA resistance were all found in PncA, not in other genes. This supports a previous suggestion that a PCR–single-strand conformational polymorphism method may be used in this setting for detection of PZA resistance, given that the WHO aim is to identify 90% of resistance (342). This would be advantageous because it is relatively cheap, rapid and reliable and might facilitate correct PZA usage in the drug regimen from start of therapy (343).

5.4.2 Performance of pyrazinamide susceptibility detection methods

Determination of resistance to pyrazinamide has been plagued by the poor performance of susceptibility testing methods. This study sought to explain some of the limitations of these methods with view of the mutations underlying the resistance. As expected, the methods employed to determine pyrazinamide resistance in this study varied in their performance characteristics. Patterns of resistance misclassification were identified associated with particular mutations for some methods.

5.4.3.1 MGIT

The MGIT assay has been considered as the gold standard approach here for determination of PZA susceptibility because mutations deemed likely to cause resistance, His51Arg, Gln10Arg, His71Arg, Gln10Pro, Ile6Ser and, His57Leu in PncA, have been found in every sample that was resistant by the MGIT assay and never in those that have been determined as sensitive. Mutations were identified in samples deemed MGIT sensitive and, as discussed with regards to the mutations Lys48Thr and perhaps Pro62Ser in PncA, these mutations may confer low level or intermediate resistance that is not detected by the MGIT assay.

5.4.3.2 Wayne's test

The Wayne's assay falsely identified resistance in only one of the samples in this set, giving it a 97.8% specificity. It failed to identify nine truly resistant samples, making the sensitivity only 60.9%. This problem with sensitivity occurred with five of eleven instances of the Gln10Arg mutation in PncA, one of two occurrences of the His71Arg mutation in PncA and three of seven occurrences of the His51Arg mutation in PncA coupled with R381V in RpsA. The impact of these mutations on PZAse activity is unclear, however the Wayne's assay suggests the enzyme may have retained some activity in some of the samples with these mutations. Three of these false sensitive samples reported by the Wayne's assay contained mixed infections, though only one, MP313, showed heteroresistance relative to pyrazinamide; four other heterogeneous samples with homogeneous PZA resistance mutations were correctly identified by the Wayne's assay. The sample heterogeneity was not a consistent issue amongst the samples where the Wayne's test falsely reported sensitivity. Given few samples with other mutations found in this dataset, it is not clear whether this represents a significant association between these mutations and poor Wayne's test performance. A larger data set would be required to provide statistical power to validate this.

5.4.3.3 MODS

Using the MGIT assay as the gold standard for PZA DST, the sensitivity of MODS in this sample set was high, at 91.3%. For two of eleven samples with the Gln10Arg mutation in PncA, MODS did not detect resistance that was predicted by the other methods. There were particular issues with its specificity which was only 73.9%: MODS reported false resistance in 12 of 46 samples considered sensitive. MODS indicated resistance in two of four samples with the Lys48Thr mutation in PncA, a mutation that was not associated with resistance by any of the sequence-based tools or MGIT. The mutation was seen in these samples in the absence of identifiable sample heterogeneity. In samples containing this mutation, growth in PZA-broth has been found to be comparable to only 20% of the growth in PZA-free broth (336) and it may convey intermediate resistance. Here, this may have been enough growth to have indicated resistance by MODS in these two samples. MODS also falsely predicted resistance in the one sample with a Pro62Ser mutation in PncA and Arg212Arg mutation in RpsA, and one of six samples with only the Arg212Arg mutation in RpsA. For these other mutations that were associated with MGIT sensitivity but MODS resistance, phenotypic characterisation of MIC was not available. Given that resistance by MODS is determined by visualisation of microscopic growth of a strain in PZA-broth, it might identify intermediate resistance in samples where the MIC may be too low for resistance to be identified by MGIT. That said, MODS also identified seven of 35 samples without mutations to be pyrazinamide resistant. The poor apparent specificity of PZA MODS as used in this study is a limiting factor to its widespread use for clinical DST. This is unfortunate owing to the low cost of testing with this method but with a better understanding of the clinical susceptibility of strains, the role of MODS in the landscape of PZA resistance detection might be different in the future. Used in combination with the Wayne's test, the two methods would only concordantly identify 12 resistant isolates though this approach would omit all 'false resistance' indicated by MODS. Neither of these approaches, individually or in combination, provide acceptable rates of accurate PZA detection should MGIT resistance be considered true resistance.

5.4.3.4 Sequence-based tools

Collectively, the sequence based tools predicted all of the mutations considered to cause resistance to pyrazinamide. PhyResSE was the only individual tool to offer complete concordance with the MGIT assay which was considered determinant of certain resistance. Mykrobe did not predict resistance to pyrazinamide. KVarQ did not predict resistance from Gln10Arg in PncA in the samples MP247 and MP342, the latter of which was also predicted to be sensitive by the Wayne's assay, yet it did predict resistance caused by this mutation in other samples. Other methods were also inconsistent with determining resistance in samples with this mutation, however each method predicted sensitivity associated with this mutation in a different sample. The algorithms for detection of mutations in a sample are variable across the tools and it seems likely that the KVarQ algorithm excluded the mutations in MP247 and MP342 as seen here. Algorithms need to be improved in such a way to optimise sensitivity and specificity of the predictions that they make.

Mutations other than those associated with MGIT resistance could also be detected. KVarQ predicted resistance in samples MP276, MP320, MP027 and MP202 caused by the Lys48Thr mutation in PncA, unlike the other predictions tools. On evaluation in this study, this mutation was considered to potentially be associated with a resistant or intermediate pyrazinamide susceptibility phenotype. KVarQ also predicted resistance in the two samples, MP027 and MP050, caused by Pro62Ser in PncA, in agreement with the MODS phenotypic method. Similarly, the role of these mutations in PZA resistance is unclear. As discussed previously, further evaluation of the role of these mutations in resistance is necessary and will improve the performance of sequence-based pyrazinamide prediction through more appropriate inclusion of mutations in databases. SNV-calling determined further mutations, including those in RpsA that were not predicted to cause resistance by any of the tools and not considered to cause resistance after evaluation of the literature. Given that these mutations are present in resistance-associated genes, the exclusion of these mutations from the

databases used to call resistance is both necessary and occurring appropriately, at least in view of this dataset.

5.4.3 Evaluation of MODS for susceptibility prediction in rifampicin and isoniazid

The MODS assay was conducted directly on broth cultures directly from decontaminated sputum and should represent some of the diversity of TB strains within the sputum sample, though this is biased by culture. Whole genome sequencing of multiple colony sweeps across solid media culture from each sample was used to analyse some of the diversity within the samples. By comparing the results from MODS and this WGS approach, the capacity for MODS to characterise resistance, including heteroresistance, in samples could therefore be evaluated, though a bias may have been introduced during solid media growth that was also conducted prior to WGS. The performance of PZA MODS has already been evaluated in section 5.4.3.3 so this section of the thesis will focus on RIF and INH MODS. For all drugs evaluated, RIF, INH, and PZA, MODS detected resistance in all but one sample where heteroresistance was apparent. It only failed to detect the heteroresistance caused by the *inhA* promoter mutation in MP313 which seems to be a problem with identifying resistance caused by that mutation rather than heterogeneity, as discussed below.

There were five instances of discordance between MODS and the results of any sequence-based DST method for rifampicin resistance. In five samples, MODS did not detect resistance but the resistance-causing mutations, Leu430Pro, Asp435Tyr, Asp535Val and Leu452Pro, were identified by all of the software tools. This discordance between MODS and WGS-based methods for high confidence resistance-associated mutations indicates that MODS lacks sensitivity. This may not be caused by the mutation in the sample since these mutations are also seen in other samples that were correctly identified to be resistant by MODS. In sample MP005, no rifampicin resistance-associated mutation was identified but MODS detected rifampicin

resistant. As the gold standard method, MGIT testing, is not available in this location the phenotypic susceptibility results cannot be verified.

For most samples, there was concordance between INH MODS and predictions made by the sequence-based tools. MODS consistently determined that the samples were sensitive to isoniazid where INH resistance-conferring mutations were not reported. MODS failed to recognise resistance in eight out of nine samples that had only the *inhA* promoter mutation in isoniazid resistance-associated regions, including sample MP313 which was heteroresistant. The one sample where resistance was reported by MODS had the same lineage background as other samples with this mutation that were reported to be sensitive by MODS. Walker *et al* previously found that 73 of 79 isolates with this mutation were phenotypically resistant according to MGIT analysis (224), supporting the role of this mutations in causing resistance. This mutation has previously been associated with low-level or intermediate isoniazid resistance when using other drug susceptibility testing methods particularly when compared to resistance caused by the Ser315Thr mutations in KatG (344–348). Though a standard, 0.4 $\mu\text{g}/\text{mL}$ concentration of isoniazid was used for MODS, it may be above the minimum inhibitory concentration (MIC) associated with the *inhA* promoter mutation. Many studies have used both 0.1 and 0.4 $\mu\text{g}/\text{mL}$ isoniazid for MODS or simply the lower concentration to determine INH resistance (272,274–276,349). Moore *et al* (274) concluded that use of only 0.4 $\mu\text{g}/\text{mL}$ isoniazid did not reduce performance of the DST method in a study conducted in Peru, however they did not identify the mutations conferring resistance in these strains. When the higher concentration was used by Huang *et al* (281), many samples were sensitive by the MODS assay but resistant by the proportion method that uses only 0.2 $\mu\text{g}/\text{mL}$ INH. The sample with both the Ser315Thr mutation in KatG and the *inhA* promoter mutation was phenotypically resistant in this study, concordant with previous assertions that the presence of both mutations results in a higher MIC (344). This study indicates that MODS with 0.4 $\mu\text{g}/\text{mL}$ isoniazid may fail to identify samples with the *inhA* promoter mutation causing intermediate resistance to isoniazid. This is supported by results of a meta-analysis of MODS sensitivity and specificity that found lower MODS sensitivity when using 0.4 $\mu\text{g}/\text{mL}$

isoniazid than 0.1 $\mu\text{g}/\text{mL}$ (278). This mutation is also associated with, though not predictive of, ethionamide co-resistance (345).

5.4.4 Evaluation of sequence-based tools for prediction of drug susceptibility

5.4.4.1 All whole genome sequencing analyses

Whole genome sequencing and the combination of PhyResSE, Mykrobe and KVarQ for rifampicin, isoniazid and pyrazinamide susceptibility prediction collectively predicted all but one instance of resistance to the drugs. The one sample, MP005, that was RIF resistant by MODS but did not show any mutations may have been due to poor MODS specificity as rifampicin resistance-associated mutations are well-described. Alternatively, the possibility exists that further mechanisms for rifampicin resistance are yet to be discovered. WGS was able to identify well-described resistance-conferring mutations in MODS sensitive samples in four instances for rifampicin and eight instances for isoniazid. Resistance-causing mutations for isoniazid were otherwise concordant with the susceptibility phenotypes offered by MODS. Relative to detection of resistance by MGIT, WGS approaches were able to collectively detect all mutations that were conferred resistance to PZA. It is not clear how the mutations in MGIT sensitive samples might relate to clinical drug susceptibility. A better understanding of the genotype-phenotype relationship related to these mutations is necessary for conclusive WGS-based diagnosis.

Resistance prediction based on whole genome sequencing has the advantage of determining susceptibility to all drugs with a single test, including for little-evaluated second-line drugs. Ten XDR or pre-XDR strains were predicted here using the combined results of the sequence-based tools. Drug susceptibility prediction of second-line antibiotics is limited by the unclear relationship between genotype and phenotype and the predictions made here could not be validated as no phenotypic DST was available. These pre-XDR and XDR strains would not have been identified in the many locations that lack DST for second-line drugs. In the Peruvian setting, only 48% of the

confirmed rifampicin-resistant *M. tuberculosis* is tested for resistance against second-line drugs. It appears that resistance characterisation beyond rifampicin testing improves the rate of treatment success in Peru given that treatment succeeds for only 35% of RR/MDR-TB patients yet 65% of XDR-TB patients for whom the diagnosis would have been confirmed (269). As such, identifying resistance in patients is essential for treating them with the appropriate drugs in the appropriate regimen and, in the case of pre-XDR TB monitoring them for progression of the strains to further drug resistance. Whole genome sequencing could offer the necessary complete profiling of drug resistance in settings such as this one in Peru, progressing the diagnosis and treatment capacity within the setting if implemented.

5.4.4.2 Individual sequence-based tools

Individually, no WGS-based tool predicted all of the resistance-conferring mutations that were reported collectively.

- Mykrobe does not report resistance to PZA, with the authors citing that all mutations relating to this drug resistance being poorly predictive (212). It did not report both mutations related to INH resistance found in sample MP063; the *inhA* promoter mutation was not identified in this sample only but the Ser315Thr mutation in KatG was reported. It also did not identify the resistance-causing Ile491Phe mutation in RpoB. The Gly406Ala, Gly406Ser and Tyr319Ser mutations in EmbB and the C491T mutation in the *rrs* gene were reported to cause resistance by the other sequence-based tools but not Mykrobe. It is unclear why these omissions were made. Mykrobe features an unusual way of mutation reporting, with the replacement amino acid in a substitution denoted by X rather than the specific amino acid. It also reports 'minor alleles' where a few or often one read amongst those reported has a resistant allele but the clinical significance of this is not discussed explicitly in either the report or the publication describing the tool.
- PhyResSE predicted all mutations that caused resistance to rifampicin, isoniazid and pyrazinamide, aside from in one sample with a very small proportion of reads causing heteroresistance as discussed above. It did not predict resistance against second-line drugs in all samples as

identified by other tools, including the C491T mutation in gene *rrs* and one of eight instances of the A1401G mutation in gene *rrs*. This latter sample was heteroresistant at the 1401 nucleotide in *rrs* with only three of 35 reads with the resistant allele. There were a number of mutations for which PhyResSE was the only method to identify the mutation. This may be a reflection of over-prediction of resistance-associated with mutations that do not confer phenotypic resistance. It could be that Mykrobe and PhyResSE are failing to detect resistance-conferring mutations. Nonetheless, it goes to show the importance of a clear understanding of genotype and phenotype yet again.

- KVarQ failed to identify mutations in quite a list of samples: two of 11 samples with the Gln10Arg mutation in PncA; one of 14 samples with the Ser450Leu mutation in RpoB though this sample was heteroresistant; the only sample with the Ile491Phe mutation in RpoB; one of 37 samples with the Ser315Thr mutation in KatG; all 11 samples with the Tyr319Ser mutation in EmbB and both Gly406Ser mutations in EmbB (though it did identify the mutation Gly406Ala in the same gene); one of three samples with the Ala90Val mutation in GyrA though this sample was heteroresistant and one in eight samples with the A1401G mutation in gene *rrs* which was also heteroresistant. For some of these omitted samples, the mutation appears to have not been predicted due to heteroresistance at the given allele, for others the mutation may not be present in the underlying database. KVarQ did predict resistance associated with the Lys48Thr and Pro62Ser with Arg212Arg mutations in PncA.

Whilst it appears that all resistance-associated mutations in this small sample set have been identified, the databases of resistance mutations and algorithms used by each tool to determine resistance are inconsistent. Inclusion or omissions of mutations in the database is more problematic for those associated with resistance to pyrazinamide, ethambutol and the other second-line drugs identified. This is probably owing to the complicated picture of susceptibility for these drugs where the presence of a mutation may, or may not, indicate clinical resistance. Although the algorithms used to call SNPs occasionally impact the resistance mutations that are identified, this mainly

affects the numbers of reads associated with each resistance allele. This has a notable impact on the identification of heteroresistance, particularly by KVarQ that failed to identify any resistance in three of six heteroresistant samples. As the understanding of the genotype-phenotype relationship of drug susceptibility improves, our ability to compile databases with more clinically impactful predictions will also improve. Software that reports well-validated MIC and dosage recommendations based on the mutations and linked to clinical outcome data identified is feasible and may hopefully soon be possible. This could open up opportunities for treatment that could otherwise be diminished by the binary resistant/susceptible approach to drug susceptibility, especially benefitting patients with predicted XDR-TB.

5.4.4.3 Characterisation of heterogeneity using sequence-based tools

Mixed cultures or clinical samples potentially contain multiple strains of *M. tuberculosis* with heterogeneity in their lineage, drug resistance profile or elsewhere throughout their genomes though the latter is insignificant to diagnosis. KVarQ identified 14 samples with 'mixed coverage' which included samples that have mixed lineages, mixed drug susceptibility profiles and mixed alleles in other parts of the genome that have no impact on either lineage or drug resistance. Neither PhyResSE nor Mykrobe offered any indication of mixed lineages. MGplacer was used specifically to identify mixed lineages and found 11 samples with heterogeneous lineages and specified the lineages present. This study is the first to use MGplacer in contemporary samples (206,210). Only Mykrobe and PhyResSE predicted heteroresistance as KVarQ did not report allele frequencies but all of the samples with heteroresistance were identified by KVarQ to have 'mixed coverage'. Seven heteroresistant samples were identified. Mykrobe failed to detect heteroresistance associated with pyrazinamide in one sample since it does not predict resistance to this drug. PhyResSE failed to detect heteroresistance in one distinct sample. For this, it seems likely that the frequency of resistant alleles was too low to be reported using the tool's prediction algorithm since Mykrobe reported only three of 25 reads were contained the resistant allele.

Whole genome sequencing allowed for characterisation of mixed lineage samples with very low ratios of minor to major strains using MGplacer but not such low ratios of resistant to sensitive reads at heteroresistant alleles. MGplacer defines each lineage with multiple SNPs; for heteroresistance was determined based on one mutation position. It would be expected that a greater depth of sequencing per sample would result in identification of heterogeneity in further samples that contain mixed infections, especially for samples with lower proportions of the minor strain(s). 'Deep-sequencing' of mixed TB strains to a very high and very expensive depth of coverage has previously been shown to detect minor strains that comprised 0.64% to 35.83% of the population in the sample (350). Such deeper sequencing might help to explain the presence of low frequency drug resistance alleles detected in many samples by Mykrobe, particularly associated with rifampicin resistance. These samples tend to have only one read with the resistant allele whilst the others are susceptible. Such results may occur due to biases or sequencing errors or they may represent the true diversity of strains within the sample that would be more apparent with deeper sequencing (351). Should the latter be the case, these low frequency resistance alleles might be amplified during first-line drug treatment (165). In the publication introducing the Mykrobe software package, Bradley et al. mention 'minor alleles' and their useful predictive value for determining clinically relevant resistance (212). They do not define the allele frequency that constitutes a 'minor allele', however, so it is unclear what the relevance of these results are, as Bradley et al. determine it.

5.4.5 Conclusions

Going forwards, this study highlighted a number of areas where further research and considerations should be made. First, a better understanding of the phenotypes and clinical outcomes associated with the Lys48Thr and Pro62Ser mutations in PncA is required to confirm their place in the sensitive-intermediate-resistant classification of susceptibility. In turn, this will inform an improved understanding of the performance of all of the pyrazinamide susceptibility determination methods employed here, particularly MODS

which appeared over-sensitive relative to MGIT and the sequence-based tools for which the performance could easily be improved given modifications to the databases underlying them. The performance of MODS for detecting isoniazid resistance was compromised in samples with the mutation in the *inhA* promoter region and might be improved with reconsideration of the critical concentration of INH used in this assay. Use of samples with multiple heterogeneous *M. tuberculosis* strains had no consistent impact on the phenotypic DST methods. Heterogeneity negatively influenced predictions by KVarQ which should be addressed with alternative algorithms for prediction of mutations. Mykrobe would be improved by inclusion of pyrazinamide prediction, a clear explanation of the basis for 'minor alleles' and their impact and reporting of allele substitutions. All tools would be improved with a clinically and phenotypically validated database of mutations, particularly for predicting ethambutol, streptomycin, fluoroquinolone and aminoglycoside resistance. The presence of multiple lineages in a sample was not reported by any of the sequence-based tools, though 'mixed coverage' was described by KVarQ, yet doing so would be advantageous to epidemiological analysis and surveillance efforts.

6. Discussion

6.1 Genomics in the diagnosis of tuberculosis

Tuberculosis is the world's leading cause of death from an infectious disease(1). It infected an estimated 10.4 million people in 2016 alone (1). Control of the disease epidemic is threatened by both the vast number of infected individuals and antibiotic resistance that is being acquired into strains and spread between people. There are already people infected with strains resistant to so many antibiotics that they are incurable within their national tuberculosis programme and potentially with any anti-tuberculosis therapy available anywhere in the world (97,110–112). Halting and reversing this epidemic is the mission set out by the END TB programme and is of such importance that it is integrated into the Sustainable Development Goals. Achieving the aims of a 90% reduction of deaths and an 80% reduction of incidence by 2030 relative to 2015 is dependent on improving diagnosis of tuberculosis and getting more people into effective treatment regimens (1). This requires diagnostic tools with improved technical capacity to diagnose TB that will be accessed by more people who have the disease (216). Access to diagnostic tools is greatly linked to national development and income status. In the most developed and wealthiest countries, the future objective is towards developing and using tools with the ultimate technical capacity for diagnosis. Recently this has begun to include whole genome sequencing (WGS) (219–222,238,352). In low income countries, compromises are made in diagnostic capacity and outcome to fit with lower financial budgets (129). Here, future objectives tend towards detecting disease amongst as many people as possible that have suspected active infection and offering DST to the individuals most at risk of drug resistant tuberculosis. Diagnostic innovation is needed in both settings to meet the diagnostic goals and ultimately, the goals in global health.

Throughout this thesis, my primary aim has been improving TB diagnosis by reaching ultimate technical performance of genomics as a diagnostic tool, ideally using unbiased direct sequencing of sputum. In the duration of this

PhD, the role of genomics in TB diagnosis has been developed from infancy to clinical reality and is seeking evaluation for endorsement by the World Health Organisation (1). In September 2013, a number of studies had used whole genome sequencing of solid-cultured isolates in research. The exciting new development was the proof-of-principle that sequencing could be undertaken in a much quicker time frame using early positive liquid culture to characterise drug resistance and epidemiological type of the MTBC strain (238). Since then, a number of improvements have been made to the technical capacity of clinical genomics, addressing limitations in

- Sputum processing and DNA extraction
- Genotype-phenotype relationship associated with drug resistance
- Detecting and characterising genetic heterogeneity in clinical samples
- Sequencing platform performance
- Bioinformatic tools

Now, whole genome sequencing is used in well-resourced reference laboratories for retrospective species identification and epidemiological analyses. It has been used prospectively to also achieve drug susceptibility testing predictions (219,222). On-going efforts seek to further address these areas of limitations.

6.1.1 Direct sequencing of tuberculosis

Obtaining sufficient mycobacterial DNA from sputum in order to sequence the genomes of MTBC in a clinically relevant time frame has been a major challenge for integration of genomics into prospective diagnosis of tuberculosis. WGS is typically undertaken after selective culture of the microorganisms of interest, but owing to the slow replication rate of mycobacteria, this is prohibitively time-consuming when diagnosing TB. Instead, sequencing would ideally be conducted using DNA obtained from sputum in a 'direct' manner, circumventing the need for MTBC culture.

The study presented in chapter three was the first published to undertake DNA extraction and genomic sequencing in a direct manner for the diagnosis of tuberculosis (353). Low coverage of the MTBC genomes was obtained from smear-positive sputum. This was sufficient to detect TB and to identify the

species and lineages of the aetiological mycobacteria but did not offer the depth of sequencing necessary for drug susceptibility testing.

A subsequent publication by Brown *et al* (354), described specifically capturing *M. tuberculosis* DNA using biotinylated RNA baits from SureSelect (Agilent, USA) from sputum. This resulted in 20 genomes from 24 smear-positive sputa with 90% genome coverage at $\geq 20\times$ depth that were suitable for drug susceptibility prediction; two low coverage genomes from two smear-negative culture-positive sputa and two high quality and five low quality genomes ($<1\times$ depth of coverage) from a total of eight smear-negative culture-negative sputa. Possible resistance-conferring mutations were identified in 88% of instances of phenotypic resistance to a drug and in 94% of instances of phenotypic drug susceptibility no known resistance mutation was identified. Detection of TB from smear positive sputa was attributed to dead bacilli, which may be likely when sputum is obtained from patients who are completing anti-TB treatment but it is not clear whether this was the case in this publication. The diagnostic approach they proposed took one week, less than any WHO-endorsed approach to susceptibility testing for drugs other than rifampicin, and cost US\$350 per sample at the time of their investigation. This same methodological approach was used more recently when prospectively diagnosing a patient. It provided clarity on inconclusive results obtained from LPAs and allowed the treatment regimen to be refined such that it might be better tolerated, less toxic, more likely to be completed and still effective (220).

Votintseva *et al* (355) demonstrated a faster and cheaper direct-WGS approach. They sequenced DNA obtained directly from smear-positive sputa after decongesting and decontaminating it, washing it with saline and removal of human DNA using the MolYsis Basic5 kit (Life technologies, USA). Sufficient reads were obtained from Illumina MiSeq sequencing to conduct DST in only 62% of samples giving results concordant with phenotypic DST. They estimated that this method could be achieved in 44 hours and cost £96 per sample with the Illumina MiSeq or in 16 hours, costing £196 per sample with the Illumina MiniSeq. They also demonstrated the potential to undertake the approach with the Oxford Nanopore Technologies (ONT) R9.4 MinION

sequencing platform in 12.5 hours from a mixture of culture-negative sputum DNA and *M. bovis* BCG DNA, though not directly from a clinical sample.

It remains to be seen whether tuberculosis may be diagnosed with a universal diagnostic approach that might also detect other pathogens in sputum. As observed in the recent studies summarised above, the limitation continues to be the small proportion of MTBC-derived reads amongst the abundance of read data obtained from other organisms, whether host or other microbes. It seems likely that in most sputum samples investigated in this study, the problem lies with the abundance of human DNA in the extracts as illustrated by the high proportion of human-derived reads in most samples in chapters three and four. Though in some samples an overwhelming majority of reads are neither derived from MTBC nor host DNA, this is not typical and may not reflect the composition of samples with a high proportion of human-derived reads, after human DNA has been depleted. Tentative evidence to support this comes from the results of samples sequenced by method B that employed decontamination of sputum prior to DNA extraction. Sequencing revealed a similar proportion of MTBC-associated reads as in samples extracted by methods E and F where the microbiome was not decontaminated and only a small amount of human DNA, if any, may have been depleted (according to the results of chapter four). It seems that there could be a high ratio of MTBC-derived DNA to DNA derived from other microflora in samples with typically high proportions of human DNA after that has been depleted. The MoLYsis approach used by Votintseva *et al* (355) fulfils the objective to deplete human DNA and may offer the opportunity to sequence only the microflora of a sputum sample, including the MTBC.

Whether sufficient coverage of the MTBC present could be obtained from microbiomes depleted of human DNA is unclear, as is the ability to detect other (co-)infections from a minimally sequenced metagenome. It would depend on the abundance of MTBC-derived reads relative to reads derived from the rest of the microflora. Certainly in the samples with high proportions of microflora other than tuberculosis, both here and in the work by Votintseva *et al* (355), the coverage of MTBC has been low, possibly too low for drug susceptibility predictions. Should depletion of human DNA alone not be

adequate to both characterise the microbiome and tuberculosis, partial enrichment of MTBC using an approach targeted amplification and/or MTBC capture, as used by Brown *et al* (354) may be more feasible. Deep metagenomic sequencing was recently used to prospectively rule out *Mycobacterium tuberculosis* as the cause of meningitis, instead identifying a helminth, *Taenia solium*, and after anti-helminthic treatment the patient's case improved (356). Metagenomic sequencing was undertaken from cerebrospinal fluid not sputum but it still marks development in this field. Further investigation of these sample processing methods should be undertaken, keeping in mind the lessons from chapter four about study design when intending to illustrate depletion of human DNA from small numbers of sputum samples, which is highly heterogeneous within the sample and between patients.

As yet no ideal sputum processing method has been developed that could rapidly and cheaply characterise tuberculosis directly from sputum (357), let alone achieve this with a universal diagnostic method. Significant progress towards these goals has been made and this will likely continue in the future (358).

6.1.2 Drug susceptibility testing from genomic sequence data

Molecular methods of drug susceptibility testing have long been plagued by our limited understanding of the genetic basis of clinical resistance for drugs other than rifampicin and isoniazid. Strong concordance between DST predictions and susceptibility phenotypes for both these drugs and pyrazinamide was demonstrated in chapter five. There were only a few cases of discordance between the genetic and phenotypic methods and these could be attributed to the performance of the MODS and/or Wayne's assays which are not the gold standard DST approach for any of these drugs. Ideally those discrepancies for RIF and INH should be assessed by phenotypic re-characterisation with the gold standard liquid culture assay. Novel codon substitutions in a relatively small number of PZA resistant strains indicate that there is yet more to understand about the basis of pyrazinamide susceptibility, as does the presence of the Lys48Thr mutation in PncA where

its role in susceptibility is not clear. Further studies to characterise resistance-associated mutations are needed for better understanding of their role in susceptibility.

Specific to genomic diagnosis, prediction of DST relies on variation detection algorithms and comparison to databases of known resistance mutations. When this PhD commenced in September 2013, this DST prediction relied on custom analysis pipelines and limited resistance databases (84,359) but now there are several open-access user-friendly available tools for this purpose (86,212,213,285,360,361). As illustrated in chapter five, the predictions made by these tools are highly variable as no common standardised prediction algorithms nor verified database of resistance mutations are available for use. The key issues here were the inclusion of particular mutations in the database used by the bioinformatic tool and their predicted role in drug susceptibility, their reporting of mutations with previously unreported amino acid substitutions, detection of mutations at sites displaying heteroresistance and low coverage samples. These tools also provided information about epidemiological types but did not identify strains in mixed lineage samples. The performance of the tools has recently been compared, coming to similar conclusions about their variability as presented here, further defining the sources of variations and making recommendations for development of the tools (346,362). This includes standardisation of bioinformatic methods, making the results from each tool comparable and reliance on resistance mutations with clearly understood clinical implications. A clinically validated database for interpretation of resistance-associated mutations is in development by the ReSeqTB initiative for this purpose (363). It must be maintained as new mutations will be discovered (223) and new drugs will be included in treatment regimens for which DST will be required. It is clear that WGS has a role in the future of clinical DST, indeed its clinical benefit has already been established (219–222,238,352).

6.1.3 Heterogeneity analysis using whole genome sequencing

Heterogeneity of tuberculosis in sputum may be derived from infection with multiple distinct strains of MTBC or diversification of a single strain that has evolved within the host. With the former scenario, mixed lineages of TB may be apparent and in both scenarios hetero-resistance, or heterozygosity at genomic sites known to be associated with drug resistance, may occur. This heterogeneity can be reflected in culture from sputum analysis of multiple isolates or from liquid culture that was inoculated with sputum. With any culturing step, bias in the strains and loss or gain of mutations may occur (229).

Much heterogeneity, both with mixed lineages and heteroresistance, was observed in the samples sequenced from Peru (chapter five) and may reflect high frequency of TB transmission within the study population. Phenotypic DST can detect resistance in as few as 1% of strains (148) and whilst it is clear that this cannot be achieved with the routine depth of sequencing of 20-100x genome coverage, the threshold for detecting true diversity in a sample is unclear. Heterozygosity may be observed even from 'clonal' isolates and may arise from sequencing bias (351) so it is unclear whether low frequency variation at resistance-associated sites is a true reflection of in-host heterogeneity. Analysis of heterogeneity was useful in this work as a lens with which to evaluate the diagnostic tools in the investigation whilst detecting this heteroresistance is also useful for patient management and for understanding whether resistance may have been transmitted or acquired in-host.

Strain typing in chapter five was achieved using phylogenetic placement such that strains with multiple lineages could be identified. Construction of phylogenetic trees and assessing transmission of TB between people, as ideally desired when using genomics for outbreak investigation or surveillance, is especially challenging in heterogeneous samples (364,365) owing to the dependence on SNPs. This may be circumvented with a Bayesian approach to phylogeny as demonstrated in a recent analysis of Ebola virus

transmission (366). Studies of mixed samples might also benefit from the longer reads available through ONT sequencing approaches as for other pathogens (367) or deep sequencing to fully resolve in-host diversity (350).

6.1.4 The future of diagnostic genomics

The use of genomics in diagnosis is now well established for strain typing and epidemiological analysis in reference laboratories and actionable drug susceptibility predictions in prospective analyses. There is still much room for improvement in direct sequencing methods, refinement of sequencing techniques, bioinformatic analyses and interpretation of results, as described in this thesis, but also plenty of hope that this can be achieved.

The WHO expects to assess ‘molecular sequencing’ for DST in 2018 (1). It is not certain whether this means whole genome sequencing or which kind of processing may be used to generate the DNA extract for sequencing. Nevertheless, it includes developing a list of high confidence mutations and assessing its accuracy relative to phenotypic methods, potentially as an alternative to them (1). It could be imagined that WGS might be integrated into a diagnostic algorithm in a similar manner to that described by Witney *et al* (368) in an intermediate-level or reference laboratory with highly developed infrastructure. Without significant automation of the process, highly skilled staff would be required to carry out the laboratory preparation of samples and undertake sequencing, which may potentially be deemed a high-level biosafety risk. A standardised diagnostic analysis pipeline and mutation database would be required and the role of newly discovered mutations in drug susceptibility would need to be determined with phenotypic DST (369). Sequencing from early positive MGIT cultures has produced results in a median of nine days in the hands of Pankhurst *et al* (370); direct sequencing produced results in five days in the hands of Nimmo *et al* (220) and this could potentially be reduced even further with the methods employed by Votintseva *et al* (355). At this stage, direct sequencing comes with compromises in sensitivity of diagnosis given challenges with reliably recovering the whole genome. The performance of genomics for the diagnosis of tuberculosis is variable, depending on the methods used.

The technical superiority of genomics has been narrowed somewhat in recent years. Until 2016, the only rapid DST available was for rifampicin and isoniazid susceptibility (155,159) and WGS provided DST for all other drugs quicker than any other method (222). Since the endorsement of the rapid line probe assays for FLQs and SLIDs (160), resistance to these drugs can now be achieved rapidly too. Similarly, the release of the GeneXpert Ultra assay (157) has raised the bar for sensitivity and specificity detection, and the LF-LAM assay has improved diagnosis of TB in HIV-positive individuals (138). These are undoubtedly significant developments but go to show how competitive genomic diagnosis must be in order for its performance to be worth the large investment in money and resources to implement this tool. It is certainly not one that will be available everywhere in the immediate future. Targeted genotypic tools may come to market earlier and cheaper.

6.2 Diagnosis in developing countries and the role of genomics

All of the work presented in this thesis was all conducted in collaboration with developing countries; at least a third of my research time was spent in The Gambia. There were issues, delays, communication issues and setbacks that would have been unlikely to occur had these studies been undertaken entirely in the United Kingdom, but that the work was carried in these settings is exemplary to the kind of cutting-edge research that is feasible in such locations. Testament to this, the MRC Unit, The Gambia is currently setting up a regional centre of excellence focusing on genomics and has produced a number of recent publications using genomics (208,371–375). It has been argued that genomics may be integrated into the diagnostic laboratory networks of other developing countries, should logistical, financial and ethical issues be circumvented (376). From my own experience that seems feasible in the not-too-distant future, not for routine diagnosis, but for TB outbreaks or unusual cases of antibiotic resistance against many drugs (264). Of course, there is a broad spectrum of countries that might be considered ‘developing’ and, even regions in the same country have vastly different

diagnostic capacities. In some settings, the high demands of TB genomics would not be feasible and diagnosis is currently challenged to offer any DST at all. There are plenty of diagnostic methods other than genomics, at lower cost and requiring fewer resources, that may make up the entirety of the diagnostic algorithm in lower income, lower resource settings, as discussed in this thesis. These provide highly beneficial diagnostic capacity but with some technical limitations. Innovation to understand these approaches, using studies such as the one presented in chapter five, provide the opportunity to improve them.

6.3 Conclusions

The work presented throughout this thesis has highlighted the inherent difficulties faced for diagnosing tuberculosis, both in countries where cutting-edge technology may be implemented and those that rely only on more traditional methods. Achieving the ultimate END TB goals to reduce the mortality and incidence of disease associated with tuberculosis depends on meeting targets of diagnosing 90% of people who develop TB then successfully treating 90% of them (342). Improvements to the technical capacity of diagnostic tools is critical for this. Despite recent developments documented by other researchers, it is clear that further work, such as presented in this thesis, remains necessary to unleash the full potential of genomics for diagnosing tuberculosis, to improve existing diagnostic tools and to create new ones.

References

1. World Health Organization. Global Tuberculosis Report 2017. 2017.
2. Daniel TM. The history of tuberculosis. *Respir Med*. 2006;100(11):1862–70.
3. Herzog BH. History of tuberculosis. *Respiration*. 1998;65(1):5–15.
4. Roberts CA, Buikstra JE. The bioarchaeology of tuberculosis: a global perspective on a re-emerging disease. Gainesville, Fl: University press of Florida; 2003.
5. Young DB. Blueprint for the white plague. *Nature*. 1998;393(6685):515–6.
6. Grigg ER. The arcana of tuberculosis with a brief epidemiologic history of the disease in the USA. *Am Rev Tuberc*. 1958;78(2):151–72.
7. Krause AK. Tuberculosis and public health. New York Tuberculosis and Health Association. 1927.
8. Laennec RT. De l'auscultation médiate: ou traité du diagnostic des maladies des poumons et du coeur. Vol. 2. Brosson; 1819.
9. Laennec RTH, Forbes J. A Treatise on the Diseases of the Chest, and on Mediate Auscultation. Samuel S. and William Wood; 1838.
10. Villemin JA. Études sur la tuberculose: preuves rationnelles et expérimentales de sa spécificité et de son inoculabilité. J.-B. Baillière et fils; 1868.
11. Budd W. The Nature and the Mode of Propagation of Phthisis. *Bristol Medico-Chirurgical J*. 1967;82(3):68.
12. Cambau E, Drancourt M. Steps towards the discovery of Mycobacterium tuberculosis by Robert Koch, 1882. *Clin Microbiol Infect*. 2014;20(3):196–201.
13. Koch R. Die aetiologie der tuberculose, a translation by Berna Pinner and Max Pinner with an introduction by Allen K. Krause. *Am Rev Tuberc*. 1932;25:285–323.
14. Wilson LG. The historical decline of tuberculosis in Europe and America: its causes and significance. *J Hist Med Allied Sci*. 1990;45(3):366.
15. Davies RPO, Tocque K, Bellis MA, Rimmington T, Davies. PDO.

- Historical declines in tuberculosis in England and Wales: improving social conditions or natural selection? *Int J Tuberc Lung Dis.* 1999;3(12):1051–4.
16. Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M. Drivers of tuberculosis epidemics: The role of risk factors and social determinants. *Soc Sci Med.* 2009;68(12):2240–6.
 17. Vynnycky E, Fine PEM. Lifetime Risks, Incubation Period, and Serial Interval of Tuberculosis | *American Journal of Epidemiology* | Oxford Academic. *Am J Epidemiol.* 2000;152(3):247–63.
 18. Houston A, Macallan DC. Extrapulmonary tuberculosis. *Medicine (Baltimore).* 2014;42(1):18–22.
 19. Lee JY. Diagnosis and treatment of extrapulmonary tuberculosis. *Tuberc Respir Dis (Seoul).* 2015;78(2):47–55.
 20. Tsukamura M. Numerical Classification of Slowly Growing Mycobacteria. *Int J Syst Bacteriol.* 1976;26(4):409–20.
 21. Smith NH, Kremer K, Inwald J, Dale J, Driscoll JR, Gordon S V., et al. Ecotypes of the Mycobacterium tuberculosis complex. *J Theor Biol.* 2006;239(2):220–225.
 22. de Jong BC, Antonio M, Gagneux S. Mycobacterium africanum-review of an important cause of human tuberculosis in West Africa. *PLoS Negl Trop Dis.* 2010;4(9):e744.
 23. Gehre F, Kumar S, Kendall L, Ejo M, Secka O, Ofori-Anyinam B, et al. A Mycobacterial Perspective on Tuberculosis in West Africa: Significant Geographical Variation of M. africanum and Other M. tuberculosis Complex Lineages. *PLoS Negl Trop Dis.* 2016;10(3):e0004408.
 24. Affolabi D, Faihun F, Sanoussi N, Anyo G, Shamputa IC, Rigouts L, et al. Possible outbreak of streptomycin-resistant Mycobacterium tuberculosis Beijing in Benin. *Emerg Infect Dis.* 2009;15(7):1123–5.
 25. Viana-niero C, Gutierrez C, Sola C, Filliol I, Rastogi N, Boulahbal F. Genetic Diversity of Mycobacterium africanum Clinical Isolates Based on IS6110-Restriction Fragment Length Polymorphism Analysis, Spoligotyping, and Variable Number of Tandem. *J Clin Microbiol.* 2001;39(1):57–65.
 26. De Jong BC, Hill PC, Aiken A, Jeffries DJ, Onipede A, Small PM, et al. Clinical presentation and outcome of tuberculosis patients infected by

- M. africanum* versus *M. tuberculosis*. *Int J Tuberc Lung Dis*. 2007;11(4):450–6.
27. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *J Infect Dis*. 2008;198(7):1037–43.
 28. Meyer CG, Scarisbrick G, Niemann S, Browne ENL, Chinbuah MA, Gyapong J, et al. Pulmonary tuberculosis: Virulence of *Mycobacterium africanum* and relevance in HIV co-infection. *Tuberculosis*. 2008;88(5):482–9.
 29. Grosset J, Decroix G, Sors C. Tuberculosis due to *Mycobacterium africanum* in African negroes in the Paris area. *Rev Tuberc Pneumol*. 1971;35:430–436.
 30. Grosset J, Sangare S, Rist N, Meyer L. Cultural and Biochemical Characteristics of Tubercle Bacilli Isolated From 230 Cases of Tuberculosis in Mali. *Bull Int Union Tuberc*. 1974;49:177–187.
 31. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat Genet*. 2013;45(2):172–9.
 32. van Soolingen D, van der Zanden a G, de Haas PE, Noordhoek GT, Kiers a, Foudraine N a, et al. Diagnosis of *Mycobacterium microti* infections among humans by using novel genetic markers. *J Clin Microbiol*. 1998;36(7):1840–5.
 33. Niemann S, Richter E, Dalügge-Tamm H, Schlesinger H, Graupner D, Königstein B, et al. Two cases of *Mycobacterium microti*-derived tuberculosis in HIV-negative immunocompetent patients. *Emerg Infect Dis*. 2000;6(5):539–42.
 34. Aranaz A, Cousins D, Mateos A, Domínguez L. Elevation of *Mycobacterium tuberculosis* subsp. *caprae* Aranaz et al. 1999 to species rank as *Mycobacterium caprae* comb. nov., sp. nov. *Int J Syst Evol Microbiol*. 2003;53(6):1785–9.
 35. Cousins D V., Bastida R, Cataldi A, Quse V, Redrobe S, Dow S, et al. Tuberculosis in seals caused by a novel member of the *Mycobacterium tuberculosis* complex: *Mycobacterium pinnipedii* sp. nov. *Int J Syst Evol*

- Microbiol. 2003;53(5):1305–14.
36. Clarke C, Helden P Van, Miller M, Parsons S. Animal-adapted members of the Mycobacterium tuberculosis complex endemic to the southern African subregion. *J S Afr Vet Assoc.* 2016;1–7.
 37. Pittius NCG van, Perrett KD, Michel AL, Keet DF, Hlokwe T, Streicher EM, et al. Infection of African Buffalo (*Syncerus Caffer*) By *Oryx Bacillus*, a Rare Member of the Antelope Clade of the Mycobacterium Tuberculosis Complex. *J Wildl Dis.* 2012;48(4):849–57.
 38. Coscolla M, Lewin A, Metzger S, Maetz-Rennsing K, Calvignac-Spencer S, Nitsche A, et al. Novel Mycobacterium tuberculosis complex isolate from a wild chimpanzee. *Emerg Infect Dis.* 2013;19(6):969–76.
 39. Aliyu G, El-Kamary SS, Abimiku A, Brown C, Tracy K, Hungerford L, et al. Prevalence of Non-Tuberculous Mycobacterial Infections among Tuberculosis Suspects in Nigeria. *PLoS One.* 2013;8(5):e63170.
 40. Adjemian J, Olivier KN, Seitz AE, Holland SM, Prevots DR. Prevalence of nontuberculous mycobacterial lung disease in U.S. medicare beneficiaries. *Am J Respir Crit Care Med.* 2012;185(8):881–6.
 41. Wells W. Airborne contagion and air hygiene. An ecological study of droplet infection. Cambridge (MA): Harvard University Press; 1955.
 42. Shaw JB, N. W-W. Infectivity of pulmonary tuberculosis in relation to sputum status. *Am Rev Tuberc.* 1954;69:724–32.
 43. Gadkowski LB, Stout JE. Cavitory pulmonary disease. *Clin Microbiol Rev.* 2008;21(2):305–33.
 44. Rodrigo T, Caylà JA, García de Olalla P, Galdós-Tangüis H, Jansà JM, Miranda P, et al. Characteristics of tuberculosis patients who generate secondary cases. *Int J Tuberc Lung Dis.* 1997;1(4):352–7.
 45. Kwan C, Ernst JD. HIV and tuberculosis: A deadly human syndemic. *Clin Microbiol Rev.* 2011;24(2):351–76.
 46. Lönnroth K, Castro KG, Chakaya JM, Chauhan LS, Floyd K, Glaziou P, et al. Tuberculosis control and elimination 2010-50: cure, care, and social development. *Lancet.* 2010;375(9728):1814–29.
 47. Walker TM, Crook DW, Peto TEA, Conlon CP. Whole-genome sequencing identifies nosocomial transmission of extra-pulmonary *M. tuberculosis*. *Qjm.* 2016;109(12):819–20.
 48. Cosivi O, Grange JM, Daborn CJ, Raviglione MC, Fujikura T, Cousins

- D, et al. Zoonotic tuberculosis due to *Mycobacterium bovis* in developing countries. *Emerg Infect Dis*. 1998;4(1):59–70.
49. Dürr S, Müller B, Alonso S, Hattendorf J, Laise CJM, van Helden PD, et al. Differences in Primary Sites of Infection between Zoonotic and Human Tuberculosis: Results from a Worldwide Systematic Review. *PLoS Negl Trop Dis*. 2013;7(8).
 50. Barry CE, Boshoff H, Dartois V, Dick T, Ehrt S, Flynn J, et al. The spectrum of latent tuberculosis: rethinking the goals of prophylaxis. *Nat Rev Microbiol*. 2009;7(12):845–55.
 51. Israel HL, Hetherington HW, Ord. JG. A study of tuberculosis among students of nursing. *J Am Med Assoc*. 1941;117(10):839–44.
 52. Walzl G, Ronacher K, Hanekom W, Scriba TJ, Zumla A. Immunological biomarkers of tuberculosis. *Nat Rev Immunol*. 2011;11(5):343–54.
 53. Verrall AJ, G. Netea M, Alisjahbana B, Hill PC, van Crevel R. Early clearance of *Mycobacterium tuberculosis*: A new frontier in prevention. *Immunology*. 2014;141(4):506–13.
 54. Orme IM. A new unifying theory of the pathogenesis of tuberculosis. *Tuberculosis*. 2014;94(1):8–14.
 55. Dhar N, McKinney JD. Microbial phenotypic heterogeneity and antibiotic tolerance. *Curr Opin Microbiol*. 2007;10(1):30–8.
 56. Orme IM, Robinson RT, Cooper AM. The balance between protective and pathogenic immune responses in the TB-infected lung. *Nat Immunol*. 2014;16(1):57–63.
 57. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, et al. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet*. 2011;43(5):482–6.
 58. Gill WP, Harik NS, Whiddon MR, Liao RP, Mittler JE, Sherman DR. A replication clock for *Mycobacterium tuberculosis*. *Nat Med*. 2009;15(2):211–4.
 59. Tiemersma EW, van der Werf MJ, Borgdorff MW, Williams BG, Nagelkerke NJD. Natural history of tuberculosis: Duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: A systematic review. *PLoS One*. 2011;6(4):e17601.
 60. Jamwal S V., Mehrotra P, Singh A, Siddiqui Z, Basu A, Rao KVS.

- Mycobacterial escape from macrophage phagosomes to the cytoplasm represents an alternate adaptation mechanism. *Sci Rep.* 2016;6(February):1–9.
61. Vergne I, Chua J, Singh SB, Deretic V. Cell biology of *Mycobacterium tuberculosis* phagosome. *Annu Rev Cell Dev Biol.* 2004;20:367–94.
 62. Armstrong BYJA, Hart APDA. Response of cultured macrophages to *Mycobacterium tuberculosis*, with observations on fusion of lysosomes with phagosomes. *J Exp Med.* 1971;134(3):713–40.
 63. Chackerian A a, Alt JM, Perera T V, Dascher CC, Behar SM. Dissemination of *Mycobacterium tuberculosis* Is Influenced by Host Factors and Precedes the Initiation of T-Cell Immunity Dissemination of *Mycobacterium tuberculosis* Is Influenced by Host Factors and Precedes the Initiation of T-Cell Immunity. *Infect Immun.* 2002;70(8):4501–9.
 64. Wolf AJ, Desvignes L, Linas B, Banaiee N, Tamura T, Takatsu K, et al. Initiation of the adaptive immune response to *Mycobacterium tuberculosis* depends on antigen production in the local lymph node, not the lungs. *J Exp Med.* 2008;205(1):105–15.
 65. Kwon YS, Cho YN, Kim MJ, Jin HM, Jung HJ, Kang JH, et al. Mucosal-associated invariant T cells are numerically and functionally deficient in patients with mycobacterial infection and reflect disease activity. *Tuberculosis.* 2015;95(3):267–74.
 66. Iliopoulos A, Psathakis K, Aslanidis S, Skagias L, Sfikakis PP. Tuberculosis and granuloma formation in patients receiving anti-TNF therapy. *Int J Tuberc Lung Dis.* 2006;10(5):588–90.
 67. Wallis RS, Doherty TM, Onyebujoh P, Vahedi M, Laang H, Olesen O, et al. Biomarkers for tuberculosis disease activity, cure, and relapse. *Lancet Infect Dis.* 2009;9(3):162–72.
 68. Allix-Béguec C, Fauville-Dufaux M, Stoffels K, Ommeslag D, Walravens K, Saegerman C, et al. Importance of identifying *Mycobacterium bovis* as a causative agent of human tuberculosis. *Eur Respir J.* 2010;35(3):692–4.
 69. World Health Organization. Guidelines on the management of latent tuberculosis infection. Geneva: World Health Organization. 2015.
 70. Pasipanodya JG, Gumbo T. A new evolutionary and pharmacokinetic-

- pharmacodynamic scenario for rapid emergence of resistance to single and multiple anti-tuberculosis drugs. *Curr Opin Pharmacol*. 2011;11(5):457–63.
71. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet*. 2013;45(7):784–90.
 72. Colangeli R, Arcus VL, Cursons RT, Ruthe A, Karalus N, Coley K, et al. Whole genome sequencing of *Mycobacterium tuberculosis* reveals slow growth and low mutation rates during latent infections in humans. *PLoS One*. 2014;9(3):e91024.
 73. Lillebaek T, Norman A, Rasmussen EM, Marvig RL, Folkvardsen DB, Andersen ÅB, et al. Substantial molecular evolution and mutation rates in prolonged latent *Mycobacterium tuberculosis* infection in humans. *Int J Med Microbiol*. 2016;306(7):580–5.
 74. Mcgrath M, Gey van pittius NC, Van helden PD, Warren RM, Warner DF. Mutation rate and the emergence of drug resistance in *Mycobacterium tuberculosis*. *J Antimicrob Chemother*. 2014;69(2):292–302.
 75. Shah NS, Auld SC, Brust JCM, Mathema B, Ismail N, Moodley P, et al. Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. *N Engl J Med*. 2017;376(3):243–53.
 76. Helden P Van, Lukey PT, Duncan K. In Situ Detection of *Mycobacterium tuberculosis* Transcripts in Human Lung Granulomas Reveals Differential Gene Expression in Necrotic Lesions *Gael. Society*. 2002;70(11):6330–8.
 77. Walter ND, Dolganov GM, Garcia BJ, Worodria W, Andama A, Musisi E, et al. Transcriptional Adaptation of Drug-tolerant *Mycobacterium tuberculosis* During Treatment of Human Tuberculosis. *J Infect Dis*. 2015;212(6):990–8.
 78. Turapov O, O'Connor BD, Sarybaeva AA, Williams C, Patel H, Kadyrov AS, et al. Phenotypically adapted *Mycobacterium tuberculosis* populations from sputum are tolerant to first line drugs. *Antimicrob Agents Chemother*. 2016;60(4):AAC.01380-15.
 79. Zhang Y. Persisters, persistent infections and the Yin-Yang model.

- Emerg Microbes Infect. 2014;3(October 2013):1–10.
80. Zhang Y, Yew WW, Barer MR. Targeting persisters for tuberculosis control. *Antimicrob Agents Chemother.* 2012;56(5):2223–30.
 81. Rodrigues L, Machado D, Couto I, Amaral L, Viveiros M. Contribution of efflux activity to isoniazid resistance in the *Mycobacterium tuberculosis* complex. *Infect Genet Evol.* 2012;12(4):695–700.
 82. Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, et al. Dynamic population changes in *mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *J Infect Dis.* 2012;206(11):1724–33.
 83. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet.* 2014;46(3):279–86.
 84. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet.* 2013;45(10):1183–9.
 85. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat Genet.* 2013;45(10):1255–60.
 86. Coll F, McNERney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* 2015;7(1):51.
 87. Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic determinants of drug resistance in *mycobacterium tuberculosis* and their diagnostic value. *Am J Respir Crit Care Med.* 2016;194(5):621–30.
 88. Shekar S, Yeo ZX, Wong JCL, Chan MKL, Ong DCT, Tongyoo P, et al. Detecting novel genetic variants associated with isoniazid-resistant *Mycobacterium tuberculosis*. *PLoS One.* 2014;9(7):e102383.
 89. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant *M. tuberculosis* strains identifies compensatory mutations in RNA polymerase. *Nat Genet.* 2012;44(1):106–10.
 90. Schrag SJ, Perrot V, Levin BR. Adaptation to the fitness costs of

- antibiotic resistance in *Escherichia coli*. *Proc Biol Sci*. 1997;264(1386):1287–91.
91. Feuerriegel S, Köser CU, Niemann S. Phylogenetic polymorphisms in antibiotic resistance genes of the mycobacterium tuberculosis complex. *J Antimicrob Chemother*. 2014;69(5):1205–10.
 92. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Karen R, Victor TC, et al. Genomic Analysis Identifies Targets of Convergent Positive Selection in Drug Resistant *Mycobacterium tuberculosis*. *Nat Genet*. 2013;45(10):1–19.
 93. Bantubani N, Kabera G, Connolly C, Rustomjee R, Reddy T, Cohen T, et al. High rates of potentially infectious tuberculosis and multidrug-resistant tuberculosis (MDR-TB) among hospital inpatients in KwaZulu Natal, South Africa indicate risk of nosocomial transmission. *PLoS One*. 2014;9(3):e90868.
 94. Otero L, Krapp F, Tomatis C, Zamudio C, Matthys F, Gotuzzo E, et al. High prevalence of primary multidrug resistant tuberculosis in persons with no known risk factors. *PLoS One*. 2011;6(10):e26276.
 95. Wang S, Ph D, Jiang G, Xia H, Zhou Y, Zhao B, et al. National Survey of Drug-Resistant Tuberculosis in China. *N Engl J Med*. 2012;366(23):2161–70.
 96. Grandjean L, Gilman RH, Martin L, Soto E, Castro B, Lopez S, et al. Transmission of Multidrug-Resistant and Drug-Susceptible Tuberculosis within Households: A Prospective Cohort Study. *PLoS Med*. 2015;12(6):e1001843.
 97. Klopper M, Warren RM, Hayes C, van Pittius NCG, Streicher EM, Müller B, et al. Emergence and spread of extensively and totally drug-resistant tuberculosis, South Africa. *Emerg Infect Dis*. 2013;19(3):449–55.
 98. Nsofor CA, Jiang Q, Wu J, Gan M, Liu Q, Zuo T, et al. Transmission is a Noticeable Cause of Resistance Among Treated Tuberculosis Patients in Shanghai, China. *Sci Rep*. 2017;7(1):7691.
 99. Yang C, Luo T, Shen X, Wu J, Gan M, Xu P, et al. Transmission of multidrug-resistant *Mycobacterium tuberculosis* in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect Dis*. 2017;17(3):275–84.

100. Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniowski F. Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLOS Med.* 2016;13(10):e1002137.
101. Moonan PK, Teeter LD, Salcedo K, Ghosh S, Ahuja SD, Flood J, et al. Transmission of multidrug-resistant tuberculosis in the USA: a cross-sectional study. *Lancet Infect Dis.* 2013;13(9):777–84.
102. Kendall EA, Fofana MO, Dowdy DW. Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. *Lancet Respir Med.* 2015;3(12):963–72.
103. Odone A, Calderon R, Becerra MC, Zhang Z, Contreras CC, Yataco R, et al. Acquired and transmitted multidrug resistant tuberculosis: The role of social determinants. *PLoS One.* 2016;11(1):e0146642.
104. Who, The World Health Organization. Treatment of tuberculosis: guidelines. 4Th Ed. 2010;160.
105. World Health Organization. Implementing tuberculosis diagnostics: A policy framework. Geneva World Heal Organ. 2015;39.
106. World Health Organization. The use of molecular line probe assays for the detection of resistance to second-line anti-tuberculosis drugs [Internet]. 2016. Available from: <http://www.who.int/tb/WHOPolicyStatementSLLPA.pdf>
107. World Health Organization WH, Global Tuberculosis Programme. WHO treatment guidelines for drug-resistant tuberculosis: 2016 update. 2016.
108. World Health Organization. Global tuberculosis report 2016. 2016.
109. Banerjee R, Allen J, Westenhouse J, Oh P, Elms W, Desmond E, et al. Extensively Drug-Resistant Tuberculosis in California, 1993–2006. *Clin Infect Dis.* 2008;47(4):450–7.
110. Migliori G, De Iaco G, Besozzi G, Centis R, Cirillo D. First tuberculosis cases in Italy resistant to all tested drugs. *Euro Surveill.* 2007;12(20).
111. Velayati AA, Masjedi MR, Farnia P, Tabarsi P, Ghanavi J, ZiaZarifi AH, et al. Emergence of new forms of totally drug-resistant tuberculosis bacilli: Super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran. *Chest.* 2009;136(2):420–5.
112. Udwardia ZF, Amale RA, Ajbani KK, Rodrigues C. Totally Drug-

- Resistant Tuberculosis in India. *Clin Infect Dis*. 2006;54(December):579–81.
113. Migliori GB, Centis R, D'Ambrosio L, Spanevello A, Borroni E, Cirillo DM, et al. Totally drug-resistant and extremely drug-resistant tuberculosis: The same disease? *Clin Infect Dis*. 2012;54(9):1379–80.
 114. Migliori GB, Sotgiu G, Gandhi NR, Falzon D, DeRiemer K, Centis R, et al. Drug resistance beyond extensively drug-resistant tuberculosis: Individual patient data meta-analysis. *Eur Respir J*. 2013;42(1):169–79.
 115. Ahmed MM, Velayati AA, Mohammed SH. Epidemiology of multidrug-resistant, extensively drug resistant, and totally drug resistant tuberculosis in Middle East countries. *Int J Mycobacteriology*. 2016;5(3):249–56.
 116. WHO. Drug-resistant TB: Totally drug-resistant TB FAQ [Internet]. [cited 2017 Dec 4]. Available from: <http://www.who.int/tb/areas-of-work/drug-resistant-tb/totally-drug-resistant-tb-faq/en/%0D>
 117. World Health Organization. Policy guidance on DST of second-line antituberculosis drugs. World Health Organisation. 2008.
 118. Falzon D, Gandhi N, Migliori GB, Sotgiu G, Cox HS, Holtz TH, et al. Resistance to fluoroquinolones and second-line injectable drugs: Impact on multidrug-resistant TB outcomes. *Eur Respir J*. 2013;42(1):156–68.
 119. Dobler CC, Korver S, Batbayar O, Nyamdulam B, Oyuntsetseg S, Tsolmon B, et al. Multidrug-resistant tuberculosis in patients for whom first-line treatment failed, Mongolia, 2010–2011. *Emerg Infect Dis*. 2015;21(8):1451–4.
 120. Pietersen E, Peter J, Streicher E, Sirgel F, Rockwood N, Mastrapa B, et al. High frequency of resistance, lack of clinical benefit, and poor outcomes in capreomycin treated South African patients with extensively drug-resistant tuberculosis. *PLoS One*. 2015;10(4):e0123655.
 121. Dheda K, Gumbo T, Gandhi NR, Murray M, Theron G, Udwadia Z, et al. Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. *Lancet Respir Med*. 2014;2(4):321–38.
 122. World Health Organization. Guidelines for treatment of drug-susceptible tuberculosis and patient care, 2017 update. Geneva: World Health Organization. 2017.
 123. World Health Organization. Companion handbook to the WHO

- guidelines for the programmatic management of drug-resistant tuberculosis. 2014.
124. World Health Organization. Guidelines for the programmatic management of Multidrug-resistant Tuberculosis. World Health Organization. 2011.
 125. Dheda K, Limberis JD, Pietersen E, Phelan J, Esmail A, Lesosky M, et al. Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. *Lancet Respir Med*. 2017;5(4):269–81.
 126. New TB drugs [Internet]. [cited 2017 Dec 6]. Available from: <http://www.newtbdrugs.org/>
 127. New TB drug regimens [Internet]. [cited 2017 Dec 6]. Available from: <http://www.newtbdrugs.org/pipeline/regimens>
 128. Koser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, et al. Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. *PLoS Pathog*. 2012;8(8).
 129. Martin LJ, Roper MH, Grandjean L, Gilman RH, Coronel J, Caviedes L, et al. Rationing tests for drug-resistant tuberculosis – who are we prepared to miss? *BMC Med*. 2016;14(1):30.
 130. Menzies D. Interpretation of repeated tuberculin tests: Boosting, conversion, and reversion. *Am J Respir Crit Care Med*. 1999;159(1):15–21.
 131. Menzies D, Gardiner G, Farhat M, Greenaway C, Pai M. Thinking in three dimensions: A web-based algorithm to aid the interpretation of tuberculin skin test results. *Int J Tuberc Lung Dis*. 2008;12(5):498–505.
 132. Pai M, Denkinger CM, Kik S V., Rangaka MX, Zwerling A, Oxlade O, et al. Gamma interferon release assays for detection of *Mycobacterium tuberculosis* infection. *Clin Microbiol Rev*. 2014;27(1):3–20.
 133. Farhat M, Greenaway C, Pai M, Menzies D. False-positive tuberculin skin tests: What is the absolute effect of BCG and non-tuberculous mycobacteria? *Int J Tuberc Lung Dis*. 2006;10(11):1192–204.
 134. World Health Organization. Use of tuberculosis interferon-gamma release assays (IGRAs) in low- and middle-income countries: policy

- statement. 2011.
135. Collins LF, Geadas C, Ellner JJ. Diagnosis of latent tuberculosis infection: Too soon to pull the plug on the tuberculin skin test. *Ann Intern Med.* 2016;164(2):122–4.
 136. World Health Organization (WHO). Chest Radiography in Tuberculosis. 2016.
 137. Cohen R, Muzaffar S, Capellan J, Azar H, Chinikamwala M. The validity of classic symptoms and chest radiographic configuration in predicting pulmonary tuberculosis. *Chest.* 1996;109(2):420–3.
 138. World Health Organization. The use of lateral flow urine lipoarabinomannan assay (LF-LAM) for the diagnosis and screening of active tuberculosis in people living with HIV: policy. 2015.
 139. Ehrlich P. Referate aus den Verein fur innere Medicin zu Berlin. *Dtsch Medizinische Wochenschrift.* 1882;9:246–249.
 140. Gandhi NR, Moll A, Sturm AW, Pawinski R, Govender T, Lalloo U, et al. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lancet.* 2006;368(9547):1575–80.
 141. WHO. Same-day diagnosis of tuberculosis by microscopy: policy statement. 2011.
 142. Bonnet M, Ramsay A, Gagnidze L, Githui W, Guerin RJ, Varaine F. Reducing the number of sputum and thresholds for positivity: An smear microscopy samples examined opportunity to optimise. *Int J Tuberc Lung Dis.* 2007;11(9):953–8.
 143. Liippo KK, Kulmala K, Tala EO. Focusing tuberculosis contact tracing by smear grading of index cases. *Am Rev Respir Dis.* 1993;148(1):235–6.
 144. Gopi PG, Chandrasekaran V, Subramani R, Santha. T. Association of conversion & cure with initial smear grading among new smear positive pulmonary tuberculosis patients treated with Category I regimen 123, no. 6 (2006): 807. *Indian J Med Res.* 2006;123(6):807.
 145. Steingart KR, Henry M, Ng V, Hopewell PC, Ramsay A, Cunningham J, et al. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. *Lancet Infect Dis.* 2006;6(9):570–581.
 146. World Health Organization. Use of liquid TB culture and drug susceptibility testing (DST) in low and medium income settings:

- summary report of the expert group meeting on the use of liquid culture media, Geneva, 26 March 2007. 2007.
147. Pfyffer GE, Welscher HM, Kissling P, Cieslak C, Casal MJ, Gutierrez J, et al. Comparison of the mycobacteria growth indicator tube (MGIT) with radiometric and solid culture for recovery of acid-fast bacilli. *J Clin Microbiol.* 1997;35(2):364–8.
 148. Folkvardsen DB, Svensson E, Thomsen V, Rasmussen EM, Bang D, Werngren J, et al. Can molecular methods detect 1% isoniazid resistance in mycobacterium tuberculosis? *J Clin Microbiol.* 2013;51(5):1596–9.
 149. Middlebrook G, Reggiardo Z, Tigertt WD. Automatable radiometric detection of growth of *Mycobacterium tuberculosis* in selective media. *Am Rev Respir Dis.* 1977;115(6):1066–9.
 150. Canetti G, Fox W, Khomenko A, Mahler HT, Menon NK, Mitchison DA, et al. Advances in techniques of testing mycobacterial drug sensitivity, and the use of sensitivity tests in tuberculosis control programmes. *Bull World Health Organ.* 1969;41(1):21–43.
 151. Canetti G, Froman S, Grosset J, Hauduroy P, Langerova M, Mahler HT, et al. *Mycobacteria: Laboratory Methods for Testing Drug Sensitivity and Resistance.* Bull Wld Hlth Org. 1963;29(5):565–78.
 152. Vareldzis BP, Grosset J, Kantor I De, Crofton J, Laszlo A, Felten M, et al. Drug-resistant tuberculosis: laboratory issues. *Tuber Lung Dis.* 1994;75:1–7.
 153. Van Deun A, Aung KJM, Bola V, Lebeke R, Hossain MA, De Rijk WB, et al. Rifampin drug resistance tests for tuberculosis: Challenging the gold standard. *J Clin Microbiol.* 2013;51(8):2633–40.
 154. World Health Organization. Updated interim critical concentrations for first-line and second-line DST (as of May 2012). 2012.
 155. World Health Organization. Automated real-time nucleic acid amplification technology for rapid and simultaneous detection of tuberculosis and rifampicin resistance: Xpert MTB/RIF System. Policy statement. World Health. 2011.
 156. World Health Organization. Using the Xpert MTB/RIF assay to detect pulmonary and extrapulmonary tuberculosis and rifampicin resistance in adults and children. 2013.
 157. World Health Organization. WHO Meeting Report of a Technical

- Expert Consultation: Non-inferiority analysis of Xpert MTB/RIF Ultra compared to Xpert MTB/RIF. Geneva: World Health Organization; 2017. 2017.
158. World Health Organization. The use of loop-mediated isothermal amplification (TB-LAMP) for the diagnosis of pulmonary tuberculosis RAPID. 2016.
 159. World Health Organization. Molecular line probe assays for rapid screening of patients at risk of multidrug-resistant tuberculosis (MDR-TB). Policy statement. 2008.
 160. World Health Organization. WHO Guideline: The use of molecular line probe assays for the detection of resistance to isoniazid and rifampicin. 2016.
 161. World Health Organization. Noncommercial culture and drug-susceptibility testing methods for screening patients at risk for multidrug-resistant tuberculosis Policy statement. 2011.
 162. Angeby KAK, Klintz L, Hoffner SE. Rapid and inexpensive drug susceptibility testing of *Mycobacterium tuberculosis* with a nitrate reductase assay. *J Clin Microbiol*. 2002;40(2):553–5.
 163. Kontsevaya I, Nikolayevskyy V, Kovalyov A, Ignatyeva O, Sadykhova A, Simak T, et al. Tuberculosis cases caused by heterogeneous infection in Eastern Europe and their influence on outcomes. *Infect Genet Evol*. 2017;48:76–82.
 164. Black PA, de Vos M, Louw GE, van der Merwe RG, Dippenaar A, Streicher EM, et al. Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates. *BMC Genomics*. 2015;16(1):857.
 165. Hingley-Wilson SM, Casey R, Connell D, Bremang S, Evans JT, Hawkey PM, et al. Undetected multidrug-resistant tuberculosis amplified by first-line therapy in mixed infection. *Emerg Infect Dis*. 2013;19(7):1138–41.
 166. Pérez-Lago L, Lirola MM, Navarro Y, Herranz M, Ruiz-Serrano MJ, Bouza E, et al. Co-infection with Drug-Susceptible and Reactivated Latent Multidrug-Resistant *Mycobacterium tuberculosis*. *Emerg Infect Dis*. 2015;21(11):2098–100.
 167. Feuerriegel S, Köser CU, Richter E, Niemann S. *Mycobacterium canettii*

- is intrinsically resistant to both pyrazinamide and pyrazinoic acid. *J Antimicrob Chemother.* 2013;68(6):1439–50.
168. Maiga M, Siddiqui S, Diallo S, Diarra B, Traoré B, Shea YR, et al. Failure to recognize nontuberculous mycobacteria leads to misdiagnosis of chronic pulmonary tuberculosis. *PLoS One.* 2012;7(5):e36902.
 169. Barnes PF, Cave MD. Molecular Epidemiology of Tuberculosis. *N Engl J Med.* 2003;349(12):1149–56.
 170. Genewein A, Telenti A, Bernasconi C, Schopfer K, Bodmer T, Mordasini C, et al. Molecular approach to identifying route of transmission of tuberculosis in the community. *Lancet.* 1993;342(8875):841–4.
 171. Witney AA, Bateson ALE, Jindani A, Phillips PPJ, Coleman D, Stoker NG, et al. Use of whole-genome sequencing to distinguish relapse from reinfection in a completed tuberculosis clinical trial. *BMC Med.* 2017;15(1):71.
 172. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain Identification of Mycobacterium tuberculosis by DNA Fingerprinting : Recommendations for a Standardized Methodology. *J Clin Microbiol.* 1993;31(2):406–9.
 173. Cave MD, Eisenach KD, McDermott PF, Bates JH, Crawford JT. IS6110: conservation of sequence in the Mycobacterium tuberculosis complex and its utilization in DNA fingerprinting. *Mol Cell Probes.* 1991;5(1):73–80.
 174. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PLoS One.* 2009;4(11):e7815.
 175. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The Epidemiology of Tuberculosis in San Francisco--A Population-Based Study Using Conventional and Molecular Methods. *N Engl J Med.* 1994;330(24):1703–9.
 176. Van Soolingen D, Hermans PWM, De Haas PEW, Soll DR, Van Embden JDA. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: Evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol.* 1991;29(11):2578–86.

177. Yang Z, De Haas P, Wachmann C, Van Soolingen D, Van Embden J, Andersen Å. Molecular epidemiology of tuberculosis in Denmark in 1992. *1995*;33(8):2077–81.
178. Joseph B V., Soman S, Radhakrishnan I, Hill V, Dhanasooraj D, Ajay Kumar R, et al. Molecular epidemiology of Mycobacterium tuberculosis isolates from Kerala, India using IS6110-RFLP, spoligotyping and MIRU-VNTRs. *Infect Genet Evol.* 2013;16:157–64.
179. Jonsson J, Hoffner S, Berggren I, Bruchfeld J, Ghebremichael S, Pennhag A, et al. Comparison between RFLP and MIRU-VNTR genotyping of mycobacterium tuberculosis strains isolated in stockholm 2009 to 2011. *PLoS One.* 2014;9(4):e95159.
180. Żaczek A, Ziólkiewicz M, Wojtasik A, Dziadek J, Sajduda A. IS6110-based differentiation of Mycobacterium tuberculosis strains. *Polish J Microbiol.* 2013;62(2):201–4.
181. Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, et al. Transmission of tuberculosis in New York City--an analysis by DNA fingerprinting and conventional epidemiologic methods. *N Engl J Med.* 1994;330(24):1710–6.
182. Kamerbeek J, Schouls L, Kolk a, Agterveld M Van, Kuijper S, Bunschoten a, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology . *Simultaneous Detection and Strain Differentiation of Mycobacterium tuberculosis for Diagnosis and Epidemiology. Jcm.* 1997;35(4):907–14.
183. Jagielski T, Van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. Current methods in the molecular typing of mycobacterium tuberculosis and other Mycobacteria. *Biomed Res Int.* 2014;2014:645802.
184. Frothingham R, Meeker-O'Connell WA. Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats. *1998*;144:1189–96.
185. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsche-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. *J Clin Microbiol.* 2006;44(12):4498–510.
186. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable

- association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci U S A*. 2004;101(14):4871–6.
187. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, De Jong BC, Narayanan S, et al. Variable host–pathogen compatibility in MTB. 2006;103(8):2869–73.
 188. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial lineages causing pulmonary and extrapulmonary Tuberculosis, Ethiopia. *Emerg Infect Dis*. 2013;19(3):460–3.
 189. Hartl D, Clark A. *Principles of Population Genetics*. Vol 116. Sunderland: Sinauer Associates,; 1997.
 190. Gagneux S, Small PM. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis*. 2007;7(5):328–37.
 191. Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet*. 2010;42(6):498–503.
 192. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodtkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med*. 2011;364(8):730–9.
 193. World Health Organization. High-priority target product profiles for new tuberculosis diagnostics. 2014.
 194. Lessem E. *The Tuberculosis Diagnostics Pipeline*. 2015.
 195. World Health Organization. *The Paradigm Shift*. 2015.
 196. United National Development Programme. *Human development reports: Gambia* [Internet]. 2016 [cited 2017 Dec 20]. Available from: <http://hdr.undp.org/en/countries/profiles/GMB>
 197. Determining the Impact of Enhanced Case Finding on Tuberculosis Notification in The Gambia (ECF) [Internet]. *clinicaltrials.gov*. [cited 2017 Dec 19]. Available from: <http://clinicaltrials.gov/show/NCT01660646>
 198. The Gambia MU. *The Gambian Survey of Tuberculosis Prevalence (GAMSTEP)* [Internet]. 2014 [cited 2017 Nov 19]. Available from: [http://www.moh.gov.gm/sites/default/files/GAMSTEP REPORT.pdf](http://www.moh.gov.gm/sites/default/files/GAMSTEP%20REPORT.pdf)

199. Kent PT, P. KG. Public health mycobacteriology: a guide for the level III laboratory. US Department of Health and Human Services, Public Health Service, Centers for Disease Control; 1985.
200. Caviedes L, Lee T, Gilman RH, Sheen P, Spellman E, Lee EH, et al. Rapid, efficient detection and drug susceptibility testing of *Mycobacterium tuberculosis* in sputum by microscopic observation of broth cultures. *J Clin Microbiol.* 2000;38(3):1203–8.
201. Wayne LG. Simple pyrazinamidase and urease tests for routine identification of mycobacteria. *Am Rev Respir Dis.* 1974;109:147–51.
202. De Jong BC, Antonio M, Awine T, Ogungbemi K, De Jong YP, Gagneux S, et al. Use of spoligotyping and large sequence polymorphisms to study the population structure of the *Mycobacterium tuberculosis* complex in a cohort study of consecutive smear-positive tuberculosis cases in the gambia???. *J Clin Microbiol.* 2009;47(4):994–1001.
203. Lim YW, Schmieder R, Haynes M, Willner D, Furlan M, Youle M, et al. Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. *J Cyst Fibros.* 2013;12(2):154–64.
204. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357–9.
205. Chan JZ-M, Sergeant MJ, Lee OY-C, Minnikin DE, Besra GS, Pap I, et al. Metagenomic analysis of tuberculosis in a mummy. *N Engl J Med.* 2013;369(3):289–90.
206. Kay GL, Sergeant MJ, Zhou Z, Chan JZ-M, Millard A, Quick J, et al. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun.* 2015;6:6717.
207. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics.* 2012 Oct 15;28(20):2678–9.
208. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ.* 2014;2:e585.
209. Matsen F a, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics.* 2010;11(1):538.

210. Zheming Zhou. MGplacer [Internet]. [cited 2017 Aug 2]. Available from: <https://sourceforge.net/projects/mgplacer/>
211. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012 Mar 1;22(3):568–76.
212. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic resistance predictions from genome sequence data for *S. aureus* and *M. tuberculosis*. *bioRxiv.* 2015;6:18564.
213. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: A web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol.* 2015;53(6):1908–14.
214. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *J Stat Softw.* 2010;36(3):1–48.
215. World Health Organization. Implementing the End TB Strategy: The Essentials. Geneva: World Health Organization World Health Organization. 2015.
216. Pai M, Schito M. Tuberculosis diagnostics in 2015: landscape, priorities, needs, and prospects. *J Infect Dis.* 2015;211(Suppl 2):S21-8.
217. Kik S V, Denking C, Casenghi M, Vadnais C, Pai M. Tuberculosis diagnostics: which target product profiles should be prioritised? *To. Eur Respir J.* 2014;44(2):537–40.
218. Wells WA, Boehme CC, Cobelens FGJ, Daniels C, Dowdy D, Gardiner E, et al. Alignment of new tuberculosis drug regimens and drug susceptibility testing: A framework for action. *Lancet Infect Dis.* 2013;13(5):449–58.
219. Pankhurst LJ, del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: A prospective study. *Lancet Respir Med.* 2016;4(1):49–58.
220. Nimmo C, Doyle R, Burgess C, Williams R, Gorton R, McHugh TD, et al. Rapid identification of a *Mycobacterium tuberculosis* full genetic drug resistance profile through whole genome sequencing directly from sputum. *Int J Infect Dis.* 2017;62:44–6.
221. Arnold A, Witney AA, Vergnano S, Roche A, Cosgrove CA, Houston A,

- et al. XDR-TB transmission in London: Case management and contact tracing investigation assisted by early whole genome sequencing. *J Infect.* 2016;73(3):210–8.
222. Witney AA, Gould KA, Arnold A, Coleman D, Delgado R, Dhillon J, et al. Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. *J Clin Microbiol.* 2015;53(5):1473–83.
223. Walker TM, Kohl TA, Omar S V., Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: A retrospective cohort study. *Lancet Infect Dis.* 2015;15(10):1193–202.
224. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: An observational study. *Lancet Respir Med.* 2014;2(4):285–92.
225. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Med.* 2013;10(2):e1001387.
226. Stucki D, Ballif M, Bodmer T, Coscolla M, Maurer AM, Droz S, et al. Tracking a tuberculosis outbreak over 21 years: Strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *J Infect Dis.* 2015;211(8):1306–16.
227. Michel AL, Müller B, van Helden PD. *Mycobacterium bovis* at the animal-human interface: A problem, or not? *Vet Microbiol.* 2010;140(3–4):371–81.
228. Stucki D, Ballif M, Egger M, Furrer H, Altpeter E, Battegay M, et al. Standard genotyping overestimates transmission of *Mycobacterium tuberculosis* among immigrants in a low incidence country. *J Clin Microbiol.* 2016;54(May):JCM.00126-16.
229. Metcalfe J, Streicher E, Theron G, Colman R, Penaloza R, Allender C, et al. *Mycobacterium tuberculosis* Subculture Results in Loss of Potentially Clinically Relevant Heteroresistance. *Antimicrob Agents Chemother.* 2017;61(11):e00888–17.
230. Shamputa IC, Rigouts L, Eyongeta LA, Abdullah N, Aila E, Deun A

- Van, et al. Genotypic and Phenotypic Heterogeneity among Mycobacterium tuberculosis Isolates from Pulmonary Tuberculosis Patients. *J Clin Microbiol.* 2004;42(12):5528–36.
231. Warren RM, Victor TC, Streicher EM, Richardson M, Beyers N, Gey van Pittius NC, et al. Patients with active tuberculosis often have different strains in the same sputum specimen. *Am J Respir Crit Care Med.* 2004;169(5):610–4.
232. Cohen T, Wilson D, Wallengren K, Samuel EY, Murray M. Mixed-strain Mycobacterium tuberculosis infections among patients dying in a Hospital in KwaZulu-Natal, South Africa. *J Clin Microbiol.* 2011;49(1):385–8.
233. Wang JY, Hsu HL, Yu MC, Chiang CY, Yu FL, Yu CJ, et al. Mixed infection with Beijing and non-Beijing strains in pulmonary tuberculosis in Taiwan: Prevalence, risk factors, and dominant strain. *Clin Microbiol Infect.* 2011;17(8):1239–45.
234. Parrish N, Carrol K. Importance of improved TB diagnostics in addressing the extensively drug-resistant TB crisis. *Futur Med.* 2008;3(4):405–13.
235. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med.* 2012;366(24):2267–75.
236. Pai M. Transforming the diagnosis of tuberculosis: an editorial board member's opinion at the 15th year of Expert Review of Molecular Diagnostics. *Expert Rev Mol Diagn.* 2015;15(3):295–8.
237. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. *Lancet Infect Dis.* 2013;13(2):137–46.
238. Koser CU, Bryant JM, Becq J, Torok ME, Ellington MJ, Marti-Renom MA, et al. Whole-Genome Sequencing for Rapid Susceptibility Testing of *M. tuberculosis*. *N Engl J Med.* 2013;369(3):290–2.
239. Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P. Metagenomics for pathogen detection in public health. *Genome Med.* 2013;5(9):81.
240. Nakamura S, Maeda N, Miron IM, Yoh M, Izutsu K, Kataoka C, et al.

- Metagenomic diagnosis of bacterial infections. *Emerg Infect Dis*. 2008;14(11):1784–6.
241. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA*. 2013;309(14):1502–10.
 242. National human genome research institute. Cost per raw megabase of DNA sequencing [Internet]. [cited 2017 Dec 20]. Available from: https://www.genome.gov/images/content/cost_per_megabase2.jpg
 243. Aminian M, Couvin D, Shabbeer A, Hadley K, Vandenberg S, Rastogi N, et al. Predicting *Mycobacterium tuberculosis* Complex Clades Using Knowledge-Based Bayesian Networks. *Biomed Res Int*. 2014;2014:1–11.
 244. Coll F, Preston M, Afonso Guerra-Assunção J, Hill-Cawthorn G, Harris D, Perdigão J, et al. PolyTB: A genomic variation map for *Mycobacterium tuberculosis*. *Tuberculosis*. 2014;94(3):346–54.
 245. NCBI. Sequence read archive [Internet]. [cited 2017 Nov 15]. Available from: <http://www.ncbi.nlm.nih.gov/Traces/sra/>
 246. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006 Nov 1;22(21):2688–90.
 247. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;12(1):87.
 248. Behr MA, Warren S, Salamon H, Hopewell PC, de Leon AP, Daley C, et al. Transmission of *Mycobacterium tuberculosis* from patients smear-negative for acid-fast bacilli. *Lancet*. 1999;353(9151):444–9.
 249. Burgess a L, Fitzgerald DW, Severe P, Joseph P, Noel E, Rastogi N, et al. Integration of tuberculosis screening at an HIV voluntary counselling and testing centre in Haiti. *AIDS*. 2001;15(14):1875–9.
 250. Hargreaves NJ, Kadzakumanja O, Whitty CJM, Salaniponi FML, Harries AD, Squire SB. ‘Smear-negative’ pulmonary tuberculosis in a DOTS programme: Poor outcomes in an area of high HIV seroprevalence. *Int J Tuberc Lung Dis*. 2001;5(9):847–54.
 251. Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, et al. Characterization of the viral microbiome in patients with severe

- lower respiratory tract infections, using metagenomic sequencing. *PLoS One*. 2012;7(2):e30875.
252. Fischer N, Rohde H, Indenbirken D, Günther T, Reumann K, Lütgehetmann M, et al. Rapid metagenomic diagnostics for suspected outbreak of severe pneumonia. *Emerg Infect Dis*. 2014;20(6):1072–5.
 253. Lim YW, Evangelista JS, Schmieder R, Bailey B, Haynes M, Furlan M, et al. Clinical insights from metagenomic analysis of sputum samples from patients with cystic fibrosis. *J Clin Microbiol*. 2014;52(2):425–37.
 254. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, et al. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One*. 2009;4(10).
 255. Young JC, Chehoud C, Bittinger K, Bailey A, Diamond JM, Cantu E, et al. Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. *Am J Transplant*. 2015;15(1):200–9.
 256. Bouwman AS, Kennedy SL, Muller R, Stephens RH, Holst M, Caffell AC, et al. Genotype of a historic strain of *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A*. 2012;109(45):18511–6.
 257. Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, et al. A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA. *PLoS One*. 2013;8(10):e76096.
 258. Hunter SJ, Easton S, Booth V, Henderson B, Wade WG, Ward JM. Selective removal of human DNA from metagenomic DNA samples extracted from dental plaque. *J Basic Microbiol*. 2011;51(4):442–6.
 259. Hemalatha D, Rajendhran J, Gunasekaran P. A simple method for selective removal of human DNA from oral microbiome for functional metagenomic applications. *J Sci Ind Res (India)*. 2013;72(9–10):572–6.
 260. Garcia-Garcerà M, Garcia-Etxebarria K, Coscollà M, Latorre A, Calafell F. A New Method for Extracting Skin Microbes Allows Metagenomic Analysis of Whole-Deep Skin. *PLoS One*. 2013;8(9):e74914.
 261. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Moller N, et al. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol*. 2014;52(1):139–46.
 262. Zhou L, Pollard AJ. A novel method of selective removal of human DNA improves PCR sensitivity for detection of *Salmonella Typhi* in

- blood samples. *BMC Infect Dis.* 2012;12(1):164.
263. Barnes HE, Liu G, Weston CQ, King P, Pham LK, Waltz S, et al. Selective Microbial Genomic DNA Isolation Using Restriction Endonucleases. *PLoS One.* 2014;9(10):e109061.
264. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD, et al. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerg Infect Dis.* 2016;22(2):331–4.
265. Global tuberculosis report 2016 summary. 2016.
266. WHO. Use of high burden country lists for TB by WHO in the post-2015 era. Geneva: World Health Organization. 2015.
267. Barletta F, Otero L, De Jong BC, Iwamoto T, Arikawa K, Van Der Stuyft P, et al. Predominant Mycobacterium tuberculosis families and high rates of recent transmission among new cases are not associated with primary multidrug resistance in Lima, Peru. *J Clin Microbiol.* 2015;53(6):1854–63.
268. Bonilla CA, Crossa A, Jave HO, Mitnick CD, Jamanca RB, Herrera C, et al. Management of Extensively Drug-Resistant Tuberculosis in Peru: Cure Is Possible. *PLoS One.* 2008;3(8):e2957.
269. World Health Organization. Peru WHO TB profile [Internet]. 2017 [cited 2017 Dec 19]. Available from: https://extranet.who.int/sree/Reports?op=Replet&name=%2FWHO_HQ_Reports%2FG2%2FPROD%2FEXT%2FTBCountryProfile&ISO2=PE&LAN=EN&outtype=pdf
270. Sanchez-Padilla E, Merker M, Beckert P, Jochims FT, Dlamini H, Kahn P, et al. Detection of Drug-Resistant Tuberculosis by Xpert MTB/RIF in Swaziland. *N Engl J Med.* 2015;372(12):1181–2.
271. Ocheretina O, Escuyer VE, Mabou M, Royal-mardi G, Collins S, Vilbrun SC, et al. Correlation between Genotypic and Phenotypic Testing for Resistance to Rifampin in Mycobacterium tuberculosis Clinical Isolates in Haiti : Investigation of Cases with Discrepant Susceptibility Results. 2014;9(3).
272. Caviedes L, Lee T, Gilman RH, Sheen P, Spellman E, Lee EH, et al. Rapid, efficient detection and drug susceptibility testing of Mycobacterium tuberculosis in sputum by microscopic observation of broth cultures. *J Clin Microbiol.* 2000;38(3):1203–8.

273. Moore D a J, Moore D a J, Mendoza D, Mendoza D, Gilman RH, Gilman RH, et al. Microscopic Observation Drug Susceptibility Assay, a Rapid, Reliable Diagnostic Test for Multidrug-Resistant Tuberculosis Suitable for Use in Resource-Poor Settings. *Microbiology*. 2004;42(10):4432–7.
274. Moore DAJ, Evans CAW, Gilman RH, Caviedes L, Coronel J, Vivar A, et al. Microscopic-observation drug-susceptibility assay for the diagnosis of TB. *N Engl J Med*. 2006;355(15):1539–50.
275. Mello FCQ, Arias MS, Rosales S, Marsico AG, Pavón A, Alvarado-Gálvez C, et al. Clinical evaluation of the microscopic observation drug susceptibility assay for detection of Mycobacterium tuberculosis resistance to isoniazid or rifampin. *J Clin Microbiol*. 2007;45(10):3387–9.
276. Shiferaw G, Woldeamanuel Y, Gebeyehu M, Girmachew F, Demessie D, Lemma E. Evaluation of microscopic observation drug susceptibility assay for detection of multidrug-resistant Mycobacterium tuberculosis. *J Clin Microbiol*. 2007;45(4):1093–7.
277. Arias M, Mello FCQ, Pavon A, Marsico AG, Alvarado-Galvez C, Rosales S, et al. Clinical Evaluation of the Microscopic-Observation Drug-Susceptibility Assay for Detection of Tuberculosis. *Clin Infect Dis*. 2007;44(5):674–780.
278. Minion J, Leung E, Menzies D, Pai M. Microscopic-observation drug susceptibility and thin layer agar assays for the detection of drug resistant tuberculosis: a systematic review and meta-analysis. 2010;3099(10):1–11.
279. Ghiraldi LD, Campanerut PAZ, Sposito FLE, Sato DN, Leite CQF, Hirata M, et al. Evaluation of the microscopic observation drug susceptibility assay for detection of Mycobacterium tuberculosis resistance to pyrazinamide. *Clin Microbiol Infect*. 2011;17(12):1792–7.
280. Wu X, Lu W, Shao Y, Liu Q, Shi X, Wang X, et al. The indirect microscopic observation drug susceptibility assay demonstrated high concordance with the indirect MGIT method for pyrazinamide susceptibility testing. *J Antimicrob Chemother*. 2015;70(8):2295–9.
281. Huang Z-K, Luo Q, Jiang B-X, Li W-T, Xu X-M, Xiong G-L, et al. Performance of the microscopic observation drug susceptibility assay in pyrazinamide susceptibility testing for Mycobacterium tuberculosis. *Chinese Med journal*. 2013;126:4334–9.

282. Huang ZK, Luo Q, Xiong GL, Li WT, Xu XM, Li J. Evaluation of indirect drug susceptibility testing using the MODS assay for the detection of XDR-TB in China. *Int J Tuberc Lung Dis.* 2014;18(4):461–5.
283. Huang Z, Li G, Chen J, Li W, Xu X, Luo Q, et al. Evaluation of MODS assay for rapid detection of *Mycobacterium tuberculosis* resistance to second-line drugs in a tertiary care tuberculosis hospital in China. *Tuberculosis.* 2014;94(5):506–10.
284. Nikolayevskyy V, Kranzer K, Niemann S, Drobniewski F. Whole genome sequencing of *Mycobacterium tuberculosis* for detection of recent transmission and tracing outbreaks: A systematic review. *Tuberculosis.* 2016;98:77–85.
285. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics.* 2014;15(1):881.
286. McDermott W., Tompsett R. Activation of Pyrazinamide and Nicotinamide in Acidic Environments in Vitro. *Am Rev Tuberc Pulm Dis.* 1954;70(4):748–54.
287. Mitchison DA. The action of antituberculosis drugs in short-course chemotherapy. *Tubercle.* 1985;66(3):219–25.
288. Zumla AI, Gillespie SH, Hoelscher M, Philips PPJ, Cole ST, Abubakar I, et al. New antituberculosis drugs, regimens, and adjunct therapies: Needs, advances, and future prospects. *Lancet Infect Dis.* 2014;14(4):327–40.
289. Ibrahim M, Andries K, Lounis N, Chauffour A, Truffot-Pernot C, Jarlier V, et al. Synergistic activity of R207910 combined with pyrazinamide against murine tuberculosis. *Antimicrob Agents Chemother.* 2007;51(3):1011–5.
290. Tasneen R, Tyagi S, Williams K, Grosset J, Nuermberger E. Enhanced bactericidal activity of rifampin and/or pyrazinamide when combined with PA-824 in a murine model of tuberculosis. *Antimicrob Agents Chemother.* 2008;52(10):3664–8.
291. Zhang Y, Shi W, Zhang W, Mitchison D. Mechanisms of Pyrazinamide Action and Resistance. *Microbiol Spectr.* 2013;100(2):130–4.
292. Andries K. A Diarylquinoline Drug Active on the ATP Synthase of *Mycobacterium tuberculosis*. *Science (80-).* 2005;307(5707):223–7.

293. Diacon AH, Donald PR, Pym A, Grobusch M, Patientia RF, Mahanyele R, et al. Randomized pilot trial of eight weeks of bedaquiline (TMC207) treatment for multidrug-resistant tuberculosis: Long-term outcome, tolerability, and effect on emergence of drug resistance. *Antimicrob Agents Chemother.* 2012;56(6):3271–6.
294. Zhang Y, Mitchison D. The curious characteristics of pyrazinamide: A review. *Int J Tuberc Lung Dis.* 2003;7(1):6–21.
295. Cheng SJ, Thibert L, Sanchez T, Heifets L, Zhang Y. *pncA* mutations as a major mechanism of pyrazinamide resistance in *Mycobacterium tuberculosis*: Spread of a monoresistant strain in Quebec, Canada. *Antimicrob Agents Chemother.* 2000;44(3):528–32.
296. Tabshis M, Weed WJ. Lack of significant in vitro sensitivity of *Mycobacterium tuberculosis* to pyrazinamide on three different solid media. *Am Rev Tuberc Pulm Dis.* 1953;67(3):391–5.
297. Peterson ND, Rosen BC, Dillon NA, Baughn AD. Uncoupling environmental pH and intrabacterial acidification from pyrazinamide susceptibility in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother.* 2015;59(12):7320–6.
298. Zhang Y, Wade MM, Scorpio A, Zhang H, Sun Z. Mode of action of pyrazinamide: Disruption of *Mycobacterium tuberculosis* membrane transport and energetics by pyrazinoic acid. *J Antimicrob Chemother.* 2003;52(5):790–5.
299. Zhang Y, Permar S, Sun Z. Conditions that may affect the results of susceptibility testing of *Mycobacterium tuberculosis* to pyrazinamide. *J Med Microbiol.* 2002;51(1):42–9.
300. Scorpio A, Zhang Y. Mutations in *pncA*, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nat Med.* 1996;2(6):662–7.
301. Shi W, Zhang X, Jiang X, Ruan H, Barry CE, Wang H, et al. Pyrazinamide inhibits trans-translation in *Mycobacterium tuberculosis*: a potential mechanism for shortening the duration of chemotherapy. *Science* (80-). 2012;333(6049):1630–2.
302. Dillon NA, Peterson ND, Feaga HA, Keiler KC, Baughn AD. Anti-tubercular Activity of Pyrazinamide is Independent of trans-

- Translation and RpsA. *Sci Rep*. 2017;7(1):6135.
303. Muto A, Fujihara A, Ito KI, Matsuno J, Ushida C, Himeno H. Requirement of transfer-messenger RNA for the growth of *Bacillus subtilis* under stresses. *Genes to Cells*. 2000;5(8):627–35.
 304. Thibonnier M, Thiberge JM, De Reuse H. Trans-translation in *Helicobacter pylori*: Essentiality of ribosome rescue and requirement of protein tagging for stress resistance and competence. *PLoS One*. 2008;3(11):e3810.
 305. Keiler KC. Biology of *trans*-Translation. *Annu Rev Microbiol*. 2008;62(1):133–51.
 306. Ramirez-Busby SM, Valafar F. Systematic review of mutations in pyrazinamidase associated with pyrazinamide resistance in mycobacterium tuberculosis clinical isolates. *Antimicrob Agents Chemother*. 2015;59(9):5267–77.
 307. Whitfield MG, Soeters HM, Warren RM, York T, Sampson SL, Streicher EM, et al. A global perspective on pyrazinamide resistance: Systematic review and meta-analysis. *PLoS One*. 2015;10(7):e0133869.
 308. World Health Organization. Guidelines for surveillance of drug resistance in tuberculosis. 2015.
 309. Tan Y, Hu Z, Zhang T, Cai X, Kuang H, Liu Y, et al. Role of *pncA* and *rpsA* gene sequencing in detection of pyrazinamide resistance in mycobacterium tuberculosis isolates from southern China. *J Clin Microbiol*. 2014;52(1):291–7.
 310. Aono A, Chikamatsu K, Yamada H, Kato T, Mitarai S. Association between *pncA* gene mutations, pyrazinamidase activity, and pyrazinamide susceptibility testing in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother*. 2014;58(8):4928–30.
 311. Whitfield MG, Warren RM, Streicher EM, Sampson SL, Sirgel FA, Van Helden PD, et al. *Mycobacterium tuberculosis pncA* polymorphisms that do not confer pyrazinamide resistance at a breakpoint concentration of 100 micrograms per milliliter in MGIT. *J Clin Microbiol*. 2015;53(11):3633–5.
 312. Hirano K, Takahashi M, Kazumi Y, Fukasawa Y, Abe C. Mutation in *pncA* is a major mechanism of pyrazinamide resistance in *Mycobacterium tuberculosis*. *Tuber Lung Dis*. 1997;78(2):117–22.

313. Zhang S, Chen J, Shi W, Liu W, Zhang W, Zhang Y. Mutations in panD encoding aspartate decarboxylase are associated with pyrazinamide resistance in *Mycobacterium tuberculosis*. *Emerg Microbes Infect.* 2013;2(6):e34.
314. Shi W, Chen J, Feng J, Cui P, Zhang S, Weng X, et al. Aspartate decarboxylase (PanD) as a new target of pyrazinamide in *Mycobacterium tuberculosis*. *Emerg Microbes Infect.* 2014;3(8):e58.
315. Ilina EN, Shitikov EA, Ikryannikova LN, Alekseev DG, Kamashev DE, Malakhova M V., et al. Comparative Genomic Analysis of *Mycobacterium tuberculosis* Drug Resistant Strains from Russia. *PLoS One.* 2013;8(2):e56577.
316. Simons SO, Mulder A, Van Ingen J, Boeree MJ, Van Soolingen D. Role of rpsA gene sequencing in diagnosis of pyrazinamide resistance. *J Clin Microbiol.* 2013;51(1):382.
317. Alexander DC, Ma JH, Guthrie JL, Blair J, Chedore P, Jamieson FB. Gene sequencing for routine verification of pyrazinamide resistance in *Mycobacterium tuberculosis*: A role for pncA but not rpsA. *J Clin Microbiol.* 2012;50(11):3726–8.
318. Akhmetova A, Kozhamkulov U, Bismilda V, Chingissova L, Abildaev T, Dymova M, et al. Mutations in the pncA and rpsA genes among 77 *Mycobacterium tuberculosis* isolates in Kazakhstan. *Int J Tuberc Lung Dis.* 2015;19(2):179–84.
319. Rodwell TC, Fink L, Catanzaro D, Jackson RL, Pettigrove M. A Multinational Analysis of Mutations and Heterogeneity in PZase , RpsA , and PanD Associated with Pyrazinamide Resistance in M / XDR *Mycobacterium tuberculosis*. 2017;(February):1–9.
320. Dillon NA, Peterson ND, Rosen BC, Baughn AD. Pantothenate and pantetheine antagonize the antitubercular activity of pyrazinamide. *Antimicrob Agents Chemother.* 2014;58(12):7258–63.
321. Njire M, Tan Y, Mugweru J, Wang C, Guo J, Yew WW, et al. Pyrazinamide resistance in *Mycobacterium tuberculosis*: Review and update. *Adv Med Sci.* 2016;61(1):63–71.
322. Chang KC, Yew WW, Zhang Y. Pyrazinamide susceptibility testing in *Mycobacterium tuberculosis*: A systematic review with meta-analyses. *Antimicrob Agents Chemother.* 2011;55(10):4499–505.

323. Salfingert MAX, Heifets LB. Determination of Pyrazinamide MICs for *Mycobacterium tuberculosis* at Different pHs by the Radiometric Method. 1988;32(7):1002–4.
324. Hoffner S, Ängeby K, Sturegård E, Jönsson B, Johansson A, Sellin M, et al. Proficiency of drug susceptibility testing of *Mycobacterium tuberculosis* against pyrazinamide: the Swedish experience. 2013;17(March):1486–90.
325. Cunha J. Testes de sensibilidade à pirazinamida em todos os isolados do complexo *Mycobacterium tuberculosis* - - uma análise crítica Pyrazinamide susceptibility testing in all isolates of the *Mycobacterium* - a critical analysis tuberculosis complex -. *Fisioterapia*. 2012;18(4):188–9.
326. Sharma B, Pal N, Malhotra B, Vyas L, Rishi S. Comparison of MGIT 960 & pyrazinamidase activity assay for pyrazinamide susceptibility testing of *Mycobacterium tuberculosis*. *Indian J Med Res*. 2010;132(7):72–6.
327. Miotto P, M. CA, Feuerriegel S, Casali N, Drobniewski F, Rodionova Y, et al. *Mycobacterium tuberculosis* Pyrazinamide Resistance Determinants: a Multicenter Study. 2014;5(5):1–10.
328. Zimic M, Sheen P, Quiliano M, Gutierrez A, Gilman RH. Peruvian and globally reported amino acid substitutions on the *Mycobacterium tuberculosis* pyrazinamidase suggest a conserved pattern of mutations associated to pyrazinamide resistance. *Infect Genet Evol*. 2010;10(2):346–9.
329. Werngren J, Sturegård E, Juréen P, Ängeby K, Hoffner S, Schönd T. Reevaluation of the critical concentration for drug susceptibility testing of *Mycobacterium tuberculosis* against pyrazinamide using wild-type MIC distributions and *pncA* gene sequencing. *Antimicrob Agents Chemother*. 2012;56(3):1253–7.
330. Somoskovi A, Dormandy J, Parsons LM, Kaswa M, Goh KS, Rastogi N, et al. Sequencing of the *pncA* gene in members of the *Mycobacterium tuberculosis* complex has important diagnostic applications: Identification of a species-specific *pncA* mutation in '*Mycobacterium canettii*' and the reliable and rapid predictor of pyrazinamide res. *J Clin Microbiol*. 2007;45(2):595–9.
331. Stoffels K, Mathys V, Fauville-Dufaux M, Wintjens R, Bifania P.

- Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother*. 2012;56(10):5186–93.
332. Zhang H, Deng J, Bi L, Zhou Y, Zhang Z, Zhang C. Characterization of *Mycobacterium tuberculosis* nicotinamidase / pyrazinamidase. 2008;275:753–62.
333. Maslov DA, Zaïchikova M V., Chernousova LN, Shur K V., Bekker OB, Smirnova TG, et al. Resistance to pyrazinamide in Russian *Mycobacterium tuberculosis* isolates: PncA sequencing versus Bactec MGIT 960. *Tuberculosis*. 2015;95(5):608–12.
334. Sheen P, Couvin D, Grandjean L, Zimic M, Dominguez M, Luna G, et al. Genetic Diversity of *Mycobacterium tuberculosis* in Peru and Exploration of Phylogenetic Associations with Drug Resistance. *PLoS One*. 2013;8(6).
335. Zimic M, Fuentes P, Gilman RH, Gutiérrez AH, Kirwan D, Sheen P. Pyrazinoic acid efflux rate in *Mycobacterium tuberculosis* is a better proxy of pyrazinamide resistance. *Tuberculosis*. 2012;92(1):84–91.
336. Sheen P, Lozano K, Gilman RH, Valencia HJ, Loli S, Fuentes P, et al. PncA gene expression and prediction factors on pyrazinamide resistance in *Mycobacterium tuberculosis*. *Tuberculosis*. 2013;93(5):515–22.
337. Sheen P, Ferrer P, Gilman RH, López-Llano J, Fuentes P, Valencia E, et al. Effect of pyrazinamidase activity on pyrazinamide resistance in *Mycobacterium tuberculosis*. *Tuberculosis*. 2009;89(2):109–13.
338. Gumbo T. New susceptibility breakpoints for first-line antituberculosis drugs based on antimicrobial pharmacokinetic/pharmacodynamic science and population pharmacokinetic variability. *Antimicrob Agents Chemother*. 2010;54(4):1484–91.
339. Rodloff A, Bauer T, Ewig S, Kujath P, Müller E. Susceptible, intermediate, and resistant - the intensity of antibiotic action. *Dtsch Ärzteblatt Int*. 2008;105(39):657–62.
340. Morlock GP, Crawford JT, Butler WR, Brim E, Sikes D, Mazurek GH, et al. Phenotypic Characterization of pncA Mutants of *Mycobacterium tuberculosis*. *Aac*. 2000;44(9):2291–5.
341. Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D,

- et al. Two new rapid SNP-typing methods for classifying mycobacterium tuberculosis complex into the main phylogenetic lineages. *PLoS One*. 2012;7(7):e41253.
342. World Health Organization. The End TB Strategy. 2015.
343. Sheen P, Méndez M, Gilman RH, Peña L, Caviedes L, Zimic MJ, et al. Sputum PCR-single-strand conformational polymorphism test for same-day detection of pyrazinamide resistance in tuberculosis patients. *J Clin Microbiol*. 2009;47(9):2937–43.
344. Ramaswamy S V, Reich R, Dou S-J, Jasperse L, Pan X, Wanger A, et al. Single nucleotide polymorphisms in genes associated with isoniazid resistance in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother*. 2003;47(4):1241–50.
345. Lavender C, Globan M, Sievers A, Billman-jacobe H, Fyfe J. Molecular Characterization of Isoniazid-Resistant *Mycobacterium tuberculosis* Isolates Collected in Australia. *Antimicrob Agents Chemother*. 2005;49(10):4068–74.
346. Phelan J, O’Sullivan DM, Machado D, Ramos J, Whale AS, O’Grady J, et al. The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. *Genome Med*. 2016;8(1):132.
347. Zhang Y, Telenti A. Genetics of drug resistance in *Mycobacterium tuberculosis*. *Molecular genetics of mycobacteria*, ASM Press, Washington, DC. 2000. 235-254 p.
348. Kim S-Y, Park Y-J, Kim W-I, Lee S-H, Chang CL, Kang S-J, et al. Molecular analysis of isoniazid resistance in *Mycobacterium tuberculosis* isolates recovered from South Korea. *Diagn Microbiol Infect Dis*. 2003;47(3):497–502.
349. Park WG, Bishai WR, Chaisson RE, Dorman SE. Performance of the Microscopic Observation Drug Susceptibility Assay in Drug Susceptibility Testing for *Mycobacterium tuberculosis*. *J Clin Microbiol*. 2002;40(12):4750–2.
350. Gan M, Liu Q, Yang C, Gao Q, Luo T. Deep whole-genome sequencing to detect mixed infection of mycobacterium tuberculosis. *PLoS One*. 2016;11(7):e0159029.
351. Schirmer M, Ijaz UZ, D’Amore R, Hall N, Sloan WT, Quince C. Insight

- into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 2015;43(6):1–16.
352. Feasey NA, Pond M, Coleman D, Solomon AW, Cosgrove CA, Delgado R, et al. Moxifloxacin and pyrazinamide susceptibility testing in a complex case of multidrug-resistant tuberculosis. *Int J Tuberc Lung Dis.* 2011;15(3):417–20.
353. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ.* 2014;2:e585.
354. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid whole-genome sequencing of mycobacterium tuberculosis isolates directly from clinical samples. *J Clin Microbiol.* 2015;53(7):2230–7.
355. Votintseva AA, Bradley P, Pankhurst L, Del C, Elias O, Loose M, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *J Clin Microbiol.* 2017;55(5):1285–98.
356. Khan M, Banerji D, Langelier C, Ph D, Bryson- C. Metagenomics for chronic meningitis: clarifying interpretation and diagnosis. *BioRxiv.* 2017;213561.
357. McNerney R, Clark TG, Campino S, Rodrigues C, Dolinger D, Smith L, et al. Removing the bottleneck in whole genome sequencing of *Mycobacterium tuberculosis* for rapid drug resistance analysis: a call to action. *Int J Infect Dis.* 2017;56:130–5.
358. Thoendel M, Jeraldo PR, Greenwood-Quaintance KE, Yao JZ, Chia N, Hanssen AD, et al. Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *J Microbiol Methods.* 2016;127:141–5.
359. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. *PLoS Med.* 2009;6(2):e1000002.
360. Iwai H, Kato-Miyazawa M, Kirikae T, Miyoshi-Akiyama T. CASTB (the comprehensive analysis server for the *Mycobacterium tuberculosis* complex): A publicly accessible web server for epidemiological

- analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. *Tuberculosis*. 2015;95(6):843–4.
361. Faksri K, Tan JH, Chaiprasert A, Teo Y-Y, Ong RT-H. Bioinformatics tools and databases for whole genome sequence analysis of *Mycobacterium tuberculosis*. *Infect Genet Evol*. 2016;45:359–68.
 362. Schleusener V, Köser CU, Beckert P, Niemann S, Feuerriegel S. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Sci Rep*. 2017;7:46327.
 363. Starks AM, Aviles E, Cirillo DM, Denkinger CM, Dolinger DL, Emerson C, et al. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. *Clin Infect Dis*. 2015;61(suppl 3):S141–6.
 364. Tarashi S, Fateh A, Mirsaeidi M, Siadat SD, Vaziri F. Mixed infections in tuberculosis: The missing part in a puzzle. *Tuberculosis*. 2017;107:168–74.
 365. Hatherell H-A, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med*. 2016;14(1):21.
 366. Maio N De, Worby CJ, Wilson DJ, Stoesser N. Bayesian Reconstruction of Transmission within Outbreaks using Genomic Variants. *BioRxiv*. 2017;213819.
 367. Martin J, Phan HTT, Findlay J, Stoesser N, Pankhurst L, Navickaite I, et al. Covert dissemination of carbapenemase-producing *Klebsiella pneumoniae* (KPC) in a successfully controlled outbreak: long- and short-read whole-genome sequencing demonstrate multiple genetic modes of transmission. *J Antimicrob Chemother*. 2017;72:3025–34.
 368. Witney AA, Cosgrove CA, Arnold A, Hinds J, Stoker NG, Butcher PD. Clinical use of whole genome sequencing for *Mycobacterium tuberculosis*. *BMC Med*. 2016;14(1):46.
 369. Schön T, Miotto P, Köser CU, Viveiros M, Böttger E, Cambau E. *Mycobacterium tuberculosis* drug-resistance testing: challenges, recent developments and perspectives. *Clin Microbiol Infect*. 2017;23(3):154–60.
 370. Votintseva AA, Pankhurst LJ, Anson LW, Morgan MR, Gascoyne-Binzi

- D, Walker TM, et al. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *J Clin Microbiol.* 2015;53(4):1137–43.
371. Senghore M, Otu J, Witney A, Gehre F, Doughty EL, Kay GL, et al. Whole-genome sequencing illuminates the evolution and spread of multidrug-resistant tuberculosis in Southwest Nigeria. *PLoS One.* 2017;12(9):e0184510.
372. Senghore M, Bayliss SC, Kwambana-Adams BA, Foster-Nyarko E, Manneh J, Dione M, et al. Transmission of *Staphylococcus aureus* from humans to green monkeys in the Gambia as revealed by whole-genome sequencing. *Appl Environ Microbiol.* 2016;82(19):5910–7.
373. Kwambana-Adams BA, Asiedu-Bekoe F, Sarkodie B, Afreh OK, Kuma GK, Owusu-Okyere G, et al. An outbreak of pneumococcal meningitis among older children (≥ 5 years) and adults after the implementation of an infant vaccination programme with the 13-valent pneumococcal conjugate vaccine in Ghana. *BMC Infect Dis.* 2016;16:575.
374. Roca A, Bojang A, Bottomley C, Gladstone RA, Adetifa JU, Egere U, et al. Effect on nasopharyngeal pneumococcal carriage of replacing PCV7 with PCV13 in the Expanded Programme of Immunization in The Gambia. *Vaccine.* 2015;33(51):7144–51.
375. Heyderman RS, Bentley SD, Pluschke G, Senghore M, Antonio M, Chaguzza C, et al. Region-specific diversification of the highly virulent serotype 1 *Streptococcus pneumoniae*. *Microb Genomics.* 2015;1(2).
376. Mahomed S, Naidoo K, Dookie N, Padayatchi N. Whole genome sequencing for the management of drug-resistant TB in low income high TB burden settings: Challenges and implications. *Tuberculosis.* 2017;107:137–43.

Supplementary materials

Table S1: Sample characteristics and basic outcomes from metagenomic analysis of samples extracted by methods E and F in chapter three

Seq. ref. no.	ZN smear grade	DNA concentration ($\mu\text{g}/\text{mL}$)	Total reads
E1	+++	564	303988
E2	+++	324	2943076
E3	+++	4.4	821772
E4	++	3780	1275312
E5	++	30	2018802
E6	++	970	831798
E7	+	0.588	188620
E8	+	32.4	3278340
F1	+++	27.8	989442
F2	+++	2.28	2170640
F3	++	71	1617808
F4	++	250	1204408
F5	++	7.7	1537676
F6	++	48.8	2411708
F7	+	25	2818238
F8	+	0.629	1851892

Table S2: Demographic and phenotypic data for samples in chapter five

Sample	Sex	Age	Treated	RIF	INH	PZA			Clinical drug susceptibility
				MODS	MODS	MGIT	Wayne	MODS	
MP051	M	43	N	S	S	S	S	S	DS
MP059	M	22	N	S	S	S	S	S	DS
MP069	F	50	N	S	S	S	S	S	DS
MP082	M	23	N	S	S	S	S	S	DS
MP124	F	38	N	S	S	S	S	S	DS
MP126	M	n/a	N	S	S	S	S	S	DS
MP167	F	56	N	S	S	S	S	S	DS
MP186	F	20	N	S	S	S	S	S	DS
MP199	F	70	N	S	S	S	S	S	DS
MP200	F	19	N	S	S	S	S	S	DS
MP203	M	20	N	S	S	S	S	S	DS
MP210	M	37	N	S	S	S	S	S	DS
MP284	F	41	N	S	S	S	S	S	DS
MP306	M	61	N	S	S	S	S	S	DS
MP307	M	29	n/a	S	S	S	S	S	DS
MP315	M	57	Y	S	S	S	S	S	DS
MP019	M	53	Y	S	S	S	S	R	DS
MP035	M	27	N	S	S	S	S	R	DS
MP038	M	57	N	S	S	S	S	R	DS
MP111	M	30	N	S	S	S	S	R	DS
MP176	M	38	Y	S	S	S	S	R	DS
MP248	M	35	N	S	S	S	S	R	DS
MP268	M	18	N	S	S	S	S	R	DS
MP092	M	61	Y	S	S	S	R	S	DS

MP175	M	30	N	S	S	R	S	R	DS
MP032	M	28	N	S	S	R	R	R	DS
MP040	M	17	N	S	S	R	R	R	DS
MP013	M	42	N	S	R	S	S	S	DS
MP041	M	29	N	S	R	S	S	S	DS
MP073	M	57	N	S	R	S	S	S	DS
MP118	F	75	N	S	R	S	S	S	DS
MP169	M	31	N	S	R	S	S	S	DS
MP026	M	18	N	S	R	S	S	R	DS
MP027	M	46	Y	S	R	S	S	R	DS
MP028	M	39	N	S	R	S	S	R	DS
MP050	M	22	N	S	R	S	S	R	DS
MP313	M	33	N	R	S	R	S	R	RR
MP031	M	47	N	R	S	R	R	R	RR
MP294	M	36	N	R	S	R	R	R	RR
MP005	F	83	Y	R	R	S	S	S	MDR
MP042	F	23	N	R	R	S	S	S	MDR
MP054	M	53	N	R	R	S	S	S	MDR
MP063	M	23	Y	R	R	S	S	S	MDR
MP120	M	18	N	R	R	S	S	S	MDR
MP182	M	67	n/a	R	R	S	S	S	MDR
MP275	M	20	N	R	R	S	S	S	MDR
MP276	F	37	N	R	R	S	S	S	MDR
MP308	n/a	n/a	n/a	R	R	S	S	S	MDR
MP320	F	50	n/a	R	R	S	S	S	MDR
MP360	M	21	N	R	R	S	S	S	MDR
MP202	F	50	Y	R	R	S	S	R	MDR
MP190	F	19	N	R	R	S	S	S	MDR

MP078	M	31	N	R	R	R	S	R	MDR
MP177	M	36	Y	R	R	R	S	R	MDR
MP249	F	n/a	N	R	R	R	S	R	MDR
MP342	M	66	N	R	R	R	S	R	MDR
MP364	M	28	n/a	R	R	R	S	R	MDR
MP367	M	31	n/a	R	R	R	S	R	MDR
MP369	M	18	N	R	R	R	S	R	MDR
MP123	M	37	N	R	R	R	R	S	MDR
MP170	M	21	Y	R	R	R	R	S	MDR
MP075	M	40	N	R	R	R	R	R	MDR
MP096	M	38	N	R	R	R	R	R	MDR
MP241	M	32	N	R	R	R	R	R	MDR
MP247	n/a	n/a	N	R	R	R	R	R	MDR
MP282	F	68	Y	R	R	R	R	R	MDR
MP287	M	20	N	R	R	R	R	R	MDR
MP341	M	36	N	R	R	R	R	R	MDR
MP366	M	15	N	R	R	R	R	R	MDR

Red font indicates that the result is associated with resistance; green font indicates sensitivity.

Table S3: Results of drug susceptibility testing for rifampicin in chapter five Red font indicates that the result is associated with resistance; green font indicates sensitivity. The number of resistant and sensitive reads were not reported by KVarQ. Mykrobe reported mutations with only the position and reference allele, not the allele present in the sample.

Sample	MODS	Mykrobe (R/S)	PhyResSE (R/S)	KVarQ
MP284	S	—	—	—
MP032	S	—	—	—
MP035	S	—	—	—
MP038	S	—	—	—
MP040	S	—	—	—
MP041	S	—	—	—
MP051	S	—	—	—
MP069	S	—	—	—
MP111	S	—	—	—
MP124	S	—	—	—
MP126	S	—	—	—
MP167	S	—	—	—
MP169	S	—	—	—
MP175	S	—	—	—
MP199	S	—	—	—
MP203	S	—	—	—
MP210	S	—	—	—
MP306	S	—	—	—
MP307	S	—	—	—
MP315	S	—	—	—
MP082	S	Phe425X in RpoB: 1/30	—	—

MP118	S	Gly426X in RpoB: :1/17 Ser431X in rpoB: 1/18	—	—
MP248	S	Gly426X in RpoB: 1/33 Asp335X in rpoB: 1/32	—	—
MP059	S	Gly426X in RpoB: 1/24	—	—
MP200	S	Gly442X in RpoB: 1/23	—	—
MP028	S	Gly442X in RpoB: 2/66	—	—
MP073	S	Asn438X in RpoB: 2/47	—	—
MP026	S	Pro439X in RpoB: 2/39	—	—
MP268	S	Gln432X in RpoB: 1/33	—	—
MP176	S	Arg447X in RpoB: 2/46 Ser450X in rpoB: 2/41	—	—
MP186	S	Ser428X in RpoB: 1/23 M434X in rpoB: 1/27	—	—
MP092	S	Ser450X in RpoB 1/23	—	—
MP019	S	Leu430X in RpoB: 28/0	Leu430Pro in RpoB: 41/0	Leu430Pro in RpoB
MP013	S	Asp435X in RpoB: 41/0	Asp435Tyr in RpoB: 58/0	Asp435Tyr in RpoB
MP027	S	Asp435X in RpoB: 49/0	Asp435Val in RpoB: 81/0	Asp435Val in rpoB

MP050	S	Leu452X in RpoB: 30/0 Phe425X in RpoB: 2/22	Leu452Pro in RpoB: 41/0	Leu452Pro in rpoB
MP005	R	His445X in RpoB: 2/57	—	—
MP054	R	—	Ile491Phe in RpoB: 81/0	—
MP031	R	Ser450X in RpoB: 5/12 Gly442X in RpoB: 2/20	Ser450Leu in RpoB: 11/32	—
MP042	R	His445X in RpoB: 33/0	His445Leu in RpoB: 53/1	His445Leu in rpoB
MP063	R	Asp435X in RpoB: 14/0	Asp435Val in RpoB: 31/0	Asp435Val in rpoB
MP123	R	Asp435X in RpoB: 20/0	Asp435Val in RpoB: 36/0	Asp435Val in rpoB
MP078	R	Asp435X in RpoB: 25/0	Asp435Val in RpoB: 36/0	Asp435Val in rpoB
MP177	R	Asp435X in RpoB: 27/1	Asp435Val in RpoB: 42/0	Asp435Val in rpoB
MP202	R	Asp435X in RpoB: 30/0	Asp435Val in RpoB: 44/1	Asp435Val in rpoB
MP249	R	Asp435X in RpoB: 30/0	Asp435Val in RpoB: 47/0	Asp435Val in rpoB
MP096	R	Asp435X in RpoB: 34/0	Asp435Val in RpoB: 51/0	Asp435Val in rpoB
MP170	R	Asp435X in RpoB: 37/0	Asp435Val in RpoB: 55/1	Asp435Val in rpoB
MP247	R	Asp435X in RpoB: 42/0	Asp435Val in RpoB: 57/0	Asp435Val in rpoB
MP342	R	Asp435X in RpoB: 21/0	Asp435Val in RpoB: 39/0	Asp435Val in rpoB
MP287	R	Asp435X in RpoB: 37/0	Asp435Val in RpoB: 64/0	Asp435Val in rpoB
MP364	R	Asp435X in RpoB: 38/0	Asp435Val in RpoB: 61/0	Asp435Val in rpoB

MP276	R	Asp435X in RpoB: 57/0	Asp435Val in RpoB: 86/2	Asp435Val in rpoB
MP341	R	Asp435X in RpoB: 16/0 Phe425X in RpoB: 2/14	Asp435Val in RpoB: 34/0	Asp435Val in rpoB
MP182	R	Asp435X in RpoB: 29/2 Thr427X in RpoB: 1/33	Asp435Val in RpoB: 49/3	Asp435Val in rpoB
MP320	R	Asp435X in RpoB: 25/0 Gly42X in RpoB: 1/25	Asp435Val in RpoB: 41/0	Asp435Val in rpoB
MP275	R	Ser450X in RpoB: 31/0 Asp435X in RpoB: 1/32	Ser450Leu in RpoB: 46/2	Ser450Leu in rpoB
MP360	R	Ser450X in RpoB: 28/0 Phe425X in RpoB: 2/31	Ser450Leu in RpoB: 50/0	Ser450Leu in rpoB
MP075	R	Ser450X in RpoB: 18/0 Phe433X in RpoB: 1/14	Ser450Leu in RpoB: 26/0	Ser450Leu in rpoB Ile491Val in rpoC
MP282	R	Ser450X in RpoB: 24/0 Gly426X in RpoB: 1/26	Ser450Leu in RpoB: 46/0	Ser450Leu in rpoB
MP313	R	Ser450X in RpoB: 22/2	Ser450Leu in RpoB: 29/2	Ser450Leu in rpoB
MP241	R	Ser450X in RpoB: 39/0	Ser450Leu in RpoB: 48/0	Ser450Leu in rpoB
MP367	R	Ser450X in RpoB: 13/0	Ser450Leu in RpoB: 20/1	Ser450Leu in rpoB
MP369	R	Ser450X in RpoB: 19/0	Ser450Leu in RpoB: 29/0	Ser450Leu in rpoB
MP294	R	Ser450X in RpoB: 23/0	Ser450Leu in RpoB: 34/0	Ser450Leu in rpoB

MP308	R	Ser450X in RpoB: 27/0	Ser450Leu in RpoB: 41/0	Ser450Leu in rpoB
MP366	R	Ser450X in RpoB: 27/0	Ser450Leu in RpoB: 52/0	Ser450Leu in rpoB
MP190	R	Ser450X in RpoB: 35/0	Ser450Leu in RpoB: 58/0	Ser450Leu in rpoB
MP120	R	Ser450X in RpoB: 20/1	Ser450Leu in RpoB: 34/0	Ser450Leu in rpoB Val483Gly in rpoC

Table S4: Results of drug susceptibility testing methods for isoniazid in chapter five

Sample	MODS	Mykrobe (R/S)	PhyResSE (R/S)	KVarQ
MP035	S	—	—	—
MP038	S	—	—	—
MP051	S	—	—	—
MP059	S	—	—	—
MP069	S	—	—	—
MP082	S	—	—	—
MP092	S	—	—	—
MP111	S	—	—	—
MP124	S	—	—	—
MP126	S	—	—	—
MP167	S	—	—	—
MP186	S	—	—	—
MP199	S	—	—	—
MP200	S	—	—	—
MP203	S	—	—	—
MP210	S	—	—	—
MP284	S	—	—	—
MP306	S	—	—	—

MP307	S	—	—	—
MP315	S	—	—	—
MP248	S	<i>fabG1</i> promoter mutation -16nt (A>X): 1/20	—	—
MP176	S	<i>fabG1</i> promoter mutation -8nt (T>X): 1/33	—	—
MP040	S	<i>inhA</i> promoter mutation -15nt (C>X): 15/0	<i>inhA</i> promoter mutation -15nt (C>T): 35/1	<i>inhA</i> promoter mutation -15nt (C>T)
MP032	S	<i>inhA</i> promoter mutation -15nt (C>X): 27/0	<i>inhA</i> promoter mutation -15nt (C>T): 52/1	<i>inhA</i> promoter mutation -15nt (C>T)
MP313	S	<i>inhA</i> promoter mutation -15nt (C>X): 27/8	<i>inhA</i> promoter mutation -15nt (C>T): 34/11	<i>inhA</i> promoter mutation -15nt (C>T)
MP294	S	<i>inhA</i> promoter mutation -15nt (C>X): 30/0	<i>inhA</i> promoter mutation -15nt (C>T): 49/1	<i>inhA</i> promoter mutation -15nt (C>T)
MP031	S	<i>inhA</i> promoter mutation -15nt (C>X): 32/0	<i>inhA</i> promoter mutation -15nt (C>T): 43/0	<i>inhA</i> promoter mutation -15nt (C>T)
MP268	S	<i>inhA</i> promoter mutation -15nt (C>X): 35/0	<i>inhA</i> promoter mutation -15nt (C>T): 56/0	<i>inhA</i> promoter mutation -15nt (C>T)
MP175	S	<i>inhA</i> promoter mutation -15nt (C>X): 50/0	<i>inhA</i> promoter mutation -15nt (C>T): 78/0	<i>inhA</i> promoter mutation -15nt (C>T)
MP019	S	<i>inhA</i> promoter mutation -15nt (C>X): 32/0	<i>inhA</i> promoter mutation -15nt (C>T): 45/0	<i>inhA</i> promoter mutation -15nt (C>T)
MP367	R	<i>inhA</i> promoter mutation -15nt (C>X):11/1	<i>inhA</i> promoter mutation -15nt (C>T): 20/1	<i>inhA</i> promoter mutation -15nt (C>T)
MP063	R	Ser315X in KatG: 13/0	Ser315Thr in KatG: 23/0 <i>inhA</i> promoter mutation -15nt (C>T): 26/0	Ser315Thr in KatG <i>inhA</i> promoter mutation -15nt (C>T)

MP054	R	Ser315X in gene KatG: 49/0	Ser315Thr in KatG: 63/0	Ser315Thr in KatG
MP050	R	Ser315X in gene KatG: 22/0	Ser315Thr in KatG: 29/0	Ser315Thr in KatG
MP241	R	Ser315X in gene KatG: 31/0	Ser315Thr in KatG: 40/0	Ser315Thr in KatG
MP177	R	Ser315X in gene KatG: 40/0	Ser315Thr in KatG: 54/0	Ser315Thr in KatG
MP073	R	Ser315X in gene KatG: 44/0	Ser315Thr in KatG: 56/0	Ser315Thr in KatG
MP169	R	Ser315X in gene KatG: 62/0	Ser315Thr in KatG: 91/0	Ser315Thr in KatG
MP027	R	Ser315X in gene KatG:49/0	Ser315Thr in KatG: 66/0	—
MP075	R	Ser315X in KatG: 12/0	Ser315Thr in KatG: 26/0	Ser315Thr in KatG
MP041	R	Ser315X in KatG: 17/0	Ser315Thr in KatG: 28/0	Ser315Thr in KatG
MP118	R	Ser315X in KatG: 18/7	Ser315Thr in KatG: 25/16	—
MP247	R	Ser315X in KatG: 22/0	Ser315Thr in KatG: 37/0	Ser315Thr in KatG
MP123	R	Ser315X in KatG: 24/0	Ser315Thr in KatG: 38/0	Ser315Thr in KatG
MP202	R	Ser315X in KatG: 25/0	Ser315Thr in KatG: 33/0	Ser315Thr in KatG
MP078	R	Ser315X in KatG: 28/0	Ser315Thr in KatG: 38/0	Ser315Thr in KatG
MP341	R	Ser315X in KatG: 28/0	Ser315Thr in KatG: 51/0	Ser315Thr in KatG
MP366	R	Ser315X in KatG: 28/0	Ser315Thr in KatG: 45/0	Ser315Thr in KatG
MP013	R	Ser315X in KatG: 30/0	Ser315Thr in KatG: 39/0	Ser315Thr in KatG
MP042	R	Ser315X in KatG: 30/0	Ser315Thr in KatG: 43/0	Ser315Thr in KatG

MP120	R	Ser315X in KatG: 30/0	Ser315Thr in KatG: 42/0	Ser315Thr in KatG
MP026	R	Ser315X in KatG: 32/0	Ser315Thr in KatG: 55/0	Ser315Thr in KatG
MP182	R	Ser315X in KatG: 33/0	Ser315Thr in KatG: 49/0	Ser315Thr in KatG
MP170	R	Ser315X in KatG: 34/0	Ser315Thr in KatG: 48/0	Ser315Thr in KatG
MP249	R	Ser315X in KatG: 35/0	Ser315Thr in KatG: 55/2	Ser315Thr in KatG
MP190	R	Ser315X in KatG: 37/0	Ser315Thr in KatG: 58/0	Ser315Thr in KatG
MP096	R	Ser315X in KatG: 38/0	Ser315Thr in KatG: 51/0	Ser315Thr in KatG
MP028	R	Ser315X in KatG: 67/0	Ser315Thr in KatG: 103/0	Ser315Thr in KatG
MP005	R	Ser315X in KatG: 72/0	Ser315Thr in KatG: 106/0	Ser315Thr in KatG
MP369	R	Ser315X in KatG:10/0	Ser315Thr in KatG: 19/0	Ser315Thr in KatG
MP282	R	Ser315X in KatG:16/0	Ser315Thr in KatG: 25/1	Ser315Thr in KatG
MP308	R	Ser315X in KatG:20/0	Ser315Thr in KatG: 29/1	Ser315Thr in KatG
MP360	R	Ser315X in KatG:22/0	Ser315Thr in katG: 38/0	Ser315Thr in KatG
MP364	R	Ser315X in KatG:24/0	Ser315Thr in KatG: 37/0	Ser315Thr in KatG
MP275	R	Ser315X in KatG:28/0	Ser315Thr in KatG: 43/0	Ser315Thr in KatG
MP342	R	Ser315X in KatG:30/0	Ser315Thr in KatG: 43/0	Ser315Thr in KatG
MP287	R	Ser315X in KatG:35/0	Ser315Thr in KatG: 56/0	Ser315Thr in KatG
MP320	R	Ser315X in KatG:42/0	Ser315Thr in KatG: 60/0	Ser315Thr in KatG

MP276	R	Ser315X in KatG:46/0	Ser315Thr in KatG: 71/0	Ser315Thr in KatG
--------------	---	-------------------------	----------------------------	----------------------

Red font indicates that the result is associated with resistance; green font indicates sensitivity. The number of resistant and sensitive reads are not reported by KvarQ. Mykrobe reported mutations with only the position and reference allele, not the allele present in the sample.

Table S5: Results of drug susceptibility testing for ethambutol, streptomycin, fluoroquinolones and aminoglycosides *Red font indicates that the result is associated with resistance; green font indicates association with sensitivity; black font indicated that no susceptibility prediction was made. Fractions indicate the number of reads in which the resistance-associated allele (numerator) and sensitive alleles (denominator) were reported by genomic prediction tools. The number of resistant and sensitive reads is not reported by KVarQ.*

Sample	Ethambutol			Streptomycin			Fluoroquinolones			Aminoglycosides		
	Mykrobe	PhyResSE	KVarQ	Mykrobe	PhyResSE	KVarQ	Mykrobe	PhyResSE	KVarQ	Mykrobe	PhyResSE	KVarQ
MP005	—	—	—	—	—	—	Pro86 X in GyrA: 2/58	—	—	—	—	—
MP013	—	—	—	—	—	—	—	—	—	—	—	—
MP019	—	—	—	—	—	—	—	—	—	—	—	—
MP026	—	—	—	—	—	—	—	—	—	—	—	—
MP027	Met306 X in EmbB: 56/0	Met306V al in EmbB: 72/0	Met306V al in EmbB	—	—	—	Asp94 X in GyrA: 46/0	Asp94Gly in GyrA: 74/0	Asp94Gly in GyrA	A1401G in gene <i>rrs</i> : 68/0	A1401G in gene <i>rrs</i> : 91/0	A1401 G in gene <i>rrs</i>

MP02 8	—	—	—	—	—	C491T in gene <i>rrs</i>	—	—	—	—	—	—
MP03 1	—	—	—	—	—	—	—	—	—	—	—	—
MP03 2	—	—	—	—	—	—	—	—	—	—	—	—
MP03 5	—	—	—	—	—	—	—	—	—	—	—	—
MP03 8	Met306 X in EmbB: 2/22	—	—	A514X in gene <i>rrs</i> : 32/0	A514C in gene <i>rrs</i> : 52/2	A514C in gene <i>rrs</i>	—	—	—	—	—	—
MP04 0	—	—	—	—	—	—	—	—	—	—	—	—
MP04 1	—	—	—	—	—	—	—	—	—	—	—	—
MP04 2	—	Gly406S er in EmbB: 52/1	—	—	—	—	—	—	—	—	—	—

MP05 0	—	Gly406Ser in EmbB: 41/0	—	Lys43Arg in RpsL: 32/6	Lys43Arg in RpsL: 42/8	Lys43Arg in RpsL	Asp94X in GyrA: 29/0	Asp94Gly in GyrA: 40/1	Asp94Gly in GyrA	A1401G in gene <i>rrs</i> : 38/0	A1401G in gene <i>rrs</i> : 56/0	A1401G in gene <i>rrs</i>
MP05 1	—	—	—	—	—	—	—	—	—	—	—	—
MP05 4	—	—	—	—	—	—	—	—	—	—	—	—
MP05 9	—	—	—	—	—	—	—	—	—	—	—	—
MP06 3	—	—	—	—	—	—	—	—	—	—	—	—
MP06 9	—	—	—	—	—	—	—	—	—	—	—	—
MP07 3	Met306X in EmbB: 2/46	—	—	—	—	—	—	—	—	—	—	—
MP07 5	—	—	—	—	—	—	—	—	—	—	—	—

MP078	—	Tyr319Ser in EmbB: 39/0	—	—	—	—	Ser91X in GyrA: 20/0	Ser91Pro in GyrA: 33/0	Ser91Pro in GyrA	A1401 G in gene <i>rrs</i> : 31/0	A1401 G in gene <i>rrs</i> : 49/0	A1401 G in gene <i>rrs</i>
MP082	—	—	—	—	—	—	—	—	—	—	—	—
MP092	—	—	—	—	—	—	—	—	—	—	—	—
MP096	—	Tyr319Ser in EmbB: 31/0	—	—	—	—	—	—	—	—	—	—
MP111	—	—	—	—	—	—	—	—	—	—	—	—
MP118	—	—	—	—	—	—	—	—	—	—	—	—
MP120	—	—	—	—	—	—	—	—	—	—	—	—
MP123	—	Tyr319Ser in EmbB: 34/0	—	—	—	—	—	—	—	—	—	—

MP12 4	—	—	—	—	—	—	—	—	—	—	—	—
MP12 6	—	—	—	—	—	—	Asp94 X in GyrA: 29/0	Asp94Gly in GyrA: 47/0	Asp94Gly in GyrA	—	—	—
MP16 7	—	—	—	—	—	—	—	—	—	—	—	—
MP16 9	—	—	—	—	—	—	—	—	—	—	—	—
MP17 0	—	Tyr319Ser in EmbB: 46/0	—	C517X in gene <i>rrs</i> : 1/32	—	—	—	—	—	—	—	—
MP17 5	—	—	—	—	—	—	—	—	—	—	—	—
MP17 6	—	—	—	Lys43Arg in RpsL: 60/0	Lys43Arg in RpsL: 77/0	Lys43Arg in RpsL	—	—	—	—	—	—
MP17 7	—	Tyr319Ser in EmbB: 45/0	—	—	—	—	Ala90 X in GyrA: 19/0	Ala90Val in GyrA: 30/0	Ala90Val in GyrA	—	—	—

MP18 2	—	—	—	—	—	—	—	—	—	—	—	—
MP18 6	—	—	—	—	—	—	—	—	—	A1401 G in gene <i>rrs</i> : 3/25	—	—
MP19 0	Met306 X in EmbB: 1/32	—	—	—	—	C491T in gene <i>rrs</i>	—	—	—	—	—	—
MP19 9	—	—	—	—	—	—	—	—	—	—	—	—
MP20 0	—	—	—	—	—	—	—	—	—	—	—	—
MP20 2	Met306 X in EmbB: 26/0	Met306V al in EmbB: 43/0	Met306V al in EmbB	—	—	—	Pro86 X in GyrA: 1/23 Asp94 X in GyrA: 26/0	Asp94Gly in GyrA: 40/0	Asp94Gly in GyrA	A1401 G in gene <i>rrs</i> : 28/0	A1401 G in gene <i>rrs</i> : 53/0	A1401 G in gene <i>rrs</i>
MP20 3	—	—	—	—	—	C491T in gene <i>rrs</i>	—	—	—	—	—	—

MP210	—	—	—	—	—	—	—	—	—	—	—	—
MP241	—	—	—	—	—	—	Ala90X in gyrA: 20/4	Ala90Val in GyrA: 28/6	—	—	—	—
MP247	—	Tyr319Ser in EmbB: 49/0	—	—	—	—	—	—	—	—	—	—
MP248	—	—	—	—	—	C491T in gene <i>rrs</i>	—	—	—	—	—	—
MP249	—	Tyr319Ser in EmbB: 43/0	—	—	—	—	—	—	—	—	—	—
MP268	—	—	—	—	—	—	—	—	—	—	—	—
MP275	Met306X in EmbB: 19/0	Met306Ile in EmbB: 26/2	Met306Ile in EmbB	Lys43Arg in RpsL: 21/1	Lys43Arg in RpsL: 38/1	Lys43Arg in RpsL	—	—	—	—	—	—

MP276	Met306X in EmbB: 37/0	Met306Ile in EmbB: 56/0	Met306Val in EmbB	—	—	—	Asp94X in GyrA: 32/0	Asp94Gly in GyrA: 52/0	Asp94Gly in GyrA	A1401G in gene <i>rrs</i> : 60/0	A1401G in gene <i>rrs</i> : 81/0	A1401G in gene <i>rrs</i>
MP282	Met306X in EmbB: 30/0	Met306Ile in EmbB: 42/0	Met306Ile in EmbB	Lys43Arg in RpsL: 41/0	Lys43Arg in RpsL: 51/0	Lys43Arg in RpsL	Ala90X in gyrA: 18/0	Ala90Val in GyrA: 30/0	Ala90Val in GyrA	A1401G in gene <i>rrs</i> : 32/0	A1401G in gene <i>rrs</i> : 38/0	A1401G in gene <i>rrs</i>
MP284	—	—	—	—	—	—	—	—	—	—	—	—
MP287	—	Tyr319Ser in EmbB: 53/0	—	—	—	—	—	—	—	—	—	—
MP294	—	—	—	—	—	—	—	—	—	—	—	—
MP306	—	—	—	—	—	—	—	—	—	—	—	—
MP307	—	—	—	—	—	—	—	—	—	—	—	—

MP308	Met306X in EmbB: 41/0	Met306Ile in EmbB: 57/0	Met306Ile in EmbB	Lys43Arg in RpsL: 32/0	Lys43Arg in RpsL: 54/0	Lys43Arg in RpsL	—	—	—	—	—	—
MP313	—	—	—	—	—	—	Asp89X in GyrA: 1/32	—	—	—	—	—
MP315	—	—	—	—	—	—	—	—	—	—	—	—
MP320	Met306X in EmbB: 21/0	Met306Val in EmbB: 39/0	Met306Val in EmbB	—	—	—	Asp94X in GyrA: 23/0	Asp94Gly in GyrA: 40/2*	Asp94Gly in GyrA	A1401G in gene <i>rrs</i> : 38/0	A1401G in gene <i>rrs</i> : 50/0	A1401G in gene <i>rrs</i>
MP341	—	Tyr319Ser in EmbB: 34/0	—	—	—	—	—	—	—	—	—	—
MP342	—	Tyr319Ser in EmbB: 46/0	—	—	—	—	—	—	—	—	—	—

MP360	—	Gly406Ala in EmbB: 33/0	Gly406Ala in EmbB	—	—	—	—	—	—	—	—	—
MP364	—	Tyr319Ser in EmbB: 49/0	—	—	—	—	—	—	—	—	—	—
MP366	—	—	—	—	—	—	—	—	—	—	—	—
MP367	—	—	—	—	—	—	—	—	—	—	—	—
MP369	—	—	—	—	—	—	—	—	—	—	—	—

Red font indicates that the result is associated with resistance; green font indicates sensitivity. The number of resistant and sensitive reads are not reported by KvarQ. Mykrobe reported mutations with only the position and reference allele, not the allele present in the sample.