The role of whole-genome sequencing technology in the control and treatment of *Mycobacterium tuberculosis* infection

Timothy M. Walker

Oriel College, Oxford

Nuffield Department of Medicine

A thesis submitted for the degree of Doctor of Philosophy

Hilary Term 2015

The role of whole-genome sequencing technology in the control and treatment of *Mycobacterium tuberculosis* infection

A thesis submitted for the degree of Doctor of Philosophy, Hilary Term 2015.

Timothy M. Walker, Oriel College, Oxford.

In 2013 an estimated 9 million patients were diagnosed with tuberculosis across the globe, leading to 1.5 million deaths. In the UK, just under 8,000 cases were notified. Where resources allow, tuberculosis control is based on the identification of outbreaks, and the timely diagnosis and appropriate treatment of infected patients. However, current methods for identifying tuberculosis outbreaks are limited in their specificity, whilst the definitive diagnostic tests remain culture-dependent and can hence take weeks before producing a result. Whole-genome sequencing (WGS) technology is now affordable, rapid and accurate, and in this thesis I explore its potential both for detecting transmission and for identifying the genetic variation underlying drug resistance.

Understanding the degree of *M. tuberculosis* genetic diversity within and between epidemiologically related individuals is a prerequisite to using WGS to identify *Mycobacterium tuberculosis* transmission. In chapter 3 I outline how this diversity is rarely greater than 5 nucleotide variants and also describe how the pattern of genetic diversity within an outbreak relates to the epidemiologically recognised transmission patterns. In chapter 4 I apply the findings from chapter 3 to all tuberculosis cases in Oxfordshire over a 6-year period to show that although most patients with tuberculosis were born in a high-incidence country, the odds of transmission among UK-born patients are in fact greater. These findings have contributed to the decision by Public Health England to invest in the routine whole-genome sequencing of *M. tuberculosis* from 2015. In chapter 5 I explore whether the potential utility of future sequence data can be increased by also predicting phenotypic drug susceptibility. I therefore devise an algorithm to characterise relevant genetic variation associated with phenotypic drug resistance or susceptibility.

I conclude that WGS has a significant contribution to make towards improving patient outcomes and decreasing onward transmission of disease.

TABLE OF CONTENTS

1 PREFACE

1.1 ACKNOWLEDGEMENTS

There are many people I would like to thank who have contributed to the work outlined in this thesis and who have supported me throughout. To start with, I would not have done any of this work had Matt Scarborough not told me one evening at hand-over for a Churchill hospital night shift to “go and speak to Derrick”. I did, and was subsequently enthused into applying for the UK-CRC fellowship.

I started my time in research learning basic laboratory skills in the NDCLS laboratories with the help, and to the bemusement, of Dai in particular. Tonya, a much harder taskmaster, taught me to hold a pipette properly. I then spent 6 months commuting to Birmingham where I learnt basic mycobacteriology skills under the supervision of Grace Smith, Jason Evans, Sarah Gardiner and the BMS staff at the Birmingham reference laboratory. I am grateful for their patience, support and flexibility.

The work in chapter 3 could not have been done without the friendly co-operation of the TB nurses from around the Midlands who were all receptive to my calls for help in reconstructing the epidemiology. All were generous with their time and data. Ruth Harrell played a huge role in liaising with the TB nurse at the Birmingham chest clinic, in with particular Cathy Brown. I am also very grateful to Martin Dedicoat, and to Philip Monk who was the enthusiast-in-chief for WGS and provided great insight into how epidemiological practice might be served by this new technology. Without Philip, Martin and Grace smoothing the

way for me it would not have been possible to do the study. I am grateful to Danny Wilson and David Eyre for their technical support as I was learning to use computers.

The work in chapter 4 was made possible through Chris Conlon, Noel McCarthy and the Oxford TB nurses, Lynne Parker, Karen Bennett and Sheila Churchill. The Public Health England Tuberculosis Section contributed data from their Enhanced Tuberculosis Surveillance system and all the UK mycobacterial reference laboratories provided samples from Oxfordshire patients dispersed around the country. I am grateful to Marcus Morgan for his help in the Oxford TB laboratory, and to David Wyllie for his generous advise and input on storage of sensitive patient data.

Jessica Hedge made a huge contribution to chapter 5 with her meticulous work on homoplasy, as did Carlos who I am sure has magic in his fingers. I am also very grateful to collaborators Melinda Munang, Zam Iqbal, Phelim Bradley, Stefan Niemann, Silke Feuerriegel, Thomas Kohl, Roland Diel, Philip Supply, Nazir Ismail and Shaheed Vally Omar for their intellectual inputs and for contributing samples and data to what has been a very large study.

The fantastic colleagues / friends I have shared an office with have made my time in research lots of fun. I am lucky to have worked in such a collaborative and entertaining atmosphere, so thank you Nicole, Claire, David, Bernadette, Nick, John, Anna, Ana, and Amy. Thanks also to Ali, for being Ali, to Helen and to Betty.

Overall however, the trinity of Tim Peto, Derrick Crook and Sarah Walker have been central to all the work I have done and each has been enormously

generous with their ideas, time, enthusiasm and/or red pen. Exhaustive and exhausting discussions and arguments with each, but in particular with Tim, have formed most of the ideas developed in this thesis. It has been a special learning environment.

On a personal level I am grateful to Pablo, Lizzie, James, Abi, and Maria, my wonderful friends in east Oxford and elsewhere who have kept me sane. And to Katherine, who has supported me enthusiastically and whose patience I somehow still have not managed to break.

If one can dedicate a thesis, I would like to dedicate it to my mother Ela and to my late father Nick who made me what I am and would have been very proud.

1.2 FUNDING

This thesis was supported by a Medical Research Council Research Training Fellowship (MRC/J011398/1).

Additional funding sources included the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre and the UK Clinical Research Collaboration (UKCRC) Modernising Medical Microbiology Consortium via the UKCRC translational Infection Research Initiative supported by the Medical Research Council, Biotechnology and Biological Sciences Research Council, and the National Institute for Health Research on behalf of the Department of Health (grant G0800778) and the Wellcome Trust (087646/Z/08/Z), as well as Wellcome Trust Sanger Institute core funding (grant 098051).

1.3 Ethics Statements

The Public Health Act (1984) requires that cases of *Mycobacterium tuberculosis* infection are reported to local authorities, who in turn have legislative cover to undertake necessary follow-up under the 2003 Health Protection Agency Act and the 2002 Statutory Instrument 1438. Consequently, all the work outlined in this thesis was performed within an ethical framework in Public Health England (formerly the Health Protection Agency) for service delivery evaluation.

1.4 DECLARATIONS AND ATTRIBUTIONS

Studies like those detailed in this thesis are by their nature collaborative and involve a large number of people. The bioinformatics pipeline that assembled the raw sequence data and filtered for variant calling was created and run by a specialist bioinformatics team. I had no role in this. However, each chapter is based on work that I did under the supervision of Professors Tim Peto, Sarah Walker and Derrick Crook. All figures and tables were made by me unless stated otherwise in sections 1.4.1 to 1.4.3 below, or in the individual figure / table legends.

1.4.1 CHAPTER 3 ATTRIBUTIONS

Sarah Walker and Tim Peto initiated the study design before I joined the group, identifying the need to explore within-host diversity through both cross-sectional and longitudinal sampling, as well as between-host diversity in family / household, and community clusters. I defined these groups and selected the samples for each. I cultured the samples and extracted the DNA. I travelled to meet with the relevant TB nurses across the Midlands to discuss each community cluster, and worked closely with Ruth Harrell who did the same for the outbreaks in Birmingham. I used existing python scripts within the MMM group to analyse the whole-genome sequencing (WGS) data, draw phylogenetic trees, and explore the sequence data manually to characterise the patters of nucleotide variants in each cluster. I designed the transmission networks under the supervision of Tim Peto. The thresholds of 5 and 12 SNPs were a product of discussions with Tim Peto who drew the original version of Figure 11. Tim Peto also wrote the STATA programme for Figure 13, whereas

Daniel Wilson wrote the R script from which I produced Figure 12. Ruth Harrell produced component B and C of Figure 15. I wrote the chapter and the published manuscript the chapter is based on.

1.4.2 CHAPTERS 4 ATTRIBUTIONS

Derrick Crook suggested I undertake a population study of Oxfordshire. I designed the geographical boundaries of the study, the period to be studied and identified all cases to be included. I retrieved and cultured the samples and extracted the DNA. I obtained the epidemiological data through interviews with other health professionals and co-analysed it with the genetic data. I decided the analysis should be based on a dichotomy between birth in high or low incidence countries, rather than the usual 'UK born / non-UK born' dichotomy, or on the basis of ethnicity, as has been done in other studies. I built the logistic regression model under the supervision of Sarah Walker and Tim Peto. Maeve Lalor generated Figure 19 on the basis of data I provided. I wrote the chapter and the published manuscript the chapter is based on.

1.4.3 CHAPTER 5 ATTRIBUTIONS

Georgia Kapatai started a small study to predict phenotypic resistance from 300 UK based sequences I had used for the study in chapter 3. She sought to identify known resistance determinants from *de novo* assembled sequences, but she left the group before completing the study. I took over from her, and together with Tim Peto designed a far larger, more ambitious study. The study design was an iterative process of many long conversations over one to two years over which the scope and size of the study steadily expanded as new samples became available. The design of the algorithm was a synthesis of

discussion between myself and Tim Peto, with valuable input from Sarah Walker as well. I extracted DNA for almost all the UK based samples included in this study. The other samples were donated by collaborators.

Carlos Del Ojo Elias wrote the python script that extracted the relevant sequence data from VCF (variant calling format) files for me. Zam Iqbal and Phelim Bradley ran the samples through Cortex to identify the indels. I selected the 23 candidate genes based on a catalogue of resistance determinants supplied by Silke Feuerriegel. I learnt to manipulate large data sets in STATA, and wrote the scripts to run the analysis under the supervision of Tim Peto and Sarah Walker. I created a phylogenetic tree of 3651 samples in RaxML from which Jessica Hedge identified and quantified all the homoplasic loci. I incorporated Jess's output into my analysis using STATA. Sarah Walker designed the suggested workflow in Figure 32 and Kate Niehaus created Figure 29 from data I provided. I wrote the chapter as well as the manuscript (now ready for submission) that this chapter is based on.

1.5 PAPERS ARISING

1. Walker TM, Ip CL, 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 Feb;13(2):137–46.

2. Walker TM, Monk P, Smith EG, Peto TEA. Contact investigations for outbreaks of Mycobacterium tuberculosis: advances through whole genome sequencing. Clin Microbiol Infect. 2013 Sep;19(9):796–802.

3. Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. The Lancet Respiratory Medicine. 2014 Apr;2(4):285–92.

4. A manuscript based on the work in Chapter 5 is being submitted to a journal for consideration.

1.6 ABBREVIATIONS

BCG, Bacille Calmette-Guerin

CTAB, cetyltrimethylammonium bromide

DNA, Deoxyribonucleic Acid

DST, Drug Susceptibility Testing

ETS, Enhanced Tuberculosis Surveillance

GTR, General time-reversible (model)

H, Isoniazid

HIV, Human Immunodeficiency Virus

IGRA, Interferon-γ Release Assays

IPT, Isoniazid Preventive Therapy

IUATLD, International Union Against Tuberculosis and Lung Disease

LJ, Loewenstein-Jensen

LPA, Line-probe assay

LTBI, Latent tuberculosis infection

MDG, Millennium Development Goals

MDR-TB, Multi Drug Resistant Tuberculosis

MGIT, Mycobacterial Growth Indicator Tube

MIC, Minimum Inhibitory Concentration

MIRU-VNTR, Mycobacterial Interspersed Repetitive Unit – Variable Number Tandem Repeat

MMM, Modernising Medical Microbiology

MODS, Microscopic Observation Direct Susceptibility Test

MRC, Medical Research Council (MRC)

NHS, National Health Service

NICE, National Institute for Health and Care Excellence

ONS, Office of National Statistics

OUH, Oxford University Hospitals

PAS, Para-aminosalicylic Acid

PCR, Polymerase-catalase Chain Reaction

PHE, Public Health England (formerly the HPA, Health Protection Agency)

R, Rifampicin

RFLP, Insertion Sequence-based Restriction Fragment Length Polymorphisms

S, Streptomycin

SNP, Single Nucleotide Polymorphism

T, Thiacetazone

TST, Tuberculin Skin Test

WHO, World Health Organization

XDR-TB, Extensively Drug Resistant Tuberculosis

Z, Pyrazinamide

2 BACKGROUND AND LITERATURE REVIEW

2.1 HISTORY

The Henle-Koch postulates, 50 years in their conception and laid out by Robert Koch in 1890, state that (1) a “parasite” must be found in each instance of disease in conjunction with the expected pathology, that (2) the parasite must not be found as a bystander in unrelated disease, and that (3) the parasite must evoke disease when isolated from an afflicted host, cultured and inoculated into a hitherto healthy host. On this basis, Koch argued, the parasite could be considered as the causative organism of disease, and not as an artefact of the disease itself.[1]

When Koch addressed the Berlin Physiological Society in 1882, it was to report on an organism he had successfully stained, visualized through microscopy and isolated in culture, which satisfied his postulates as the causative agent of tuberculosis.[2] He had conducted 13 animal experiments to demonstrate how pathology followed host inoculation, how the causative organism could be isolated in culture from lesions identified on necropsy, and how that culture could in turn cause disease in additional hosts upon fresh inoculation.

Just 18 years earlier in his 1864 paper to the Boston Society for Medical Observation, “Is consumption ever contagious, or communicated by one person to another in any manner?”, Henry Bowditch argued that although Galen and Aristotle both feared contagion from consumption, latter day opinion “assured that contagion, …, is, at least in this country, a delusion”.[3] Koch’s findings thus constituted a paradigm shift from the accepted wisdom that consumption was a

diathesis, less a contagion, as acknowledged in his 1906 Nobel lecture that prior to his discovery, "…even in its most dangerous form, pulmonary phthisis, was not considered infectious".[4]

Phthisis (pulmonary tuberculosis) was the cause of one in four deaths in London between 1790-1810,[5] but its contribution to the overall death rate declined over the subsequent 100 years, as it did in other western European settings, largely prior to Koch's discovery (Figure 1).[6] Reasons for this decline are disputed but hypotheses range from improvements in living standards to the isolation of 'consumptives' in Poor Law infirmaries or sanatoria,[5,6] policies whose unforeseen consequences will have been to interrupt transmission.

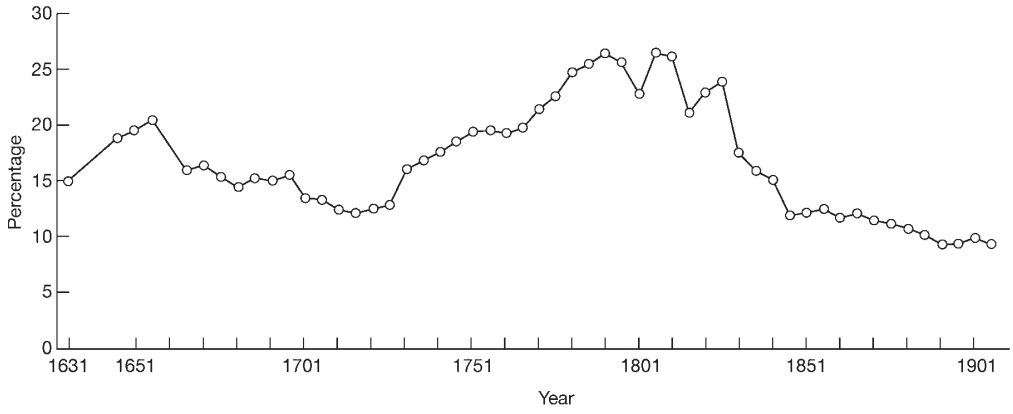


Figure 1: Percentage of phthisis deaths to total deaths from all causes, London 1631-1910. Taken from Leonard G Wilson, Commentary: Medicine, population, and tuberculosis, International Journal of Epidemiology 2005; 34: 521-524[5]

A genome-sequence based analysis of the evolutionary history of *Mycobacterium tuberculosis* has suggested the pathogen emerged around 70,000 years ago in Africa and was spread around the globe as *Homo sapiens* migrated to other continents.[7] It is argued that *M. tuberculosis* adapted to increases in population density by more rapid onward transmission between

hosts. Industrialization, urbanization and population growth will thus have provided favourable conditions for the spread of tuberculosis in Europe from the early to mid 18th century.[8] Although the detection of ancient DNA has confirmed the presence of *M. tuberculosis* in Africa, Asia and the Americas prior to European colonisation,[9] the modern epidemic curves on these continents appear to have lagged behind that in Europe.[10] A likely explanation is the re-introduction of the disease from European colonisers at the height of the European pandemic. Troop and population movements associated with colonial conquest will have contributed to outbreaks of infectious disease, including tuberculosis.[11,12] For example, it was observed that tuberculosis was more common among British troops than “natives” in India in the late 19th century.[13] The increase in incidence among Indians over the subsequent decades may well therefore relate to a surge in British troop numbers after the 1857 Indian mutiny, further facilitated by the building of the railways that will have provided efficient channels of transmission for disease.[14-16]

Today, tuberculosis remains a disease of poverty in high and low/middle income countries alike, with the global burden felt most acutely in Africa, Asia and South America where many of the world’s two billion people infected with latent or active tuberculosis can be found (Figure 2).[17]

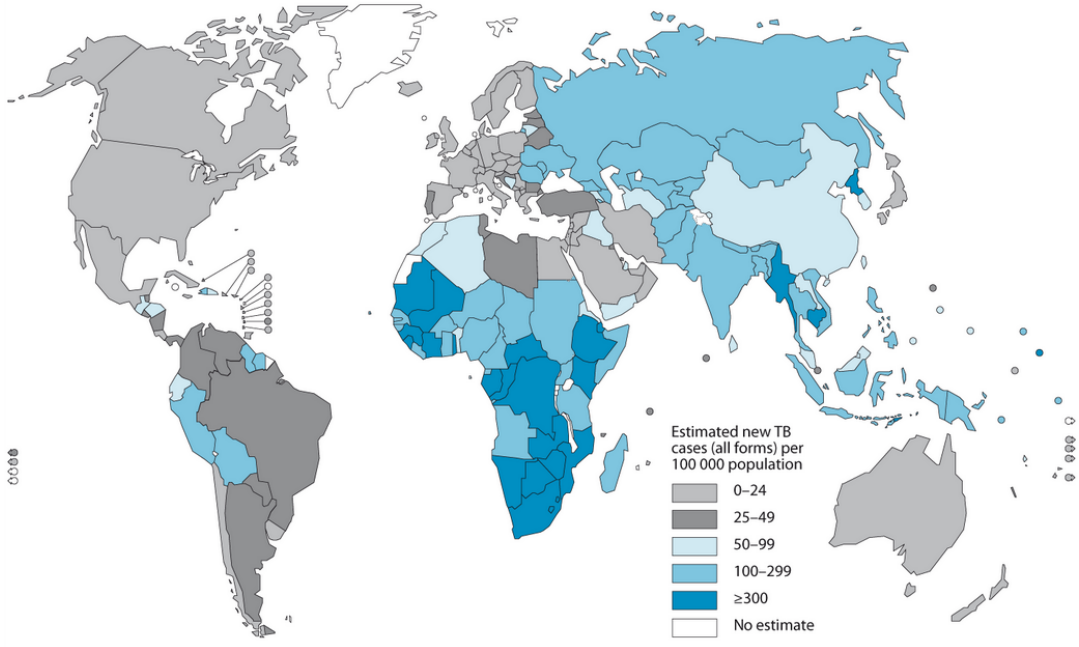


Figure 2 Estimated tuberculosis incidence rates, 2012. Taken from the World Health Organization Global Tuberculosis Report 2013 [18]

2.2 PATHOGENESIS AND IMMUNOLOGY

M. tuberculosis is transmitted via pulmonary aerosols and phagocytosed by alveolar macrophages within each new host.[19] Although a small subset of hosts are thought to clear the infection at this early stage, in the great majority of hosts the ensuing adaptive immune response leads to the formation of granulomas.[20,21] The primary pulmonary granuloma and associated draining lymph nodes are known as the "Ghon complex",[22] which can be radiologically apparent and may be associated with symptoms if primary disease is manifested.[23,24] Granulomas can subsequently evolve in one of three ways. In necrotising granulomas the bacilli survive in a hypoxic environment made up of macrophages, monocytes and neutrophils, surrounded by an outer layer of lymphocytes and fibroblasts. The centre, described as "caseous", is filled with

necrotic material from dead cells. Alternatively, granulomas also exist in a non-necrotising state, predominantly composed of macrophages, or lastly in a fibrotic state as a conglomeration of fibroblasts (Figure 3).[20,25,26]

Although the bacilli may be contained within granulomas for the host's entire lifetime in what is termed "latent" tuberculosis infection (LTBI), the infection can also re-activate at a later date, either spontaneously or as a result of an immunosuppressive episode. The blockade of Tumour Necrosis Factor (TNF) with monoclonal antibodies for a number of autoimmune diseases has been associated with reactivation of tuberculosis and represents one such mechanism, albeit iatrogenic, through the inhibition of normal macrophage function.[27,28]

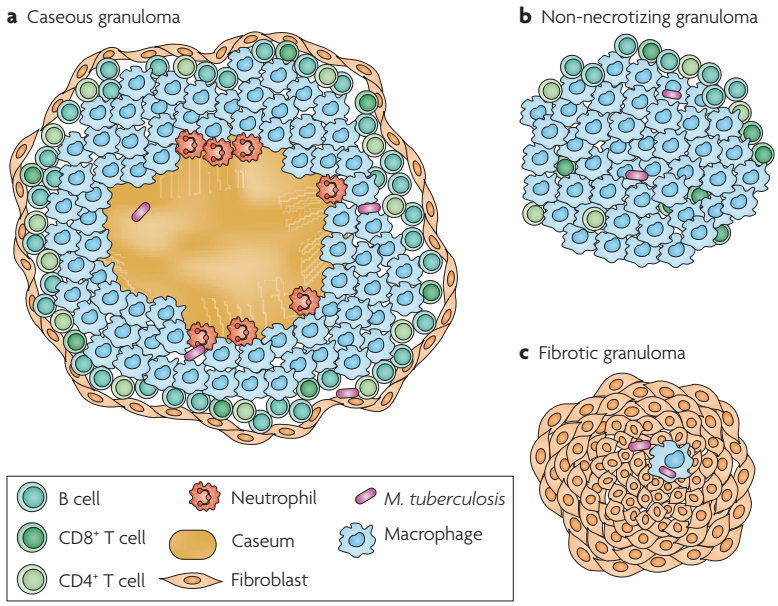


Figure 3: Tuberculous granulomas, taken from Clifton E. Barry et. al., The spectrum of latent tuberculosis: rethinking the biology and intervention strategies, Nature Review Microbiology; 7, 845-855 (2009).[25] Non-necrotising granulomas are usually associated with active disease and fibrotic granulomas with latent infection, whilst necrotising granulomas are associated with either state.

Reactivated tuberculosis can affect almost any anatomical site.[29] It can manifest as post-primary disease in the upper lobes of the lung as an exudative pneumonia, cavities can form as pulmonary granulomas breakdown to release bacilli into major airways, or the bacilli can migrate from the lungs via the lymphatics or vasculature to other organs.[30] In the case of tuberculous meningitis, the bacilli disseminate to the meninges or brain parenchyma to form what are known as Rich foci. When these foci rupture into the subarachnoid space, meningitis ensues.[31] In other forms of disease such as spinal tuberculosis, also known as “Pott’s disease”, there is direct infiltration into bone from adjacent lymph nodes, or via a vascular route.[32] If widespread, such a vascular dissemination of bacilli from a pulmonary focus to multiple organs, resulting in a large number of necrotic lesions, is known as “miliary” tuberculosis.[33,34] The term dates back to 1700 when Jean-Jacques (John Jacob) Manget likened the lesions on autopsy to millet seeds.[35]

Of all the forms of disease, only pulmonary tuberculosis is transmissible and will thus ultimately ensure the evolutionary success of the organism. The hyper-conservation of T-cell epitopes in *M. tuberculosis* has been used to argue that this facilitates immune recognition, leading to either latent infection or cavity formation. Both may advantage the pathogen as periods of latency could be a means of transmitting to future generations, whereas cavity formation can aid the onward transmission when active disease develops.[36]

The role of T-cells in determining the course of the disease is best illustrated by the impact of CD4+ T-cells diminishing in number or function, as in the context of the Human Immunodeficiency Virus (HIV).[37] Perhaps the most dramatic

example to date of rapid progression to disease is the outbreak of extensively drug resistant (XDR) tuberculosis among HIV co-infected patients in Tugela Ferry, KwaZulu Natal, South Africa, where 52 of 53 patients died within a median of 16 days (IQR 6-37) of diagnosis.[38] Both in vitro and in vivo data indicate that *M. tuberculosis* specific CD4+ T-cells are relatively more susceptible to HIV than, for example, cytomegalovirus specific T-cells, leaving HIV-infected hosts preferentially at risk of tuberculosis.[39] Indeed, the role of CD4+ T-cells in the formation of granulomas is suggested by an association between the CD4 count and the frequency of cavitation in pulmonary disease, and consequently the diminished infectiousness of HIV co-infected patients.[40] If CD4+ T-cells thus contribute towards disease control in latency as well as cavity formation in active disease, their relative deficiency leaves patients susceptible to a more rapid progression to disease, but renders them relatively less infectious.

The memory T-cell responses evoked by the Bacille Calmette-Guerin (BCG) vaccine have been shown to protect infants from disease, especially meningitis, but have been less successful at protecting adults from pulmonary disease.[41] Named after Albert Calmette and Camille Guerin, BCG is a live strain of *Mycobacterium bovis* attenuated through passage. It was first used in 1920, is part of the WHO expanded immunization programme, and has been given to more than 4 billion recipients.[42] Although other vaccines are in development and have been tested in clinical trials, none has yet proven an adequate adjunct or replacement for BCG.[43]

T-cell responses are also the basis for both tuberculin skin tests (TST) and interferon-γ release assays, each of which are designed to identify patients infected with tuberculosis but not manifesting clinical disease. The TST involves an intradermal injection of tuberculin purified protein derivative with the extent of any delayed-type hypersensitivity reaction manifesting as induration of the skin to be read between 48 and 72 hours later. The test was developed in 1907 by Charles Mantoux on the back of work by Robert Koch and Clemens von Pirquetto,[44] with a similar test based on multiple adjacent punctures of the skin developed by Frederick Heaf in 1951.[45] Although the latter is less operator-dependent, it is also less specific because the increased dose of tuberculin involved is more likely to evoke a hypersensitivity response among patients exposed to non-tuberculous mycobacteria.[46] Both versions of the TST suffer from a lack of specificity in BCG vaccinated individuals due to cross-reactivity, as well as poor sensitivity in patients with impaired cellular immunity. The newer interferon-γ release assays (IGRA) instead measure T-cell responsiveness to the *M. tuberculosis* specific antigens (ESAT-6 and CFP-10) encoded by the region of difference 1 (RD1) that was lost through passage in BCG.[47]

2.3 Clinical features

The classic constitutional symptoms of tuberculosis include fevers, night sweats and weight loss, all typically reported over a period of weeks to months.[48] In addition, the disease will be associated with localized symptoms according to the form the disease takes. Pulmonary disease will thus normally present with a cough or haemoptysis and eventually breathlessness; central nervous system

disease with a headache, neck stiffness and altered consciousness; and Pott's disease with back pain, sometimes a gibbus, and eventually paralysis. Some patients with pulmonary disease may be asymptomatic despite having culture positive disease.[49]

Based on data from the pre-chemotherapy era, the estimated case fatality rate of smear positive and smear negative pulmonary tuberculosis in the absence of treatment is approximately 70% and 20% respectively.[50] Some extra-pulmonary manifestations of disease such as miliary tuberculosis or tuberculous meningitis are universally fatal if left untreated,[35,48] whereas the natural history of other forms of disease such as intra-thoracic lymph node disease can be more benign.[51]

The disease is transmitted by patients with pulmonary disease, of whom those with cavities that function as open reservoirs from which large numbers of bacilli can be coughed up and aerosolised, are theoretically the most infectious of all.[52] However, the infectiousness of different patients can vary and some highly infectious patients do not have cavities.[53,54] In patients in whom LTBI reactivates as active disease, about half manifest as pulmonary disease, of whom about half in turn are sputum smear-positive and hence considered most infectious.[55]

2.4 MICROBIOLOGY

The Mycobacteriaciae, a genus of Actinobacteria, are characterized by a hydrophobic, mycolic acid (lipid) rich cell wall that is resistant to decolourisation with alcohol or weak acids after staining with carbol fuchsin.[56,57] As of 20th

August 2014 there were 169 species of mycobacteria with standing in the nomenclature, 11 new species having been approved in 2013 and one in 2014.[58] As part of the Runyon classification scheme, the species have traditionally been divided into rapid and slow growers according to their ability to grow visible colonies within seven days in culture.[59] It is among the slow growers that the most prolific mammalian pathogens are found, including *Mycobacterium leprae* (the causative organism of leprosy), and the sub-species within the *Mycobacterium tuberculosis complex* (the organisms causing clinical tuberculosis). Although *M. tuberculosis sensu strictu* is the most common human pathogen within the complex, *M. bovis*, *M. africanum*, *M. canettii*, *M. microti*, *M. caprae* and *M. pinnipedii* can all cause similar disease in a variety of mammalian hosts.[60] In nutrient rich conditions the replication time for *M. tuberculosis* is about 20 hours.[61]

2.5 MICROSCOPY

The characteristic microbiological features of *M. tuberculosis* provide diagnostic targets for laboratory based testing. It was through the application of dye and the use of slide microscopy that *M. tuberculosis* bacilli were first visualized. Robert Koch's initial staining protocol involved prolonged exposure to methylene blue and then Bismarck brown, a technique that was modified after Paul Ehrlich's observation that the bacilli were "acid-fast" (resistant to decolourisation), Franz Ziehl's use of carbolic acid and Edward Rindfleisch's heating of the slide to facilitate uptake of dye through the thick cell wall. Friedrich Neelsen then combined fuchsin and carbolic acid in solution.[62] Although this "Ziehl-Neelsen" technique remains in use, fluorescent light

microscopy is now the method of choice for the examination of primary clinical samples in many laboratories.[63] Because the auramine dye used in fluorescent light microscopy in place of carbofuchsin provides a stronger contrast against a counter stained background, the microscopy time has more than halved and the sensitivity increased by an average of 10% over the conventional Ziehl-Neelsen method.[64] More recent advances of importance to high burden settings have demonstrated how auramine based microscopy can now even be automated.[65] Because of their similar cell wall, the presence of any species of mycobacteria in sufficient quantity will result in positive smear microscopy, but in the right clinical context a microscopy result can nevertheless prove a valuable diagnostic indicator. Indeed, in many low-income settings it is the only available test.[64]

2.6 CULTURE

It was only in the 1950s that *M. tuberculosis* culture was widely adopted for diagnostic purposes. Although it remained common to culture *M. tuberculosis* in live guinea pigs until the early 1970s,[66] culture media were also widely used. Initially egg-based in the form of Loewenstein-Jensen (LJ) slopes, liquid culture media containing antibiotics selective for mycobacteria (Middlebrooks broth) proved a significant advance in the early 1970s.[67] Performing microscopy on *M. tuberculosis* cultures meant visualization of the characteristic serpentine cords of *M. tuberculosis*, apparent when sufficient biomass is at hand, thereby differentiating *M. tuberculosis* from other mycobacteria (although *M. chelonae* and *M. kansasii* can produce a similar appearance).[68,69]

Culturing mycobacteria also facilitated the development of biochemical and phenotypic tests by which to identify *M. tuberculosis*. In 1956 Kiyoshi Konno suggested that the production of niacin in culture was unique to *M. tuberculosis* among the mycobacteria.[70] In fact it is *M. tuberculosis*' inability to metabolize niacin that leads to its accumulation, a characteristic shared by a few other species including *M. simiae*,[71,72] but the observation nevertheless led the first of a range of biochemical tests to identify *M. tuberculosis*.[70,71] Ernest Runyon's eponymously named phenotypic classification system, based on speed of growth, colonial morphology and pigmentation, was introduced three years later, in 1959.[70]

In 1977 Gardner Middlebrook developed 7H12 liquid medium containing C14 labelled palmitic acid in which mycobacterial growth could be detected by radiometric readings.[73] This approach detected of *M. tuberculosis* in clinical samples faster than solid media, and was further updated in the late 1990's with the introduction of the fully automated BACTEC 960 system that dispensed with the need for radioactive material or manually operated instruments. Since then the 7ml of Middlebrook's broth contained in the BACTEC Mycobacterial Growth Indicator Tube (MGIT) has become the initial culture medium of choice in many laboratories.[74] More recent innovations have been proposed to further reduce the time to culture positivity through the addition of ascorbic acid to standard culture media and the use of a micro-aerophilic environment, but these have yet to be widely implemented.[75]

2.7 PHENOTYPIC DRUG SUSCEPTIBILITY TESTING

Phenotypic drug susceptibility testing aims to identify whether a particular strain of *M. tuberculosis* is either sensitive or resistant to a given drug. Georges Canetti and colleagues sought to define the terms 'sensitive' and 'resistant' in their 1963 WHO bulletin:

> " "Sensitive" strains are those that have never been exposed to the main anti-tuberculosis drugs ("wild" strains) and that respond to these drugs, generally in a remarkably uniform manner. "Resistant" strains are those that differ from sensitive strains in their capacity to grow in the presence of higher concentrations of a drug. This definition of resistance is based on the laboratory response; strains that are resistant in this sense do not *necessarily* fail to respond to the usual doses of the drug in the lesions of the patient. However, a diminished clinical response is likely to occur whenever resistance is demonstrated in the laboratory...."[76]

There are three traditional types of phenotypic drug susceptibility testing (DST) endorsed by the WHO, which are broadly interpreted according to the principle set out by Canetti *et. al*.. The absolute concentration method estimates the minimum inhibitory concentration (MIC) of drug on growth of *M. tuberculosis* through serial two-fold dilutions of a drug, incorporated into agar or liquid medium. Resistance is assessed according to the number of colonies growing above a critical concentration of the drug (e.g. more than 20 colonies).[77] The critical concentration is set at a level where the drug inhibits 95% or more of previously unexposed wild-type strains.[78] The resistance ratio method adds a known susceptible control isolate to the equation, allowing a ratio of the MICs of the test and control isolates to be calculated. A ratio of less than 2, or more than 8 signifies susceptibility or resistance, respectively. Both these approaches are dependent on the size of the inoculum, upon which the third method, the proportion method, is less dependent. It measures the proportion of colonies

growing on both drug containing and drug free media, using a set threshold (commonly 1%) for the inferred proportion of resistance organisms present in the original clinical sample of interest. In liquid culture, where colonies cannot be counted, the proportion method involves inoculating drug containing media with the sample of interest, and control media with a 100-fold dilution, and calculating the ratio of the two growth indices.[77]

What each of these three approaches share is their slow turnaround time and, in less than expert hands, poor reproducibility. A number of newer DST methods have been endorsed by the WHO for use in reference laboratories.[79] The Microscopic Observation Direct Susceptibility (MODS) test uses a micro-titre plate in which *M. tuberculosis* can be cultured from clinical specimens simultaneously to DST for a range of drugs. An inverted light microscope is used to identify the characteristic serpentine cords *M. tuberculosis* forms from an early stage in liquid culture.[69] Colorimetric micro-titre plates use the colour change that results from the reduction of indicators such as alamar blue or resazurin in the presence of mycobacterial growth at different concentrations of drug to determine MICs.[80] The reduction of nitrate by *M. tuberculosis* growth can also be used to trigger a colour change.[79]

2.8 TREATMENT

In classical times fresh air and high altitude had been considered therapeutic for consumptives.[81] The idea re-emerged in 1854 with Hermann Brehmer's founding of a sanatorium in Silesia with the controversial aim of curing patients. Although initially decried as heretical, the approach became popular and the sanatorium movement lasted over a century. First conceived of to provide cold,

fresh mountain air and exercise-based therapy, later sanatoria focused on rest and gave less credence to the importance of altitude.[82] There may nevertheless have been a beneficial effect from the exposure to sunlight that might have increased vitamin D levels, contributing to healthy macrophage activity.[83] However, their main impact on public health may have been through the isolation of infectious cases away from the wider community rather than to alter the natural history of disease in infected individuals themselves.[84]

Drug treatments were also pioneered in the years prior to 1945, including gold and sulphone compounds[85] but the introduction of streptomycin (S) in 1945 heralded the start of the modern chemotherapy era for tuberculosis.[67] The efficacy of streptomycin, which is administered by intra-muscular injection and inhibits DNA translation by binding 16s ribosomal RNA,[86] for the treatment of pulmonary tuberculosis was assessed in the 1946-8 Medical Research Council (MRC) study, the first ever randomized control trial of its kind (streptomycin vs. bed rest). Although the study found a clear mortality benefit in the treatment arm as well as evidence of radiological and clinical improvement over 6 months, the improvements slowed over the duration of the study and by 6 months 35 of 41 patients produced culture positive specimens highly resistant to streptomycin.[87] A parallel MRC investigation into the efficacy of streptomycin in tuberculous meningitis also demonstrated some benefit, this time with comparatively little drug resistance emerging.[88]

In response to reports from Sweden about the apparent efficacy of a new oral drug and the rapid evolution of streptomycin resistance observed in the first

1948 trial, the MRC conducted follow-up studies into the use of para-aminosalicylic acid (PAS) as monotherapy, and in combination with streptomycin. The mechanism of action of PAS remains poorly understood.[86] Published in 1950 and 1952 these studies demonstrated that PAS was efficacious, albeit not as efficacious as streptomycin, but also that combination therapy reduced the emergence of drug resistance in a dose dependent manner: 40-50% of patients given lower-dose treatment (5g and 10g daily respectively) developed streptomycin resistance after four months of therapy, compared to just 15% at the a higher dose (20g).[89,90]

Isoniazid (H) and pyrazinamide (Z), both orally administered pro-drugs, were introduced in 1952. Isoniazid is converted to its active form by the catalase/peroxidase enzyme encoded by the katG gene. It inhibits mycolic acid synthesis by interfering with the inhA derived NADH-dependent enoyl-ACP reductase. Pyrazinamide is activated by the pncA-encoded pyrazinamidase/nicotinamidase enzyme, and accumulates within bacilli under acidic conditions, inhibiting membrane transport. It can therefore act against non-replicating bacilli.[86] Early reports of both *in vitro* and *in vivo* resistance to both these drugs led to warnings about the use of any anti-tuberculosis drugs in general as monotherapy.[91] Isoniazid was subjected to an MRC trial comparing it to the established combination of streptomycin and PAS. Although clinical outcomes were similar between the two groups, 10%, 50% and 70% of patients on isoniazid monotherapy developed resistance after 1, 2 and 3 months of therapy respectively.[92] The trial was expanded to include an additional arm receiving streptomycin and isoniazid combination therapy. Patients in both combination therapy arms then fared better than those on

isoniazid monotherapy, with 4%, 9% and 13% of patients on combination therapy developing resistance to isoniazid 1, 2 and 3 months into their treatment.[93] The authors drew a series of still apposite conclusions, that no drug should be used as monotherapy; that drug susceptibility testing should be performed where possible to avoid functional monotherapy in the event of pre-existing resistance to one drug; and that the history of the patient or their contacts should be explored for evidence of previous treatment or drug resistance prior to the commencement of therapy where it is not reasonable to await the results of susceptibility testing.

The problem of drug resistance was quantified in 1955-6 when the MRC conducted the first national survey, reporting a 2.5%, 2.6% and 1.3% prevalence of resistance to streptomycin, PAS and isoniazid respectively. The higher prevalence of resistance to the first two drugs was considered a consequence of their earlier introduction and more widespread use. But it was also noted that patients with drug resistant strains were more likely to have been contacts of other patients treated with and resistant to the relevant drug, suggesting the presence of both primary and acquired drug resistance in the community.[94]

Concern also remained about the optimal duration of therapy, especially in patients with cavitatory disease. This led the MRC to set up a further study in 1955 (published in 1962), randomising patients to between six months and four years of PAS+H or PAS+H plus six weeks of daily streptomycin. Not only did the triple therapy arm perform better, but patients receiving chemotherapy for up to two years suffered fewer relapses.[95] The need for three drugs was also

investigated in a parallel study conducted by the International Union Against Tuberculosis and Lung Disease (IUATLD), published in 1964. Here patients were treated with streptomycin, PAS and isoniazid for 28 weeks and then with either PAS and isoniazid, or just isoniazid for a further 24 weeks. The success led to a three month regimen of in-patient based triple therapy followed by 9 months follow-on PAS and isoniazid being widely practiced throughout Europe thereafter, despite the high cost of both PAS and of hospitalisation, and the high drop-out rate from the study.[96,97]

The high financial cost of both PAS and streptomycin led to a search for alternative drugs to partner isoniazid. The anti-tuberculosis activity of thiacetazone (T), a pro-drug that is activated by ethA and inhibits mycolic acid cyclopropanation, was established in 1946, as were its toxic side effects seen at higher doses.[98] The drug was tested at lower doses in a series of trials in Kenya, Uganda and Tanzania between 1960 and 1970 as a cheaper alternative to PAS.[99,100] These established the efficacy of thiacetazone as a partner for isoniazid and that 8 weeks of initial daily streptomycin improved outcomes in patients with more advanced cavitatory disease.[100] The findings were further replicated at the Madras (now Chennai) Chemotherapy Centre where home therapy was pioneered over lengthy sanatoria stays. Patients treated at home with isoniazid and PAS (later thiacetazone) had the same outcome as in-patients (90% remained culture negative after 5 years follow-up) with no additional risk to household contacts.[101,102] Compliance was tested using urine ferric chloride testing (for PAS) as well as stock count, but the problem of how to efficiently maximise compliance was the catalyst for new studies exploring both intermittent and “short-course” therapy.[67]

The introduction of orally administered ethambutol (E) (active against cell wall arabinogalactan synthesis) and rifampicin (R) (which binds the β-subunit of RNA-polymerase),[86] and the revisiting of pyrazinamide (originally released in 1952), allowed new, shorter and potentially better tolerated regimens to be explored. A 6-month combination of isoniazid plus ethambutol (HE) was initially trialled by Japanese researchers in 1966, achieving a 97.6% culture conversion rate compared to 88.8% of patients in the control group receiving PAS and isoniazid.[103] Further studies conducted by the MRC demonstrated the effectiveness of one-year courses of HR, HE and H+PAS, each with streptomycin as an initial supplement.[104] Between 1972 and 1974 the MRC reported on the progress of their study comparing six month regimens of SHR, SHZ, SHT, and SH, to the established East African regimen of 2STH/16TH over 18 months. At 18 months follow-up the relapse rates were 3% for the rifampicin and 8% for the pyrazinamide containing arms, comparing favourably with the standard 18-month regimen (3%). The thiacetazone and dual therapy arms fared significantly worse.[105]

These studies were rapidly followed by a further MRC trial in East Africa exploring the benefits of initial intensive therapy within the 6-month regimens, and of the added value, if any, of streptomycin to a regimen containing both isoniazid and rifampicin. Six-month regimens of HR, SHR, SHRZ/TH and SHRZ/SHZ were compared, with the results demonstrating little difference across the study arms, and consequently little benefit from adding streptomycin to H+R. Pre-existing isoniazid resistance was however identified as a predictor of poor treatment outcome.[106] A follow-up study published in 1980 demonstrated the importance of pyrazinamide in the initial intensive treatment

phase, as well as the benefit of continuing the treatment for 8 months in total, with no patients having been given two months of SHRZ and 6 months of follow-on TH having relapsed at 19 months.[107] The importance of pyrazinamide was once again underlined in 1982 by a Hong Kong / MRC collaboration that showed equivalence between the 6-month regimens containing HRSZE, HRSZ, and HREZ, but not that containing HRSE.[108] In 1984 the British Thoracic Society then compared 2HRE/7HR with 2HRSZ/4HR or 2HREZ/4HR, demonstrating equivalence across the regimens and again providing assurance that 6 months of therapy was sufficient, recommending either 2HRSZ/4HR or 2HREZ/4HR as standard regimen.[109]

Motivated by the cheaper cost of ethambutol and the danger of blood borne virus transmission from the reuse of needles on administration of streptomycin in some low-income settings, the Pasteur Institute in Madagascar compared 2HRZE/4HR to 2HRZS/4HR. The ethambutol-containing regimen was just as effective without loss of compliance, as had been feared.[110,111] This new, fully oral four drug, six-month regimen was subsequently recommended by the WHO as the preferred option for treatment naive patients.[112] Attempts to either prolong this regimen to eight months to spare the use of rifampicin (2HREZ/6HE), or to reduce it to four months substituting isoniazid or ethambutol with moxifloxacin (M) or gatifloxacin (G) respectively have since failed.[113-115]

2.8.1 SURGICAL TREATMENTS

Surgical treatments for tuberculosis were pioneered in the early 19th century. These evolved from the drainage of cavities to so called ‘collapse therapies’

aimed at starving the pathogen of oxygen.[116] Carlo Forlanini was the first to widely publicise artificial pneumothorax therapy in 1882, the same year as Koch's discovery.[117] It is estimated that over 100,000 patients received this treatment over the next quarter of a century. Thoracoplasty (the removal of ribs) was also introduced in the late 19th century, with some survival benefit suggested from observational studies, as was induced phrenic nerve palsy and extra pleural pneumonolysis (the implantation of material in the extra pleural space to collapse the upper lobe of the lung). Surgical lobectomies or pneumonectomies were widely performed from the 1930s into the 1950s and 1960s, but became less frequent as the evidence of drug treatments of tuberculosis accumulated.[116]

With the emergence of XDR tuberculosis, surgery is again being considered an important anti-tuberculosis therapy today. Indications for surgical resection including persistence of sputum smear and culture positivity despite optimal drug therapy, localised disease, and adequate respiratory reserve.[118]

2.9 Epidemiology

2.9.1 Global

Over the course of the 20th century the incidence of tuberculosis across much of the northern hemisphere fell steadily but the burden of disease in many southern, low-income countries remained high.[75,119,120] Whilst this had led to tuberculosis being relatively neglected as a research and public health priority in high-income countries, in 1990 it was still estimated to be responsible for 2.6 million deaths a year in "developing" countries.[121]

The emergence of the HIV pandemic in the 1980s resulted in an increase in incidence and mortality from tuberculosis in sub-Saharan Africa and a greater awareness in Europe and North America.[122,123] At a similar time the steady downward trend in incidence in eastern Europe since World War Two was reversed as countries transitioned from Soviet / communist rule towards market economies, with the consequent economic decline impacting on healthcare systems and the availability of drugs.[124] These factors not only contributed to an increase in the global incidence of disease, but also to the prevalence of drug resistant / multi-drug resistant (MDR) disease.[124,125] Against this background the World Health Organization (WHO) declared tuberculosis a global public health emergency in 1993, predicting 30 million deaths from the disease over the following decade in the absence of urgent intervention.[126,127]

Following the WHO's emergency declaration a plan was outlined to achieve 70% case detection and 85% cure rate for smear positive cases by 2000. It involved encouraging governments to commit political will and adequate resources to control tuberculosis by investing in the acquisition of drugs, in passive case finding among symptomatic patients, in standardized, directly observed treatment, and in reliable reporting and monitoring systems.[128] This became known as the "Directly Observed Therapy, Short-course" (DOTS) strategy and by 2000 had been adopted in 148 countries.[129] The year 2000 also saw the introduction of the millennium development goals (MDG), of which number 6 is to reverse the rising incidence of tuberculosis by 2015.[130]

The 2014 World Health Organization report estimates 9 million people were diagnosed with tuberculosis in 2013, with 1.5 million deaths attributable to the disease. 1.1 million (13%) patients were co-infected with HIV. About 480,000 patients are estimated to have had MDR disease, with an estimated 210,000 attributable deaths (44%). The majority of cases were in low-income countries, with 56% in the WHO South-East Asian and Western Pacific regions (24% in India alone), and 29% in the African region. Although the WHO European region stretches to the eastern border of Russia, it accounts for just 4% of global cases.[131] Within the European region 84% of incident cases were in territories formerly under Soviet influence, or in Turkey.[132] Overall these findings demonstrate a steady but slow decline in global incidence, in line with the MDG target. However, it is also reported that only 45% of estimated MDR-TB cases are being detected, threatening long-term control.[131]

2.9.2 National

A similar pattern has been described in the UK. After a steady rise in incidence through the 1990s, rates have stabilised since 2005,[133] but remain among the highest in western Europe, measuring 12.3 cases per 100,000 population in 2013.[134] The majority of cases in the UK were identified in urban areas, with London alone accounting for 38% of notified cases in 2013, an incidence of 36 per 100,000 population (Figure 4).[134] Incidence varied within London as well, with the highest rates reported in the boroughs of Brent and Newham (100 and 117 per 100,000 respectively).[135] The incidence among London's more than 3,000 homeless is 300 per 100,000.[136]

In 2013 almost 60% of UK cases were aged between 15 and 44, and 72% of patients were born outside of the UK. The incidence among the non-UK born population was 70 per 100,000, compared to 4 per 100,000 in the remaining UK born population. 60% of patients born outside of the UK were born in South Asia. 52% of patients were reported to have had a pulmonary focus. 59% of all cases had culture confirmed disease, rising to 71% among those with pulmonary infection. 7% had an organism resistant to a first line drug, and 1.6% had MDR disease.[134]

2.9.3 Individual risk factors

At the level of the individual, about 5% of hosts who do not manifest symptoms of primary infection progress to disease within the first 18 months of infection and a further 5% at some subsequent point in their lifetime.[29] This expected 10% lifetime risk of reactivation from LTBI rises to 10% per year in individuals infected with HIV.[17]

Risk factors for acquisition of the disease are multiple. Lönnroth *et. al.* distinguish between proximate and up-stream risk factors.[137] The former are factors that either increase the risk of exposure to the disease or impede the immune response. These include exposure within domiciliary spaces or workspaces with limited air circulation and a high background prevalence of the disease, such as prisons. But they also include HIV, diabetes, malnutrition, smoking and in-door air pollution. Up-stream risk factors are the social determinants of some proximate risk factors, and include socio-economic status and long-term demographic changes such as urbanisation.

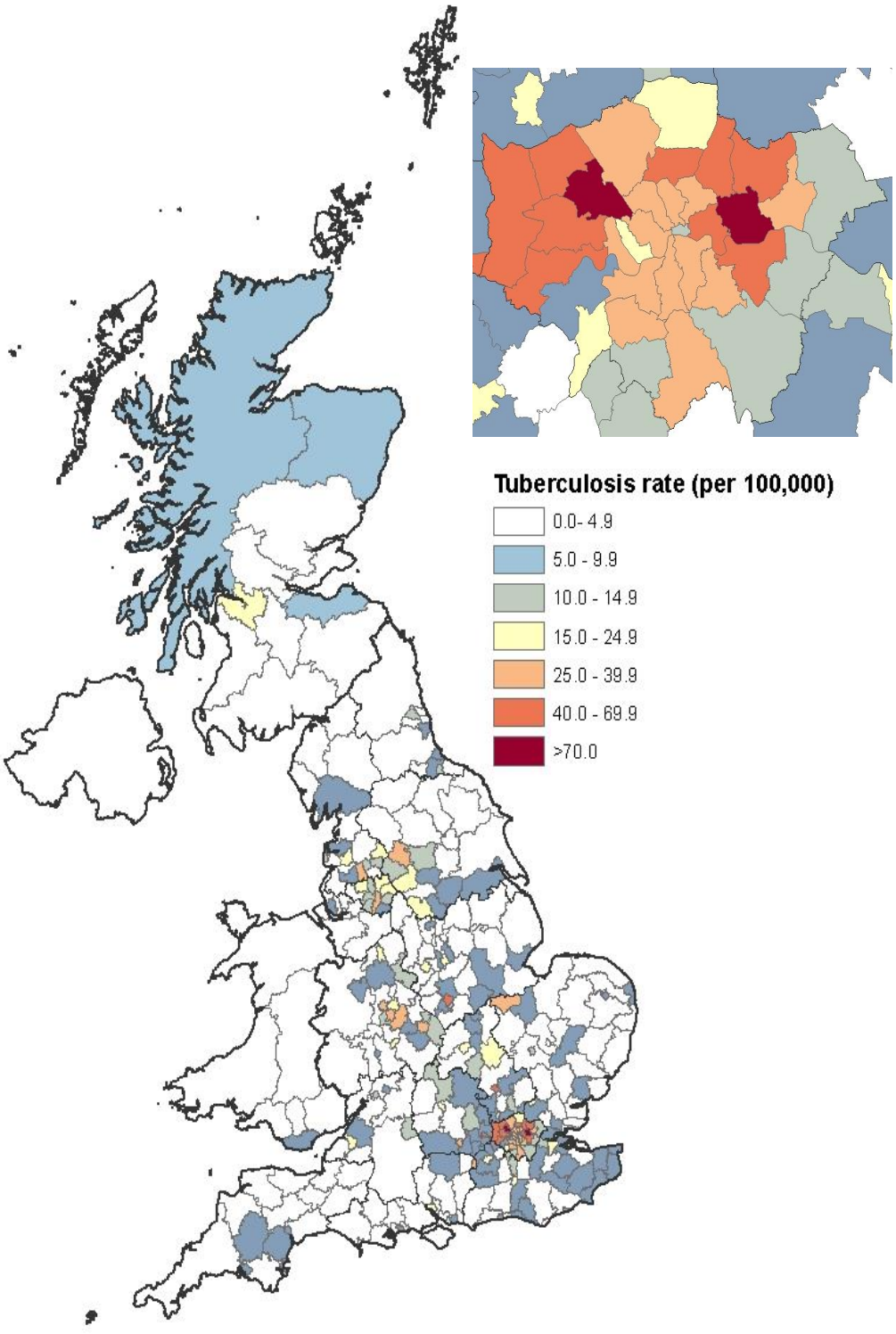


Figure 4: Three year mean tuberculosis incidence by local authority, UK (2011-2013) Taken from Public Health England's "Tuberculosis in the UK: 2014 report"[134]

2.10 TRANSMISSION

Although doubts about whether tuberculosis was a contagion were settled after 1882, it was in the late 1950s that a series of landmark papers by William Wells and Richard Riley investigated its mode of transmission. Wells and Riley created a closed-circuit ventilation system around patient isolation rooms on a tuberculosis ward in Baltimore and placed guinea pigs in the ventilation systems. A mean 156 guinea pigs were exposed at any one time, with tuberculin skin tests being performed on a monthly basis and animals with positive reactions replaced by unexposed ones. At two years 71 guinea pigs were infected, with the average time to infection being 10 days.[53,138]

Wells and Riley calculated the infectious 'quantum' (dose) contaminating air in their experiment. Each guinea pig was estimated to breathe $8ft^3$ of air per day, with 156 guinea pigs collectively breathing $12{,}480ft^3$ over 10 days. As at least one guinea pig was infected every 10 days, it was estimated that at least one quantum of tuberculosis infection capable of infecting a guinea pig must be contained within $12{,}000ft^3$ of room air. This number was related to the volumes of air breathed by nurses working on tuberculosis wards as estimated in the pre-chemotherapy era, where the average exposure time before tuberculin skin test conversion was 6-18 months. One infectious quantum was thus estimated to be contained in $24{,}000ft^3$ of air breathed by a nurse working on a tuberculosis ward for 1,200 hours a year (40 30 hour weeks) and breathing $20ft^3$ of air per hour.[138]

One notable observation from the Wells-Riley experiments was the differential infectiousness of patients. Of 22 guinea pigs infected within a given time frame,

19 were infected by just 2 of 62 patients admitted to the ward during this period.[53] When the experiment was repeated some 50 years later in Lima, Peru, a similar phenomenon was reported, with 108 of 125 guinea pigs infected by just one patient.[54] This is of interest as a meta-analysis of pooled data from 41 studies quantified the risk of infection among household contacts of patients with ‘open’ tuberculosis at 50% for developing LTBI and less than 5% for active disease.[139] Based on these numbers each patient with open tuberculosis must have the unlikely equivalent of over 20 household contacts to result in one further infectious case if each patient were equally infectious (Figure 5). Hyper-infectious individuals, as reported here as well as in different community outbreak settings,[140-143] thus offer a potential explanation.

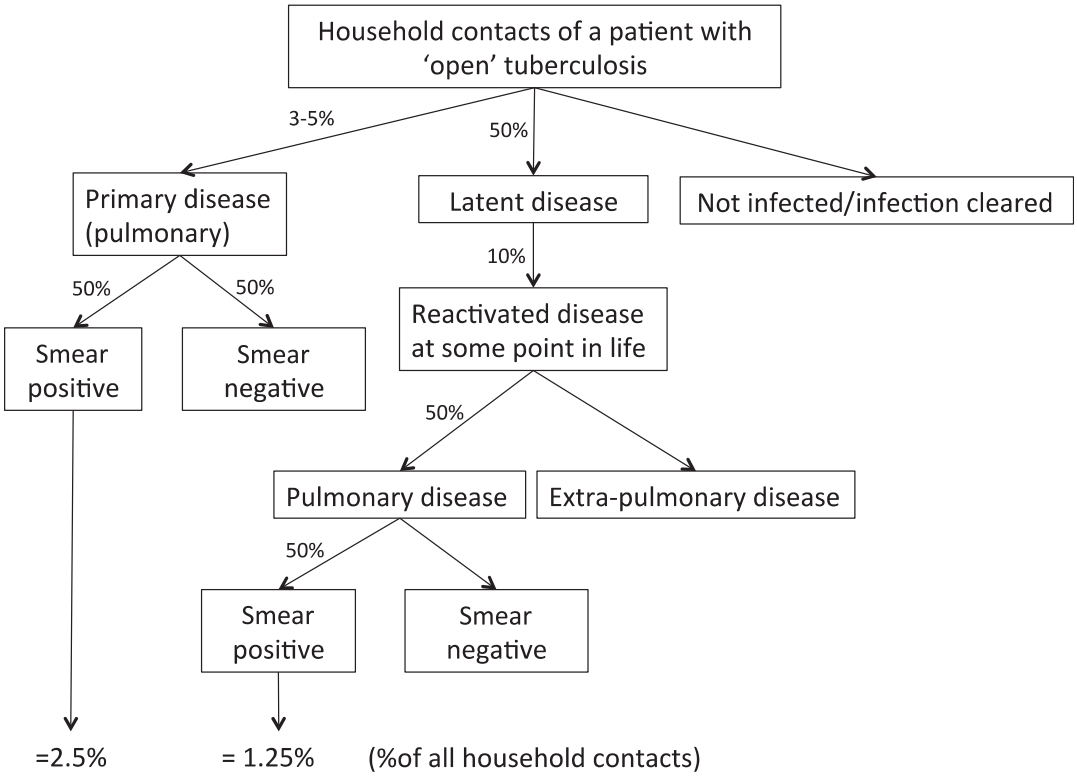


Figure 5: Proportion of household contacts likely to develop infectious tuberculosis

Taking such variability into account, Riley later defined a formal model of the infectiousness of airborne pathogens indoors over time. What has become known as the Wells-Riley equation states that the probability of transmission (P) is related to the number of infectious individuals (I) in a confined space, their breathing rate (p) and the rate with which they produce infectious quanta (q), the exposure time (t), and the ventilation rate within the defined space (Q).[144]

$$P = 1 - e^{-Iqpt/Q}$$

More recently this model was adapted to account for variation in the quantum concentration as measured by changing CO_2 levels within the indoor

space.[121] Incorporating the fraction of inhaled air that has been exhaled by another occupant of the room ($\bar{f}$), and taking account of the total number of (infectious and non-infectious) people in the room (n), the equation was adapted to

$$P = 1 - e^{-\bar{f}Iqt/n}$$

One application has been to estimate the relative annual risk of infection across different modes of public transport in Cape Town, South Africa. Here the risk for daily commuters was highest in shared taxis, which were also found to have the highest CO_2 concentrations at around 2,000 parts per million, almost twice that seen in trains and buses. Commuters using these taxis regularly had a 5% annual risk of acquiring tuberculosis.[145] The risks of regular exposure to public transport were also assessed in Lima, Peru and Houston, Texas, USA, with a similarly increased risk observed.[146,147]

The model has also been used to quantify the risk to children of household exposure to infectious adults in Cape Town townships. More than the early diagnosis of adults through active case finding, improved household ventilation was predicted to be key to reducing transmission to children. Given the overcrowded conditions, cold winters and poor housing quality in the townships, the model implicated poverty of infrastructure in transmission more than inadequacies in healthcare provision.[148] Taken to an extreme, such as in an overcrowded South African prison, the annual risk of infection approaches 90%, depending on the number of hours prisoners are locked up with infectious cases in their cells each day (Figure 6).[149] A critical threshold for infection

was however derived from a study of overcrowded township classrooms, where the same group estimated a threshold of 1.6% rebreathed CO_2 ($\bar{f}$), corresponding to 1,000 parts per million room air, for tuberculosis transmission.[150]

Despite the Wells-Riley inspired models of exposure, in practice the so-called '8 hour rule' is still commonly applied. It considers a minimum threshold of 8 hours exposure to be necessary before contact tracing is initiated and is based on a study of a patient with MDR-TB who infected up to 15 fellow passengers on a flight from Chicago to Honolulu in 1994, but not on a shorter connecting flight between Baltimore and Chicago.[151] It has since been incorporated into WHO guidelines for air travel as well as National Institute for Health and Clinical Excellence (NICE) guidelines for hospital infection control.[152,153]

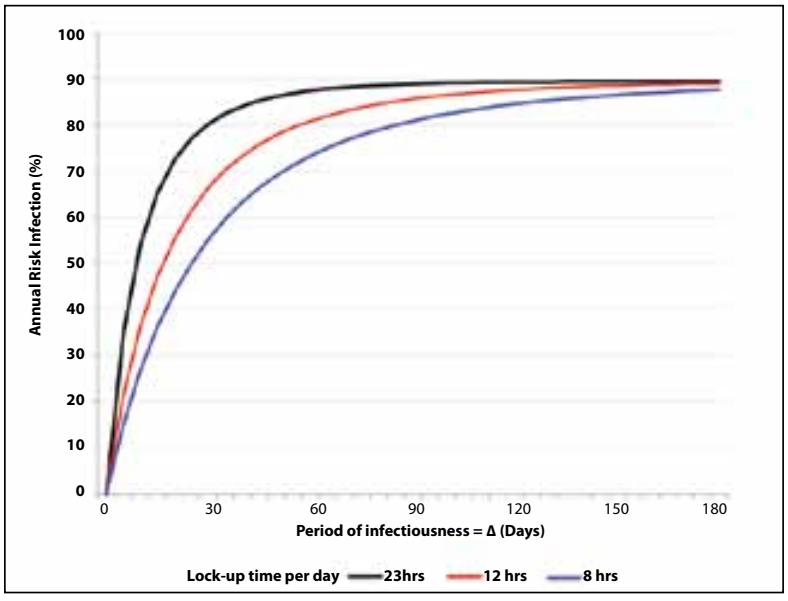


Figure 6: The annual risk of tuberculosis transmission via shared prison cell occupancy with an infectious tuberculosis case, Pollsmoor prison, Cape Town. Taken from Johnstone-Robertson *et. al.,* Tuberculosis in a South African Prison – a transmission modelling analysis, S Afr Med J 2011[149]

2.11 Public health control measures

Aside from improvements in infrastructure that make buildings and public transport less conducive to transmission, public health control measures targeted at the individual include isolation, contact tracing and case finding, prophylaxis and treatment of cases and their contacts, and vaccination.

Sanatoria served, if not by design then in effect, to isolate infectious patients from their immediate community.[84] Isolation of known cases remains an integral part of national tuberculosis guidelines, which in the UK recommend isolation of sputum smear positive hospital in-patients in single rooms until the completion of two weeks of therapy or hospital discharge.[153]

2.11.1 Case finding

Case finding can be either passive, where patients seek out medical help, or active, where patients with tuberculosis are sought within the community.[154] In the UK and other industrialised countries the use of mobile mass x-ray screening was introduced in the 1930s but phased out again by the 1970s after the realisation that most patients with active tuberculosis were seeking healthcare for their symptoms.[155] London's 'Find and Treat' mobile x-ray service remains an exception, screening the city's homeless, high-risk and vulnerable populations as these 'hard-to-reach' patients are less likely to seek healthcare promptly themselves.[136]

Targeted Contact Investigations to identify source and secondary cases within outbreaks are however standard public health practice. Guidance varies across Europe and the USA with some countries initiating contact investigations only

for potential ‘source cases’ (patients with smear-positive pulmonary tuberculosis) and others recommending contact investigations for ‘index cases’ in general, regardless of whether they are considered plausibly infectious.[156,157] The standard model for contact investigations has been to trace potentially exposed individuals across widening ‘concentric circles’ or ‘ripples in the pond’ until the rate of positive screening tests reflects the background community prevalence of disease.[158] Most contact investigations focus on household contacts first, and are extended into the wider community only if at-risk individuals are identified or if a wider outbreak is suspected. These environments include schools and workplaces, both of which are relatively structured settings in which to conduct contact investigations, but also pubs, bars or homeless shelters where attendees are more transient.[136] As investigations extended beyond the more structured settings they are increasingly dependent on contacts being named by index cases. This can pose difficulties for pubic health services if the homeless do not know the names or whereabouts of their associates, if substance misusers are reluctant to divulge the names of other misusers for fear of the law, or if recent migrants from high-incidence countries are mistrustful of the authorities, fear deportation for being undocumented, or are inhibited by social stigma.[136,159,160]

Social network analyses augment the traditional model of tracing named contacts by linking patients to one another as well as to places of social aggregation to better uncover transmission networks.[158] As such they have the potential to aid contact investigations even where patients are unwilling to name their contacts, and to link individuals who themselves may not even have been aware of social connections, and to identify super-spreaders.[141]

2.11.2 SCREENING AND PROPHYLAXIS

The desired outcome of contact investigations is the treatment of identified active cases and the prophylaxis of contacts screening positive for LTBI. To prevent reactivation of disease, NICE guidance is to offer treatment of latent disease to patients aged 35 or under, or to all healthcare workers or HIV positive patients, who screen positive on either TST (in the absence of BCG vaccination) or by IGRA. Treatment options include either 6 months of isoniazid preventive therapy (IPT) or 3 months of isoniazid and rifampicin combination therapy.[153] Guidelines in the United States prefer 9 months of IPT.[157]

The efficacy of IPT has been weighed against the risks of drug-induced hepatitis, the reason for NICE recommending an age cut-off of 35.[161] A study of 28,000 patients in the early 1980s compared 12, 24 and 52 weeks of IPT in patients with presumed LTBI and found that 24 weeks prevented most cases of tuberculosis relative to the number of cases of hepatitis caused.[162] This regimen was also found to be more cost-effective than treating for the full 52 weeks.[163] However the efficacy of IPT-based interventions does appear to be contingent on the incidence of the setting. A recent study was conducted among 78,000 South African miners to investigate the value of mass screening and treatment of LTBI. In this very high incidence setting, with a case notification rate of between 3,000 and 4,000 per 100,000 miners, IPT was only effective for the duration the therapy was given.[164]

In the UK one of the highest risk groups for LTBI are recent migrants. A health economic analysis of screening and IPT for new entrants to the UK was performed in 2011. Although screening all new entrants would identify close to

100% of LTBI in this population, restricting screening to migrants from countries with an incidence of 250 or more per 100,000 was found to be most cost effective.[165] Nevertheless, national practice in the UK is variable and the screening that is offered has been found to be inversely proportional to the local burden of disease.[166]

Although the treatment of active cases has been guided by the WHO's DOTS strategy as outlined above since 1993, the directly observed treatment component is not mandated by NICE in the UK for patients other than those considered at high-risk of non-adherence.[153] The management of patients is however monitored as part of routine quality assurance for tuberculosis control programmes in the US, and now in the UK, in a process of cohort review that involves collective, quarterly review of the management of all tuberculosis cases. The consequences of cohort review have been to improve treatment completion rates, follow-up and data collection, which in New York in particular have coincided with a dramatic drop in the incidence since inception in 1993.[167,168]

2.12 VACCINATION

BCG vaccination was introduced in the UK as a public health measure in 1953 and phased out from the routine vaccination schedule in 2005.[169] It provides 60-80% protection against disseminated disease and tuberculous meningitis to infants, although its protective capacity against pulmonary disease varies by geography, interestingly by latitude, in older patients.[41,170,171] Its use in the UK is therefore now restricted to infants of up to 12 months of age considered at high risk of tuberculosis. This includes those born in areas with an incidence

over 40 cases per 100,000, or those born to parents or grandparents themselves born in areas with an incidence in excess of 40 per 100,000.[169] Current evidence suggests that there is little benefit in repeatedly vaccinating patients over time.[171]

2.13 Molecular advances

The discovery of life's basic hereditary material, nucleic acid, dates back to the 1860s when Gregor Mendel published the results of his breeding experiments with peas, Ernst Häckel suggested the cell nucleus as the source of heritable traits, and Friedrich Miescher isolated precipitates from leukocytes nuclei that "could not be dissolved … and therefore do not belong to any known type of protein." He termed these precipitates "nuclein".[172] In 1953, 84 years after Miescher's observation, James Watson and Francis Crick, in collaboration with Rosalind Franklin and Maurice Wilkins, described the molecular structure of deoxyribonucleic acid (DNA).[173] In 1977, papers by Fred Sanger, Nicklen and Coulson, and by Allan Maxam and Walter Gilbert each described methods for sequencing short strands of DNA around 100 nucleotides in length.[174,175] There followed in 1979 the development of the first computer programmes to assemble multiple short sequences of DNA into larger contiguous sequences based on overlapping reads,[176] and in 1981 a method for fragmenting genomes into multiple pieces in an untargeted way for what has become known as 'shotgun sequencing'.[177] Kary Mullis and Fred Faloona published their method of polymerase-catalase chain reaction (PCR) for the amplification of nucleic acid sequences in 1987.[178] The first complete *M. tuberculosis* genome was sequenced using Sanger technology and published in 1998.[179]

What has become known as ‘next-generation’ sequencing, or whole-genome sequencing, technology can be traced back to the year 2000 and a report on massively parallel signature sequencing.[180] This early technology was based on the random fragmentation of DNA strands and the cloning of these on the surfaces of micro beads to achieve an amplified fluorescence-based signal indicating the nucleic acid sequence. By 2005 a number of parallel technologies were evolving, of which the cyclic-array sequencing-by-synthesis approach has since gone on to dominate. Building on the early whole-genome sequencing (WGS) techniques, the 454 Life Sciences (later Roche), and also Solexa (now Illumina) technologies were based on the clonal amplification of short fragments of DNA, using polymerases to add nucleotides to single strands of DNA.[181] Illumina platforms have come to dominate the market, producing about 90% of all sequence data in the world in 2012.[182] Although third-generation, single stranded nanopore sequencing has been heralded as having the potential to provide future advances, it remains in its experimental phase. It avoids any need for amplification and instead passes long strands of DNA through fluid filled, charged nanopores registering the changes in electrical current nucleotide-by-nucleotide.[183]

Illumina sequencing technology starts with the preparation of a sequencing library during which genomic DNA is fragmented, sheared ends are repaired, and adaptors are ligated to the denatured molecules. Unique indices are introduced to identify the DNA and subsequent sequencing reads throughout the process. Cluster generation then involves a size-selected subset of molecules being washed over a flow cell lane where each is immobilised by adaptors hybridising to the first of two types of primers fixed to the base of the

flow cell lanes. A polymerase enzyme extends the sequence from the fixed primers, thereby generating a complement to each hybridising molecule. The two complementing strands are subsequently denatured, and the hybridising strand removed. Each fixed, single stranded molecule now arches over so that its free adaptor can hybridise with the second type of fixed primer in the flow cell lane. With the formation of each arch, or 'bridge', a polymerase forms a complementary arch in an isothermic reaction to produce forward and reverse strands of complementary DNA, each fixed to the flow cell. These are denatured and new bridges are formed as part of a cyclical process leading to massive clonal amplification.

The generation of clusters through amplification is important to ensure a sufficiently intense signal can be produced from each cluster when sequencing-by-synthesis is commenced. However, as each cluster is made up of both forward and reverse strands, an inherently mixed signal would emanate upon sequencing. To avoid this the reverse strands are first cleaved and removed, exposing just the forward strands. A sequencing primer is hybridised to the 5'-end of the strands and fluorescently labelled nucleotides are washed over the flow cell in equal proportion to compete for their addition to the complementary sequence before the unsuccessful bases are washed away. The cycle is repeated base-by-base with the fluorescent tags on each emitting a colour signal that is captured by a camera. The number of cycles corresponds to the length of the reads and the sequence of fluorescent signals to the nucleotide sequence on the read. To sequence the corresponding reverse strands, bridges are again formed with the free adaptors now binding the second of the two types of fixed primers on the flow cell. A polymerase produces a

complementary strand, the reverse strand, with the forward strands subsequently being cleaved and removed this time. The reverse strands are then sequenced-by-synthesis in the same way as the forward strands were (Figure 7).[184,185]

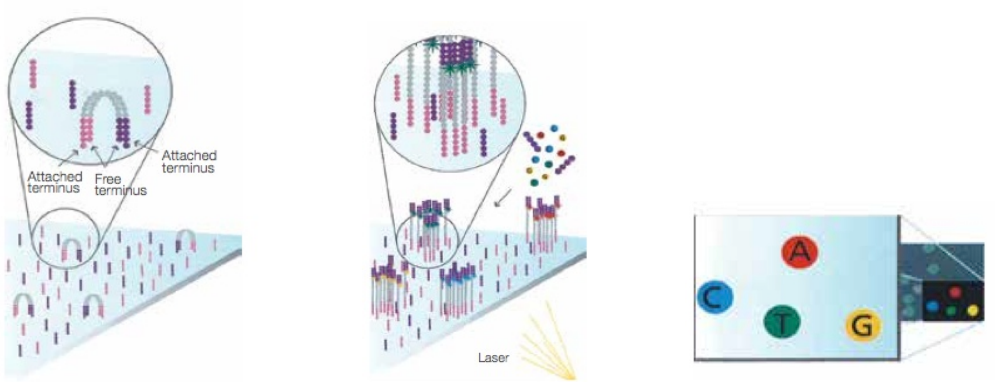


Figure 7: From left to right, the images demonstrate the process of bridge-amplification, clustering, and sequencing-by-synthesis as fluorescently labeled nucleotides are added to he flow cell and base-specific signals are emitted. Figures taken from reference [185] (Illumina technology spotlight product information).

The assembly of millions of reads produced by massively parallel sequencing on Illumina, and other platforms, poses a bioinformatic challenge. To generate a 'genome sequence' the reads can be either mapped to a pre-existing reference genome or they can be assembled, *de novo*, into multiple 'contigs' on the basis of overlapping sequence.[186,187] The closer the sequence of interest to the reference genome of choice, the greater the opportunity for identifying variant nucleotide sites, or single nucleotide polymorphisms (SNPs) by direct comparison. Where no reference is available or where it is believed to diverge significantly from the sample of interest, a *de novo* approach is preferable.[188] However, neither mapping nor *de novo* assembly can overcome the problem of repetitive sequences that exceed the read length generated.[189] Whilst raw

data are generic, approaches to the trade-off between maximising the proportion of the genome covered and optimising the precision with which nucleotide variants are called therefore differ.[190]

2.14 MOLECULAR TECHNIQUES APPLIED TO *M. TUBERCULOSIS*

Advances in molecular techniques have contributed both to the diagnosis of tuberculosis and to our understanding of its transmission and epidemiology. PCR has enabled the targeted amplification of *M. tuberculosis* specific sequences for species identification as well as the identification of SNPs underlying drug resistance,[191,192] and has also been key to the Mycobacterial Interspersed Repetitive Unit-Variable Number Tandem Repeat (MIRU-VNTR) genotyping approach.[193]

2.14.1 SPECIES IDENTIFICATION

Two commercial kits using PCR to amplify *M. tuberculosis complex* specific 16s ribosomal RNA, the Gen-Probe MTD (*Mycobacterium tuberculosis* Direct) test and the Roche Amplicor and Cobas Amplicor, were introduced and licensed by the US Food and Drug Administration in the mid 1990s.[191] To further distinguish between species within the *M. tuberculosis complex* on the basis of 16s and 23s ribosomal RNA, the RD1 deletion and SNPs within the gyrB gene, a commercial line probe assay (LPA) based on PCR and reverse hybridisation was introduced by HAIN Lifesciences in 2003.[194] Each of these tests reduces the time to diagnosis but are still dependent on an initial culture step, at least in a routine setting. The Cepheid MTB/RIF GeneXpert was introduced as a point-of-care assay for the rapid identification of *M. tuberculosis complex* (as well as markers of rifampicin resistance) from primary clinical samples without the need

for culture or complex laboratory procedures, and was first assessed in a multi-centre study between 2008-2009.[195] It has since been adopted widely in both high and low-income settings.[196]

2.14.2 DRUG RESISTANCE TARGETS

With the new PCR technique to hand, our understanding of PCR targets for the detection of drug resistance evolved rapidly during the 1990s. As isoniazid is a pro-drug, the absence of a functioning katG gene prevents its transformation into its active form.[86] The deletion of the katG gene, encoding catalase and peroxidase, was identified as a source of isoniazid resistance in 1992, with particular codon mutations at position 315 identified as conferring resistance five years later.[197,198] Subsequent mutations in inhA, necessary for fatty acid synthesis, were described as conferring isoniazid resistance in 1994.[199] In 1993 rifampicin resistance in *M. tuberculosis* was related to a number of mutations within rpoB, which encodes the β-subunit of RNA polymerase,[200] and the region determining rifampicin resistance in rpoB has since been extended to 81 consecutive codons.[201] Mutations at codon 306 in the embB gene were confirmed as conferring resistance to ethambutol in 1997, whereas mutations conferring resistance to second line drugs such as the aminoglycosides and fluoroquinolones were described in rrs, rpsL and gyrA between 1994 and 1998.[202-205]

Most of these mutations have since been incorporated onto commercial LPAs. The first of these, in 1997, was the Inno-LiPA Rif.TB (Innogenetics) assay which probed a 69-nucleotide sub-section of the rifampicin resistance determining region of rpoB.[206] The Genotype MTBDR (HAIN Lifesciences) LPA then

incorporated mutations at katG^{315} to probe for isoniazid resistance (2005), and the MTBDR*plus* additionally probed for mutations in the upstream region of inhA (2009).[207,208] The MTBDR*sl* was the first LPA to include second line drugs, probing not only for ethambutol resistance but also for resistance to fluoroquinolones and aminoglycosides other than streptomycin.[209] The most recent LPA, the AID assay (Autoimmun Diagnostika, GmbH), also includes rpsL mutations relevant to streptomycin.[210] Although not based on reverse hybridization, the Cepheid GeneXpert also uses PCR to identify mutations in the rifampicin resistance determining region.[195]

2.14.3 GENOTYPING

Over the past two decades, genotyping has been used to augment epidemiological investigations by matching isolates from patients with culture-confirmed tuberculosis. Early examples of genotyping include phage typing and comparisons between susceptibility profiles, but neither had sufficient discriminatory power.[211] The first widely used technique to identify stable, but repetitive, sequences within the *M. tuberculosis* genome is Insertion Sequence-based Restriction Fragment Length Polymorphisms (RFLP) typing, focusing in particular on insertion sequence IS6110. Although this was a useful adjunct to contact investigations, RFLP typing is technically laborious, cannot be used to distinguish isolates with low IS6110 copy numbers and produces results that are difficult to compare across laboratories.[193] It has since been superseded by 15, and now 24-locus MIRU-VNTR typing as the current standard. MIRU-VNTR typing, which is based on nucleic acid amplification of short repeats, or ‘mini-satellites’ at designated loci within the genome, is less time consuming,

can be performed regardless of the number of repeats at each locus, and produces a digital profile that can be readily compared across laboratories.[212]

With 24 different MIRU-VNTR loci being explored, the number of potential combinations is large, as demonstrated in a Belgian study that found 610 unique profiles among 802 consecutively sampled patients over a three-year period.[213] Although usual practice would be only to investigate potential epidemiological links between patients whose isolates share identical MIRU-VNTR profiles (i.e. contain the same number of repeats across all loci), where an outbreak is suspected contact investigations may be extended to also include patients with isolates that differ at a single locus. Genotyping can thus be understood as a means of corroborating or refuting proposed transmission events between patients with epidemiological connections, and of linking further cases that might usefully be included in any contact investigation. As a link to a transmission network can be ruled out by genotyping with greater certainty than it can be ruled in,[214,215] public health teams must judge how intensively to search for a possible epidemiological link between MIRU-VNTR-matched patients in the knowledge that none might exist. While scarce resources need to be used efficiently, ending an investigation prematurely risks further community transmission.

2.15 THESIS OUTLINE

In this thesis I explore how WGS technology can contribute to our understanding of tuberculosis transmission, how it can guide public health control measures and how it might influence policy decisions, as well as how the sequence data can be used to identify genetic variation underlying drug

resistance. In Chapter 3 I examine the within-host and between-host genetic diversity seen in *M. tuberculosis* infection and estimate the levels of genetic relatedness suggestive of recent transmission, comparing predictions based on WGS data to those made by the current standard, 24-locus MIRU-VNTR typing. I also explore the potential of WGS to elucidate the structure of an outbreak, including the direction of transmission within that outbreak. In chapter 4 I apply the metrics for transmission established in chapter 3 to a population study of Oxfordshire over a six-year period to estimate the amount of disease that is locally transmitted, and the proportion that is most likely acquired overseas. Understanding the balance between these scenarios has been a key question for Public Health England, and could help to inform decisions to allocate resources to new-entrant screening and contact tracing respectively. During the course of this thesis a decision was been made by Public Health England to implement WGS into routine practice in England, replacing MIRU-VNTR typing for the purpose of detecting transmission. The routine sequencing of every isolate in the country for disease surveillance and outbreak detection presents an opportunity to exploit the sequence data for additional analyses, at no extra cost. In chapter 5 I therefore explore the potential to predict drug susceptibility, and to identify genetic variation in the form of SNPs, insertions and deletions underlying phenotypic resistance from WGS data. I describe a two-step algorithmic approach, first deriving a set of plausible resistance-determining alleles from a training data set and then testing these against an independent data set as a means of validation.

3 CALIBRATING WGS DATA AS AN EPIDEMIOLOGICAL TOOL

3.1 INTRODUCTION

Controlling the spread of *Mycobacterium tuberculosis* can be challenging even in well-resourced countries. The incidence of tuberculosis in the UK was 11.4 per 100,000 population in 2000, peaked at 14.4 per 100,000 in 2011, and was last reported as 12.3 per 100,000 in 2013,[55,134] with non-UK born individuals accounting for the increase in cases over this time.[216] In the UK, detection of tuberculosis outbreaks has been guided by 24-locus MIRU-VNTR genotyping since 2010.[55] Although it is thought that MIRU-VNTR typing can reliably exclude transmission between cases whose genotypes differ, epidemiological data is still required to confirm outbreaks among cases whose genotypes match.[214,215] Collecting this epidemiological data is difficult where patients are unwilling or unable to volunteer information, as is commonly the case in some of the social groups most at risk of tuberculosis.[159,217] Even where genotyping does lead to outbreak detection, it offers no additional insights into the underlying pattern of transmission.

As an alternative to MIRU-VNTR typing, WGS promises to detect microevolution within MTB lineages as they are transmitted between hosts.[189,211,218,219] Because backward mutations are rare,[220] the pattern of accumulated mutations can in theory suggest direction of transmission within an outbreak. This principle was outlined in 2010 by Schürch *et. al.* who sequenced three patient isolates from a large IS6110-RFLP-based outbreak in the Netherlands. The pattern of cumulative acquisition of SNPs suggested a direction of transmission that corroborated the epidemiological

data.[218] Prior to the work presented in this chapter, WGS had otherwise only been applied to a single outbreak of tuberculosis in Vancouver, Canada, where the authors combined their analysis of the sequence data with a social network analysis. The study's authors demonstrated the greater resolution of WGS over MIRU-VNTR data and argued that MIRU-VNTR had failed to show that two independent outbreaks were actually co-evolving within the same community.[141] However, no measure of the degree of relatedness that could suggest transmission had occurred was offered. The work presented in this chapter was the first attempt at exploring this question.

The *M. tuberculosis* isolates chosen for this chapter were selected from across the Midlands region of the UK. This includes both Birmingham and Leicester, where all the global clades are represented among an ethnically diverse population,[221,222] and where the annual tuberculosis incidence has been as high as 50 to 70 cases per 100,000 population.[55] The main aim was to estimate the genetic diversity between epidemiologically related strains and investigate whether similar levels of diversity could be applied as a threshold to guide community outbreak investigations.

3.2 METHODS

3.2.1 SAMPLE SELECTION

I pre-specified four groups of isolates within which to investigate within host and between host diversity that could be considered compatible with transmission: (i) cross-sectional diversity within individuals (pulmonary vs. non-pulmonary isolates obtained within 6 months of each other), (ii) longitudinal diversity within individuals (patients with multiple pulmonary isolates more than 6 months

apart), (iii) diversity between individuals in known household clusters, and (iv) MIRU-VNTR-based community clusters. Figure 8 shows an outline of the selection process, detailing the number of missing isolates and those isolates common to multiple groups. As 24-locus MIRU-VNTR was only introduced into routine service in the UK in 2010,[55] clusters were matched at either 15 or 24 MIRU-VTNR loci according to the typing protocol at the time of referral. As this study had funding to sequence between 400 and 500 samples, I attempted to sequence specific numbers of isolates within each group.

I first searched the archive database of over 13,000 frozen cultures (1994-2011) held at the Public Health England (PHE) West Midlands Public Health Laboratory for the Midlands, South Yorkshire and Humberside for all isolates meeting the above criteria. From these groups, specific isolates were selected for sequencing as below. Where initially selected samples could not be found or did not grow, replacements were selected in the same way as the initial samples. In some cases the freezer location was not listed in the database so the isolate could not be retrieved; in other cases there was no vial present in the specified freezer location; other vials contained non-viable bacteria, which was unsurprising given the age of some of the samples.

(i) I chose cross-sectional paired isolates at random, until I had obtained 50 pairs of DNA preparations ready for sequencing (in total I identified 74 pairs for locating in freezers and growing). As no data were available on likely within-host diversity at the time the choices were made, the target sample size was 50 to ensure that if no SNPs were observed between any pair, then the upper

97.5% confidence limit around the observation of 0% with one or more SNPs was 7%.

(ii) I selected longitudinal samples within individuals to maximise the time period between first and last isolate for each patient, as I considered longitudinal diversity over time to be most relevant to onward transmission. I also included any intervening samples from these patients for sequencing. No data were available on likely diversity over time when I chose the samples, so I made the arbitrary decision to sequence 100 isolates within this group, a similar number to those previously used to estimate molecular clock rates in other species.[223]

(iii) I aimed to sequence all isolates from all household outbreaks known to the surveillance laboratory (93 isolates in total). I defined these outbreaks on the basis of a shared address and a matching 15 or 24-locus MIRU-VNTR profile.

(iv) I selected community clusters on the basis of the samples sharing MIRU-VNTR profiles in order to specifically investigate what additional benefits WGS might provide over and above MIRU-VNTR. Ten reasonably sized (6-47 patients) 15 or 24-locus defined MIRU-VNTR-based community clusters were identified by the public health teams as containing some cases where direct case-to-case transmission was supported, and others where it was uncertain (207 isolates in total). In addition, 46 isolates from 18 patients that were sequenced as part of groups (i), (ii), and (iii), also belonged to a very large MIRU-VNTR-defined cluster containing over 280 patients.[224] I analysed these 18 patients as an 11th community cluster even though selection for sequencing was not based on this cluster membership. Three clusters were characterised

by school outbreaks, six by community-based substance abuse, one by a regionally dispersed ethnic group, and one by *Mycobacterium bovis* infection. To relate school clusters to their community, all available local 24-locus MIRU-VNTR matching cases were also sequenced. To investigate potential clustering across 24-locus MIRU-VNTR types, four clusters were extended to include isolates mismatching at 2 or fewer loci (see Table 1 p.71, for details of each cluster).

Of note, both failure to be located and failure to grow were strongly related to duration of storage. As expected, the longer the duration of storage the more likely an isolate could not be located, and, if it was located, that it failed to grow. For example, among the cross-sectional isolates the mean time since original isolation among the missing isolates was 8 years and among the isolates that failed to re-grow it was 9 years. This contrasted to 5 years for the successfully cultured isolates. Among the longitudinal isolates the mean time was 10 years for missing isolates and 8 years for those that failed to re-grow, compared with 6 years for successfully cultured isolates. For groups (i), (ii) and (iii) missing data from one isolate sometimes meant that other sequences were also excluded (eg if one of two longitudinal isolates from a patient in (ii) could not be located or failed to grow, then the patient could not contribute any data to (ii)).

All missing data was assumed to be completely at random in the analysis, with sequenced cases assumed to represent the underlying population. This assumption was felt to be reasonable by the PHE reference laboratory, where they have not seen any marked variation in the type of cases over the last decade, so length of storage is therefore not a plausible confounder. The global

lineages of sequenced strains reflect those prevalent in the Midlands, as would be expected from essentially random sampling.

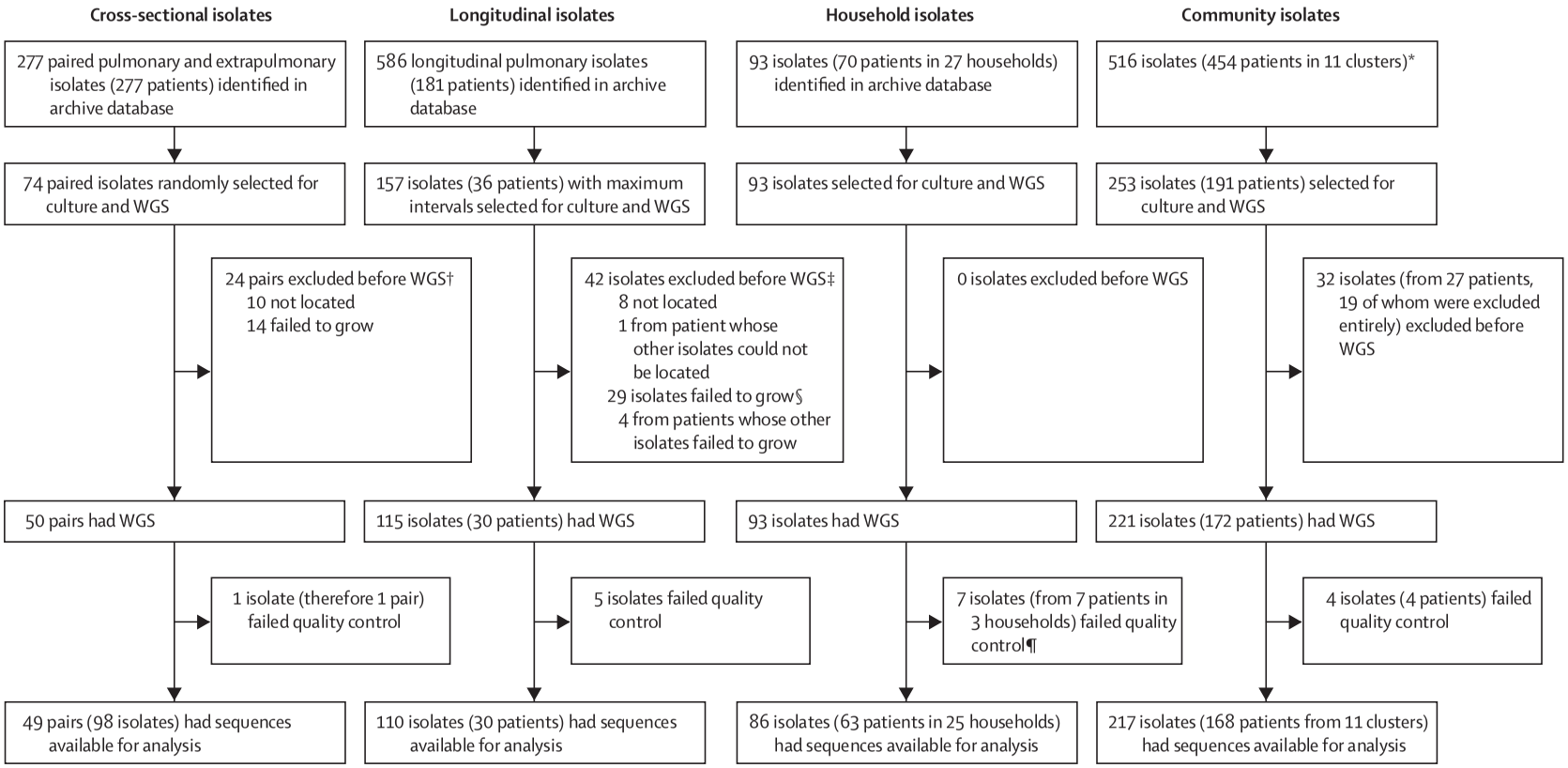


Figure 8: Sample selection. The cross-sectional and community analyses datasets overlapped by 14 isolates (eight patients); the longitudinal and community analyses by 23 (five); the longitudinal, household, and community analyses by 26 (seven); and the household and community analyses by 32 (29). † Mean time since original isolation was 8 years for missing isolates and 9 years for isolates that failed to regrow, compared with 5 years for successfully cultured isolates. ‡ Mean time since original isolation was 10 years for missing isolates and 8 years for isolates that failed to regrow, compared with 6 years for successfully cultured isolates. § One patient excluded because all his or her isolates failed to grow. ¶ Only two households were excluded. * Cluster 9 contained >280 patients and was not sequenced in its entirety.

3.3 CULTURE AND DNA EXTRACTION

I cultured samples from frozen stocks, first in MGIT until ample growth was visible, and then on LJ slopes. All cultures were incubated at 37°C until I perceived mature growth.

I used the cetyltrimethylammonium bromide (CTAB) method to extract DNA for the samples. This method is based on that described by van Soolingen *et. al*. in 1991.[225] The procedure was carried out in the category 3 safety laboratory at the West Midlands Public Health Laboratory in Birmingham.

1. Harvest two loops of culture from the LJ and suspend in 400ul of 1xTE in a 1.5ml screw-cap micro-centrifuge tube.
2. Place the tube in a covered water bath at 80°C for 20 minutes.
3. Add 50ul of lysozyme (10mg/ml) to the tube, vortex for 10 seconds, and incubate overnight at 37°C.
4. The following day, add 75ul of 10% SDS / proteinase K solution, gently invert, and incubate for 30 minutes at 65°C.
5. Add 100ul of 5 Mol NaCl
6. Add 100ul of CTAB/NaCl solution warmed to 65°C and vortex until the tube content appears opaque.
7. Incubate for 10 minutes at 65°C
8. Add 750ul of chloroform/isoamyl alcohol and invert gently for 10 seconds.
9. Centrifuge for 8 minutes at 12,000g

10. Transfer the supernatant to a fresh micro-centrifuge tube using a pipette small enough to avoid picking up any of the pelleted debris (e.g. a 200ul tip).
11. Add 450ul of 0.6 volume isopropanol.
12. Gently manually rock each tube to encourage a nucleic acid precipitate to form.
13. Incubate at -20°C for 30 minutes
14. Centrifuge for 15 minutes at 12,000g to pellet the DNA
15. Remove supernatant, leaving 20ul of liquid covering the pellet
16. Add 1ml of 70% ethanol chilled in the -20°C freezer and rock the tube carefully to wash the DNA precipitate.
17. Centrifuge for 5 minutes at 12,000g.
18. Remove supernatant, leaving 20ul of liquid covering the pellet
19. Centrifuge for 1 minute at 11,000g
20. Remove the remaining supernatant with a 20ul pipette.
21. Dry the pellet at room temperature for 15 minutes or until the ethanol has evaporated completely.
22. Re-suspend the pellet in 50-100ul of 1xTE, depending on pellet size.

I quantified DNA concentrations with the Tecan Infinite 2500 Pro using PicoGreen, a fluorescent dye that intercalates with DNA.[226] Concentrations of 18ng/ul in 100ul were considered acceptable, with 36ul being used for sequencing.

3.4 WHOLE GENOME SEQUENCING AND BIOINFORMATICS PIPELINE

Isolates were sequenced at the Wellcome Trust Sanger Institute in Hinxton, Cambridge using Illumina HiSeq platforms (Illumina, San Diego, California, USA). Libraries were prepared by staff at the sequencing-centre.

Raw sequencing reads in the form of FASTQ files were returned to Oxford where paired-end reads were mapped to the H37Rv *M. tuberculosis* reference genome (GenBank NC_000962.2) using Stampy v1.0.22,[227] but without Burrows-Wheeler Aligner pre-mapping, and using an expected substitution rate of 0.01. Repetitive sequences were identified by self-self-BLAST, accounting for 7% of the reference genome. As repetitive regions cannot be reliably mapped given the current length of sequencing reads, these regions were masked to further analysis.

A consensus of greater than or equal to 75% of reads was required to support high confidence nucleotide variant calls made using SAMtools mpileup[228] which had to be homozygous under a diploid model. Only variants supported by 5 or more reads, including one in each direction, were accepted.[229,230][176,177] Sites where minority variants represented more than 10% of read depth were defined as mixed and no variant was called. Nucleotide calls at sites in the top 97·5 percentile for depth were excluded. Consistency of the sequencing, assembly and data filtering process was evaluated by re-sequencing 59 patients' isolates and the H37Rv reference genome on different flow cells as technical replicates.

Pipeline outputs include BAM files, which are reads mapped to the reference genome, variant calling format (VCF) files, detailing the quality scores and filtering results at each nucleotide position in the genome, and FASTA files, representing the final consensus sequence for each sample.

3.5 SEQUENCE DATA ANALYSIS

I used bespoke python scripts, created by members of the Modernising Medical Microbiology (MMM) consortium, to identify variant calls and observed pairwise SNP distances across selected FASTA files.

The micro-evolutionary mutation rate, known as the molecular clock, was estimated by maximum likelihood from pairs of isolates within individuals and families under a coalescent model[231] assuming a Poisson distribution for the accumulation of mutations. Confidence intervals were obtained by parametric bootstrap. I constructed phylogenetic trees from concatenated variable sites across clustered genomes using Maximum-likelihood (PhyML 3.0)[232] in Seaview,[233] and rooted using other isolates in the collection. I manually inspected uncalled sites where variants had been identified in other samples for evidence of minority variants suggestive of mixed infection. I did not take insertion and deletions, known as “indels”, into account in this analysis.

After the analysis described here was complete (and published) the pipeline parameters were updated, removing the depth-filter that excluded nucleotide calls at sites in the top 97.5 percentile of depth. When I re-ran the analysis on the results of this updated pipeline, the outcome did not change. However, for the purposes of the data presented here, nucleotides at variant sites initially

excluded due to excessive high-quality read depth were manually re-inserted as an additional filtering step.

I assigned samples to different *M. tuberculosis* lineages on the basis of MIRU-VNTR profiles, following a schema approved by the Health Protection Agency (now Public Health England).[222]

3.5.1 Epidemiological investigation

Clinical, demographic and microbiological data was available for all isolates. I gathered epidemiological data through interviews with public health teams and supplemented it by case record review for every new index case. UK guidelines for contact tracing recommend screening household contacts, 'at risk' individuals, and any other named contacts in the community where the index case is considered infectious.[153] I graded epidemiological relationships within clusters as 'linked' (patients were known to have shared time and space with each other or a third party), 'possibly linked' (known to have shared space, but not at the same time), or as having 'no known link' (no known shared space). I summarised epidemiological relationships, SNP differences, and differences in isolation date in network diagrams representing the links between patients in each cluster in a way that maximises epidemiological and genetic proximity. I represented patients as 'nodes' and links between patients as 'edges'. To avoid double counting, I chose the most parsimonious number of connecting 'edges' (number of nodes (n) minus one) needed to represent the most plausible transmission network within each cluster. I constructed networks starting with the first patient to be diagnosed. One isolate per patient was used except where indicated (Figure 9, cluster 9). I first sought to draw an edge to another patient

with a ‘known’ epidemiological link. If there was more than one patient to choose from, I chose the patient with an isolate closest in SNPs, and where this failed to identify a unique edge, the patient closest in time (as judged by date of isolation of sample). If no ‘known’ epidemiological link existed, I sought ‘possible’ linkage before ‘no known’ linkage, in each instance prioritising by SNPs and time as described. The second edge and all subsequent edges were determined by the same rules until the network was complete.

Although school outbreaks have often been caused by particularly infectious individuals who are the source of many secondary cases,[140] an important question is whether so-called super-spreaders are also the source of many community outbreaks.[141] Because *M. tuberculosis* evolves by descent,[234,235] there is an *a priori* expectation that a star-like phylogenetic topology with several secondary cases branching directly from a common node would be apparent when an individual with several contacts remains infectious for some time. I assessed whether this approach might be able to contribute to the identification of super-spreaders in the community.

3.6 Results

3.6.1 Sequencing and Strain Summary

390 separate isolates from 254 patients were successfully sequenced. Mean reference genome coverage was 88.5% (range 86.5-89.5%). There were no discrepancies across the 59 technical replicates, suggesting highly specific SNP calling. The European-American and Central Asian lineages were most heavily represented in the data set, but isolates from the Beijing and East-

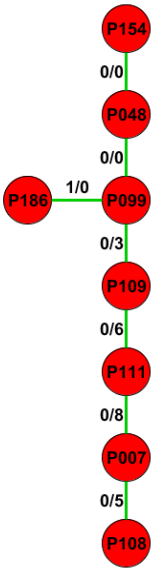
African Indian lineages were also identified, along with *M. bovis* and *M. africanum* (Figure 10).[222]

The mean number of isolates from each of the 30 longitudinally sampled patients was 3.7 (range 2-9) and the mean length of time between the first and the last isolate was 30 months (range 6-102). For the household isolates, 25 households each provided a mean of 3.4 isolates (range 2-19) to the analyses from a mean of 2.5 patients (range 2-5). The household outbreaks lasted a mean of 20 months (range 0-125). Of the community clusters, three were characterised by school outbreaks, six by community-based substance misuse, one by a regionally dispersed ethnic group, and one by *M. bovis* infection (Table 1). Mean duration of community outbreaks was 81 months (range 9-144).

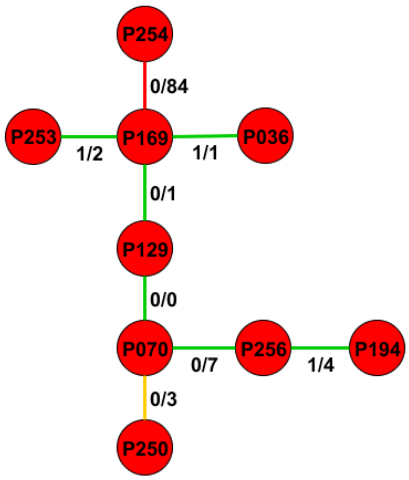
Cluster number with dominant MIRU-VNTR profile	Number of patients linked to cluster by			Number of WGS isolate	Time span across WGS isolates in months	Number of "known" epidemiological links between patients*			Number of "possible" epidemiological links*			Number of "no known" epidemiological links*		
	MIRU-VNTR 15	MIRU-VNTR 24	discordant MIRU-VNTR (22-23 loci)			≤5 SNPs	6-12 SNPs	>12 SNPs	≤5 SNPs	6-12 SNPs	>12 SNPs	≤5 SNPs	6-12 SNPs	>12 SNPs
Cluster 1 (E / School) 32433.2332514327.223423352	0	8	0	9	13	7	0	0	0	0	0	0	0	0
Cluster 2 (E / School) 32333.2512515324.234433363	6	3	0	9	92	6	0	0	1	0	0	1	0	0
Cluster 3 (C / School) 42234.2742511324.432423254	0	6	0	6	59	4	0	0	0	0	0	0	0	1
Cluster 4 (E / Substance misuse) 32433.232515322.224423542	15	31	1	54	88	8	0	0	0	0	0	38	0	0
Cluster 5 (E / Substance misuse) 32433.2432515323.241433273	7	2	0	9	138	1	0	0	6	0	0	1	0	0
Cluster 6 (E / Substance misuse) 32433.2432515324.443443153	3	19	4	29	144	20	0	0	1	1	0	1	2	0
Cluster 7 (C / Substance misuse) -2234.2742511334.432422254	0	8	2	17	104	7	0	0	1	1	0	0	0	0
Cluster 8 (B / Substance misuse) 42435.2332517333.346443584	0	6	0	6	9	1	0	0	1	0	0	3	0	0
Cluster 9 (U/E / Substance misuse) 32333.2432515314.434443183	0	16	2	46	83	7	0	0	0	0	0	7	3	0
Cluster 10 (C / Ethnic group) 42234.2742511334.432423254	2	21	0	26	102	6	0	0	0	0	0	10	0	6
Cluster 11 (V / *M. bovis)* 75553.2222415322.234323241	1	5	0	6	60	2	0	0	1	0	0	1	1	0
Total	**34**	**125**	**9**	**217**		**69**	**0**	**0**	**11**	**2**	**0**	**62**	**6**	**7**

Table 1: 11 Community Clusters. *Total number of links in a cluster is equal to the total number of patients in the cluster, minus one.

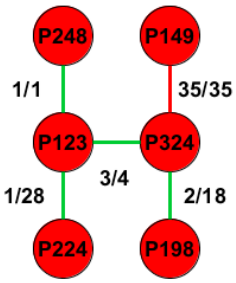
Cluster 1



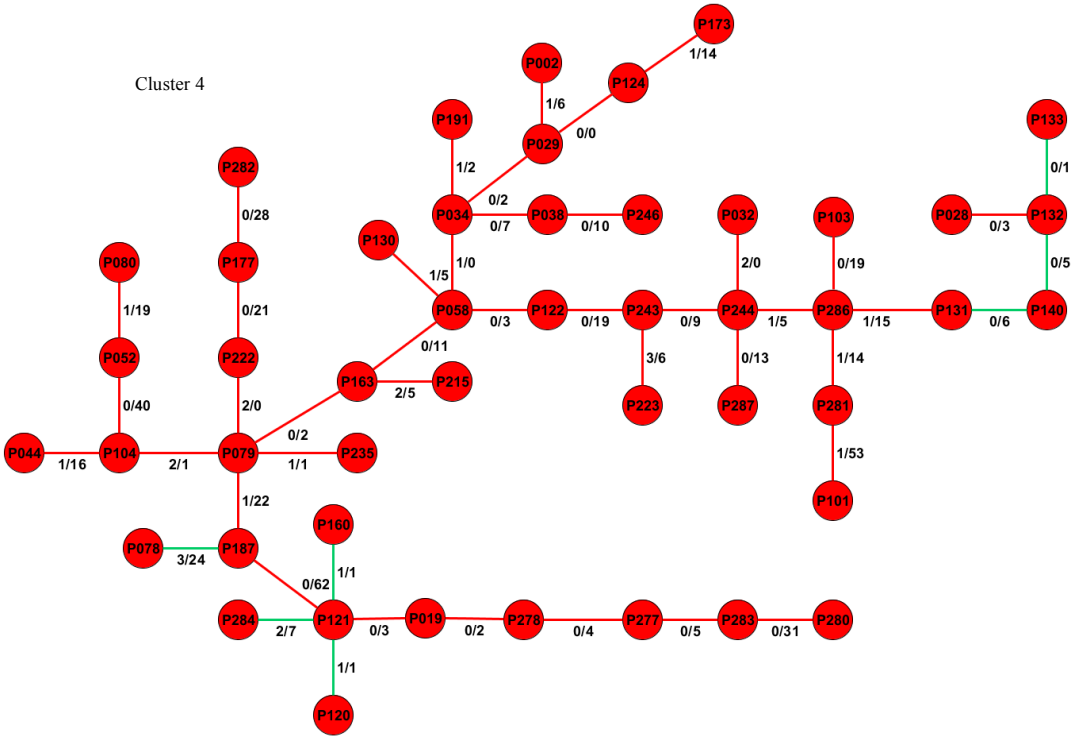
Cluster 2

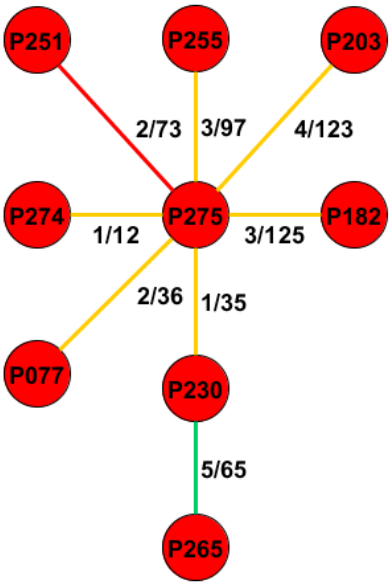


Cluster 3

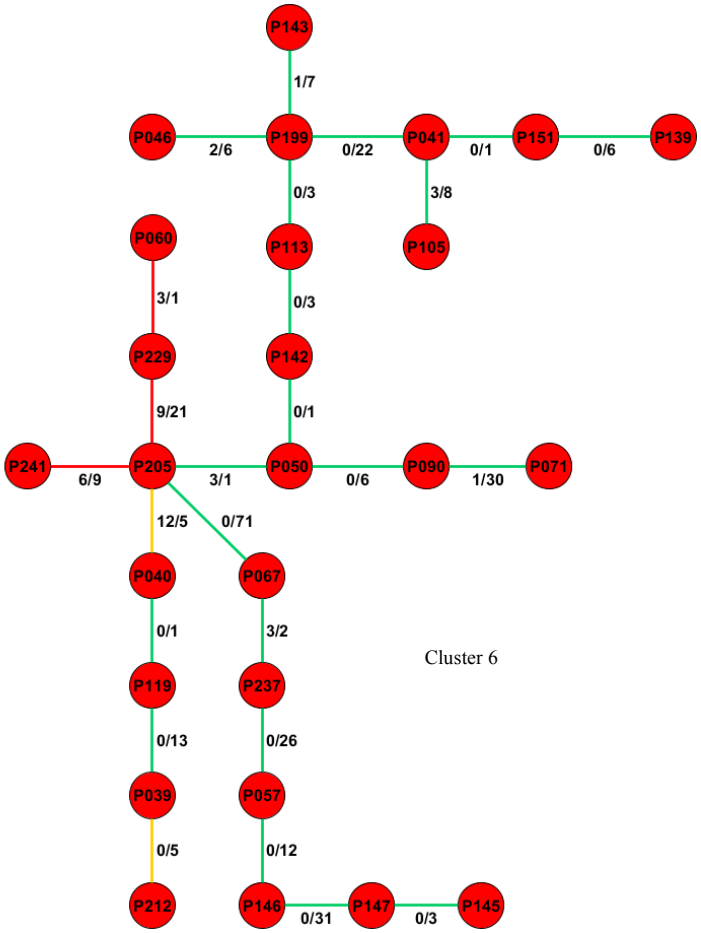


Cluster 4

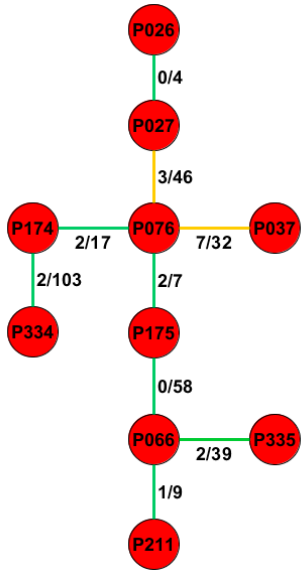




Cluster 5

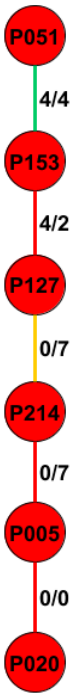


Cluster 6

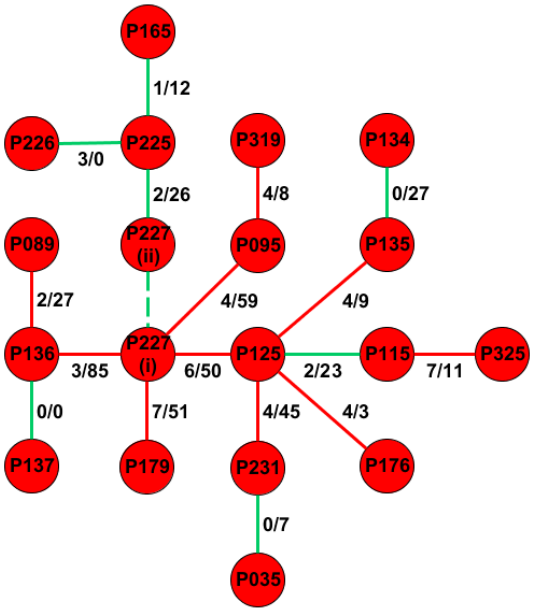


Cluster 7

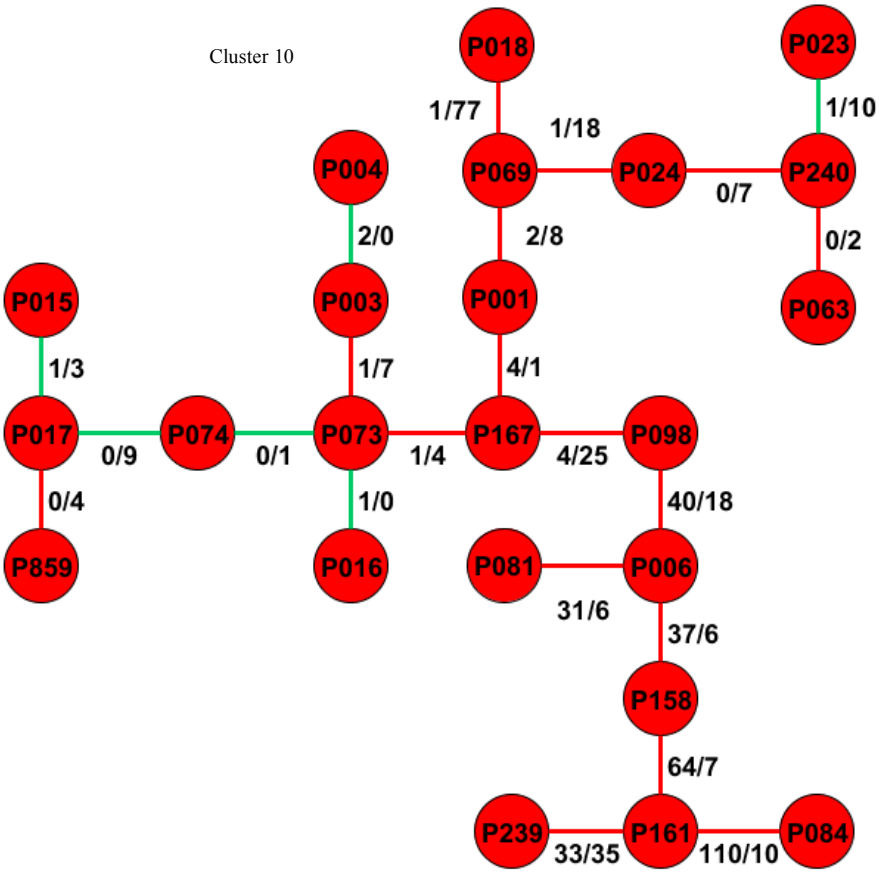
Cluster 8



Cluster 9



Cluster 10



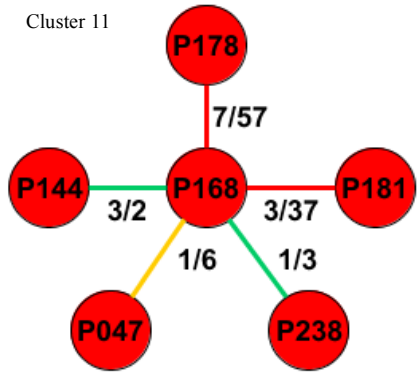


Figure 9: Eleven network diagrams, one for each cluster, showing the relationship between patients according to epidemiological linkage, SNP distance and timing of diagnosis. Nodes are coloured in red with patients represented as “P” plus their corresponding study number. Edges are coloured green where an epidemiological link was known, yellow where one was though possible, and red where no link was known. Each ‘edge’ is annotated with the number of SNPs/months separating isolates. P227 in Cluster 9 is represented by two isolates due to the large SNP distance between them.

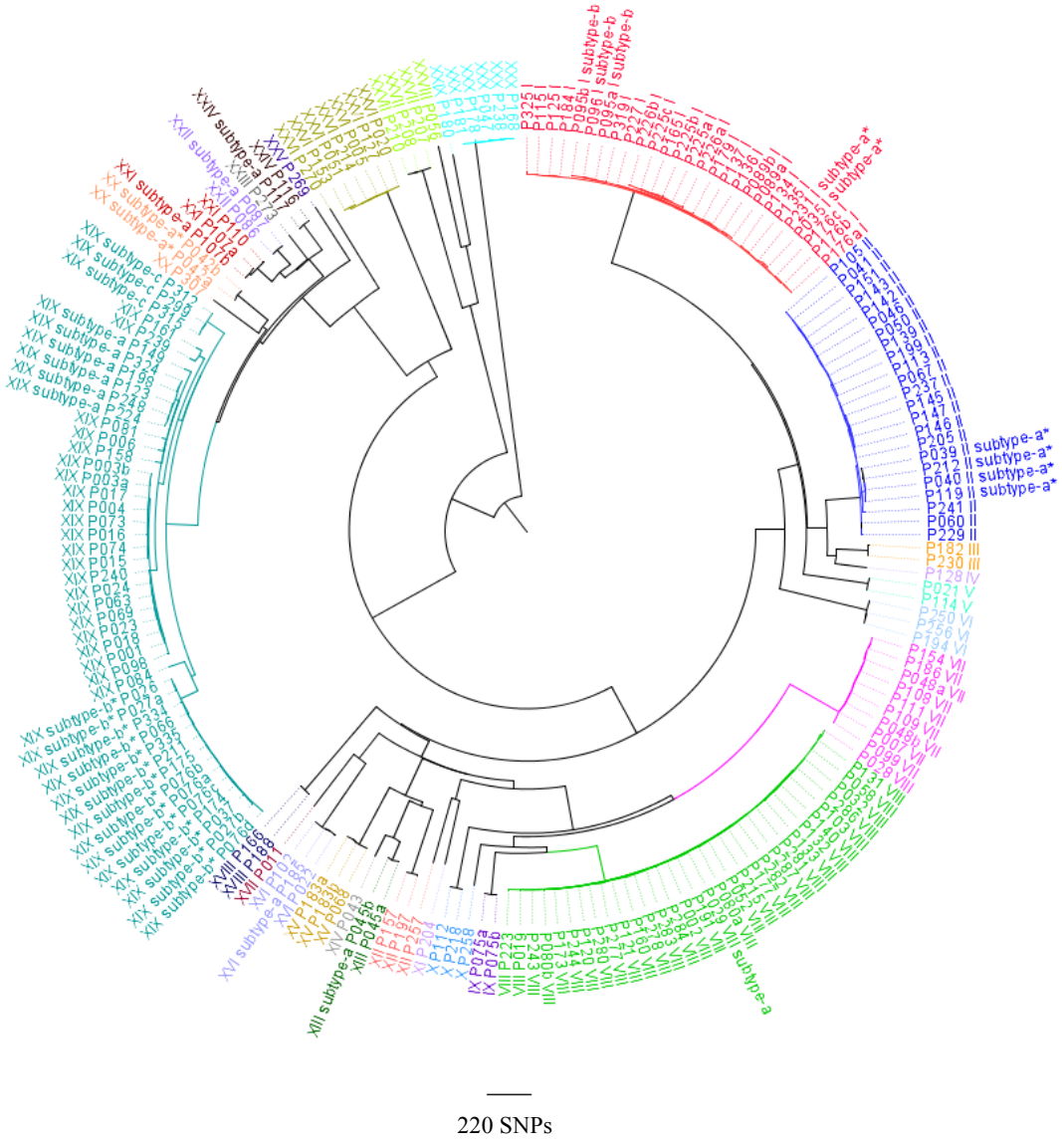


Figure 10: A maximum-likelihood phylogeny of 195 isolates typed to 24-MIRU-VNTR loci. MIRU-VNTR profiles and their subtypes are grouped by colour. Each isolate is labelled by a patient number (P) and a roman numeral indicating the MIRU-VNTR type. Each subtype listed differs from the main type at a single MIRU-VNTR locus (* 2 loci where the subtype is asterisked). Global MTB lineages are defined by MIRU-VNTR type.[222] Lineages: Beijing XXVI; European American II-VIII, X-XVI; Central Asian XVII, XIX-XXV; East-African Indian XXVII-XXIX; Unassigned I, IX-X, XVIII, XXVI; *M. bovis* XXX; *M. africanum* not shown.

3.6.2 Genetic Distance Within Individuals and Between Individuals

The greatest genomic diversity expected within individuals was estimated by sequencing paired pulmonary and non-pulmonary isolates from 49 patients, and 110 longitudinal isolates from 30 patients (Figure 11). In three cases a second infection with a different strain, rather than within-host evolution, was likely as 475, 1032 and 1096 SNPs separated isolates. In one patient with TB meningitis and a normal CT chest, paired cerebrospinal fluid and sputum isolates were 11 SNPs apart and in four individuals who developed drug resistance over seven to ten years of persistent pulmonary infection, maximum distance ranged from 6-10 SNPs. All 71 other pairwise comparisons (i.e. cross-sectional and longitudinal isolates) differed by 5 SNPs or fewer. There were significantly fewer SNPs in paired than longitudinal isolates (37 of 47 (79%) vs. 11 of 24 (46%) had 0 SNPs respectively; ranksum p=0·009).

The genetic distances between individuals in known recent transmission chains were estimated from the sequences of 86 isolates (63 individuals) from 25 household-defined outbreaks. All 38 links (number of patients minus number of outbreaks) between patients were within 5 SNPs. There was no evidence that the distribution of SNPs below this threshold differed from longitudinal isolates (ranksum p=0.16, Figure 11). Overall, after excluding differences of greater than 400 SNPs, 109 of 114 (96%) paired isolates from within individuals and household outbreaks differed by 5 SNPs or fewer, 108 of 114 (95%) by 4 SNPs or fewer and 103 of 114 (90%) by 3 SNPs or fewer.

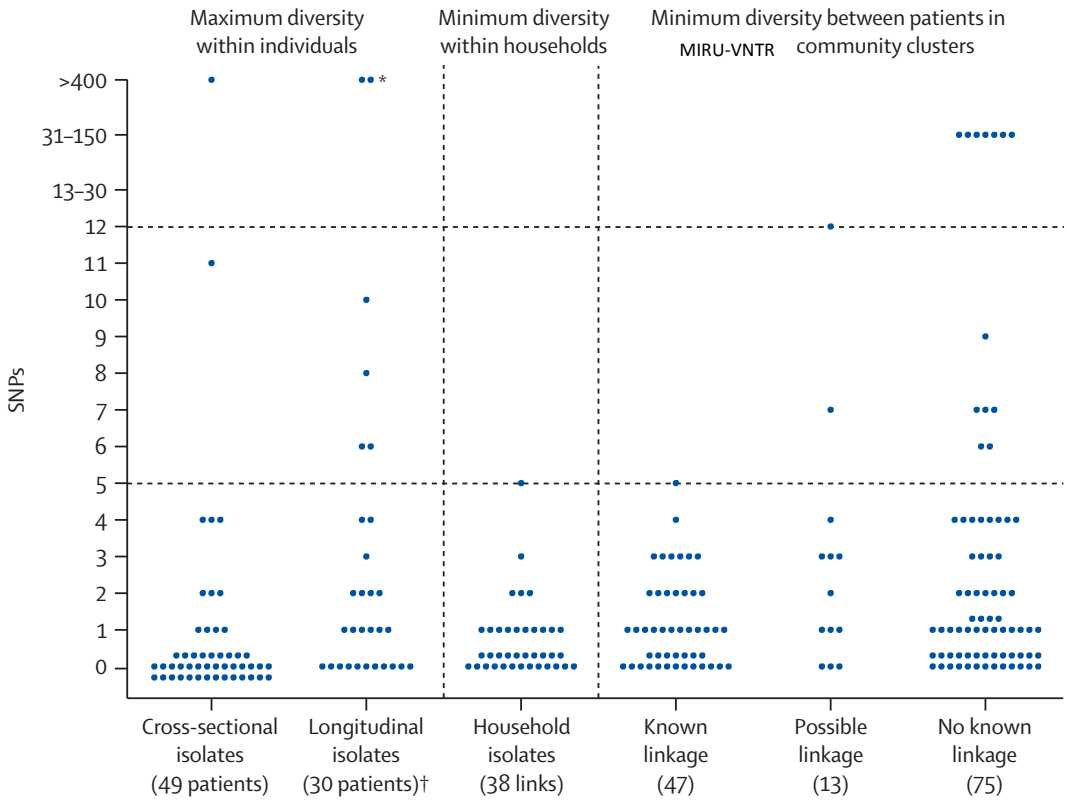


Figure 11: Time-unadjusted pairwise genetic distances in SNPs. 22 of the 38 links within the 25 household clusters also occur within community clusters (i.e., 'known linkage') but are shown with household isolates and not with community isolates. Top horizontal dashed line indicates the threshold above which direct transmission can be judged to be unlikely. The bottom horizontal dashed line indicates the threshold below which the possibility of transmission should be investigated. *Isolates had substantially different MIRU-VNTR profiles. † Pair of *M. africanum* isolates represented as 2 SNPs apart.

The rate of microevolution of *M. tuberculosis* over time was estimated from the first and last sequenced isolates of the 30 longitudinally sampled patients and from the 25 household outbreaks. A molecular clock rate of 0.5 SNPs per genome per year was inferred by maximum-likelihood (95% confidence interval 0.3-0.7) (Figure 12). All longitudinally sampled patients received anti-tuberculosis drug therapy. Although HIV testing was not systematically performed until 2011 (only eight results were available, all negative), UK rates of HIV-tuberculosis co-infection are relatively low, declining steadily from 9% in

2003 to 4.9% in 2010.[55] There was weak evidence that the initial diversity and molecular clock might vary between household outbreaks and longitudinal outbreaks (p=0.08 comparing joint vs. separate models). However, if anything the mutation rate was lower within individuals followed longitudinally (0.3 SNPs per genome per year (95% CI 0.03-0.6)) than in those in household outbreaks (0.6 SNPs per genome per year (95% CI 0.3-0.9)), and the initial diversity higher (1.2 SNPs (95% CI 0.3-1.9) vs. 0.2 SNPs (95% CI 0.008-0.7) respectively).

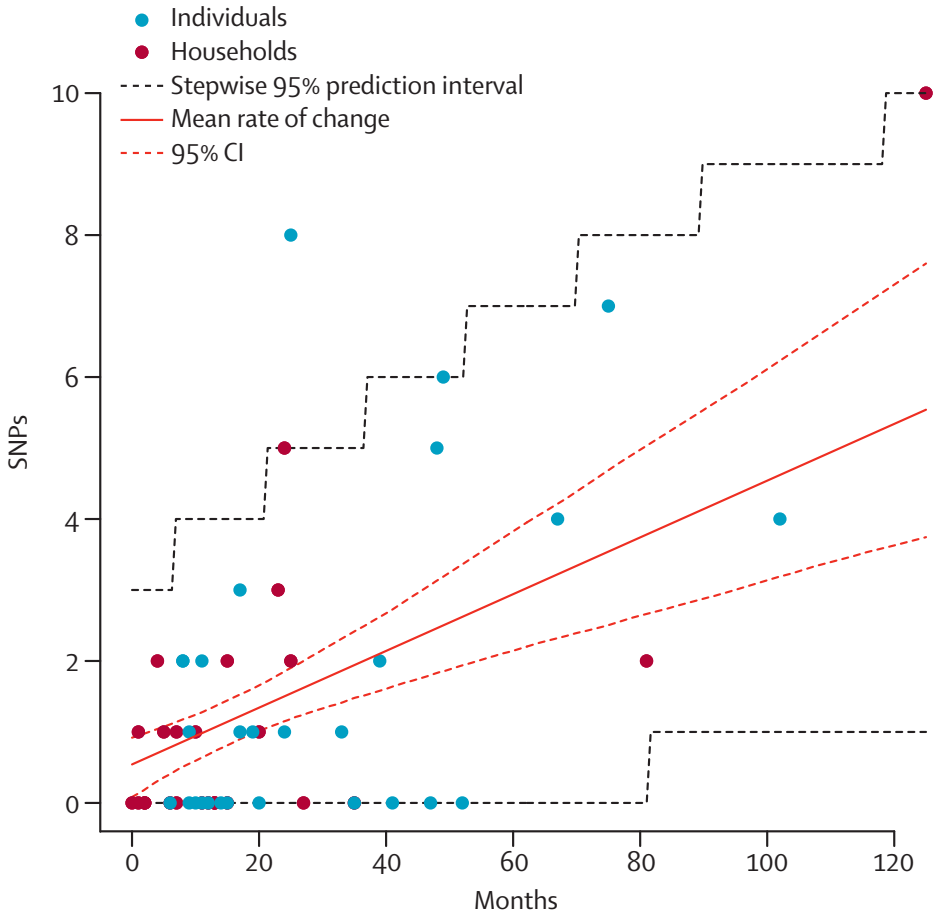


Figure 12: The rate of change in DNA sequences estimated by coalescent-based maximum likelihood from the first and last isolates from individuals with persistent open tuberculosis and from households.

I used these results to construct two thresholds against which to assess the MIRU-VNTR community clusters. I expected epidemiological linkage consistent with transmission to exist between isolates differing by 5 or fewer SNPs, and not to exist between isolates differing by more than 12 SNPs. I considered pairs differing by 6 to12 SNPs to be indeterminate.

3.6.3 Genetic distance across MIRU-VNTR-based clusters

Eleven community clusters were defined by their MIRU-VNTR profile, or up to two locus mismatches (Table 1). These clusters accounted for 217 isolates (168 patients) and had a median duration of 88 months (range 9-144). Starting from the first case in each cluster, we constructed 11 networks (one for each cluster), accounting for 157 potential transmission events (edges) (Figure 9). Within the three clusters centred on schools, 17 of 20 (85%) patients could be epidemiologically linked (Table 1), with no link confirmed in the three MIRU-VNTR-matched community isolates. Indeed, the community-based case in cluster 3 was 35 SNPs away from the school isolates. In clusters 6 and 7, 27 of 34 (79%) patients could also be epidemiologically linked, whereas in the remaining six clusters, including one associated with a recent African immigrant community (cluster 10), only 25 of 103 (24%) patients could be epidemiologically 'linked' (Table 1).

None of the 69 epidemiologically linked patients and only 2 of 13 (15%) possibly epidemiologically linked patients were separated by more than 5 SNPs (7 and 12 SNPs respectively). Conversely, 13 of 75 (17%) epidemiologically 'unlinked' patients were separated by over 5 SNPs, and 7 of 75 (9%) by more than 12 SNPs (three-way comparison Fisher's exact test $p<0.0001$). However, 22 of 69

potential transmissions known to be epidemiologically linked featured in both household and community outbreaks. Excluding these, the number of epidemiologically linked patients differing by 5 SNPs or fewer was 47 (p=0.003).

Of interest, 62 patients without epidemiological links were separated by fewer than 5 SNPs and seven by 6-12 SNPs. The ability to identify cryptic outbreaks was most evident in cluster four, where fewer than 5 SNPs separated 38 individuals with a background of substance misuse for whom contact tracing had been difficult (Table 1). The ability to rule transmission out was particularly evident in cluster 10 where more than 30 SNPs separated five individuals from a recent immigrant community and one British born individual from the next nearest patient. However, in this cluster, isolates from ten patients across two cities 45 miles apart, with no known epidemiological links, could also be genetically linked by 5 SNPs or fewer, demonstrating its potential as a “rule in” to trigger wider contact investigations.

To explore the potential for 24-locus MIRU-VNTR typing to miss genuine transmission, we assessed the proportion of sequenced isolates matching at 22 or 23 MIRU-VNTR loci that were genetically linked by 5 SNPs or fewer (Figure 13). Among 195 isolates typed at 24-loci, 14 pairs differed at 1 or 2 loci (i.e. were matched at 22 or 23 loci). Three pairs of isolates were from individuals and three from households, all within one SNP of each other. Of the four instances where mismatching isolates were included in a community cluster, two had ‘possible’ epidemiological links to the cluster (within 4 and 12 SNPs) and two had ‘no known’ link to the cluster (0 and 5 SNPs apart). The remaining four pairs had no known epidemiological links and were separated by 25, 94, 96

and 275 SNPs respectively (Fisher's exact p=0.03 comparing 6 of 6 'known' vs. 3 of 8 other links within 5 SNPs). No isolates matching at 21 MIRU-VNTR loci or fewer were genetically linked within 5 SNPs (Figure 13).

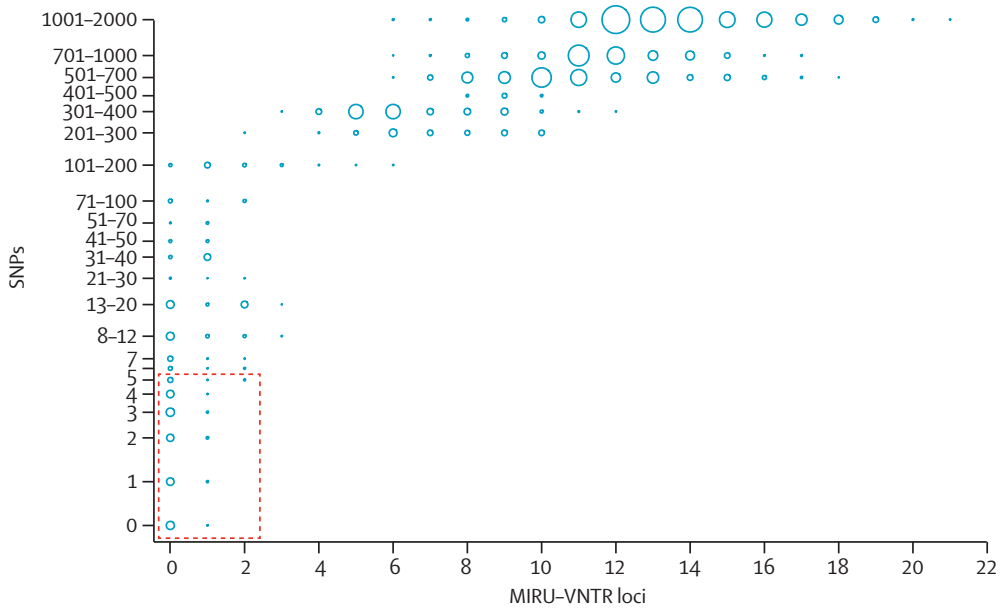


Figure 13: Comparison of all isolates with complete 24-locus MIRU-VNTR profiles. As each isolates was compared to each other isolate, the number of SNPs and MIRU-VNTR loci at which they diverge was recorded. Results are plotted on a log scale. Circle sizes are proportionate to the number of pairs diverging by a specific number of loci and SNPs. The dashed red box includes isolates that differ by five or fewer SNPs.

3.6.4 Patterns of transmission

Whilst school outbreaks are often caused by particularly infectious individuals responsible for a large proportion of secondary cases,[140] one important question is whether so-called super-spreaders are also the source of many community outbreaks.[141] A star-like phylogenetic topology was apparent in all non-school outbreaks except for cluster eight (Figure 14). The possible presence of a super-spreader was supported clinically and epidemiologically in

clusters five and seven, but there was insufficient data to confirm or refute this in the other clusters.

In cluster five, the nine isolates were sequenced in two separate HiSeq runs. Phylogenetic analysis of the first six isolates demonstrated a vacant central node in the genetic tree, consistent with a potentially still un-sequenced common source isolate. One of the three isolates from the subsequent HiSeq run matched this predicted sequence precisely (orange nodes in Figure 14). It belonged to a treatment non-compliant drug-dealer with cavitating, pulmonary, smear-positive tuberculosis. This likely super-spreader had been diagnosed early in the outbreak, had interacted with many contacts and was eventually detained under public health law in the interests of public safety.

In cluster seven, the centrally placed individual was treatment non-compliant for four years, had cavitating, pulmonary, smear-positive disease and had “known” or “possible” epidemiological links to all other infected individuals, as illustrated in Figure 15. This individual tended to drink at the homes of some of the secondary cases in the cluster, as well as in a local pub that other secondary cases also used to frequent. As such it is plausible that all other secondary cases in the cluster spent prolonged periods sharing the same air as the suspected index case, regardless of whether they had made social acquaintances with the individual.

In addition to being the first patient to be diagnosed, having clinical disease compatible with being infectious, and having confirmed or plausible epidemiological links to the secondary cases in the cluster, and being centrally placed on the phylogenetic tree, the suspected super-spreader also had a

mixture of base-calls at variant sites that were representative of the genetic variation observed between isolates from secondary cases at those sites. A plausible explanation is that a so-called cloud of genetically related but different strains may have evolved in the individual over the years of non-compliance with treatment, but that a bottle-neck at the point of transmission allowed for only one strain to infect each secondary case (Figure 15).

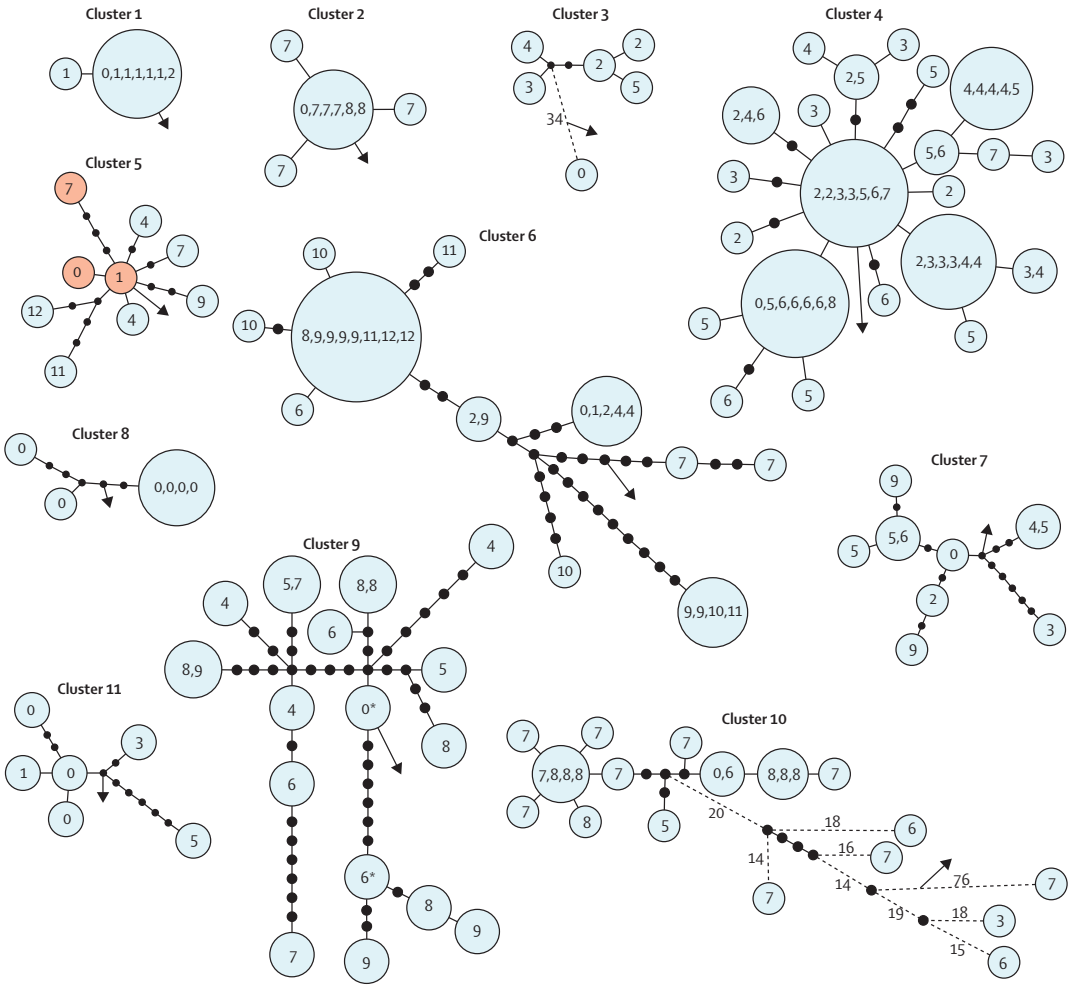


Figure 14: Genetic distances estimated with maximum likelihood. Each blue circle represents a node of people who were infected with isolates separated by 0 SNPs. Each number within a circle is one patient, and the value of the number indicates the year of the outbreak they were diagnosed in (the first patient being diagnosed at year 0 in each case). For patients with several isolates, only the closest in SNPs to the next patient is included.

As well as signalling super-spreaders, phylogenetic topology was indicative of the sampling density within MIRU-VNTR clusters. For example, fewer vacant black nodes can be seen in cluster four, where isolates from 47 of 52 patients clustered by MIRU-VNTR were sequenced, than in cluster nine, where only 18 (6%) of over 280 were sequenced (Figure 14).

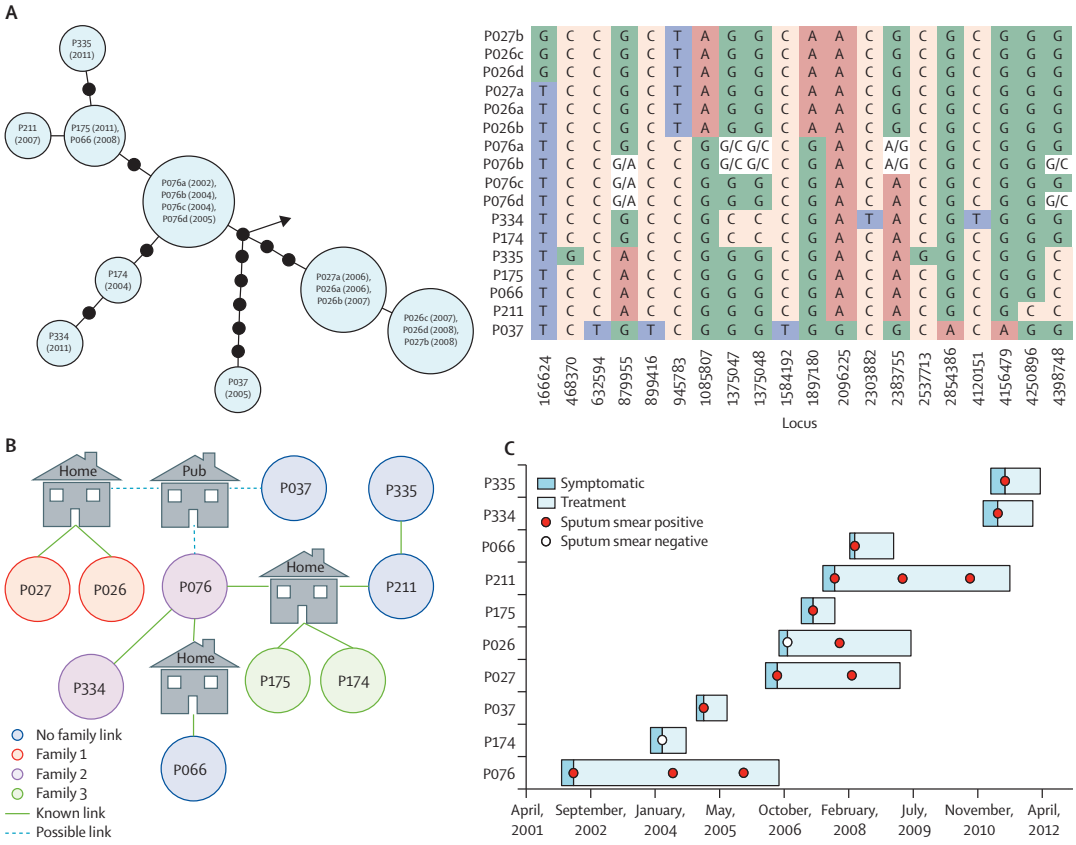


Figure 15: (A) Genetic tree and matrix of nucleotide variants. Genetic distances estimated with maximum likelihood. Each blue circle represents a node of people who were infected with isolates separated by no single nucleotide polymorphisms. Numbers within nodes are patients numbers and years of sample isolation are given in parentheses. The matrix shows nucleotide variants. (B) Time of onset of symptoms, diagnosis and treatment. (C) Sputum smear positive samples show probable infectious periods.

3.7 DISCUSSION

These findings demonstrate how WGS can delineate both the margins and the structure of tuberculosis outbreaks with unprecedented resolution. 96% of *M. tuberculosis* isolates were within 5 SNPs of another isolate taken from the same individual or from a household contact. This finding provides essential context for the diversity expected from closely related isolates in transmission chains. In a low HIV prevalence setting, I estimated a molecular clock of 0.5 SNPs per genome per year, similar to estimates in macaques,[61] predicting a maximum of 5 SNPs at three years and 10 SNPs at ten years. I then applied a metric based on two thresholds, fewer than or equal to 5 SNPs and greater than 12 SNPs, to 11 MIRU-VNTR-based community clusters to examine the potential of WGS to identify outbreaks more effectively and direct contact tracing more efficiently. Across 157 potential transmissions in these 11 clusters, all 69 epidemiologically 'linked' patients could be genetically linked within 5 SNPs, with only two of 13 patients classed as 'possibly' epidemiologically linked in the intermediate range of 6-12 SNPs. The only genetic links to exceed 12 SNPs involved 7 of 75 patients with 'no known' epidemiological links. These results suggest that within MIRU-VNTR defined clusters, WGS offers sufficient resolution to identify outbreaks where they occur and to discount them where they do not.

The idea that SNP thresholds can be set as a guide to inferring transmission was criticised soon after this data was published. In a study of 97 epidemiologically linked pairs of *M. tuberculosis* sequences from the Netherlands, the authors argue that the variability in the molecular clock over

the short time period that is relevant to transmission studies precludes any setting of arbitrary thresholds. The authors instead argue that inferences about transmission should be made on the basis of "deep sampling of a phylogenetic cluster".[236] There are a number of problems with their critique. First, their data broadly corroborates mine. The mean molecular clock is similar to that estimated in this thesis, at 0.3 SNPs per sequenced genome per year, whereas the SNP distances separating epidemiologically related pairs was 0 in 37 instances, and greater than 5 in only 3 instances (6, 7 and 8 SNPs respectively). Second, the thresholds of 5 and 12 SNPs are not designed to represent a biological truth that can be countered by their (and my own) observation that there is some variability within the molecular clock. They are instead suggested decision making guides that are (a) easy for end users to understand and (b) accommodate most of variability in microevolution that has been observed in this thesis as well as in the study from the Netherlands. Third, one can only use the phylogeny to understand transmission after the outbreak has occurred. To paraphrase Karl Marx, "phylo-geneticists have only interpreted clusters in various ways; the point, however, is to change them" [*]. If outbreaks are to be prevented then public health teams have to be able to make a decision about the relatedness of the first two sequenced isolates before additional data emerges that might enrich any phylogenetic signals, but that also represents failure of public health control. With two samples at hand, all one has to relate them to one another is genetic distance.

[*] "The philosophers have only *interpreted* the world, in various ways; the point, however, is to *change* it." The 11th of Marx's Theses on Feuerbach (1845)

There have been a number of articles published since that use WGS to study within-host genetic variation or outbreaks. To assess the maximum within-host diversity, Laura Pérez-Lago *et. al.* sequenced isolates from individuals and outbreaks where at least one isolate could be differentiated by RFLP or MIRU-VNTR patterns.[237] Despite the deliberate selection of unusual cases, 42 of 54 (78%) pairwise comparisons were within 5 SNPs. More recently, papers by Josephine Bryant *et. al.* and Afonso Guerra-Assunção *et. al.* explored the utility of WGS to distinguish relapsed infection from reinfection in high-incidence settings.[238,239] The study by Bryant *et. al.* focussed on patients enrolled in the REMoxTB trial,[114] identifying 33 of 47 pairs of isolates from patients with a second episode of disease as relapses. Of these relapsed pairs, 27 of 33 were separated by 0 SNPs, whilst 3 were separated by 1 SNP, 2 by 2 SNPs and 1 by 6 SNPs. Guerra-Assunção's study asked the same question of a population in Karonga, northern Malawi, over a 14-year period (1996-2010). They found that 51 of 66 patients with a second episode of disease had suffered a relapse of their earlier infection, of which 29 pairwise comparisons were 0 SNPs, 48 were within 5 SNPs, and the remaining three were separated by 6, 7 and 8 SNPs respectively.

Although Bryant *et. al.* [236] caution the use of SNP thresholds for the transmission, and Pérez-Lago[237] and Guerra-Assunção[239] challenge the level of my suggested thresholds, the data presented in these subsequent studies nevertheless remain consistent with my findings. Consequently, it is reasonable to assess the clinical utility of these thresholds by sequencing *M. tuberculosis* strains prospectively, as is due to start in the Midlands in early

2015. To the best of my knowledge this will be the first study to assess the added value of sequence data to public health control of tuberculosis infection.

Aside from providing SNP distances between isolates, sequence data might also resolve the structure of some transmission networks,[240] which could benefit public health if contact investigations can be focussed around the most infectious individuals. Within the eight non-school community clusters I was able to identify two possible super-spreaders by inspecting the genetic trees for nodes from which multiple lineages diverge. Although super-spreaders have previously been hypothesised, there has previously been no direct genetic evidence for them.[54,141,241] In both cases the individuals whose isolates were placed centrally on the tree were also epidemiologically the most likely source for secondary cases in their respective clusters.

The relationship between the first six sequences in cluster 5 implied the existence of a source case, or isolate that had not been sequenced. This was of particular interest not only as the prediction was confirmed by the subsequent sequencing of additional isolates, but also because the epidemiological data implied that the individual in question was the super-spreader. Current typing methods provide no insight into the existence of undiagnosed cases, which is a potential problem as patients who remain undiagnosed have many opportunities to infect secondary cases. Indeed, as the natural history of disease suggests, some patients may never seek medical attention at all.[50] Although in these specific examples the central node/source case was actually diagnosed early in each outbreak, the findings suggest that public health teams could use WGS data in real-time to target active case finding efforts where the

existence of undiagnosed individuals is predicted by the evolving genetic topology of the outbreak.

There are good theoretical reasons to be cautious when attempting to make inferences about the direction of transmission in a cluster from the phylogenetic relationships. First, because of the bottlenecks involved in sample production, any degree of genetic diversity within the donor at a single time point could lead to different, albeit closely related strains being transmitted to a recipient case and being sequenced as part of the diagnostic process, as was hypothesised by Midori Kato-Maeda.[242] Even if the full range of diversity is captured in a clinical sample, the process of culturing an isolate prior to DNA extraction and sequencing could favour one strain over another.[243] Second, incomplete sampling frames can result in inaccurate inferences if the true donor is not sampled, or if the donor is sampled only after micro-evolutionary changes that distinguish the transmitted strain from the sampled strain.[244]

Despite appropriate theoretical caution, a number of publications have since also described super-spreaders at the centre of genetic networks or phylogenetic trees. Kato-Maeda reported on nine sequenced cases from an outbreak in San Francisco where the SNP-based network corroborated the epidemiological hypothesis that a single super-spreader was responsible for the secondary cases.[242] Thomas Kohl *et. al.* argued that an individual at the centre of their tree of sequences from an outbreak in Hamburg was the index case, with epidemiological links to at least one patient within 5 of 8 secondary nodes as they appear on the tree.[245] Likewise, Stucki *et. al.* describe an outbreak in Berne where the epidemiology and genomic signals both indicate a

particular individual as the super-spreader.[246] In fact, to the best of my knowledge, there are no published examples of outbreaks where good epidemiological data was available and where the sequence data implicates a different super-spreader than the epidemiology. As the number of outbreaks sequenced increases one will be able to attribute a sensitivity and specificity to such genomic predictions, publication bias allowing. Similarly, as experience of interpreting the topology of genetic trees develops, one hopes that this will not only become a useful tool for the identification of potential super-spreaders, but also for assessing the completeness of outbreak investigations and determining the likelihood of on-going transmission.

Although MIRU-VNTR typing has been an effective way of identifying outbreaks of tuberculosis,[214,215,224,247] the genomic diversity within some MIRU-VNTR genotypes[220,248] can leave public health teams uncertain how intensively to investigate clusters where the epidemiological links are unapparent. My results, which support the recent observation that micro-evolutionary events can change a MIRU-VNTR genotype within a host,[249] add to this uncertainty. The consequences are long and costly contact tracing efforts that may ultimately be futile. Cluster 10 illustrates this dilemma best. Although all but one patient in the cluster were from the same country of origin in east Africa, and although all patients had the same 24-locus MIRU-VNTR profile, the patients were resident across four different towns in the Midlands and north of England. No patient volunteered an epidemiological link to another case that extended beyond the immediate household, yet public health teams were also aware of a prevailing social stigma that could explain such apparent reticence. It was only when the sequence data was first presented that the

public health teams discovered they had not only in some cases expended time searching for epidemiological links between genomically unrelated patients, but that there were also genomically linked cases resident in towns separated by the boundaries within which Health Protection Units operate. Whether the greater certainty afforded by WGS would have resulted in the identification of additional links in cluster 10, had the data been available during the epidemiological investigations, is a matter of speculation. However, similar data did inform on-going investigations into cluster 7, helping identify additional contacts who then received isoniazid and rifampicin prophylaxis after screening positive for LTBI by interferon-γ release assays.

A study of *M. tuberculosis* infection in nine macaques compared the molecular clock rate between latent disease and active disease with the rate observed in culture, finding no evidence that the rate (0.5 SNPs per sequenced genome per year) differed significantly between the different scenarios.[61] Within the variation I recorded, some instances were consistent with the slower mutation rate expected in latent infection, such as the 0 SNP difference over 7 years in cluster 2 (Figure 14), even though this rate is still within the margins of the credibility intervals on the mean clock rate calculated in this chapter (also 0.5 SNPs per sequenced genome per year) (Figure 12). Since the data in this chapter was published, similar clock rates have been calculated by Bryant *et. al.*, ~0.3 SNPs per genome per year based on repeatedly sequenced individuals, and by Roetzer *et. al.*, ~0.4 SNPs per genome per year based on a transmission chain in Hamburg.[236,250] Although there appears to be convergence on a clock rate across these studies, whether the clock rate is

genuinely no slower in what is commonly thought of as ‘latent infection’ still needs corroborating as the original macaque study was very small.

Along with the molecular clock, the 5 SNPs and 12 SNP thresholds are agnostic to the incidence of the study setting. As the sampling frame for my study included an ethnically diverse population in whom all main global diversity of *M. tuberculosis* were represented, and of whom many originated from high transmission countries, these results are likely to be valid in high incidence settings beyond the Midlands, UK. Indeed, the recent sequencing-based study of within host diversity in Malawi demonstrated similar SNP differences between paired sequences to those described in this chapter.[239]

However, the utility of these findings to a high incidence setting does still need to be formally assessed, as they indeed do for low-incidence settings. In theory, in a low-income, high-incidence setting where patients might remain undiagnosed, or receive inadequate treatment, there is the potential for these cases to transmit tuberculosis as super-spreaders to many recipients, leading to larger cluster sizes. Were public health teams able to conduct contact investigations in such circumstances, it is possible that they would derive more benefit from ruling out recent transmission than they do from linking cases within 5 SNPs of each other, because the number of intermediary cases within 3 years of evolution may still be large. To this end it would have been informative to characterise cluster 9 in more detail as over 280 patients were linked to it by historical RFLP and MIRU-VNTR typing. Unfortunately the resources were not available to sequence this cluster in its entirety at the time. However, in a recent study of a township in South Africa, where the incidence is

an enormous 2,000 cases per 100,000 population, RFLP-typing defined 87 independent clusters among 478 patients with a genotypic match, demonstrating numerous small clusters instead of a single uncontrolled circulating clone.[251]

Whether or not the transmission in high-incidence settings follows the pattern of many smaller or of few larger outbreaks that might render small SNP distances harder to interpret, it should nevertheless be possible to identify phylogenetic signals for potential super-spreaders. The phylogenetic topology in cluster 4, from which we sequenced 47 of 52 patient isolates, is consistent with more than one super-spreader, although the epidemiological history of this outbreak was insufficiently characterised to be able to draw any further conclusions. Where infectious individuals remain undiagnosed, the mycobacterial subpopulations evolving within these individuals over time may contribute to their identification as super-spreaders upon diagnosis, as would have been the case for the index case with the mixed-base calls observed in cluster 7.

Potential limitations of this study include the failure to sequence all isolates from the community-based clusters, as some were unavailable. However, where isolates that diverge by 5 or fewer SNPs were sequenced, any missing intermediate cases would make transmission more, not less plausible. Real-time contact investigations also have to deal with missing data, using existing information to judge the plausibility of additional, intermediary cases. Another limitation, due in part to the relatively short read lengths, is that only ~88% of each genome could be accurately mapped. This excluded repetitive segments within the reference genome such as the regions defining MIRU-VNTR.

Although additional diversity may therefore be masked from my analysis, it is unclear how much additional resolution this would provide.

The decision by PHE to pilot WGS technology for routine tuberculosis public health practice in the Midlands in 2015 has been made against a background of rapidly declining costs of WGS technologies and recent advances that have improved turnaround-times significantly.[189] Costs are now in the region of £40 per sequence, similar to the costs of MIRU-VNTR typing. I predict that WGS will likely be of benefit in complex community outbreaks where epidemiological data is difficult to obtain, and where SNP distances alone can be used to either include or exclude individuals from a transmission chain. This could by itself save public health teams time and money by preventing unnecessary investigations, as was the case in cluster 10. In some cases the phylogeny might provide information for more targeted contact investigations. Based on what we know already, the impact on how public health surveillance and intervention is practiced in the UK is likely to be substantial.

4 TUBERCULOSIS TRANSMISSION IN OXFORDSHIRE, 2007-2012.

4.1 INTRODUCTION

Having defined a metric for transmission based on genomic distance in chapter 3, I applied the method to a population study of Oxfordshire to investigate transmission patterns at a regional level. Such a study is of interest as the burden of tuberculosis in the UK is among the highest in Western Europe, with ~8000 cases diagnosed in 2013.[132,134] Although the incidence fell to 12.3 cases per 100,000 population in 2013, it had risen by over 20% from 11.4 per 100,000 in 2000 to 14.4 per 100,000 in 2011.[55,134] These additional cases have been accounted for by patients born overseas,[216] but the role of inward migration in local transmission remains unclear. Designing control measures to sustain the reverse in incidence trends requires a better understanding of the relative contributions of reactivated versus newly acquired infections to overall disease.

Population studies of this kind have been undertaken in the past, most notably by Allix-Beguec[214] and by Roetzer[252], albeit using MIRU-VNTR genotyping. Such studies would have been possible in the UK after the National Strain Typing Project introduced routine 24-locus MIRU-VNTR in 2010, but these were not done.[55] However, as chapter 3 has demonstrated, MIRU-VNTR cannot distinguish recent from historical transmission as reliably as WGS data can. Although WGS has been shown to be a more appropriate technique with which to address such questions,[141,236-238,250,253] prior to the study outlined in

this chapter, it had not yet been exploited for a population-based study to more accurately quantify transmission of tuberculosis within a defined region over a fixed period of time.

Oxfordshire has a population of 760,000 and is a low-incidence region with 8.4 cases per 100,000 population. Like the UK as a whole, most cases are restricted to a small number of urban centres.[133,254] In this chapter I present the findings of an investigation into the contribution of tuberculosis cases arising from transmission in Oxfordshire, and whether the rate of transmission events varies between those born in high-incidence and those born in low-incidence countries.

4.2 METHODS

4.2.1 CULTURE AND DNA EXTRACTION

I cultured the majority of samples from frozen stocks, first in MGIT until ample growth was visible, and then on LJ media. Where sufficient culture was apparent in archived freezer vials, I transferred culture directly onto LJ medium. Where I identified cultures for sequencing prior to archiving, I subbed these directly from MGIT to LJ, without having been frozen. All cultures were incubated at 37°C until mature growth was obtained.

The CTAB method described in the previous chapter was laborious, taking two days to process just 24 samples, and necessitating a fume cabinet because of the chloroform. I therefore developed an alternative, cleaner and more efficient method for the extraction and purification of DNA in this study. The protocol is based on the commercial Fuji Quickgene DNA extraction kit (Kurabo Bio-

Medical, Osaka, Japan). I adapted the manufacturer's protocol to optimise the DNA yield from *M. tuberculosis*, primarily by introducing a mechanical disruption step, as had been done by colleagues in my laboratory working on other pathogens. All samples processed in Oxford were subject to this method, with only the few samples not archived in Oxford and supplied by the PHE National Mycobacterial Reference Unit being processed in London using the CTAB method.

As the Quickgene method was new to our laboratory, experiments were first required to demonstrate its safety, ensuring the cultures were no longer viable before removing them from the containment level 3 laboratory for further processing. I undertook a heat-kill experiment whereby mature cultures were harvested from 36 LJ slopes, suspended in 400ul of 0.85% saline or 1XTE in a 1.5ml or 2ml screw-top micro-centrifuge tube, sonicated for 20 minutes at 35kHz (60W), and placed in a heat block set at 95°C. I repeated the experiment three times, heating the samples for 10, 30 and 120 minutes respectively. After heating I re-inoculated the cultures into MGIT and incubated these for 40 days or until the automated MGIT system indicated growth. Only among the 36 cultures heated for 120 minutes did all MGITs remain negative for growth at 40 days. 120 minutes was therefore taken forward as the preferred protocol.

The protocol was as follows:

1. Within the containment level 3 laboratory, take a 10ul plastic loop to harvest colonies from the LJ slope, and suspend them in 0.85% saline in a 1.5ml screw-cap micro-centrifuge tube.

2. Sonicate the tubes in the fume extraction cabinet for 20 minutes at a frequency of 35 kHz (60W).
3. Heat sonicated tubes in the heat block at 95°C for 2 hours. The samples are now ready to be removed form the containment level 3 laboratory.
4. To each sample add 20ul EDT (Proteinase K) and 300ul LDT buffer from the Quickgene kit
5. Transfer entire content of the sample tube to Lysing Matrix B for mechanical disruption with silica beads on the FastPrep-24 homogenizer (MP Biomedicals, Santa Ana, California, USA).
6. Run the FastPrep for 40 seconds at a frequency of 6 meters/second.
7. Repeat after a 5-minute interval.
8. Remove the lysing matrix B tube and centrifuge at 13600 rpm for 10 minutes
9. Transfer supernatant to a fresh tube, carefully avoiding transferring beads with the supernatant.
10. Place the fresh tube containing the supernatant on a heat block at 70°C for 10 minutes.
11. Add 240ul of 99% EtOH to the tube and vortex for 15 seconds.
12. Set up Quick-Gene machine: Place discard tubes in the front row and final tubes to catch the purified DNA in the back row. Filters are placed in the movable rack.
13. Transfer the entire sample content into the filter, making sure the filter is overlying the discard tube.

14. Turn the handle to operate the Quickgene machine. Air pressure is applied to the filters until all the fluid had gone through into the discard tubes. The DNA is selectively bound to the filter membrane.
15. Add 750ul WDT (supplied with the Quickgene kit) to each filter and turn the handle again until all the WDT has passed through the filter. Repeat this washing step three times in total (twice more).
16. Transfer the movable rack backwards so that the filters are overlying the final catch tubes.
17. Elute with 50-100ul of CDT (supplied with the Quickgene kit). Optimal results are obtained if the CDT is pre-warmed to 50-70°C. Once the CDT is added to each filter, wait 2-minutes before running the machine to elute the DNA through the filter and into the catch tube.

I quantified DNA concentrations with the Tecan Infinite 2500 Pro using PicoGreen, a fluorescent dye that intercalates with DNA.[226] Concentrations of 18ng/ul in 100ul were considered acceptable, with 36ul being used for sequencing.

4.3 NEXT GENERATION SEQUENCING, BIOINFORMATICS PIPELINE AND ANALYSIS

Isolates were sequenced at the Wellcome Trust Centre for Human Genetics in Oxford using Illumina HiSeq and MiSeq platforms (Illumina, San Diego, California, USA). The sequencing libraries were prepared by staff at the sequencing centres. As in the previous chapter, paired-end reads were mapped to the H37Rv *M. tuberculosis* reference genome (GenBank NC_000962.2) with Stampy v1.0.22.[227] Updating the pipeline described in chapter 3, the depth-

filter was removed, allowing sites of unusual (above the 97.5th percentile) depth to pass, increasing mean coverage from 88% to 92% of the genome. This did not impact on previously described thresholds of genetic relatedness or introduce false-positive variant calls.[253] All other described aspects of the pipeline remained the same. Mean high-quality read depth was 106 (range 25-195) and within strain standard deviations on the mean read depth varied between 9 and 55. False variant calls were assessed by re-sequencing isolates on different flow cells as technical replicates. Pairwise comparisons identified only a single false positive variant call across 202 genomes.

I applied bespoke, in-house python scripts to FASTA files to identify variant calls and quantify the observed pairwise SNP-distances across selected sequences. In contrast to chapter 3, phylogenetic trees, reconstructed using PhyML under a general time-reversible (GTR) model,[232] were based on whole-genomes rather than just variant sites. To maximise computational efficiency by sparing the PhyML software the need to impute nucleotide sequences in over 7% of the genome, invariant sites where no base was called were assumed to be the same as the reference. This included all previously masked repetitive sequences. Although the use of whole genome length sequence data allows for the nucleotide substitution rate to be calculated more accurately, this effected longer branch lengths on the phylogenetic trees but made no material difference to the topology or interpretation of trees for clusters of closely genetically related isolates.

4.3.1 CASE IDENTIFICATION AND SAMPLE SELECTION

I identified all residents of Oxfordshire postcodes with a *M. tuberculosis* culture or a clinical diagnosis of tuberculosis between 1st January 2007 and 31st December 2012 from three sources. Relevant diagnostic codes and microbiology results were identified from the Oxford University Hospitals (OUH) Patient Safety Server, a warehouse of all microbiology tests and OUH admissions. The OUH Trust provides all microbiology laboratory services, over 99% of acute care, and over 90% of specialist services in the county. This was supplemented by a review of all records kept by the Thames Valley Health Protection Unit and local specialist tuberculosis nurses. All identified cases were then checked against the national Enhanced Tuberculosis Surveillance (ETS) database. Additional demographic (age, sex, social risk factors, year of UK entry, country of birth), clinical (pulmonary vs non-pulmonary) and microbiological data (microscopy and culture results) were also derived from ETS.

I sought at least one *M. tuberculosis complex* culture for each patient with microbiologically confirmed disease. Where possible, cultures were obtained from the OUH microbiology laboratory, from a frozen archive if isolated between 2007-2010, and collected prospectively from 2011 onwards. I retrieved cultures referred to other hospitals from the mycobacterial reference laboratories in London, Birmingham, Newcastle or Edinburgh.

4.3.2 EPIDEMIOLOGICAL AND GENOMIC CLUSTER ANALYSIS

The specialist tuberculosis nurses, the lead Infectious Diseases physician for the Oxfordshire tuberculosis service, and the local Consultant for

Communicable Disease Control independently identified epidemiological linkage, defined as shared space and time, blinded to WGS, with discrepancies resolved by consensus. Assessments were based on previous national guideline-directed cluster investigations.[153,255] Genomic clusters were determined independently of the epidemiological data, and were defined where no more than 12 SNPs in isolates separated a patient from at least one other patient in the cluster. I used the upper threshold of 12 SNPs, based on the maximum within and between-host genetic diversity described in chapter 3, to maximise the sensitivity for identifying clusters. I constructed plausible transmission networks for each genomic cluster, and epidemiological cluster with sequence data available, as described in chapter 3.[253] In brief, the first patient diagnosed in each cluster was defined as the index case, with links to subsequent cases assigned first by any epidemiological linkage, then by genetic distance, and then by timing of diagnosis. Hence the total number of links in each cluster is the number of patients in that cluster minus the index case.

I calculated incidence by postcode using denominator data from the Office of National Statistics (ONS). Incidence specific to country of birth was calculated using a regional ONS denominator, as data on countries of birth by postcode were unavailable. Countries of birth with an incidence of 50 cases per 100,000 population per year were classified as high incidence, and those below this threshold were classified as low incidence.[256] I assessed the association between high incidence versus low incidence country of birth and disease characteristics or clustering by epidemiology or genomics using logistic regression in STATA version 13.1 (StataCorp, 2013). All analyses were

adjusted for age and gender, because other factors had moderate amounts of missing data.

4.4 RESULTS

390 Oxfordshire residents had a *M. tuberculosis* culture or clinical tuberculosis diagnosis. Six were excluded as their isolates were considered laboratory contaminants, leaving 384 cases. 269 had culture positive disease, 112 culture negative disease, and the status of three patients diagnosed overseas could not be ascertained. In total 22 of 269 (8%) cultures did not grow, could not be retrieved, or failed WGS quality control, leaving 247 patients with sequence data available (Figure 16).

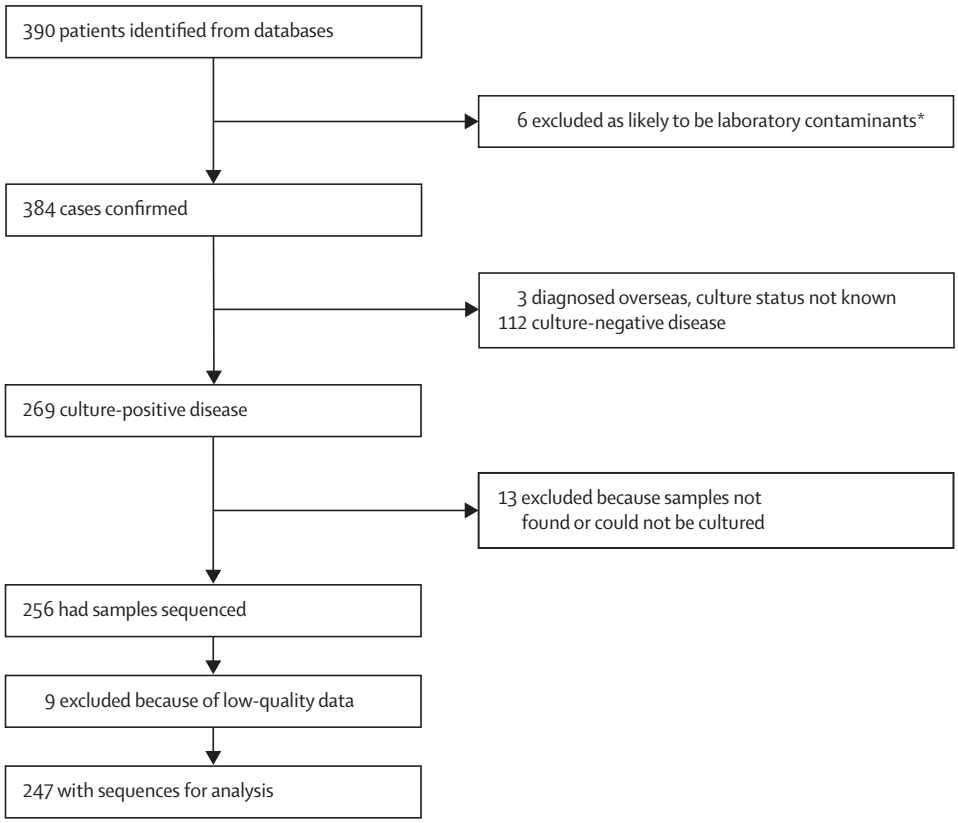


Figure 16: Flow chart of sample selection. *Three laboratory contaminants were identified previously and three by use of whole-genome sequencing.

4.4.1 EPIDEMIOLOGICAL ANALYSIS

Median patient age was 34 years (range 1-89), with 255 (67%) born in a high-incidence country, 103 (27%) in the UK and 22 (6%) in another low-incidence country (Figure 17). The place of birth was not known for four patients, including one with culture-positive disease. For non-UK born patients, a median five years (IQR 2-9) had elapsed from UK entry to tuberculosis diagnosis (Figure 18). The 6% of Oxfordshire's population born in a high-incidence country had an annual tuberculosis incidence of 109 cases per 100,000 persons compared to 3.5 per 100,000 persons for those born in a low-incidence country. Among these the incidence was 3.0 per 100,000 if UK-born, and 7.2 per 100,000 persons for those born in a low-incidence country other than the UK. Three postcodes (OX3 [n=43000], OX4 [n=62000] and OX16 [n=47000]) accounted for 20% of the population and 233 (61%) of 383 cases for whom the postcode was known (Figure 19). 178 (77%) of 232 cases residing in these postcode areas were born in high-incidence countries, versus 77 (52%) of 148 who were living elsewhere (p<0.0001). The country of origin was unknown for one patient in the three higher incidence postcodes.

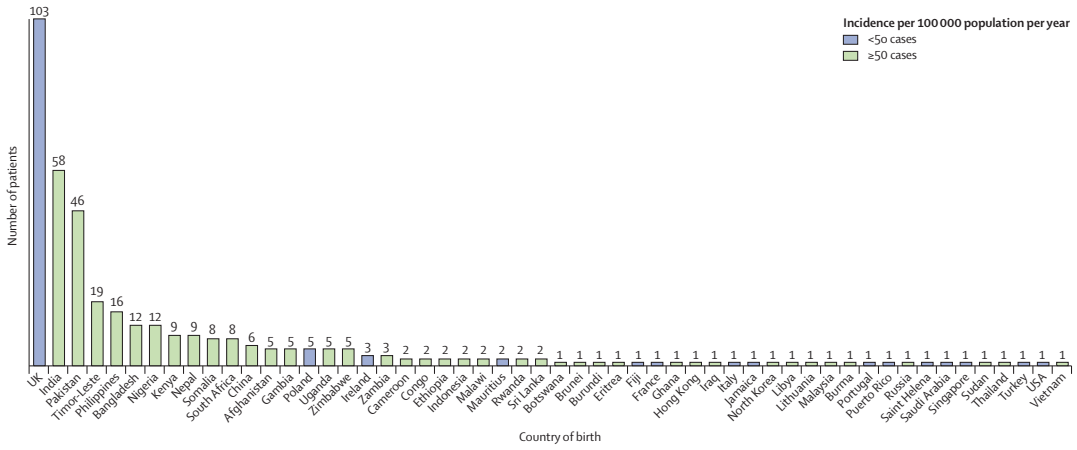


Figure 17: Country of birth of patients with tuberculosis in Oxfordshire, 2007-2012. Country of birth was not known for four patients. High and low incidences were defined according to the WHO.[256]

197 (52%) of 380 evaluable patients had pulmonary disease (the disease site was unknown for 4 patients), and those born in a low-incidence country were more likely to have pulmonary disease (odds ratio, OR=1.8; 95% confidence interval, CI 1.2-2.9; p=0·009), but possibly less likely to have culture positive disease (OR=0.6; 0.4-0.99; p=0.045). Social risk factors (alcohol or drug misuse, homelessness or time served in prison) were present in 36 (14%) of 261 evaluable patients, and were also more prevalent in those born in low-incidence countries (OR=4.4; 2.0-9.4; p<0.0001). There was no difference in the proportion of patients with available data on social risk factors from high and low-incidence countries of birth (87 (70%) of 125 cases and 174 (68%) of 255 cases, respectively; p=0.81) (Table 2). There were no significant differences in the odds of pulmonary disease, social risk factors or epidemiological linkage between UK-born patients and those born in other low-incidence countries (p>0.4) (Table 3), nor did their age differ (ranksum p=0.99) (Figure 20).

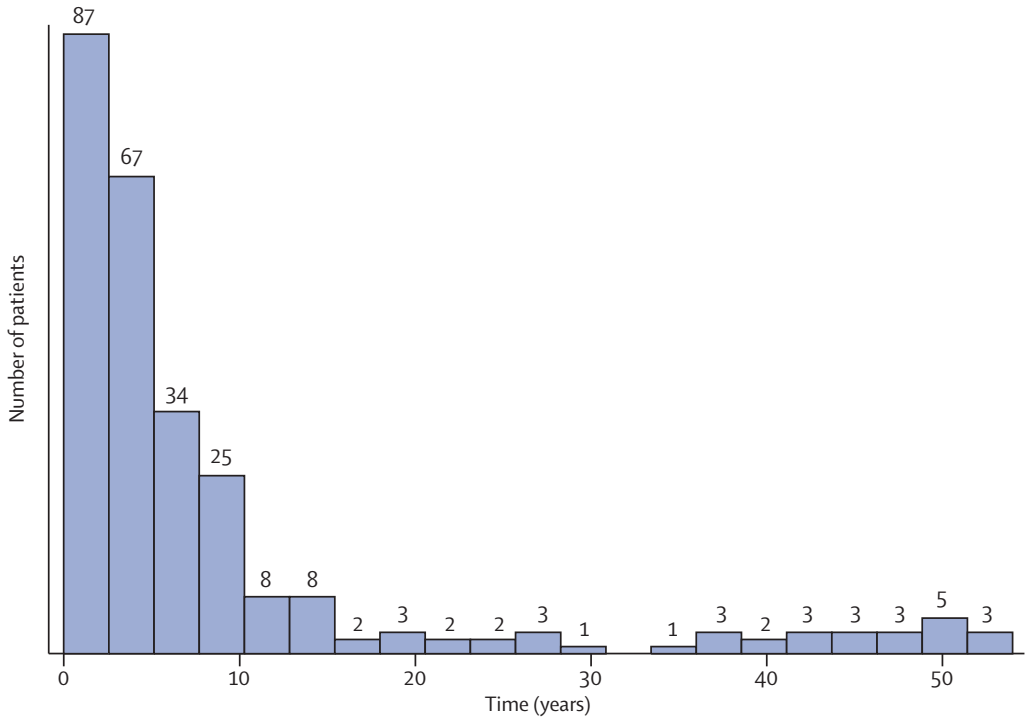


Figure 18: Time interval between entry to the UK and diagnosis of tuberculosis. Data for year of entry to the UK were not available for 12 patients.

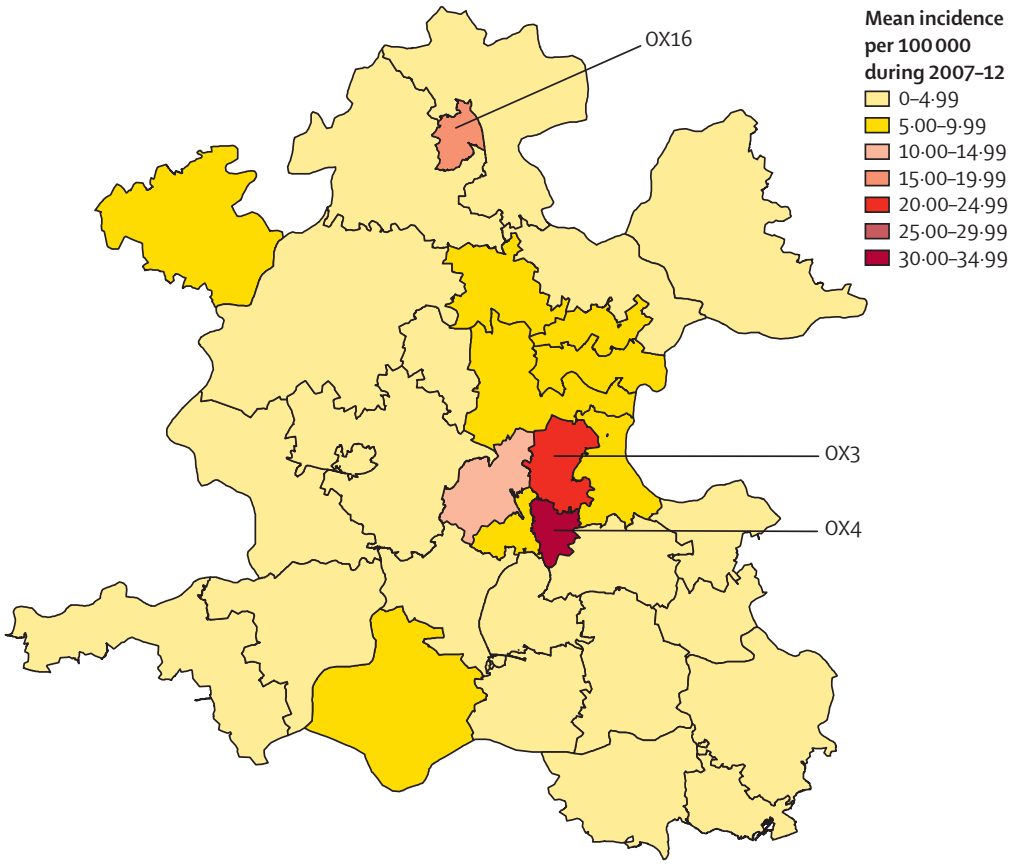


Figure 19: Mean tuberculosis incidence in Oxfordshire (2007-2012). Map based on 383 of 384 cases. The postcode for one patient was unknown. Crown copyright and database rights 2013 Ordnance Survey 100016969.

Epidemiological investigations had identified 18 clusters (E1-E18) with 46 patients, accounting for 28 potential transmission events within Oxfordshire over six years (Figure 21). All but two epidemiological links were between family members, with the remaining hypothesised transmissions occurring in a school (E8) and in the community (E10). Although MIRU-VNTR typing was introduced in 2010, it did not result in any additional epidemiological clusters being identified. Although 10 of 18 epidemiologically defined clusters involved patients born in high-incidence countries, cases born in low-incidence countries were more likely to be identified as part of an epidemiological cluster (OR=3.3; 95%

CI 1.4-7.8; p=0.006), independently of potentially confounding social risk factors (adjusted OR=3.0; 95% CI 1.2-7.2; p=0.016) (Table 2).

Children (aged under 18) were more likely to be born in a low-incidence country (p=0.001), and as expected, were more likely to have culture negative disease (p=0.003), and to be epidemiologically linked to (household or school) cluster (p<0.0001), although six of 13 UK-born patients younger than 10 years old were not epidemiologically linked to another case.

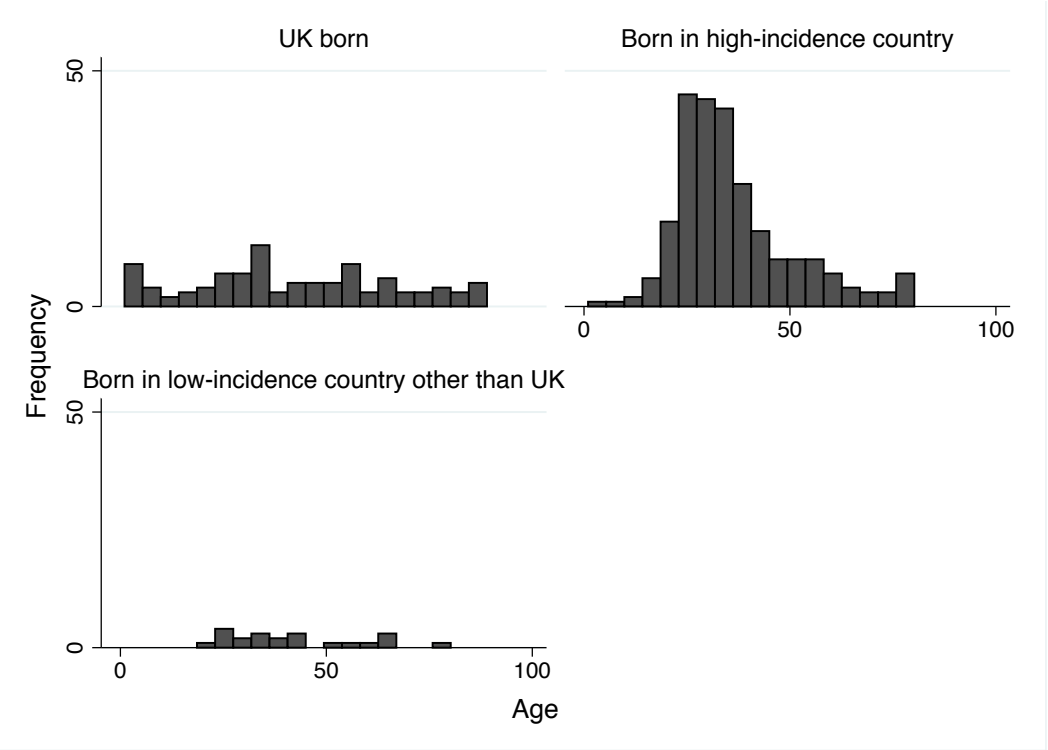


Figure 20: Age distribution of cases according to incidence in country of birth

	Patients with data available	Patients born in low-incidence countries	Patients born in high-incidence countries	Odds ratio* (95% CI)	p value
Pulmonary disease	380 (99%)	78/125 (62%)	119/254 (47%)	1·8 (1·2–2·9)	0·009
Social risk factor	261 (68%)	23/87 (26%)	13/174 (7%)	4·4 (2·0–9·4)	<0·0001
Culture positive disease	377 (98%)	81/125 (65%)	186/252 (74%)	0·6 (0·4–0·99)	0·045
Paediatric disease (age <18 years)	384 (100%)	16/125 (13%)	8/255 (3%)	4·8 (2·0–11·5)	0·001
Epidemiological cluster					
All evaluable patients	384 (100%)	25/125 (20%)	21/255 (8%)	3·3 (1·7–6·3)	<0·0001
Social risk factor data available (not adjusted for social risk factors)	261 (68%)	14/87 (16%)	11/174 (6%)	3·3 (1·4–7·8)	0·006
Social risk factor data available (adjusted for social risk factors)	261 (68%)	14/87 (16%)	11/174 (6%)	3·0 (1·2–7·2)	0·016
Whole-genome-sequencing cluster					
All evaluable patients	247 (64%)	24/74 (32%)	15/172 (9%)	5·8 (2·7–12·4)	<0·0001
Social risk factor data available (not adjusted for social risk factors)	164 (43%)	14/53 (26%)	8/117 (7%)	6·4 (2·2–18·8)	0·001
Social risk factor data available (adjusted for social risk factors)	164 (43%)	14/53 (26%)	8/117 (7%)	4·8 (1·6–14·8)	0·006

Table 2: Associations between country of birth and tuberculosis characteristics and epidemiological or genomic clustering. Data are number (%) or n/N (%), unless otherwise indicated. Denominators were the numbers of patients for whom data were available for the variables that were being compared. *For low-incidence countries versus high-incidence countries of birth and calculated with multivariable logistic regression, adjusted for age and sex (just sex for children), and for social risk factors where indicated. Social risk factors are at least one of the following: homelessness, drug or alcohol misuse, or time spent in prison.

4.4.2 GENOMIC ANALYSIS

Assessing pairwise nucleotide differences across the 247 patients with culture confirmed disease and whole-genome sequences, 39 patient isolates were within 12 SNPs of another isolate, forming 13 genomic clusters (G1-G13) with 26 plausible transmission events (Figure 21). The remaining 208 (84%) patients could not be genomically linked to another within the 6-year study. Patients born in low-incidence countries were more likely to be genomically linked to another case (OR 5.8; 95% CI 2.7-12.4; p<0.0001), even after adjustment for social risk factors (OR 4.8; 95% CI 1.6-14.8; p=0.006) (Table 2). For patients born in low-incidence countries, estimates suggested that UK-born patients were more likely to be genomically linked to a cluster but numbers were too few to exclude this finding being compatible with chance effects (OR 2.0; 95% CI 0.3-12.4; p=0.45) (Table 3).

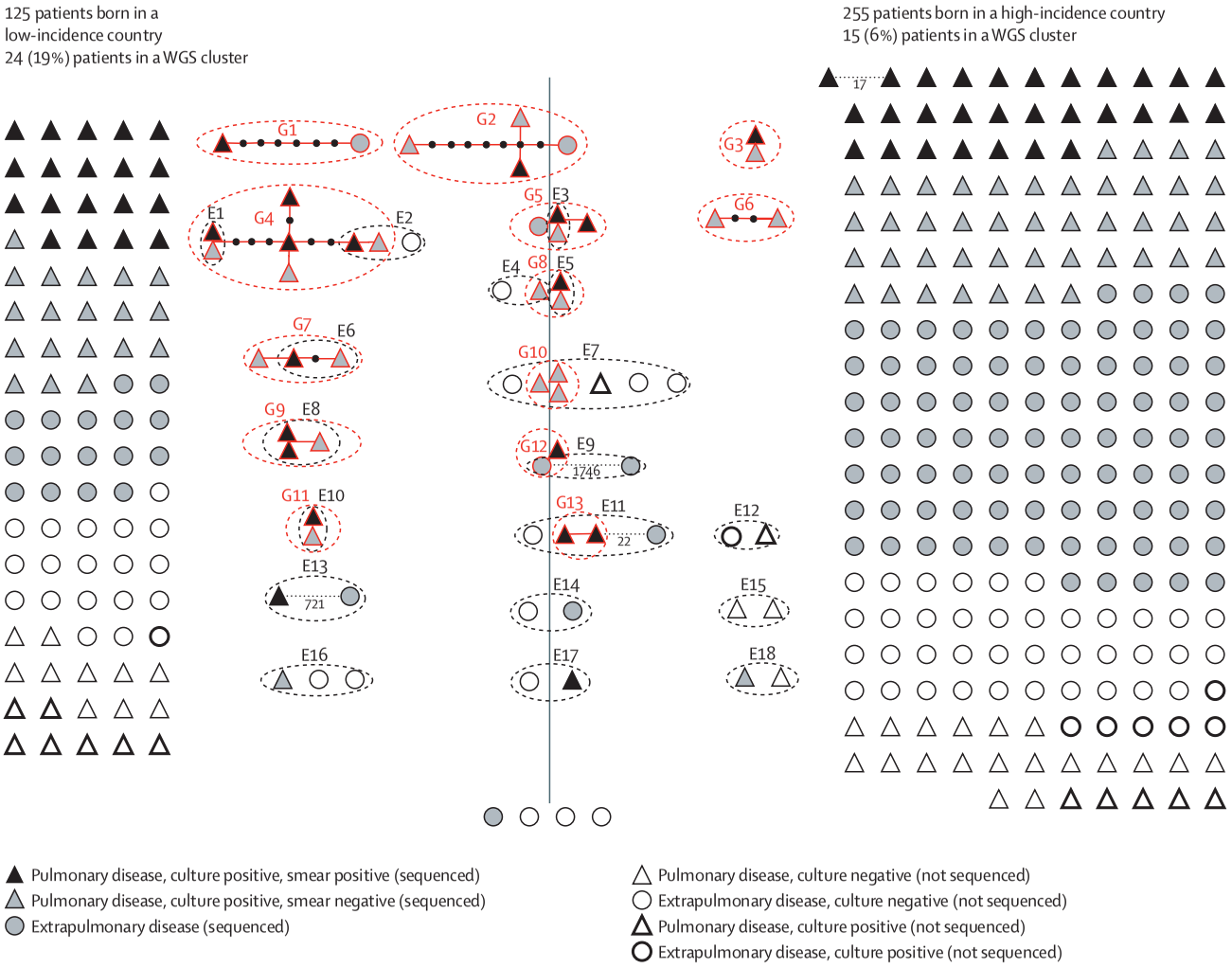

125 patients born in a
low-incidence country
24 (19%) patients in a WGS cluster
255 patients born in a high-incidence country
15 (6%) patients in a WGS cluster
G1
G2
G3
G4
G5
G6
G7
G8
G9
G10
G11
G12
G13
E1
E2
E3
E4
E5
E6
E7
E8
E9
E10
E11
E12
E13
E14
E15
E16
E17
E18
17
1746
22
721
Pulmonary disease, culture positive, smear positive (sequenced)
Pulmonary disease, culture positive, smear negative (sequenced)
Extrapulmonary disease (sequenced)
Pulmonary disease, culture negative (not sequenced)
Extrapulmonary disease, culture negative (not sequenced)
Pulmonary disease, culture positive (not sequenced)
Extrapulmonary disease, culture positive (not sequenced)

Figure 21: All cases in Oxfordshire, 2007-2012, by incidence in country of birth, and by epidemiological and genomic clustering. Patients born in low-incidence countries are on the left and those born in high-incidence countries are on the right of the figure. Four patients whose country of birth was not known are at the bottom of the figure. Each shape (triangle or circle) represents a patient. Epidemiological clusters (E1-18) are circled in black and genetic links, shown as trees with edges representing the genetic distance, are circled in red. Edges in networks are red for distances within 12 SNPs. Genetic links of interest but greater than 12 SNPs are indicated by black dashed lines, representing the SNP distances. Patients in WGS clusters who are zero SNPs apart are indicated by shapes that abut each other, whereas distances of at least 1 SNP are quantified by the number of red lines (separated by small black nodes if >1 SNP) between patients. Epidemiological or WGS clusters that include patients born in low-incidence countries and patients born in high-incidence countries cross the central vertical line.

Observed differences within genomic clusters ranged from 0-7 SNPs (median 1 SNP; interquartile range, IQR 0-2), despite a pre-defined upper limit of 12 SNPs (Figure 22). Excluding secondary cases from each genomic cluster, the median pairwise SNP difference between cases in Oxfordshire was 1106 (IQR 857-1715). No cluster was within 180 SNPs of another (Figure 23).

Outcome	N (%) with data available	Outcome in patients born in the UK	Outcome in patients born in other low-incidence countries	Odds ratio (UK-born vs other low-incidence country of birth)	95% CI	p-value
Pulmonary disease	125 (100%)	63/103 (61%)	15/22 (68%)	0.8	0·3-2·2	0·66
Social risk factor	87 (70%)	19/70 (27%)	4/17 (24%)	1·5	0·4-5·2	0·55
Culture positive disease	125 (100%)	64/103 (62%)	17/22 (77%)	0·5	0·2-1·6	0·26
Paediatric disease (age<18)	125 (100%)	16/103 (16%)	0/22 (0%)			
Epidemiological cluster	125 (100%)	22/103 (21%)	3/22 (14%)	1·2	0·3-5·0	0·76
Epidemiological cluster if data on social risk available (not adjusted for social risk)	87 (70%)	11/70 (16%)	3/17 (18%)	0·6	0·1-2·8	0·54
Adjusted for social risk	87 (70%)	11/70 (16%)	3/17 (18%)	0·55	0·1-2·6	0·45
WGS cluster	74 (59%)	22/58 (38%)	2/16 (12%)	6·6	1·3-34·4	0·033
WGS cluster if data on social risk available (not adjusted for social risk)	52 (42%)	12/39 (31%)	2/13 (15%)	2·9	0·5-16·0	0·22
Adjusted for social risk	52 (42%)	12/39 (31%)	2/13 (15%)	2·0	0·3-12·4	0·45

Note: odds ratios based on multi-variable logistic regression, adjusted for age and gender, and also for social risk factors where indicated.

Table 3: Associations between UK and other low incidence countries of birth, and disease characteristics and epidemiological or genomic clustering.

Within these 13 genomically defined clusters, 11 of 26 transmission events had been previously identified by epidemiological investigation, with none exceeding 2 SNPs. Nine of these 11 transmission events were within a household, including four between family members born in high-incidence countries (G5/E3, G8/E5, G10/E7, and G13/E11) and one between one family member born in a high-incidence country and another in the UK (G10/E7) (Figure 21). The two non-household cases identified by both epidemiology and sequencing were linked within a school (G9/E8) and in the community (G11/E10). In the retrospective review of the 15 epidemiologically unpredicted links, three were associated with the same homeless shelter (G4), one was related to time spent in the same prison (G4), and two had nearby addresses and shared cultural backgrounds (G3, G6) (Figure 21). No retrospective explanation could be proposed for the remaining nine links, including four between patients born in high and low-incidence countries (Figure 21). In all but one case (G6, two patients with smear-negative pulmonary disease), the clusters containing these possible, but epidemiologically unconfirmed transmissions involved at least one patient with smear positive pulmonary tuberculosis. Interestingly, of the seven clusters that contained household based transmissions, five also contained genomic links to non-household members not identified on contact tracing (Figure 21; Table 4).

I observed 17 epidemiologically identified but genomically unconfirmed transmission events. Of these three were genomically unrelated (22, 721 and 1746 SNPs), 12 could not be assessed due to culture negative disease and two because of sample preparation problems (Figure 21). The patient who was genomically separated from family members by 22 SNPs migrated to the UK

four years after the other cases were diagnosed, making direct transmission very unlikely. However, a distance of 22 SNPs is consistent with a dominant circulating clone in the family's region of origin as a common source.[253] A similar explanation may apply to two patients separated by 17 SNPs, born in different countries in east Africa but not epidemiologically linked (Figure 21).

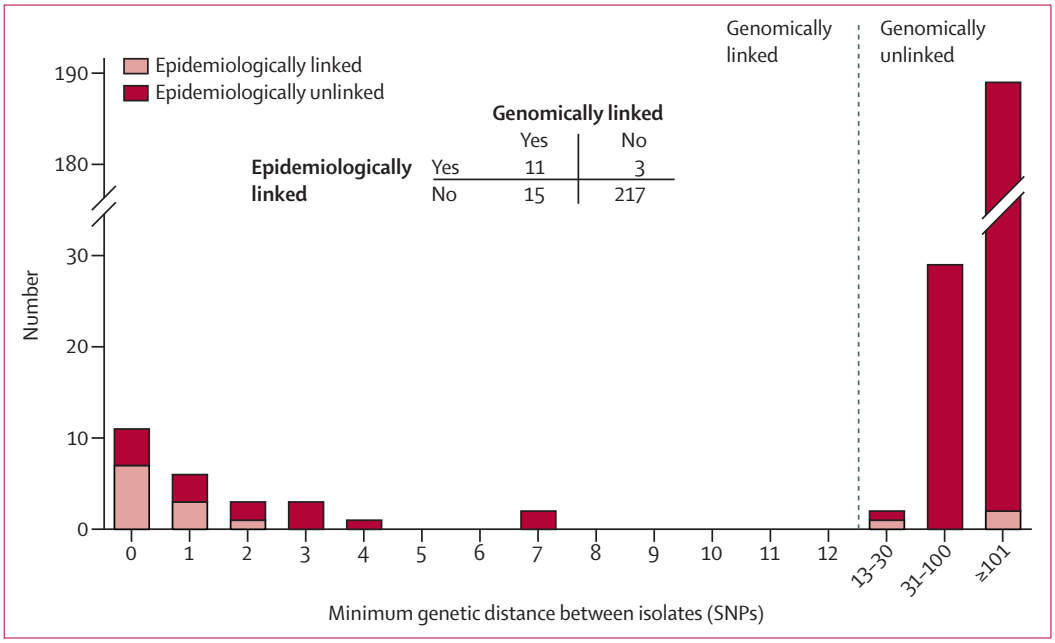


Figure 22: Minimum genetic distance between isolates, by epidemiological link.

Considering WGS with a 12 SNP threshold as the gold standard, epidemiological investigation had a sensitivity of 0.42 (95% CI 0.23-0.63) and specificity of 0.99 (0.96-1.0) for detecting transmissions. The sensitivity of epidemiological investigation was 0.46 (0.25-0.67) applying a stricter relatedness threshold of 5 SNPs (specificity 0.99 (0.96-1.0)), and 0.59 (0.33-0.82) with a 1 SNP threshold (specificity 0.98 (0.96-0.99)).

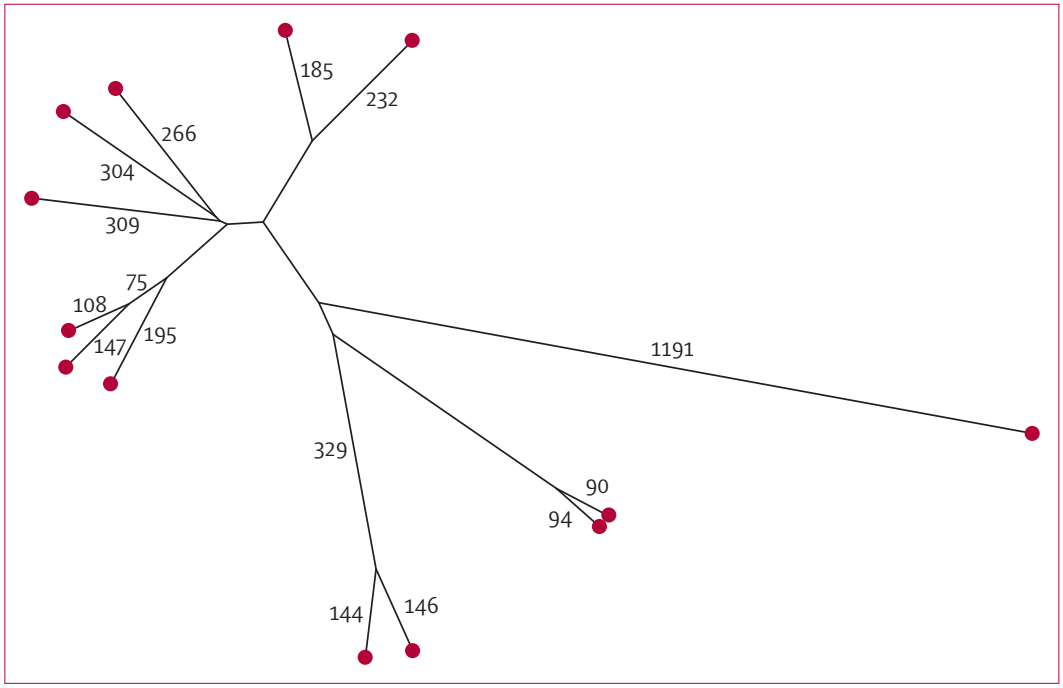


Figure 23: Phylogenetic relations between WGS clusters. Maximum-likelihood tree of 13 clusters as ascertained with whole genome sequencing are represented by red circles. SNP distances are annotated on the branches.

All Oxfordshire isolates were compared to sequences from 254 patients within epidemiologically identified clusters in the Midlands detailed in chapter 3. One Oxfordshire patient with *M. bovis* was within 2 SNPs of the nearest patient in a Midlands *M. bovis* outbreak, and one cluster (G2) of four Oxfordshire patients was within 9 SNPs of a Midlands *M. tuberculosis* cluster (clusters 11 and 5 respectively). No epidemiological links spanning these geographical boundaries were previously suspected in either case, although patients in the latter example shared social risk factors (alcohol excess).

WGS Cluster	Reconstructed patient links	SNPs	Epidemiologically linked (known to have shared time and space)	Epidemiological connection (either confirmed or hypothesised)
G 1	P347 - P136	7	No	
G 2	P44 - P33	2	No	
G 2	P33 - P119	3	No	
G 2	P119 - P249	7	No	
G 3	P406 - P460	0	No	Same country of birth / neighbourhood
G 4	P139 - P343	1	Yes	Household / family
G 4	P428 - P446	1	No	Night shelter
G 4	P376 - P428	4	No	Night shelter
G 4	P428 - P445	2	No	Night shelter
G 4	P428 - P139	3	No	
G 4	P376 - P410	0	Yes	Household / family
G 5	P386 - P387	0	Yes	Household / family
G 5	P387 - P64	0	No	
G 5	P167 - P386	1	No	
G 6	P49 - P104	3	No	Same country of birth / neighbourhood
G 7	P330 - P479	1	No	Prison
G 7	P330 - P331	2	Yes	Household / family
G 8	P486 - P96	0	Yes	Household / family
G 8	P96 - P175	0	No	
G 9	P84 - P184	0	Yes	Household / family
G 9	P84 - P27	1	Yes	School
G 10	P22 - P202	0	Yes	Household / family
G 10	P202 - P372	0	Yes	Household / family
G 11	P165 - P469	0	Yes	Social connection
G 12	P164 - P374	0	No	
G 13	P52 - P43	1	Yes	Household / family

Table 4: Links between patients are shown by WGS cluster (G1-13), together with SNP distances and results of epidemiological investigations. The number of links in each cluster is equal to the number of patients, minus one. Epidemiological connectedness was defined as shared time and space. The nature of the confirmed (known to have shared time and space) or hypothesised (not known to have shared time and space) epidemiological connection is listed.

4.5 DISCUSSION

I identified 384 patients with tuberculosis in Oxfordshire over the six years, (2007-2012) of whom 269 had microbiologically-confirmed disease and 247 could be sequenced. Of a maximum possible 246 genomically defined links between patient isolates, I found 26 (11%) that were within the defined threshold of 12 SNPs. Although more patients with tuberculosis were born in high-incidence than in low-incidence countries, it was those born in low-incidence countries, predominantly the UK, who were more likely to be part of a genomically defined cluster.

It has been previously well documented that the increase in tuberculosis incidence in the UK over the last decade is due to an increase in the number of patients born outside of the UK, largely from high-incidence countries.[216] This phenomenon is not peculiar to the UK as patients born in high-incidence countries also account for over half the cases in other western European countries.[257,258] The extent to which immigrant patients contribute to local transmission has consequently been explored in a variety of settings, with some studies demonstrating a negligible contribution and others a more substantial contribution towards local transmission (Table 5 presents a summary).[259]

Setting	Years of survey	Molecular method(s)	Major conclusion(s)	Reference
New York City, USA	1989–1992	IS*6110*-RFLP	Imported TB cases are not a major cause of recent transmission	Alland et al. (1994)
San Francisco, USA	1991–1995	IS*6110*-RFLP PGRS-RFLP	Very little disease resulted from TB transmission between the US- and foreign-born populations	Chin et al. (1998)
San Francisco, USA	1991–1995	IS*6110*-RFLP PGRS-RFLP	Transmission of TB from Mexican-born persons to other persons is uncommon	Jasmer et al. (1997)
San Francisco, USA	1991–1999	IS*6110*-RFLP LSP	The association between host's region of origin and *M. tuberculosis* strain is strong. Very limited transmission between ethnically defined, foreign-born population and USA-born individuals	Hirsh et al. (2004)
Denmark	1992–1999	IS*6110*-RFLP Spoligotyping	*M. tuberculosis* transmission between a Somalis cohort from a highly-prevalence area and Danes population is nearly nonexistent	Lillebaek et al. (2001)
Norway	1999–2001	Spoligotyping IS*6110*-RFLP	Most cases of TB are due to the import of new strains rather than to transmission within the country. The high incidence of TB among immigrants does not present a threat for the native population.	Dahle et al. (2003)
Norway	1994–2005	IS*6110*-RFLP	Twelve years of MTB importation as a result of immigration from high-incidence countries had little influence on TB transmission in the receiving low-incidence country	Dahle et al. (2007)
Germany	2003–2005	Spoligotyping IS*6110*-RFLP	No significant TB transmission from TB high-prevalence immigrant to TB low-prevalence autochthonous population	Barniol et al. (2009)
Western Sweden	1999–2002	Spoligotyping VNTR	TB transmission was not necessarily associated with the country of origin of immigrants	Brudey et al. (2004)
Hamburg, Germany	1997–2002	IS*6110*-RFLP	Recent transmission from foreigners resulted in 2.8% secondary cases in German-born individuals	Diel et al. (2004)
London, UK	1995–1997	IS*6110*-RFLP	Relatively little transmission of TB from immigrants to endogenous population	Dale et al. (2005)
The Netherlands	1993–1995	IS*6110*-RFLP	17% (95% confidence interval 9–25%) of Dutch TB cases were attributable to recent transmission from a non-Dutch source	Borgdorff et al. (1998)
Tuscany, Italy	2002–2005	Spoligotyping	12.9% of TB cases in Italian-born patients attributed to transmission from foreign-born immigrants	Garzelli et al. (2010)
Tuscany, Italy	2002–2004	Spoligotyping IS*6110*-RFLP	Frequent transmission of Beijing strains to autochthonous population	Lari et al. (2007)
The Netherlands	1993–2007	IS*6110*-RFLP	Increase in the proportion of native patients with TB attributable to recent transmission with a foreign-born index case	Borgdorff et al. (2010)
Rhode Island, USA	1995–2004	Spoligotyping VNTR-MIRU	80% of the mixed clusters of foreign- and US-born persons arose from a foreign-born source case	Vanhomwegen et al. (2011)
Madrid, Spain	2002–2004	IS*6110*-RFLP	Extensive transmission between Spanish- and foreign-born populations, caused mainly by autochthonous strains	Iñigo et al. (2007)
Almeria, Spain	2003–2006	IS*6110*-RFLP	Transmission between nationalities and between the immigrant and autochthonous populations	Martinez-Lirola et al. (2008)
Madrid, Spain	2004–2006	Spoligotyping and IS*6110*-RFLP	Marked transmission permeability among nationalities and between the immigrant and the autochthonous populations	Alonso-Rodriguez et al. (2009)
Barcelona, Spain	2003–2004	IS*6110*-RFLP and MIRU-12	Mixed clusters, reflecting transmission between Spanish- and foreign-born patients in both directions, represented 33.8% of the total clusters	Borrell et al. (2010)

Studies are listed in the order in which they are quoted in Section 3.2.

Table 5: Summary of studies of local transmission and migrant populations, taken from Garzelli C *et. al.* Infect Genet Evol 2012;12:610–8. [259]

At the one extreme Dahle *et. al.* used RFLP typing to cluster 12 years' worth of *M. tuberculosis* cultures from Norway, demonstrating that, although immigrants accounted for 69% of cases, immigration was not associated with increased transmission of disease.[260] At the other extreme, Borgdorff *et. al.*, using RFLP typing in the Netherlands, found that 17% of infections in patients of Dutch nationality over a two and a half year period in the mid-1990s were secondary to patients with non-Dutch nationality.[261] Such apparent variation in results may reflect genuine differences in transmission patterns across countries, but it might also reflect differences in analytical approaches. For example, the dichotomy drawn by Borgdorff *et. al.* was based upon nationality, whereas that of Dahle *et. al.* was based upon country of birth (Norway versus not Norway). In the Norwegian study it would have been important to quantify the number of patients born in Norway as second-generation immigrants as these patients may have a higher risk of disease, depending on the incidence in the parents' country of origin. In the Dutch study one would want to know the number of patients with Dutch nationality who are first or second-generation immigrants, as these patients might also be at higher risk than patients without a heritage link to a high-incidence country. The historical pattern of immigration is also significant. A long established and sizable immigrant community such as those of Surinamese heritage in the Netherlands might have resulted in more transmission between Dutch and non-Dutch citizens (e.g. via visiting family members, friends, newer migrants from the same community),[261] whereas the majority of patients from the more recently established Somali community in Norway might have been more homogenously categorised as non-Norwegian born.[260]

How patients are classified into meaningful groups was therefore an important consideration for my analysis. The motivation behind the study was to gain a better understanding of local epidemiology with a view to informing decision making on appropriate interventions. Because German and UK born patients would be expected to share the same risk; because Indian born British citizens would have a higher risk than UK born British citizens; and because ethnically 'black' patients from South Africa and Jamaica would have very different risks, I used the incidence in the country of birth for each patient as a proxy for the risk of developing tuberculosis instead of making a distinction based on UK birth, nationality, or ethnicity. Using the country of birth of a patient's parent might have overcome the problem of first-generation immigrant transmitting to second-generation family, but this data was not available. Nevertheless, this did not detract from the clear findings that birth in a high-incidence country was a relative risk factor for disease, but not transmission, and that control measures can be designed around this result. Better new-entrant screening would be one example.

I chose a threshold of 12 SNPs to identify plausible transmissions from pairwise comparison of sequences to reflect the maximum genetic diversity observed within hosts and between epidemiologically related hosts in the previous chapter. I could have defined a less generous SNP threshold for identifying clusters, but this would have made very little difference to the conclusions as a bimodal picture emerged with 24 genomic distances spanning 0-4 SNPs, 221 spanning over 12 SNPs, and only one spanning between 5 and 12 SNPs. This was entirely in keeping with my own and other previous observations that most patients in a transmission chain are within 5 SNPs of another patient.[236-

238,253] Reducing the relatedness threshold to 5 SNPs, and then to 1 SNP, had only a marginal impact on the sensitivity of the epidemiological investigations, which increased incrementally from 0.42 to 0.46, and to 0.59. A prospective study will help clarify the extent to which the sensitivity can be increased if the public health teams conduct their investigations in the knowledge that close genomically based links have been shown.

In this study, 67% of patients were born in a high-incidence country. These patients were less likely than those born in low-incidence countries to have pulmonary disease, and the genomic analysis showed that they were also more likely to be independent of other cases. Comparable previous studies of cities or regions that used genotyping, either RFLP or MIRU-VNTR, in combination with data from contact investigations, also observed this phenomenon, although the overall rate of transmission events was estimated to be higher. Peter Small *et. al.* used RFLP typing to quantify transmission in San Francisco between 1991 and 1992.[262] Of 473 patients, they found that 191 were linked to one of 44 different clusters by RFLP-typing, and concluded that at least 31% of their incidence was secondary to local transmission. However, only 10% of patients were epidemiologically linked to clusters, something the authors ascribe to the high proportion of homeless patients in their study. However, this may also be because of false-positive clustering by RFLP-typing. Caroline Allix-Béguec *et. al.* used 24-locus MIRU-VNTR-typing to study transmission in Brussels between 2002 and 2005.[214] Of 802 identified patients, 96 MIRU-VNTR-based clusters were defined that accounted for a total of 288 patients, giving a 23.9% clustering rate over the time period. Of these molecular links, 32% were epidemiologically confirmed. Andres Roetzer *et. al.* performed a similar study in

Schleswig-Holzstein, Northern Germany, between 2006-2010.[252] Here they calculated a clustering rate of 20.5%, after 75 of 277 isolates were assigned to one of 18 24-locus MIRU-VNTR-based clusters. An epidemiological link was described for 41 of 75 clustered cases (55%). Each of these three localities has its particular features, but, like Oxfordshire, they are all low-incidence regions within high-income countries. Although the clustering rates of 31%, 23.9% and 20.5% are all higher than the 11% observed in Oxfordshire, the percentage of links based on molecular typing that were epidemiologically confirmed was 10%, 31% and 55% respectively, lower than the maximum of 59% for genomically-defined links within a 1 SNP threshold in my study. Given the greater resolution of WGS over the other typing methods, as demonstrated in the previous chapter, these results are consistent with the expectation that WGS would result in fewer clusters being defined, but that a greater proportion of these would be epidemiologically corroborated.

The relatively low rate of transmission among patients born in high-incidence countries implies that most of these patients reactivate LTBI acquired elsewhere – possibly in their country of origin. The observation that most immigrant cases are diagnosed within the first years after arrival is a well recognised pattern and, to that extent, supports this hypothesis.[263] Although 53% of patients born in high-incidence countries had extra-pulmonary disease and will not have been infectious, unrestricted access to diagnostic and treatment services through the National Health Service (NHS) is nevertheless likely to have contributed to the relative lack of secondary cases linked to those patients with open pulmonary tuberculosis. Appropriate interventions to reduce

the overall incidence of tuberculosis might therefore best be focussed on new-entrant screening.

Until 2014, new-entrant screening in the UK had focussed on the detection of active disease at ports of entry in individuals holding visas for 6 months or longer, and from countries with an incidence of at least 40 cases per 100,000 population.[166] This has been widely criticised as an approach as it is insensitive to early infection, is not cost effective and is not practiced consistently at all borders.[264] The 2006 NICE guidance is currently being updated but recommends that screening for LTBI is performed at a local level for children from countries with an incidence of over 40 cases per 100,000 population, and for younger adults (aged 16 to 35) from countries with an incidence in excess of 500 cases per 100,000.[165] It has been argued that the recommended incidence thresholds are too high to identify latent infection in patients from south Asia who eventually account for the majority of cases of migration-associated tuberculosis.[265] Adherence to the guidelines has also been found to be both poor and inversely proportional to the incidence in UK regions.[166]

The system of screening at ports of entry has now been replaced by a programme of pre-entrant screening to identify active disease as part of the process of applying for residency, or for a minimum six-month visa, from any country with an incidence of over 40 cases per 100,000 population. Screening is for active disease only, and is based upon chest radiography and sputum smear and culture.[266] Pareek *et. al.* have shown that IGRA-based screening can be cost effective if implemented for migrants from countries with an

incidence of as low as 150 cases per 100,000 population, which would capture up to 92% of latently infected migrants to the UK.[165] Were screening to be conducted at this threshold for migrants as they register with a primary care physician it might not only result in fewer cases of reactivated tuberculosis but might also serve to promote consciousness of tuberculosis among GPs who might otherwise see very few cases in their practice. That in turn could result in improved rates of diagnosis where active disease does arise.[267]

Despite screening for new entrants, there remain a large number of tuberculosis cases among migrants to the UK. Unfettered access to healthcare in the early post-migration period may well have contributed to the comparatively little transmission that resulted from these cases. Although healthcare is freely available to all UK residents, the services appear to have been less effective in controlling disease in patients born in low-incidence countries, mainly the UK. Possible explanations might be that the excess of social risk factors in these patients led to poorer healthcare seeking behaviour, or that healthcare professionals may also investigate alternative diagnoses before considering a diagnosis of tuberculosis in this population. Both phenomena could lead to increased periods of infectivity and hence greater onward transmission.

One attempt to improve diagnostics for hard-to-reach groups (e.g. the homeless, prisoners and those accessing drug treatment services) has been the ‘Find and Treat’ initiative in London, whereby a mobile digital x-ray unit screens these vulnerable populations in their places of aggregation (hostels, prisons etc.).[268] This approach has an estimated sensitivity of 82%, a specificity of 99%, and an impressive treatment completion rate of 84%, and

manages to screen up to 10,000 individuals a year.[135,268] If the pattern of transmission observed in Oxfordshire reflects those elsewhere in the UK, it might be worth considering introducing similar services outside of London as well.

There are several limitations to this study. Like all typing methods, WGS cannot determine the source of culture-negative cases. Similarly, I was unable to assess the amount of transmission leading to latent tuberculosis, as data on interferon-γ release assay results could not be linked back to specific contact tracing investigations with confidence. However, 45 (73%) of 62 patients with sputum smear-positive disease could not be genomically linked to any other case of active disease in Oxfordshire between 2007-2012, which supports the intervention programme being relatively successful. MIRU-VNTR typing was only introduced routinely in the UK in 2010, and between 2010-12 MIRU-VNTR cluster investigations were recommended if clusters reached a defined threshold size or contained cases with defined risk factors. In the study population, no additional epidemiological links were identified when cluster investigation was conducted using this approach. Because the superior resolution of WGS has already been demonstrated,[141,250,253] I did not attempt a further comparison. Finally, were the UK-based social networks of recent migrants to span larger geographical distances than those of long-term residents, then recent migrants might more frequently be linked to trans-regional rather than regional outbreaks. However, the two genomic links that were made to the Midlands involved patients born in low-incidence, not high-incidence, countries.

This study also highlights several advantages to the future use of WGS. I identified 15 plausible but previously unrecognised transmission events within the low-incidence setting of Oxfordshire. Interestingly several of these additional transmissions were from epidemiologically identified household outbreaks to other non-household members. Had these links been identified in near-to-real-time, more intensive investigation might have found other important routes of transmission, possibly resulting in public health action.[269] By genomically linking patients in Oxfordshire and the Midlands I also demonstrate the potential for identifying previously unrecognised transmission across public health regions. The potential for identifying further links is restricted only by the size of the database for comparison, and not by geographical boundaries, so this technique could be applied to extend future contact investigations across larger regions. The consequence would have to be either greater co-ordination between public health teams that currently operate within their own Health Protection Unit boundaries, or a re-structuring of the way contact investigations are performed, by better integrating supra-regional surveillance systems with local 'shoe-leather' epidemiology.

In summary, in this chapter I have demonstrated how WGS can quantify *M. tuberculosis* transmission in an unselected, geographically restricted population over time. In this low-incidence setting, I show that the burden of disease is largely sustained by cases infected either outside of the county or the period of study, and that onward transmission within the region is associated with birth in a low-incidence rather than a high-incidence setting. Measures targeted at disease control would therefore best be focussed on screening new entrants from high-incidence settings for active and latent disease and on improving

diagnosis in and access to primary healthcare for hard-to-reach groups. As such, these findings should serve to re-emphasise the importance of maintaining access to NHS care for new entrants that could facilitate screening at GP surgeries, where indicated, as patients register. Such a programme might also serve to raise awareness of tuberculosis among GPs, including in low incidence areas of the country, resulting in earlier diagnosis of existing cases.

5 WGS FOR PREDICTING DRUG-SUSCEPTIBILITY

5.1 INTRODUCTION

Over the course of 2015, Public Health England will be piloting routine WGS for all mycobacterial cultures obtained from referring centres in the Midlands.[270] This project was initially inspired by an early recognition of the potential of WGS for tuberculosis surveillance and outbreak detection, as identified in the previous two chapters, and that there could be value in sequencing all strains for this purpose alone. However, as the same sequence data can be analysed to identify mycobacterial species and genetic variants underlying drug resistance, at no additional cost to the laboratory, the pilot study aims have since been expanded to assess WGS as a source for all three diagnostic outputs: species identification, resistance prediction, and genetic relatedness / transmission.

Drug resistance prediction is perhaps the most important challenge of the three. The World Health Organization estimated 480,000 new cases of multi-drug resistant (MDR) tuberculosis worldwide in 2013.[131] Phenotypic drug susceptibility testing (DST) for *M. tuberculosis* can take many weeks, with access to the necessary laboratory facilities limited in many countries with the greatest burden of disease.[131] Genotypic assays, based on common chromosomally mediated genetic determinants of drug resistance, take hours or days and have proven useful predictors of resistance in high and low-income countries alike.[271-273] However, as not all genetic determinants of resistance are known, molecular DST is currently considered to lack sufficient sensitivity to replace culture as the principal means of guiding patient management.[274,275]

The potential advantage of WGS over current molecular methods is that near-complete genomes can be screened for all known resistance-determinants, whilst also providing opportunities to identify novel mutations.[271,276,277] With unprecedented volumes of sequence data thus expected in the coming year, in this chapter I describe an algorithmic approach from which new genetic determinants of drug resistance can be iteratively derived from a training data set. I assess the success of the algorithmic approach against a 'validation-set' of independent isolates.

5.2 METHODS

5.2.1 SAMPLE SELECTION

I obtained 3651 *M. tuberculosis complex* sequences from the UK, Sierra Leone, South Africa, Germany and Uzbekistan in two stages. Collectively these represented all seven global clades, in addition to three *M. orygis* strains (Figure 24).[7] I first assembled a 'training set' of 2099 *M. tuberculosis* strains from isolates sequenced for chapters 3 and 4, in addition to isolates sequenced for a population study of Birmingham between 2009 and 2013, and all archived strains with phenotypic resistance to isoniazid, or line-probe assay based genotypic resistance to isoniazid, or to rifampicin in the absence of isoniazid resistance. Sequences from all previously treated, but recurrent cases of tuberculosis from Western Area and Kenema districts of Sierra Leone between 2003 and 2004 were also included,[278] as were isolates from Gauteng province, South Africa, that were sequenced as part of a WHO drug resistance surveillance project (Table 6). These South African isolates included all drug resistant isolates that had been sequenced at the time the training set was

assembled, in addition to a similar number of susceptible controls. I later assembled a validation set containing 1552 unrelated *M. tuberculosis* sequences. This included further isolates from Gauteng province, South Africa, in addition to samples from a population study of Hamburg, Germany and a drug resistance survey of Nukus, Uzbekistan (all unpublished). The South African samples again included all resistant samples that had been sequenced since those for the training set were obtained, together with a similar number of susceptible controls. Phenotypic data for one or more of eleven drugs were available for all included isolates.

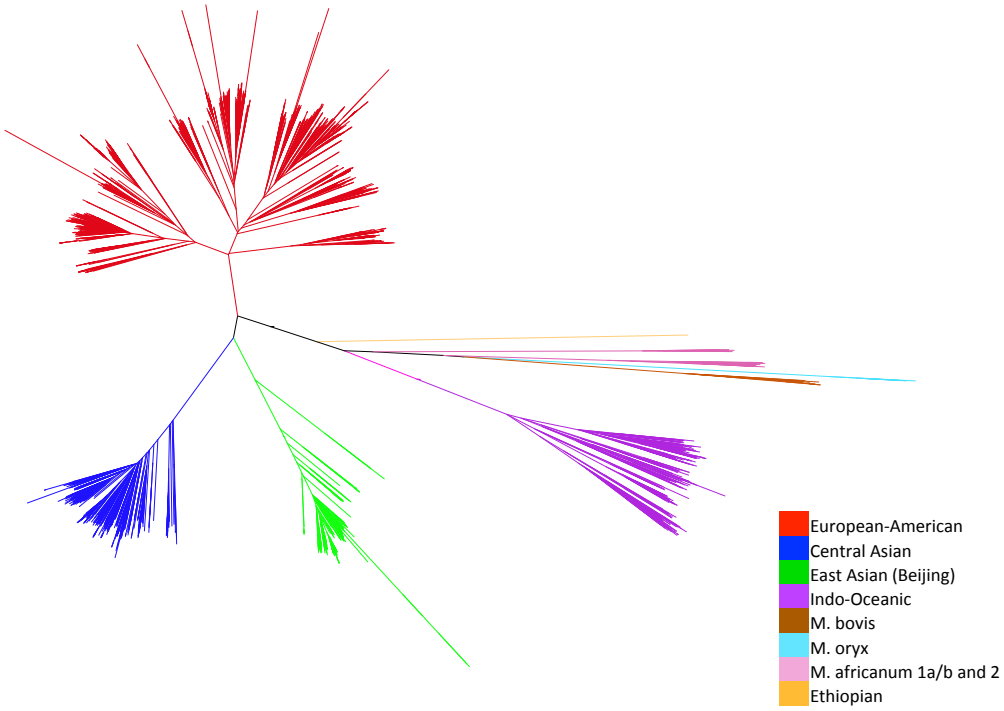


Figure 24: Maximum likelihood tree of 3651 isolates. All major global clades are represented.

Training-set	Identified	Included
Midlands, UK (unselected)	1162	1122
Midlands, UK (outbreaks)	428	412
Midlands, UK (resistance)	95	94
Oxfordshire, UK (unselected)	389	338
Gauteng province, South Africa	80	54
Kenema district, Sierra Leone	80	79
Total	2234	2009

Validation-set	Identified	Included
Hamburg, Germany	942	841
Nukus, Uzbekistan	277	261
Gauteng province, South Africa	466	450
Total	1685	1552

Table 6: Samples included in training and validation sets. Samples were excluded if they had either no available phenotypic data or where less than 88% of sites in the reference genome were called.

5.2.2 PHENOTYPING

Phenotypic DST was performed using the automated Mycobacterial Growth Indicator Tube (MGIT) 960 system (Becton Dickinson) with the WHO-endorsed proportion method at each of the source reference laboratories: UK samples were tested by the Public Health England West Midlands Public Health Laboratory, Birmingham, or at the National Mycobacterial Reference Unit, London, as part of routine patient care.[271] Cultures from Sierra Leone, Uzbekistan and Hamburg underwent phenotypic DST at the Forschungszentrum Borstel, Germany and South African cultures at the National Institute for Communicable Diseases, Johannesburg.

5.2.3 SAMPLE PREPARATION AND SEQUENCE ANALYSIS

Samples from the UK were cultured and prepared for sequencing as described in chapters 3 and 4. Those from Germany, Sierra Leone and Uzbekistan were cultured on LJ slopes and DNA extracted using the CTAB method described in chapter 3.[225] Samples from South Africa were cultured in MGIT and DNA

extracted using the Nuclisens EasyMag (Biomerieux, France) following the manufacturer's protocol. Sequencing was undertaken on Illumina (San Diego, California) platforms at the Wellcome Trust Centre for Human Genetics, Oxford, the Wellcome Trust Sanger Institute, Hinxton, the Forschungszentrum Borstel, and at the National Institute for Communicable Diseases, Johannesburg. Paired-end reads were mapped by Stampy[227] (version 1.0.17) to the H37Rv (GenBank NC000962.2) reference genome. Repetitive sections of the genome were defined by self-self BLAST and masked. Isolates with less than 88% mapped coverage of the reference genome were excluded (i.e. no lower than the mean reference genome coverage in chapter 3, where a more conservative version of the bioinformatics pipeline was used). Nucleotide calls were made with SAMtools mpileup[228] (version 0.1.18), requiring a minimum read depth of 5, including one read in each direction. Where an alternative base represented more than 10% of read depth, mixed base calls were made. These were only considered in the downstream analysis if constituting more than 90% of read depth in at least one other isolate (i.e. a non-mixed base call). Insertions and deletions ("indels") were identified using Cortex.[279] RAxML[280] (version 8.0.5) reconstructed the phylogeny under a GTRCAT model.[280] The frequency of each single nucleotide polymorphism (SNP) arising in the phylogeny was estimated using maximum-likelihood ancestral site reconstruction ("homoplasy").[281,282]

5.2.4 Algorithmic characterisation of alleles

Genome-wide association studies are ill-suited to identifying resistance-causing SNPs in clonal bacteria because of the difficulty in identifying causative alleles against a relatively monomorphic background.[283] I therefore first focussed on

23 candidate-genes and their promoter regions, each with at least one previously described drug-resistance allele (considered as SNPs in promoter regions, amino-acids in genes, or as indels). These genes were selected from a list of alleles identified in the literature by, and curated by, collaborators Stefan Niemann and Silke Feuerriegel in Borstel, Germany (Table 7). I devised an algorithm to characterise all alleles affecting these genes in the training set, and used these characterised alleles to predict DST for samples in the validation set. I then considered the contribution of other genes to phenotypic resistance.

Genes	Pubmed uid							
aphC	22646308	12654653						
eis	21300839							
embA	20427375	10639358						
embB	20427375	10639358	21300839	9257740	11854934	16641474	22646308	
embC	20427375							
embR	10639358							
fabG1	21300839	21300839	19494067	19494067				
gidB	22646308							
gyrA	19687244	21300839	21562102					
gyrB	19470506	17412727	19721073					
inhA	21110864							
iniA	10639358							
iniC	10639358							
katG	9210694	21300839	22646308					
manB	10639358							
ndh	11408244							
pncA	22646308	21300839	16848344	15616332	11641519	10681313	9692180	9056006
	9055989	8640557						
rmlD	10639358							
rpoB	11136757	7759399	8027320	8913484	9003625	10565894	10921994	14729930
	15184414	15814606	16229229	16672384	19721079	21300839	22646308	
rpsA	21835980							
rpsL	8849220	22646308						
rrs	21300839	21562102						
tlyA	16048924							

Table 7: Unique pubmed identifiers for the source papers for the 23 candidate genes, each gene having been included on the basis of at least one plausible resistance determining allele reported previously in the literature.

To check for bias resulting from the composition of sets, I repeated the analyses after switching training and validation-sets, and then again a further

100 times, randomly allocating samples to training-sets of 1825 samples and validation sets of 1826.

To algorithmically characterise alleles in the training-set, I first assumed that synonymous and lineage-defining alleles do not cause resistance, unless the latter were associated with lineage-specific resistance such as for pyrazinamide in *M. bovis*.[284] After labelling these alleles as 'benign' and setting them aside, I considered the remaining alleles within each group of genes relevant to each drug in turn (Figure 25).

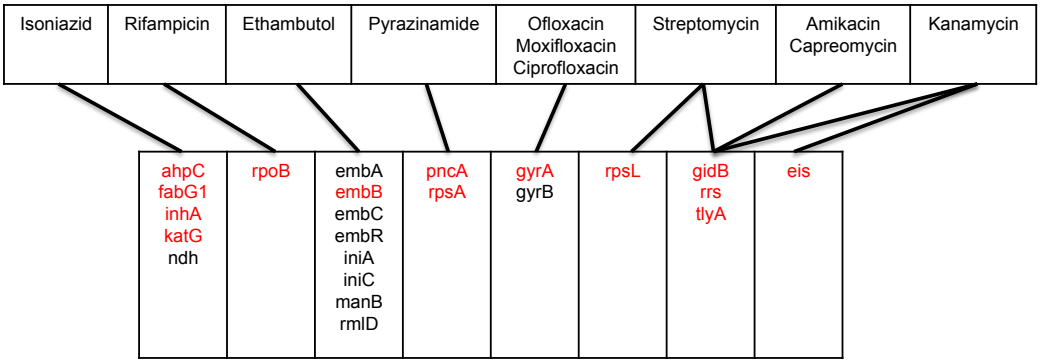
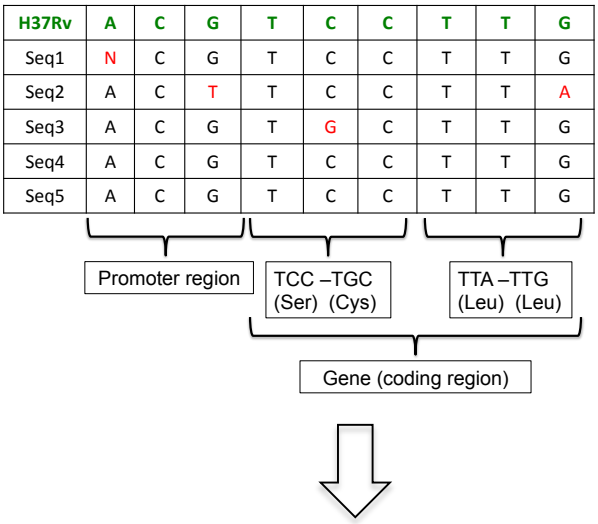


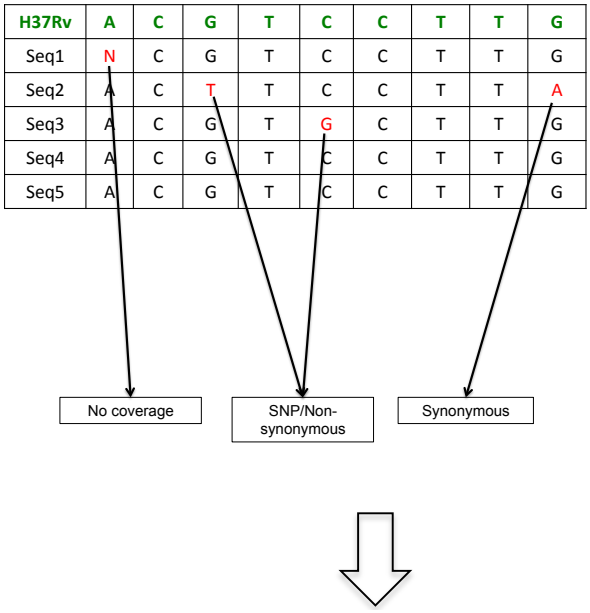
Figure 25: The subset of genes relevant to each of the 11 drugs. Those coloured red are the gene from which the algorithm was able to derive at least one resistance determining allele from the training set.

I characterised an allele as 'resistance determining' if it occurred as the only allele in at least one phenotypically resistant isolate. As alleles that do not cause resistance can clearly co-occur with those that do, I characterised alleles as 'benign' if they only occurred in phenotypically susceptible isolates, or at least where all isolates were phenotypically susceptible when an allele occurred alone. These benign alleles were then also set aside and the analysis repeated to potentially reveal further resistance-determining alleles.

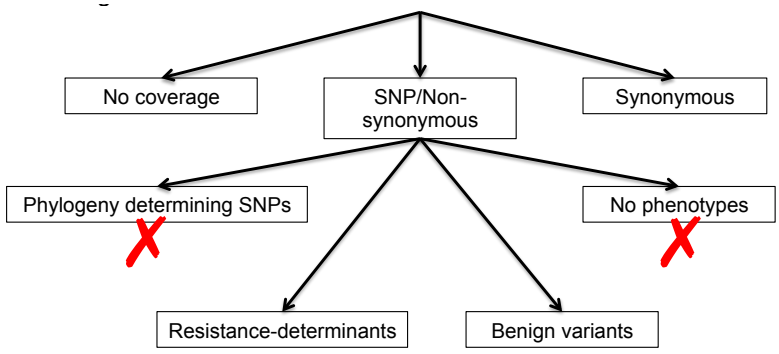
The principle behind the algorithm is illustrated by the following schematics: After each sequence in the training set is mapped to the H37Rv reference genome, every nucleotide position in the 23 genes, and 100 upstream base pairs, is inspected and identified as either the same as the reference, different from it, or 'null' (N) where no base could be called (either because of insufficient coverage or because of evidence for more than one base at that locus – a 'mixed call'). Five hypothetical sequences are shown here (Seq1 – Seq5).



Where 'N' was called at a locus in one or more sequences, but all others contained the same base call as the reference genome at that locus, the locus (be it upstream or within a gene) was set aside and not considered of further interest. Loci at which at least one sequence contained a different base call from the reference, whether in an upstream region or within a gene, were identified for further analysis. Where codons differed from the reference only on the basis of one or more sequences that had a synonymous SNP within the codon, these were also set aside.

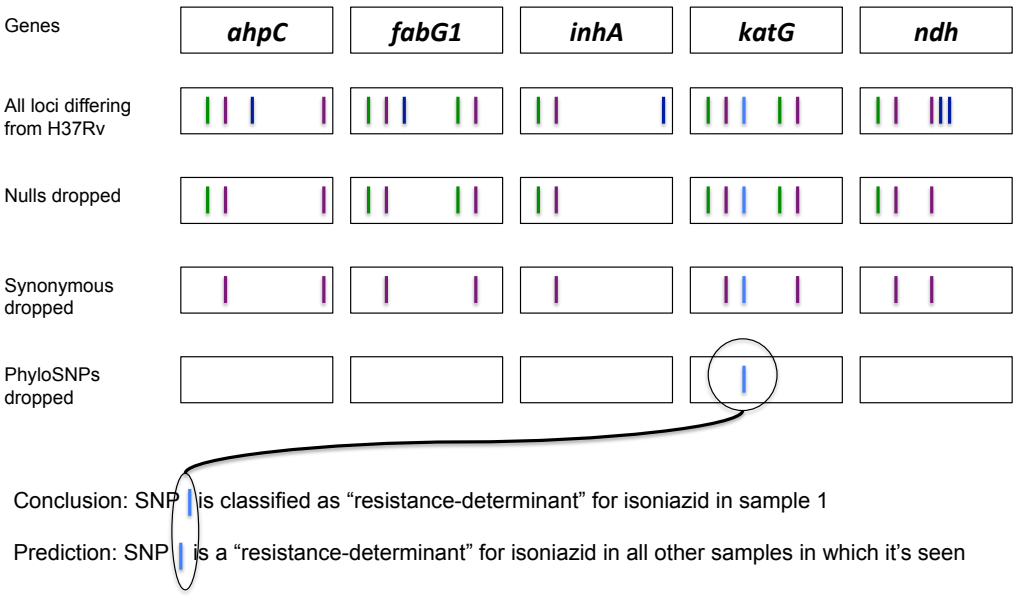


Among the remaining alleles (SNPs, amino acid substitutions and indels), some were identified as lineage, or phylogeny defining and therefore set aside. Others were only seen in sequences for which the isolate had not been phenotyped to the relevant drug. These were also set aside. The remaining alleles were then characterised as either resistance determining or as benign.



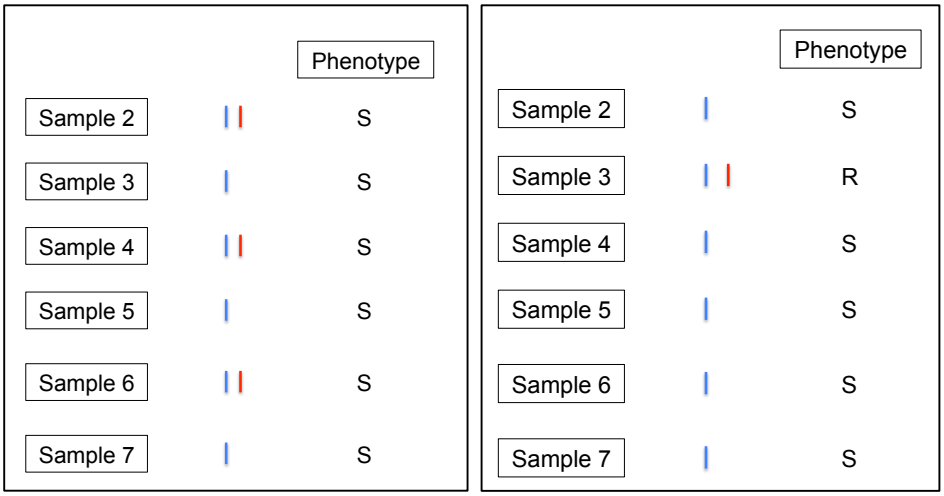
An example is given below. Here, the five of 23 genes relevant to isoniazid are shown for a hypothetical resistant isolate. Each coloured vertical line represents

a null-call, a synonymous allele, a lineage defining allele, or, as in the last row, an allele left for characterisation.

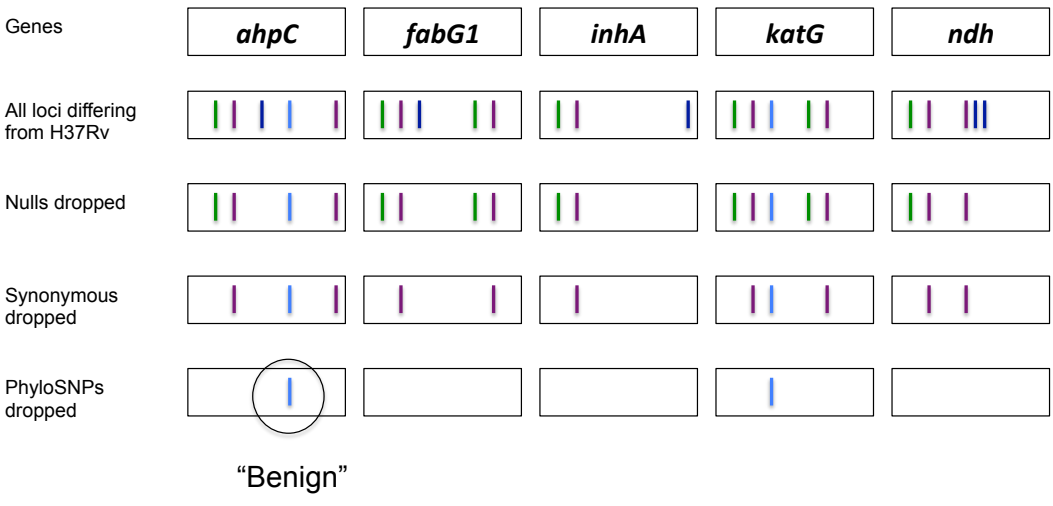


As the loci are dropped in turn according to the rule of the algorithm, only a single allele is left. Give the resistant phenotype for this sample, the allele is characterised as a ‘resistance determinant’ and all other samples in which this allele is seen are consequently predicted to be resistant to isoniazid.

Had the phenotype for this isolate been susceptible, all other isolates containing the remaining circled allele would have had to been inspected before the allele could be characterised as benign. Two scenarios could have led such a characterisation. First, if all isolates containing the allele were phenotypically susceptible, and second is all isolates containing only that allele were phenotypically susceptible. The scenarios are depicted in the panels below.



As can be seen in the right hand panel, the benign 'blue allele' occurs in one phenotypically resistant isolate together with a 'red allele'. In a second iteration of the algorithm, such benign alleles were set aside, exposing alleles such as the red allele as the only remaining allele for a resistant phenotype. This allowed additional resistance determinants to be characterised, as would be the case in the example of an isoniazid resistant strain here:



When new sequences are added to a training set, this has the potential to lead to the re-characterisation of alleles. For example, adding an eighth sample containing just the 'blue allele' to the left hand panel above would lead to the re-

characterisation of the allele as ‘resistant’, given a resistant phenotype. Resistant alleles could however only be re-characterised as benign where their original characterisation as ‘resistant’ was dependent on another allele being characterised as benign (as immediately above) *and* where the allele is otherwise always associated with a susceptible phenotype when occurring as the only allele in a sample (as in the right hand panel above).

Where phenotypic resistance could not be explained by a characterised resistance-determinant, I manually searched the sequence data for evidence of synergy between alleles, or of co-occurring compensatory alleles.

Given the binary, non-quantitative nature of the DST data, I made an assumption of ‘complete penetrance’ for each resistance determining allele. Validation-set isolates containing alleles characterised as resistance-determining in the training-set were thus predicted ‘resistant’, and those containing zero alleles, or only alleles characterised as benign, were predicted ‘susceptible’. Isolates containing uncharacterised alleles were not predicted (labelled ‘unknown’) unless co-occurring with resistance-determinants.

The predictions made for the validation set on the basis of alleles derived from the training set were then compared to the *in silico* performance of the MTBDR*plus*, MTBDR*sl* (HAIN Life-sciences, Germany) and AID (AID Diagnostika, Germany) line-probe assays (LPA). That is, I prepared a separate catalogue of resistance determining alleles based on the alleles these assays probe.[208,209,285] After that I combined all 3651 isolates and applied the algorithm to these as a whole.

As some resistant phenotypes might not be explained by alleles in the 23 genes, I next explored the remaining genome for potential explanatory alleles. As resistance-causing alleles are likely to be under positive selection pressure, these alleles are also the most likely to arise repeatedly, independently in the phylogeny.[277] Focussing the search for additional resistance determinants on homoplasic alleles, I quantified the frequency of homoplasic events for each allele in the genome and compared the frequency observed across the 23 genes and among characterised resistance determinants to that among alleles in genes elsewhere in the genome, including all open reading-frames and functional RNA molecules.[179]

Carlos Del Ojo Elias wrote the python script to extract the sequence data from the 3651 variant calling format files. Jessica Hedge identified and quantified homoplasy for each allele in the genome using ClonalFrameML[286] and R (R Development Core Team, 2010). I used STATA 13.1 (StataCorp, Texas) to manipulate the sequence data, write and run the algorithm and associate the homoplasy data for each allele in the genome with phenotypic results and allelic characterisations.

5.3 RESULTS

2099 *M. tuberculosis* isolates were sequenced as a training-set, within which 1414 independent strains could be identified by clustering isolates within 5 SNPs of another, based on the threshold defined in chapter 3. 382 (18%) were phenotypically resistant to at least one drug, 91 (4%) MDR and 4 (0.2%) extensively drug-resistant (XDR), giving a total of 701 resistant phenotypes.

The 23 genes and their promoter regions accounted for 11,473 codon and 3,838 nucleotide loci respectively, of which 10,007 of 15,311 (65%) loci were genetically identical with the H37Rv reference genome in every sequence. Of the remaining 5,304 loci, 1,682 had insufficient coverage or a mixed base-call in one or more isolates, with the same base as the reference having been called in all other isolates at these loci, and 2,466 loci differed from the reference genome in one or more isolates due only to synonymous SNPs. All these sites were set aside and not considered further (Figure 26).

After all lineage-defining alleles were then set aside, with the exception of pncA^{H57D} and rpsA^{A440T} as these were present in all *M. bovis* isolates, in genes associated with pyrazinamide resistance (see Table 7 and Figure 25). This left 991 allele/drug combinations for consideration (134 nucleotide, 774 amino-acid variants, and 83 indels) across the 11 drugs (Figure 26).

Exactly one allele remained in 74% of resistant and 12% of susceptible phenotypes, whereas zero alleles remained for 84% of susceptible and 5% of resistant phenotypes (Table 8). 112 alleles were thereby characterised as 'resistance determining' whilst 772 were characterised as 'benign'. After setting these benign alleles aside, six additional alleles were characterised as resistance determining. No further alleles could be characterised by repeating the algorithm again. 101 alleles thus remained uncharacterised, of which 60 only co-occurred with resistance determinants, compatible with a possible compensatory role (See appendix S1 for list; Figure 26, unclassified variants are in bold boxes).

Number of alleles remaining for each drug	Susceptible phenotype (%)		Resistant phenotype (%)	
0	7,566	(84)	33	(5)
1	1,111	(12)	518	(74)
2	211	(2)	130	(19)
3	61	(1)	16	(2)
4	21	(0)	3	(0)
5	6	(0)	0	(0)
6	1	(0)	1	(0)
7	0	(0)	0	(0)
8	1	(0)	0	(0)

Table 8: The number of alleles remaining in genes relevant to each drug after lineage defining and synonymous alleles have been set aside are shown by susceptible and resistant phenotypes for the 2099 training set isolates.

Overall, 146 allele/drug combinations had previously been described as resistance determining in the literature, of which 130 were characterised here. Of interest, these included 79 of 120 (66%) resistance-determinants, 31 of 772 (4%) 'benign' alleles, 20 of 71 (28%) alleles defining lineages not intrinsically drug-resistant, and 16 uncharacterised alleles (see appendix S2 for list).

The 120 resistance determining alleles, including pncA^{H57D} and rpsA^{A440T}, were spread across just 14 candidate genes (Figure 25; alleles are listed in appendix S3). At least one was present in 658 of 701 (93.9%) resistant training set phenotypes, of which 535 of 701 (76.3%) contained only that allele. 33 of 701 (4.7%) resistant phenotypes remained unexplained with zero relevant alleles in candidate genes and their upstream regions, indicating either resistance mechanisms located elsewhere in the genome, or phenotypic or labelling error. 10 of 701 (1.4%) could not be algorithmically unravelled as they contained more than one relevant allele. Six of these contained alleles previously associated with resistance in the literature.

Resistance determining alleles also occurred in 121 susceptible phenotypes. Such phenotypic variability was most evident for isolates containing $embB^{M306I}$ and $rpoB^{I491F}$, 34 of 50 and 19 of 23 of which were phenotypically susceptible to ethambutol and rifampicin respectively (see appendix S3). Although it is possible that alleles elsewhere in the genome might explain such variability through epistasis, eight ethambutol resistant isolates were 0 SNPs from at least one of ten susceptible isolates, and 3 of the rifampicin resistant isolates were also genetically indistinguishable from 3 susceptible isolates, suggesting poor phenotypic reproducibility for these alleles at least.[287]

To assess their accuracy, training set characterisations were used to predict phenotypes for an independent validation set (Table 9; Figure 27). Of 1552 validation set isolates, 449 (28.9%) were phenotypically resistant to at least one drug, 284 (18.3%) were MDR and 3 (0.2%) XDR. 58 of 120 (48.3%) alleles characterised as resistance determining in the training set, and 175 of 772 (22.7%) characterised as benign recurred in the validation set. 54 of 58 (93.1%) recurring resistance determining alleles, including 12 not known to the literature, were accurately predictive of at least one resistant validation set phenotype, of which 34 of 54 (63%) were not associated with a single susceptible phenotype (see appendix, S4).

Figure 27 shows the number of phenotypically resistant and susceptible phenotypes with which each resistance determinant is seen in both the derivation set and validation set. All alleles seen in more than one isolate are depicted by drug. Ofloxacin and amikacin are shown as representatives of the fluoroquinolones and aminoglycosides respectively. With the exception of

pyrazinamide, each drug has one or two dominant, frequently occurring alleles and a larger number of comparatively rare ones. For pyrazinamide there was no single dominant allele other than the lineage defining alleles in *M. bovis*.

89.2% of validation set phenotypes were thereby predicted resistant or susceptible with a mean 92.3% sensitivity and 98.4% specificity, using ofloxacin and amikacin as representatives of their respective drug-classes (Table 9). The presence of uncharacterised alleles prevented predictions in the remaining 10.8%. Of 94 of 1221 (7.7%) resistant phenotypes wrongly predicted susceptible, 20 of 94 (21.3%) were because of alleles characterised as benign in the training set, whilst 74 of 94 (78.7%) had zero relevant alleles, suggesting either phenotypic or labelling error, or a resistance mechanism outside of candidate genes. Of 112 of 6892 (1.6%) susceptible phenotypes wrongly predicted resistant, 55 of 112 (49%) contained alleles at $embB^{M306}$, indicative of the associated phenotypic variability. Eight however contained $katG^{S315T}$, more likely to represent labelling than phenotypic error.[271] Appendix S5 lists the susceptible phenotypes ‘falsely’ predicted resistant.

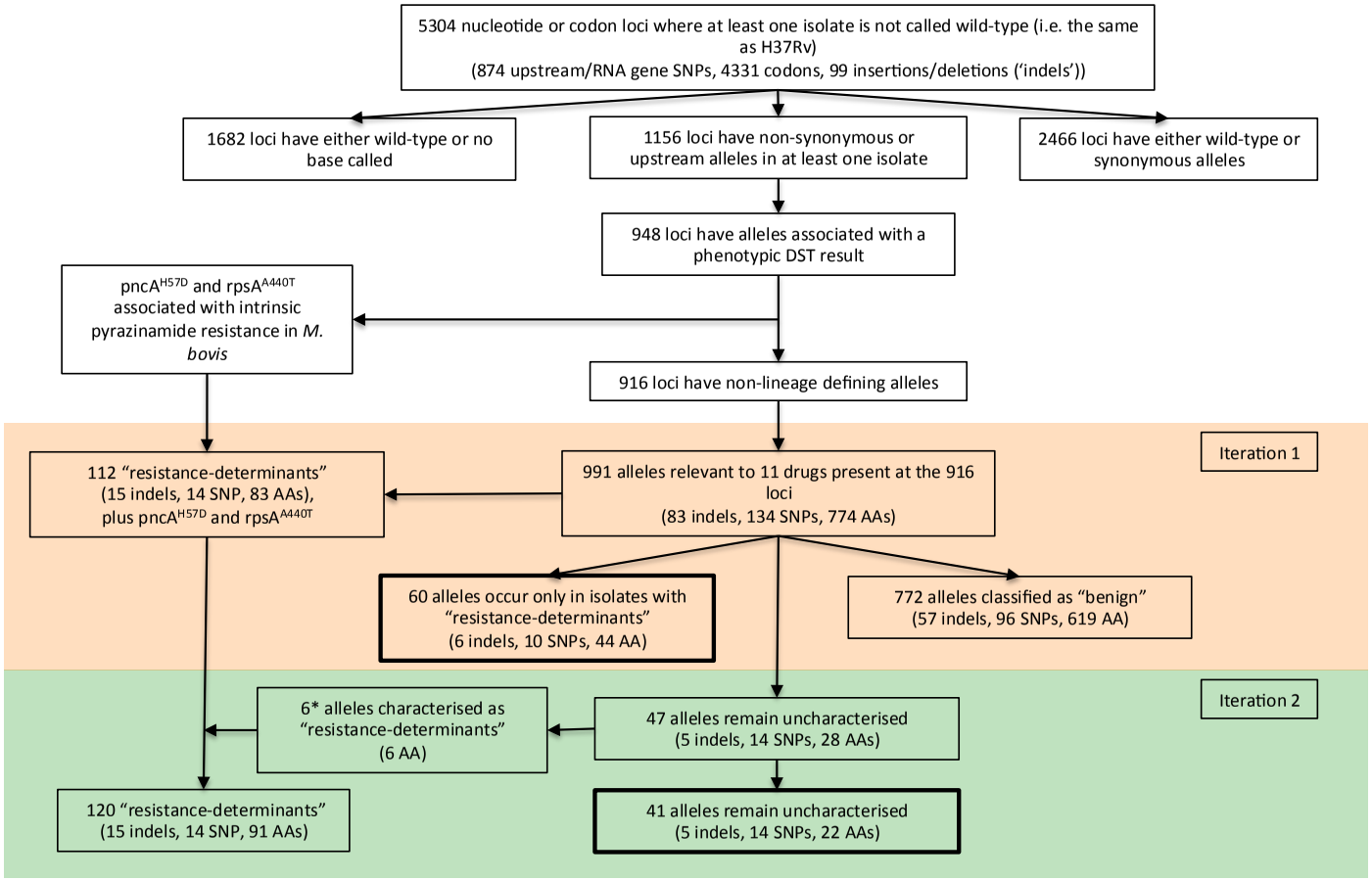


Figure 26: Algorithm for classification of genetic variation observed in derivation-set isolates.

	Phenotypically Resistant						Phenotypically Sensitive						All		Excluding Unclassfied		
	Genotype						Genotype										
	R	R_x	S_0	S_s	U	Total	R	R_x	S_0	S_s	U	Total	Sensitivity	Specificity	Sensitivity	Specificity	% Unclassifed
Isoniazid	305	5	18	1	35	364	19	0	1,065	52	52	1188	85.2	98.4	94.2	98.3	5.6
Rifampicin	263	12	8	1	16	300	9	1	1,200	4	38	1252	91.7	99.2	96.8	99.2	3.5
Ethambutol	152	6	7	1	26	192	62	5	1003	79	210	1359	82.3	95.1	95.2	94.2	15.2
Pyrazinamide	31	12	27	5	104	179	2	0	1,218	67	83	1370	24.0	99.9	57.3	99.8	12.1
Streptomycin	278	6	6	9	49	348	10	1	970	34	189	1204	81.6	99.1	95.0	98.9	15.3
Ofloxacin	2	3	4	2	0	11	0	0	489	134	38	661	45.5	100.0	45.5	100.0	5.7
Amikacin	36	16	5	0	2	59	1	2	427	38	140	608	88.1	99.5	91.2	99.4	21.3
Total	1067	60	75	19	232	1453	103	9	6372	408	750	7642	77.6	98.5	92.3	98.4	10.8

Table 9: Genotypic predictions based on: R (resistance determinant); Rx (resistance determinant as a mixed base call); S0 (zero alleles present); Ss (only sensitive alleles present); U (unclassified alleles present). Weighted mean sensitivity and specificity presented as both an overall figure and as that based on characterised alleles only (R, Rx, S0, Ss columns) without predictions for isolates classified as U as these contain uncharacterised alleles. To avoid double counting for related drugs, only ofloxacin and amikacin are included as representatives of their antibiotic classes for the calculation of the sensitivity and specificity, as these had more resistant phenotypes than their related drugs.

Number of phenotypes with allele
0
100
200
470
R (training-set)
R (validation-set)
S (training-set)
S (validation-set)
Isoniazid
katG_S315T
fabG1_C-15T
fabG1_T-8C
katG_S315N
fabG1_G-17T
inhA_S94A
inhA_I194T
inhA_I21T
katG_W191R
katG_V633A
katG_471_delG
ahpC_C-57T
Rifampicin
rpoB_S450L
rpoB_H445Y
rpoB_L452P
rpoB_H445D
rpoB_D435V
rpoB_S450W
rpoB_H445R
rpoB_H445N
rpoB_S450F
rpoB_Q432K
rpoB_D435F
rpoB_I491F
rpoB_V170F
rpoB_1396_ins
Ethambutol
embB_M306V
embB_M306I
embB_Q497R
embB_G406S
embB_G406D
embB_G406A
embB_Q497K
embB_D354A
embB_H1002R
Pyrazinamide
rpsA_A440T
pncA_A-11G
pncA_H57D
pncA_V125G
pncA_T-12C
pncA_L4S
pncA_D136N
pncA_V7L
pncA_G97D
pncA_C14R
pncA_C138R
pncA_V139L
pncA_S104R
pncA_Q141*
pncA_H57R
pncA_570_insC
pncA_494_delC
pncA_409_delT
pncA_292_insA
Streptomycin
rpsL_K43R
rpsL_K88R
rrs_A514C
gidB_202_delG
rrs_C517T
gidB_G69D
gidB_A138V
gidB_A134E
rrs_C513T
gidB_V65G
gidB_R137W
gidB_P75L
gidB_H48N
gidB_A200E
gidB_202_insG
Ofloxacin
Amikacin
gyrA_D94G
gyrA_S91P
gyrA_A90V
gyrA_D94A
rrs_A1401G

Figure 27 (above): Histogram of frequency of resistance determinants. The number of susceptible and resistant phenotypes associated with each resistance determinant are shown in black and grey for the training set and in green and orange for the validation set. Alleles probed by an LPA are highlighted in red. Alleles that are only seen once in the training set and not again in the validation set (i.e. with no additional information to validate them) are not shown. Of the fluoroquinolones and aminoglycosides, only ofloxacin and amikacin have been included as representatives of their class.

To assess whether these results were contingent on the allocation of isolates to training and validation sets, I repeated the algorithm using the validation set as the training set, and vice-versa. 93.1% of, what were now, validation set phenotypes could be predicted with a mean sensitivity and specificity of 92.1% and 97.9% respectively. I performed a further 100 iterations of the algorithm, on each occasion randomly allocating samples to training or validation sets. Over these 100 iterations, the mean proportion of predictable phenotypes was 92.7%, and the mean sensitivity and specificity 92.4% and 98.2% respectively. Detailed results of 100 iterations of the algorithm are listed in appendix S6.

Given the consistency of these results, I used the original predictions for the validation set to compare to the *in silico* performance of alleles probed by three LPAs. Excluding pyrazinamide, for which no LPA exists, these collectively predicted validation set phenotypes with a mean 81.6% sensitivity and 98.0% specificity, compared to 85.1% and 98.2% respectively for the algorithmically characterised alleles. However, unlike the LPAs, the algorithmically characterised alleles were also able to unambiguously distinguish between benign and uncharacterised alleles, allowing further improvement by restricting predictions to the 89.2% of predictable validation set phenotypes. For these the mean sensitivity and specificity, excluding pyrazinamide, was 94.6% and 98.0% respectively (Table 10).

Supplementing the derived resistance determinants with LPA alleles only marginally improved overall sensitivity to 78.5%, including pyrazinamide, and to 94.8%, excluding pyrazinamide. This was largely a consequence of including 13 additional resistance determinants in rpoB (Table 11). Of these, only $rpoB^{L430P}$ and $rpoB^{D435Y}$, both known to confer low-level resistance, occurred in training set sequences, where they were characterised as benign.[278,288]

Results for line-probe assay alleles

	Phenotypically Resistant						Phenotypically Sensitive						All		Excluding Unclassfied	
	Genotype						Genotype									
	R	R_x	S_0	S_s	U	Total	R	R_x	S_0	S_s	U	Total	Sensitivity	Specificity		
Isoniazid	304	5	0	55	0	364	17	0	0	1171	0	1188	84.9	98.6		
Rifampicin	272	14	0	14	0	300	33	0	0	1219	0	1252	95.3	97.4		
Ethambutol	132	6	0	54	0	192	53	6	0	1300	0	1359	71.9	95.7		
Streptomycin	246	2	0	100	0	348	7	1	0	1196	0	1204	71.3	99.3		
Ofloxacin	3	3	0	5	0	11	5	0	0	656	0	661	54.5	99.2		
Amikacin	37	16	0	6	0	59	2	2	0	604	0	608	89.8	99.3		
Total	994	46	0	234	0	1274	117	9	0	6146	0	6272	81.6	98.0		

Results for training-set

Isoniazid	305	5	18	1	35	364	19	0	1,065	52	52	1188	85.2	98.4	94.2	98.3
Rifampicin	263	12	8	1	16	300	9	1	1,200	4	38	1252	91.7	99.2	96.8	99.2
Ethambutol	152	6	7	1	26	192	62	5	1003	79	210	1359	82.3	95.1	95.2	94.2
Streptomycin	278	6	6	9	49	348	10	1	970	34	189	1204	81.6	99.1	95.0	98.9
Ofloxacin	2	3	4	2	0	11	0	0	489	134	38	661	45.5	100.0	45.5	100.0
Amikacin	36	16	5	0	2	59	1	2	427	38	140	608	88.1	99.5	91.2	99.4
Total	1036	48	48	14	128	1274	101	9	5154	341	667	6272	85.1	98.2	94.6	98.0

Results for combined alleles from training-set and line-probe assays

Isoniazid	306	5	18	1	34	364	19	0	1,065	52	52	1188	85.4	98.4	94.2	98.3
Rifampicin	273	13	8	0	6	300	12	3	1,200	2	35	1252	95.3	98.8	97.3	98.8
Ethambutol	152	6	7	1	26	192	65	6	1003	77	208	1359	82.3	94.8	95.2	93.8
Pyrazinamide	31	12	27	5	104	179	2	0	1,218	67	83	1370	24.0	99.9	57.3	99.8
Streptomycin	278	6	6	9	49	348	10	1	970	34	189	1204	81.6	99.1	95.0	98.9
Ofloxacin	3	3	4	1	0	11	0	0	489	134	38	661	54.5	100.0	54.5	100.0
Amikacin	37	16	5	0	1	59	2	2	427	38	139	608	89.8	99.3	91.4	99.1
Total	1080	61	75	17	220	1453	110	12	6372	404	744	7642	78.5	98.4	92.5	98.2
												Excluding pyrazinamide:	86.2	98.1	94.8	97.9

Table 10 (above): Comparison of derived training set alleles to the *in silico* performance of LPA based alleles. Results are shown for the LPA alleles, then for the derived alleles (for the same drugs), and then for the combined alleles from the LPAs and the training set. Genotypic predictions based on: R (resistance determinant); Rx (resistance determinant as a mixed base call); S_0 (zero alleles present); Ss (only benign alleles present); U (uncharacterised alleles present). Weighted mean sensitivity and specificity presented for a subset of drugs to avoid double counting. Ofloxacin and amikacin were included as representatives of their antibiotoc classes as these had more resistant phenotypes than their related drugs. Pyrazinamide excluded for line-probe comparison as no LPA exists for pyrazinamide.

	Phenotypically resistant		**Phenotypically sensitive**		
	Resistance-determinant	**Resistance-determinant as a mixed-call**	**Resistance-determinant**	**Resistance-determinant as a mixed-call**	**In training set?**
rpoB_A451G	1	0	0	0	No
rpoB_D435Y	2	0	0	0	Yes
rpoB_H445C	1	0	0	0	No
rpoB_H445L	2	0	0	0	No
rpoB_L430P	0	1	3	0	Yes
rpoB_L443F	1	0	0	2	No
rpoB_M434I	1	0	0	0	No
rpoB_N437S	1	0	0	0	No
rpoB_Q429H	2	0	1	0	No
rpoB_Q432L	1	0	0	0	No
rpoB_Q432P	2	0	0	0	No
rpoB_S441L	1	0	0	0	No
rpoB_S450Q	1	0	0	0	No

Table 11: Number and type of alleles in the rifampicin resistance determining region of rpoB that were not characterised / did not occur in training set sequences

I next reran the algorithm for all 3651 isolates, thereby increasing the number of alleles characterised as resistance determining from 120 to 232, and as benign from 772 to 1634 (Figure 28; appendix S7). Among the resistance determining alleles were 3 that had remained uncharacterised in the original training set,

and 16 that had originally been characterised as benign but that were re-characterised because of the addition of phenotypically resistant isolates containing only those alleles (see appendix S8). Eight of these 19 (42.1%) had been previously described as resistance determining in the literature. As all samples were included in this training set, no independent validation set remained, but I did make predictions for the entire set itself. 96.1% of phenotypes could be predicted with a mean sensitivity and specificity of 94.8% and 97.9% respectively (Figure 29; Table 12).

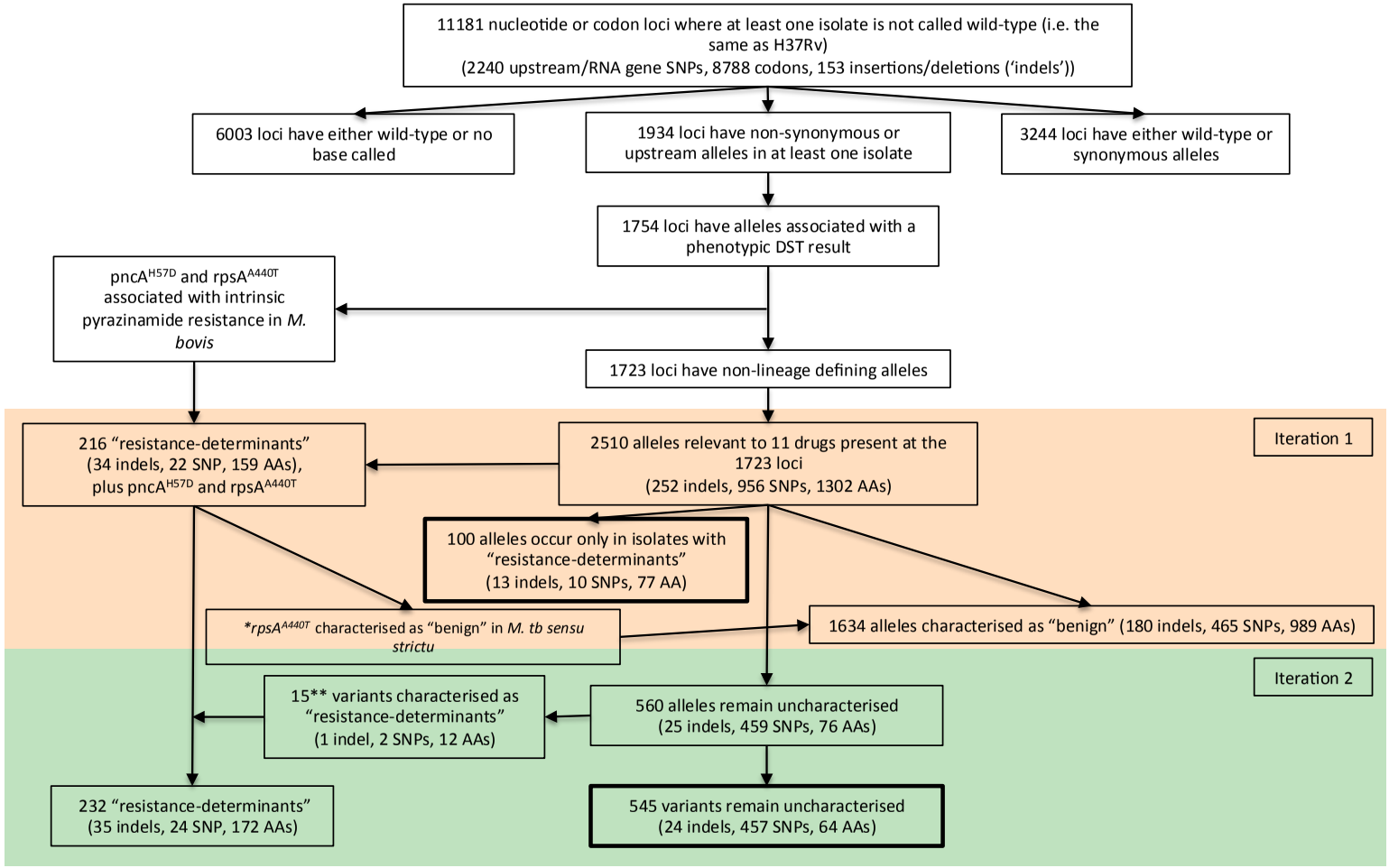

11181 nucleotide or codon loci where at least one isolate is not called wild-type (i.e. the same as H37Rv)
(2240 upstream/RNA gene SNPs, 8788 codons, 153 insertions/deletions ('indels'))
6003 loci have either wild-type or no base called
1934 loci have non-synonymous or upstream alleles in at least one isolate
3244 loci have either wild-type or synonymous alleles
1754 loci have alleles associated with a phenotypic DST result
pncA^H57D and rpsA^A440T associated with intrinsic pyrazinamide resistance in M. bovis
1723 loci have non-lineage defining alleles
Iteration 1
216 "resistance-determinants" (34 indels, 22 SNP, 159 AAs), plus pncA^H57D and rpsA^A440T
2510 alleles relevant to 11 drugs present at the 1723 loci
(252 indels, 956 SNPs, 1302 AAs)
100 alleles occur only in isolates with "resistance-determinants"
(13 indels, 10 SNPs, 77 AA)
*rpsA^A440T characterised as "benign" in M. tb sensu strictu
1634 alleles characterised as "benign" (180 indels, 465 SNPs, 989 AAs)
Iteration 2
15** variants characterised as "resistance-determinants"
(1 indel, 2 SNPs, 12 AAs)
560 alleles remain uncharacterised
(25 indels, 459 SNPs, 76 AAs)
232 "resistance-determinants"
(35 indels, 24 SNP, 172 AAs)
545 variants remain uncharacterised
(24 indels, 457 SNPs, 64 AAs)

Figure 28 (above): Algorithm as applied to all 3651 isolates. *Although in the training set rpsA^{A440T} was identified as lineage defining of *M. bovis*, which is intrinsically resistant to pyrazinamide, it is seen independently as a lone-standing (non *M. bovis*) allele in a pyrazinamide sensitive isolate in the validation set, leading to its re-characterised as benign, and leaving pncA^{H57D} as the remaining resistance determinant for *M. bovis*. 645 alleles were left uncharacterised (highlighted in boxes in bold). **Alleles occurring in phenotypically resistant isolates with the other alleles defined as 'benign' in iteration 1.

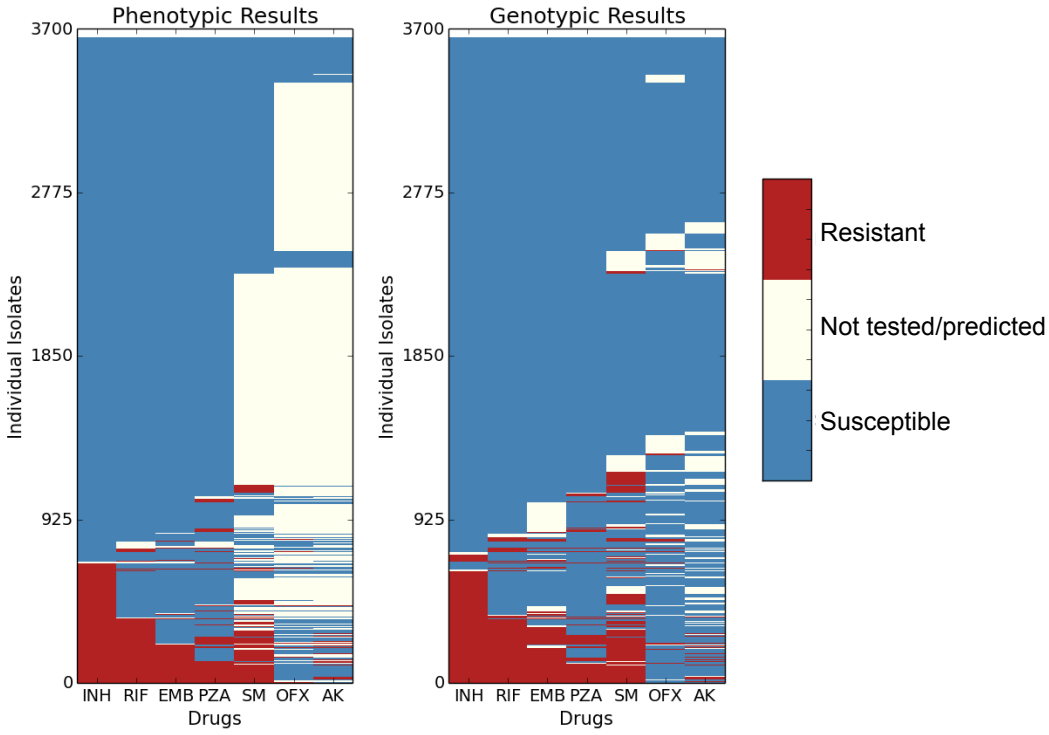


Figure 29: Antibiograms (phenotypic and genotypic) for all 3651 isolates

	Phenotypically Resistant						Phenotypically Resistant						All		Excluding Unclassfied	
	Genotype						Genotype						Sensitivity	Specificity	Sensitivity	Specificity
	R	R_x	S_0	S_s	U	Total	R	R_x	S_0	S_s	U	Total				
H: Isoniazid	621	12	36	0	10	**679**	33	2	2,666	244	15	**2960**	**93.2**	**98.8**	**94.6**	**98.8**
R: Rifampicin	385	14	9	0	2	**410**	40	4	3,033	91	25	**3193**	**97.3**	**98.6**	**97.8**	**98.6**
E: Ethambutol	217	10	8	0	16	**251**	122	8	2,353	645	248	**3376**	**90.4**	**96.1**	**96.6**	**95.8**
Z: Pyrazinamide	187	24	34	0	1	**246**	47	5	3,085	196	6	**3339**	**85.8**	**98.4**	**86.1**	**98.4**
S: Streptomycin	397	7	10	0	7	**421**	40	1	1,329	127	140	**1637**	**96.0**	**97.5**	**97.6**	**97.3**
O: Ofloxacin	18	4	5	0	1	**28**	4	0	571	120	85	**780**	**78.6**	**99.5**	**81.5**	**99.4**
A: Amikacin	42	16	5	0	1	**64**	2	2	484	75	157	**720**	**90.6**	**99.4**	**92.1**	**99.3**
Total	1867	87	107	0	38	2099	288	22	13521	1498	676	16005	**93.1**	**98.1**	**94.8**	**98.0**
Other quinilones																
Ci: Ciprofloxacin	21	1	0	0	1	23	3	0	233	42	12	290				
M: Moxifloxacin	19	1	2	0	2	24	5	0	370	108	83	566				
Other aminoglyocosides																
C: Capreomycin	41	13	7	0	1	62	35	4	482	72	126	719				
K: Kanamycin	13	0	4	0	1	18	5	0	311	80	147	543				

Table 12: Phenotypic predictions for all 3651 isolates, based on alleles derived from applying the algorithm to the same 3651 isolates. Genotypic predictions based on: R (resistance determinant); Rx (resistance determinant as a mixed base call); S_0 (zero alleles present); Ss (only benign alleles present); U (uncharacterised alleles present). Weighted mean sensitivity and specificity presented for a subset of drugs to avoid double counting. Ofloxacin and amikacin were included as representatives of their antibiotic classes as these had more resistant phenotypes than their related drugs.

To screen for additional resistance-determining alleles in the rest of the genome, I assessed all alleles across the phylogeny of all 3651 samples for homoplasy as this might be expected under positive selection pressure from drugs. I first characterised the 23 candidate-genes as a comparator for the remaining genome. Excluding indels, which were not assessed for homoplasy, and counting alleles characterised for more than one drug only once, 292 (0.8% of the 23 genes) alleles were homoplasic. These included 54 of 96 (56.3%) remaining alleles characterised as resistance-determining, 15 of 82 (18.3%) uncharacterised alleles, and 30 of 609 (4.9%) alleles characterised as benign in the original training set. Outside of the 23 candidate genes, 5,427 alleles were homoplasic, representing 0.1% of the rest of the genome. In contrast to alleles in candidate genes, I was not able to identify a correlation between the number of homoplasic emergences and phenotypic resistance, leaving candidates alleles difficult to identify (Figure 30).

To increase the probability of finding resistance-determining alleles, I sought new candidate-genes from among the 2364 of 3974 (59.6%) genes in the genome with at least one homoplasic allele. To identify the most homoplasic genes, I summed the frequency with which homoplasic alleles emerged for each gene. As expected, ten of the 14 genes providing the 120 resistance determinants in the training set were among the top 34 (1.4%) most homoplasic genes (median 67 emergences (interquartile range 12-123)). This compared to a median of 32 (IQR 8-108) across all 23 candidate-genes, and a median of 4 (IQR 2-6) in the remaining genes. I therefore searched for alleles in these ‘top 34’ most homoplasic genes to explain the 33 unexplained resistant phenotypes in the training set. Each of these phenotypes was associated with a median of 7

(IQR 5-11) alleles in these top 34 genes, too many for my algorithm to disentangle. Moreover, given that the median number for all other phenotypes in the study was 8 (IQR 4-11), these 'top 34' genes did not discriminate between resistant and susceptible phenotype.

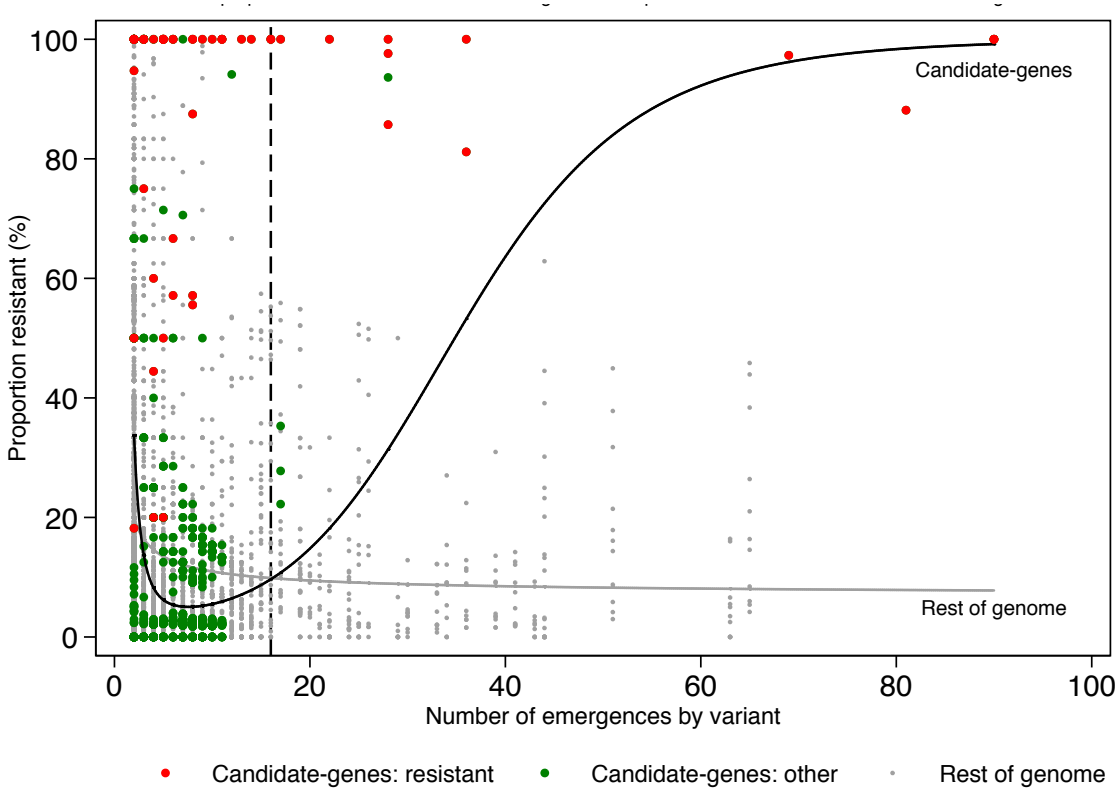


Figure 30: Alleles within the candidate-genes are shown as red (classified as resistance determinants) and green (others). Alleles in the rest of the genome are shown in grey. Curves represent the mean for alleles across candidate genes and across the rest of the genome. The variant katG^{S315T} emerged 180 times but has been set to 90, equal to the next most homoplastic variant, to fit all data to the current scale. The proportion resistant has not been changed and shape of the curve of the mean for candidate-genes remains unaltered.

A final summary of the number of training set alleles characterised as resistance determinants, as benign; the number known to the literature, or not; and the number that were homoplasic, is given in Figure 31.

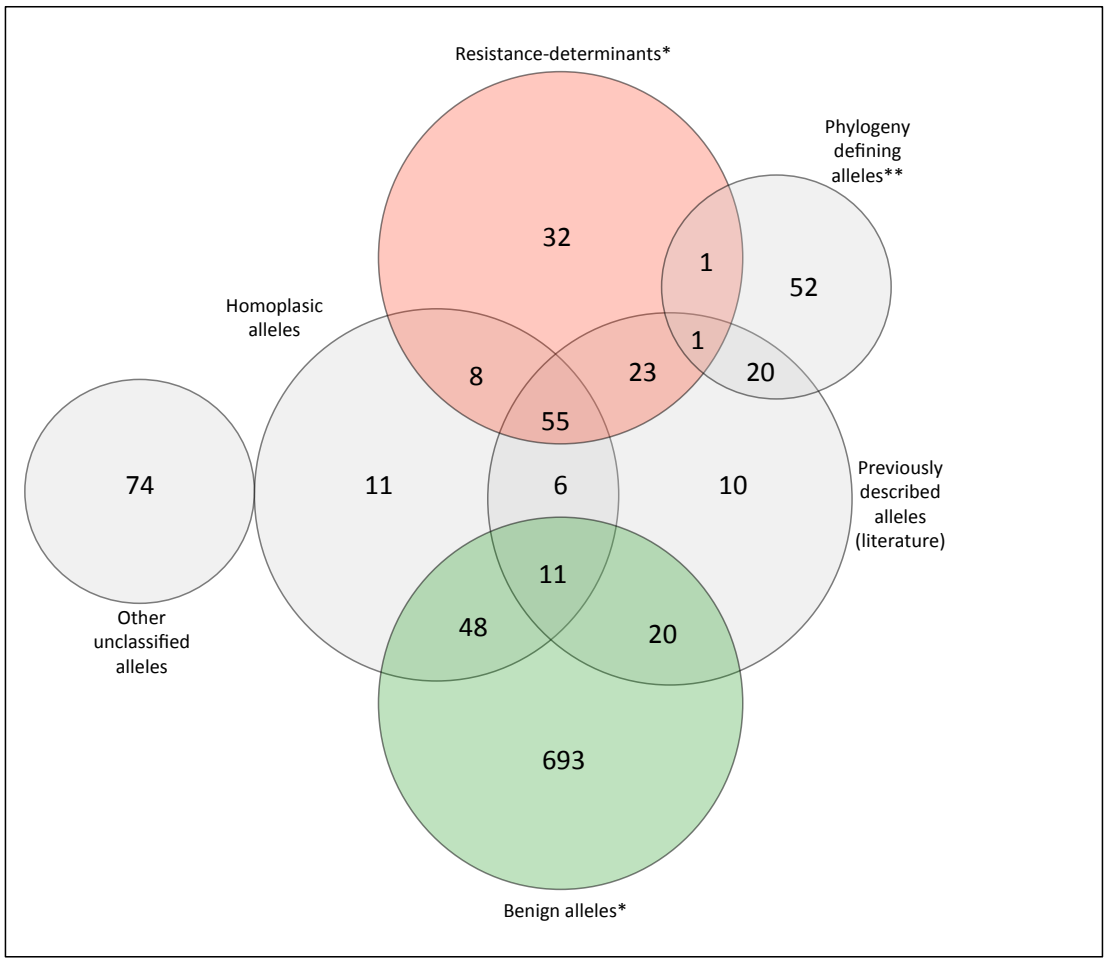


Figure 31: The number of unique alleles characterised as 'resistant' in the training set are shown in red and as benign are shown in green. The number of homoplasic alleles, alleles described in the literature, lineage-defining alleles and of uncharacterised alleles are shown in grey. Indels were not assessed for homoplasy. *Among resistance determinants and benign alleles, 15 and 55 indels, and 25 and 371 alleles seen in only one isolate respectively, were not / could not be assessed for homoplasy. ** gyrA^{A384V} defines the Indian Ocean lineage (all isolates in the lineage have this SNP) but also appears in one European American isolate; rpsAA440T defines *M. bovis* but also appears in one Central Asian isolate. Both are thereby homoplasic.

5.4 DISCUSSION

I obtained a training-set of 2099 *M. tuberculosis* sequences from the UK, Sierra Leone and South Africa to algorithmically characterise alleles across 23 candidate genes as either resistance determining or as benign. These characterised alleles predicted 89.2% of phenotypes for an independent validation set of 1552 isolates from South Africa, Germany and Uzbekistan with a sensitivity and specificity of 92.3% and 98.4% respectively.

The sampling frame was constructed opportunistically. Initially I only had access to sequences included in the training set, among which there were too few resistant phenotypes for the samples to be productively split into a training and validation set. I therefore first devised and applied the algorithm only to these samples. As the results needed validation, I then sought and obtained access to additional samples through collaborators who had sequenced them for other studies. These included further isolates from South Africa, as well as samples from two new countries, Germany and Uzbekistan. Training and validation sets were consequently mismatched in terms of size, geography and in terms of the prevalence of resistant phenotypes, potentially introducing bias. However, I also considered that having isolates from different countries in the different sets might minimise the risk of validating against isolates closely genetically related to those isolates alleles were derived from – also potentially a source of bias. Although there were samples from South Africa in both sets as defined, I considered the incidence in South Africa sufficiently high to protect against excessive serendipitous genetic clustering of samples,[131] given that they had

been randomly selected from a drug resistance survey of an entire South African province.

To address the issue of potential bias, I investigated whether the composition of training and validation sets made any material difference to the results. I randomly assigned each isolate to one of two sets, and performed 100 iterations of the algorithm. The proportion of phenotypes predicted was marginally higher on random allocation than for the original validation set (92.7% vs. 89.2%), possibly due to the effect of genetically related isolates having been split between the two sets. The sensitivity and specificity with which predictions were made was however remarkably similar.

I chose a candidate gene approach for two of reasons. First, given the proportion of resistance that could be explained by common alleles already known to the literature,[289,290] I expected any additionally identified resistance determinants to be rare. A genome-wide association study (GWAS) might therefore have lacked the power to identify these. Moreover, given that *M. tuberculosis* is a relatively clonal organism,[291] the lack of recombination would make it difficult for a GWAS to separate causative SNPs from linked genetic variation. What can therefore be identified are, at best, what Chewapreecha calls “large haplotype blocks”.[283] Second, the co-segregation of phenotypic resistance to particular drugs in tuberculosis might result in alleles causing resistance to one drug, being associated with resistance to another drug where the two resistant phenotypes frequently co-occur. Both these concerns were realised when colleagues Chieh-Hsi Wu and Daniel Wilson

subsequently used this same data set for a GWAS (data not included in this thesis).

The selection of candidate genes was clearly central to the outcome of the study. The main, publically available, on-line repository of published genetic data supposedly underlying tuberculosis drug resistance is the 'TB Dream database'.[292] Here the curators have archived all alleles reported in the literature to be associated with resistance, with curatorial effort having gone into distinguishing high from low confidence mutations. Although updates are currently in progress, the database does not yet include alleles identified since 2010. To obtain a more up-to-date database I instead relied on the 'expert opinion' of Silke Feuerriegel and Stefan Niemann at the Forschungszentrum Borstel, Germany, where they had collated an (unpublished) catalogue of resistance determinants based on alleles published in the literature that they considered likely to be causative of resistance. Their curatorial judgement was supported by some of their own, unpublished, in-house laboratory investigations. The candidate genes I chose were the 23 genes containing all the alleles in their catalogue that were relevant to the 11 drugs I had phenotypic data for. I expected the algorithm to find established alleles within these, and to possibly identify new resistance determining alleles too.

Despite the selection of these 23 genes, the training set yielded resistance determinants from just 14 genes, consistent with the other nine genes playing only a minor role in resistance to the drugs studied, or no role at all. Nevertheless, because *M. tuberculosis* lacks the genetic variation of many other bacteria[291] the degree of variation across strains was sufficiently small that

the inclusion of additional genes did not inhibit the identification of resistance determining alleles for the majority of phenotypes.

Key to the success of the algorithm was the observation that the majority of resistant phenotypes only had one non-synonymous, non-lineage defining allele, and that the majority of susceptible phenotypes had none. This facilitated a straightforward characterisation of alleles that meant compensatory alleles were unlikely to be falsely characterised as resistant, but that epistasis will have been missed by design. However, as only 10 of 701 resistant phenotypes in the training set had alleles observed but none characterised as resistance determining, and as six of these contained uncharacterised alleles previously identified in the literature, most resistance in these data appears explainable without epistasis. Such complex genetic mechanisms of resistance are therefore plausibly the exception rather than the rule.

On one level this is not surprising as the well documented utility of LPAs and the Xpert MTB/RIF depend on a straightforward relationship between single point-mutations and phenotype.[195,208,209] However, recent work has also described more complicated mechanisms of resistance to ethambutol. Safi *et. al.* demonstrate the incremental rise in MIC with the accumulation of both synonymous and non-synonymous alleles ($nuoD^{A287S}$, $Rv3792^{L198L}$, $Rv3806c^{A35E}$, $PPE30^{A184A}$, $embB^{T506N}$, $embC^{P398S}$, $embC^{Q491R}$, $embC^{L502P}$) during serial passage of *M. tuberculosis* strains exposed to different ethambutol concentrations, elegantly corroborating these findings through allelic exchange experiments. As these data were derived *in vitro*, it is of course important to investigate whether similar phenomena are replicated *in vivo*. Interestingly,

none of these mutations reported to cumulatively lead to high-level ethambutol resistance are homoplasic in my data, whereas the canonical mutations at $embB^{306}$ that confer low-level resistance arise 36 times independently in the phylogeny.[293] One possible explanation for why the positive selection pressure appears greater on variants conferring low-level resistance is a differential fitness cost, *in vivo* and *in vitro*, for the variants reported by Safi *et. al.*.

Safi *et. al.* argue they are the first to identify synonymous alleles that contribute to drug resistance in *M. tuberculosis*. Searching my data for synonymous alleles that could plausibly play a role in resistance, an allele at $fabG1^{L203L}$ stood out as the only convincing candidate. Overall this allele emerged 12 times independently in the phylogeny and was associated with isoniazid resistance in 16 of 17 isolates in which it was identified. 13 of these 16 resistant isolates contained an alternative catalogued resistance determinant predicting resistance, but three did not, with $fabG1^{L203L}$ plausibly responsible for resistance in those. If synonymous alleles were commonly the cause of drug resistance, this would significantly undermine the justification for my algorithm. However, as this was the only plausible counterexample I was able to identify, I did not consider this sufficient cause to rethink my approach to the analysis.

Whether or not this complex view of drug resistance to ethambutol can be corroborated in vivo, this and similar examples may have limited impact on diagnostic workflows that are based on non-quantitative DST. Under a dichotomous system of phenotypic reporting there is redundancy in identifying additional alleles after a threshold of predicting resistance (the breakpoint) has

been crossed. Whilst molecular assays are predictors of phenotype, and not of patient outcome, it is the ability to predict the former that they will be assessed against. However, where quantitative DST is routinely practiced and correlated to patient outcome, such complex explanations of resistance may have great utility.

Another type of genetic interaction is seen where the presence of the $gyrA^{A90G}$ allele is associated with drug resistance, but its co-existence with the $gyrA^{T80A}$ allele – found in the Uganda genotype – is associated with hyper-susceptibility. As the latter allele is not probed by the HAIN MTBDR*plus* assay, this can lead to false positive resistance predictions.[294] However, despite the existence of at least some complex, cumulative mechanisms of drug resistance and of epistasis, the success of my approach to identifying single resistance determinants suggests that these at least pose no major hurdle to predicting non-quantitative DST.

Despite the successful identification of a large number of resistance determinants and benign alleles, 10.8% of validation phenotypes were either associated with alleles that were present but uncharacterised in the derivation set, or alleles novel to the validation set. That more novel alleles were not present was a consequence of two factors: first, the majority of phenotypes were susceptible, and the majority of susceptible phenotypes were not associated with any non-synonymous, non-lineage defining alleles. Second, the majority of resistant phenotypes contained only one relevant allele, and most phenotypic drug resistance is determined by a small number of alleles that arise

repeatedly and independently across the phylogeny, and hence also across the derivation and validation sets.[271,277]

This second factor underlies the successful predictions for the validation set, even where alleles were derived from UK and sub-Saharan isolates and used to predict phenotypes for isolates from Uzbekistan. It is also the reason commercial LPAs and the Xpert MTB/RIF are diagnostically useful despite the limited number of alleles they probe.[195,208,209,285] Indeed, in one of the more successful demonstrations of genotypic predictions, Rodwell *et. al.* used Sanger sequencing and pyro-sequencing of key genes to predict phenotypes across a global collection of 417 *M. tuberculosis* isolates, demonstrating that 334 of 348 isoniazid resistance phenotypes (96%) could be explained by just three alleles ($katG^{315}$, $fabG1^{-15}$ and $fabG1^{-17}$; the latter not being probed by commercial LPAs).[208,289] However, the pattern varies across drugs and is best exemplified by the difference between isoniazid, where $katG^{S315T}$ alone explained 276 of 310 correctly predicted phenotypes, and pyrazinamide where no dominant alleles were identified. Although 12 of the 34 *pncA* resistance-determinants recurring in validation sequences were associated with pyrazinamide resistance on 44 of 45 occasions in which they were seen, their sensitivity was just 24% overall. The 32 additional resistance-determinants in *pncA* that were derived from the combined 3651 sequences indicate that many rarer alleles underlie pyrazinamide resistance,[295,296] demonstrating the LPA model to be poorly suited to pyrazinamide. However, even for drugs with canonical alleles such as $katG^{S315T}$ in isoniazid, the proportion of resistant phenotypes that are unexplained by these alleles, and hence by LPAs, still

exceed any threshold below which one might consider dispensing altogether with phenotypic DST (Table 13).

Study	Location	Study size	Isoniazid		Rifampicin	
			Sensitivity	Specificity	Sensitivity	Specificity
Hillemann 2007	Germany	125	**92**	**100**	**98.7**	**100**
Miotto 2007	Italy	173	**79.2**	**100**		
Barnard 2008	South Africa	536	**94.2**	**100**	**98.9**	**100**
Akpaka 2008	Triniad and Tobago	81	**34.6**		**95.8**	
Lacoma 2008	Spain	62	**73**	**100**	**91.7**	**100**
Causse 2008	Spain	59	**94.6**		**100**	
Evans 2009	South Africa	223	**82**	**98.9**	**94.5**	**100**
Brossier 2009	France	113	**86.3**		**100**	**100**
Anek-vorapong 2010	Thailand	214	**95.3**	**100**	**100**	**100**
Huyen 2010	Vietnam	111	**92.6**	**100**	**93.1**	**100**
Chryssanthou 2011	Sweden	604	**87.5**	**100**	**100**	**100**
Tolani 2012	India	155	**94.9**	**82**	**98.6**	**94**
Asencios 2012	Peru	95	**97.5**		**100**	
Crudu 2012	Moldova	156	**95.8**	**88.9**	**94.3**	**96**
Jin 2012	China	237	**75**	**100**	**93.5**	**100**
Maschmann 2013	Brazil	68	**60**	**100**	**82**	**94**
Felkel 2013	Nigeria	110	**86**	**100**	**83**	**100**
Ritter (AID) 2013	Switzerland, South Africa	156	**100**	**100**	**100**	**100**
Ocheretina 2013	Haiti	153			**100**	**89.5**
Rodwell 2014	India, Philippines, Moldova, South Africa	417	**96**	**97.3**	**98**	**80**
Raizada 2014	India	248	**72**	**97**	**93**	**94**
Aurin 2014	Bangladesh	300	**99.5**	**98.8**	**100**	**100**
Chen 2014	China	427	**76.5**	**95.4**	**85.9**	**93.1**
Molina-Moya 2014	Spain	65	**97.8**	**100**	**100**	**100**
		Total: 4888				
		Mean:	**87.20%**	**98%**	**96.20%**	**96.40%**

Table 13: My own literature review for the studies assessing the MTBDR*plus* LPA

The characterisation of all alleles offers advantages over LPAs and other commercial molecular assays. First, WGS data can be screened for all resistance determinants, resulting in a higher sensitivity than for the LPA based alleles alone.[208,209] Second, whereas LPAs can suggest which drugs to avoid by screening a few key resistance-determining alleles, they leave some doubt about which to give. By characterising alleles as benign, I can actively predict phenotypic susceptibility in some isolates, contrasting them from others containing uncharacterised alleles. Third, WGS based DST can be performed for additional and even novel drugs at no additional cost, contingent only on the

knowledgebase of characterised alleles. This could be helpful when designing new treatment regimens.[114,115]

No resistance determinant was identified for 43 resistant phenotypes in the training set, of which 33 contained no candidate alleles at all, suggesting either sample mislabelling, phenotypic error, or an alternative mechanisms of resistance elsewhere in the genome. It has been argued that the presence of the high-level resistance determinant $katG^{S315T}$ in isoniazid susceptible phenotypes indicates sample mislabelling, but the rate for these data, 8 of 480 (1.9%), compares favourably with previous reports,[271,297] and is too low to account for all the discordance. The imperfections in phenotypic DST are well documented,[78,298] and best illustrated in this data by the weak association between $embB^{M306I}$ and ethambutol resistance.[287] Among first-line drugs, susceptibility to pyrazinamide is also notoriously difficult to assay accurately as it is only active against *M. tuberculosis* in an acidic environment that is in turn unfavourable to the growth of *M. tuberculosis*.[296,299] Unfortunately there was little opportunity to retest phenotypes and no opportunity to perform quantitative DST, which may have at least addressed some of the problems concerning ethambutol.

Given the size of the data set, the algorithm was robust to some level of mislabelling or error in phenotypic DST for genuinely resistant isolates by implicitly up-weighting single observations of resistance over single observations of susceptibility. New samples could lead to the re-characterisation of alleles from benign to resistant, although rarely vice versa for reasons outlined in the methods section above. Nevertheless, as base-calling

from WGS data is highly reproducible,[300] the phenotypic variability around some alleles, whether due to phenotypic error or mislabelling, will become apparent in large data sets where predictions could be recast within a Bayesian analysis framework.[301]

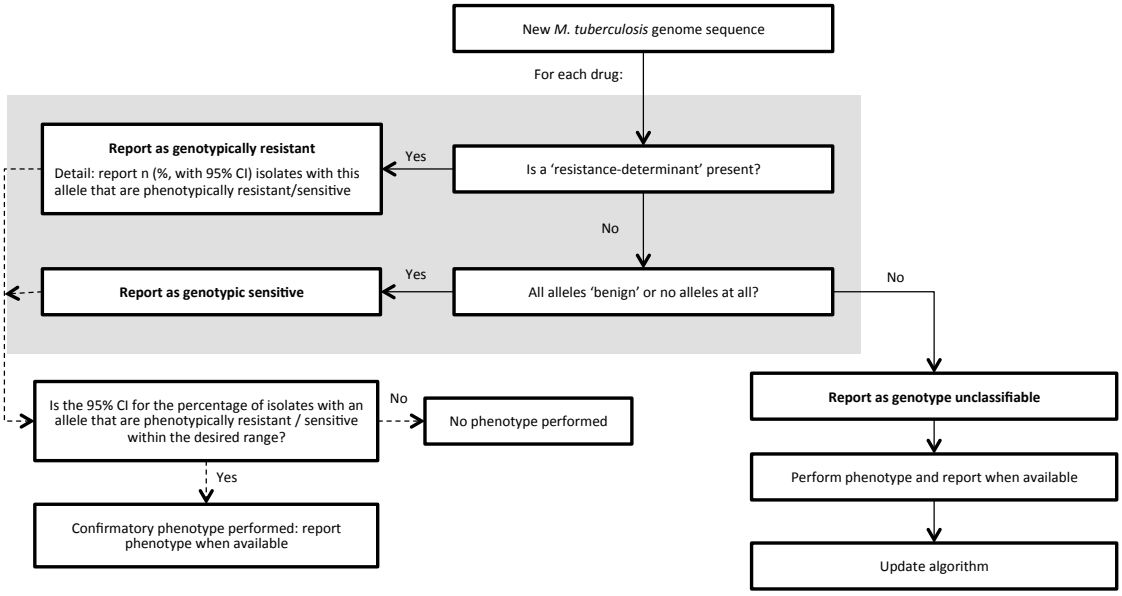


Figure 32: Suggested work flow integrating WGS based DST into routine laboratory diagnostics

Figure 32 however suggests an alternative, simpler approach to integrating the algorithmic characterisations into a routine laboratory workflow. Here, WGS based phenotypic predictions would be made for isolates containing resistance determinants, only benign alleles, or no relevant alleles. As data accrue, confidence in the characterisation of each allele will grow to the point where routine phenotyping can be restricted to isolates containing uncharacterised alleles that inhibit phenotypic prediction.

I explored the possibility of resistance mechanisms existing outside of the 23 candidate genes and their upstream regions, as defined here, by screening the

genome for homoplasy, which can represent adaptation to a positive selection pressure such as from drugs. Farhat *et. al.* also used homoplasy to signal selection pressure and associate genome-wide alleles with resistance in their study of 123 sequenced genomes.[277] Among the alleles identified were two in the ponA1 gene (at nucleotide positions 123 C→G and 1095 G→T) which they performed allelic exchange experiments on to confirm their role in rifampicin resistance. Neither of these alleles were homoplasic in my data, although other alleles in ponA1 were (Table 14). Interestingly, despite their homoplasy, none were strongly associated with phenotypic resistance unless another resistance determining alleles were already present.

Phenotypes for *ponA1* gene associated with rifampicin resistance:

Variant	Maximum number of emergences per variant	Other resistance-determinant identified in candidate-gene	Phenotypes (n)	
			Resistant	Sensitive
ponA1_S631S (CCG-TGC)	43	Yes	11	2
		No	4	420
ponA1_G473G (GGT-GGC)	2	Yes	5	0
		No	0	1
ponA1_D24N (GAC-ACC)	2	Yes	0	0
		No	0	5

Table 14: ponA1 gene alleles and their association with rifampicin resistance

The data set presented in this chapter is almost 30 times the size of that published by Farhat *et. al.*. Given the finite variation that is possible within a genome, homoplasy will inevitably accumulate with the number of different isolates sequenced. In this chapter the homoplasic sites outside of the 23 candidate genes were largely associated with drug susceptibility, suggesting either stronger signature for selection from factors other than drug pressure, or the inevitable accumulation of 'low level' homoplasy (variants emerging independently on only a small number of occasions) that flows from a larger

data set. Given the negative slope on the mean in Figure 30, both these factors may be relevant to my data. Although I therefore failed to identify further genomic variation underlying resistance by using homoplasy as a signal, I was able to show that the characterised resistance determinants were enriched for homoplasy, consistent with their algorithmic characterisation.

One of the advantages of my algorithm is the ability to characterise alleles not only as resistance determinants, but also to distinguish benign from unknown alleles. Although LPAs are able to predict ‘wild-type’ or ‘mutant’ at a small number of sites of interest,[208,209] a clean panel of wild-types only implies the absence of resistance by those mechanisms and leaves open the potential for additional, uncharacterised genetic variation. In a system that predicts phenotypes three ways, as benign, resistant or unknown, phenotypic DST can be restricted to the unknown category rather than being required for all strains that are not resistant.

Starting at the Public Health England Public Health Laboratory for the West Midlands, Public Health England is performing WGS in parallel to current workflows to assess its suitability as a one-stop diagnostic platform for mycobacterial infection. Parallel phenotypic DST will confirm the status of some, and identify further alleles, eventually, restricting phenotypic DST to isolates containing uncharacterised alleles, or where insufficient phenotypic data are available to characterise alleles confidently (Figure 32), reducing usage over time. One benefit provided at no additional cost will be to quantify second-line drug resistance prevalence in a population, or indeed any pre-existing resistance to novel drugs. As WGS data accumulates globally, our

understanding of what determines resistance at a molecular level will expand, advancing towards a system where cheaper and faster WGS based DST is in a position to gradually replace the sometimes-variable phenotypic assays.

6 CONCLUSIONS AND FUTURE WORK

In January 2015 Public Health England published their national tuberculosis strategy document motivated by incidence rates that remain among the highest in western Europe.[302] It argues that devolution of responsibilities to local 'TB control boards' would improve diagnostic and treatment services, as well as access to them, and would lead to more comprehensive contact tracing, new-entrant screening and vaccination programmes.

Whatever the merits of localism versus centralised control of this area, the desired outcomes are clearly sensible. The World Health Assembly has set a target to end the global tuberculosis epidemic by 2035, reducing the global incidence rate to fewer than 10 per 100,000 population, 15 years earlier than the WHO's target to eliminate tuberculosis by 2050 (defined as fewer than 1 case per million people per year).[131,303]

All bodies state the urgent need for research and innovation if these targets are to be met,[131,302] but it is worth considering where and how new technology can contribute to the attainment of these goals. The successful reduction in tuberculosis incidence in New York City is one of the best documented case studies in recent times. From the mid-1970s to 1992, incidence rates soared, reaching 222 per 100,000 population in Harlem.[304] Although increasing inner-city social deprivation and the HIV pandemic have been identified as factors leading to the significant rise in cases, public health interventions managed to reverse the trend in tuberculosis incidence without directly addressing these causative factors. Frieden *et. al.* argued that the decline could be largely attributed to an aggressive DOTS policy in the city, alongside better infection

control in hospitals, the closure of large-scale shelters (some accommodating over 800 individuals in a single room), and better medical services in prisons. Although they estimate that this came at the price of a ten-fold increase in the public health budget, the 6-fold decrease in caseload over the following 10 years is likely to have saved significantly more money.[304,305]

It is important to recognise that the success seen in New York City followed investment in public health services, and was not a consequence of either diagnostic advances or new drug treatments (although Frieden *et. al.* do point to the introduction of the now standard quadruple therapy regimen 6HR/2EZ as a possible contributing factor). Interventions were instead informed by an understanding that tuberculosis transmission is facilitated by overcrowding and poor treatment completion rates, not major new insights into the disease. Robin Wood has similarly argued that interrupting transmission in places of compulsory indoor aggregation such as schools and prisons would have a significant impact on tuberculosis incidence in South Africa. The data linking such poorly ventilated indoor spaces to high rates of transmission exist, with one model estimating a 90% reduction in transmission in prisons by simple measures such as reducing overcrowding and improving airflow.[150,306] Transmission in South African mines is also very well described.[164]

Although technological progress may be a necessary condition to achieving WHO targets, it is therefore far from sufficient. The roll out of the Cepheid Xpert MTB/RIF constitutes an important case study of both the advantages and limitations of expensive, high-tech diagnostic platforms. It is clear that the technology works and that it offers additional, valuable diagnostic information to

low-income settings that currently rely on smear microscopy alone.[195,307,308] However, investigations into the overall impact on patient outcome have yielded mixed results. In a multi-centre implementation study, Catharina Boehme *et. al.* found that Xpert MTB/RIF shortened time to treatment for smear negative tuberculosis in particular,[307] whereas at the Conference on Retroviruses and Opportunistic Infections 2014, Katherine Fielding and Gavin Churchyard reported how their XTEND trial evaluating the impact of GeneXpert versus two sputum smear microscopy tests in South Africa found no shortening of time to treatment, no improvement in follow-up rates, and no decrease in mortality (full report yet to be published).[309] They plausibly argue that the disappointing results are because of weak healthcare infrastructure rather than any failure on behalf of the Xpert MTB/RIF. Without adequate resources to follow-up patients or drugs to treat MDR tuberculosis, it is hard to obtain benefit from new diagnostic platforms.

The Xpert MTB/RIF also requires some basic laboratory infrastructure including an uninterrupted power supply and an ambient temperature not exceeding 30^{o}C,[196] militating against its application as a point-of-care test in low-income rural settings. In a review of implementation in 22 high burden countries, Qin *et. al.* identify that most Xpert MTB/RIF machines are based in reference or district laboratories, rather than nearer the patient,[310] with clear consequences for turnaround time and follow-up rates.

What the New York City and South African experiences both demonstrate is the clear need for investment in basic healthcare and public health infrastructure if the tuberculosis incidence is to be significantly reduced. However, whilst

measures to interrupt on-going transmission were highly effective in New York City and could be equally so in South Africa, given sufficient resources, the large reservoir of latently infected individuals will be harder to eradicate and offers the unfortunate prospect of new cases, and hence outbreaks, emerging sporadically over the decades to come.

In chapter 4 I describe the patterns of transmission in Oxfordshire, where tuberculosis is comparatively rare, where the majority of new cases are in patients born in high-incidence countries but where the risk of transmission is greater among patients born in low-incidence countries (mainly the UK). Most clusters are small and genetically unrelated to one another, consistent with the sporadic emergence of disease in latently infected individuals, or disease imported from outside of the region. Most outbreaks were linked to households, but also to a school, a prison and a homeless shelter. Whereas closing large shelters for the homeless was an obvious intervention in New York City, closing all homeless shelters is not an option, nor is closing schools or prisons. Public health minded architecture creating segregated or better-ventilated rooms would clearly be ideal, but the cost of large-scale redesign of public buildings in a low-incidence setting such as Oxfordshire would be hard to justify. More targeted interventions guided by the high-resolution data produced by WGS may therefore be preferable, and possibly more cost effective.

This is not to argue that no role exists for WGS in high-incidence settings, only that there may be more pressing interventions that are not technology driven. Of course, where uncertainty exists, WGS may be able help identify where transmission is occurring, regardless of the incidence. The results described in

chapter 4 have helped clarify how the reservoir of latent infection among individuals born in high-incidence settings has not contributed towards a large amount of local transmission in Oxfordshire. Similarly, if all tuberculosis strains in South Africa were to be sequenced, the relative contributions to the overall epidemic of transmission in the home, school, the mines, prisons or elsewhere would become more apparent.

The data underlying the SNP thresholds described in chapter 3 has been replicated in other settings, and could reasonably guide targeted public health interventions in the future.[236,238,239,242,250,253] If outbreaks can thus be identified with greater certainty, then more targeted screening, case finding and prophylaxis could contribute to the further decline in tuberculosis transmission, and eventually in its incidence too. If tuberculosis is to be eliminated, in line with WHO targets, then the reservoir of latently infected individuals cannot be allowed to contribute towards new infections, and infectious relapsed cases will require early diagnosis.

Early diagnosis requires early presentation, recognition of symptoms and efficient, accurate assays. Although the first two factors are dependent on patient behaviour and doctor awareness, technological developments are key to speeding up diagnostic assays. Whereas the Xpert MTB/RIF identifies the *M. tuberculosis complex* and the presence or absence of mutations in the rifampicin resistance-conferring region of rpoB, most other diagnostic assays currently have an intrinsic, culture-dependent delay. It might be argued that the HAIN MTBDR*plus* and MTBDR*sl* LPAs can also be performed on primary samples, but their readability is impaired by low bacillary loads, and they are

more reliably performed on cultured samples.[311,312] Nevertheless, to obtain the full diagnostic information, including species, DST to all required drugs and typing results, culture remains necessary and the delay can extend to weeks.[189]

The prospect of therefore obtaining all diagnostic information from a single test in the form of WGS data is very appealing given how a number of studies have already demonstrated that MiSeq platforms can be used to rapidly provide clinically useful data within days of sample collection for a variety of pathogens.[229,313] *M. tuberculosis* has already been successfully sequenced from MGIT culture,[275] and a recently published protocol now means this can be done reliably from the point at which MGIT cultures flag positive, potentially yielding all diagnostic information within 7 to 10 days of sample collection.[314] Work on the extraction, purification and sequencing of nucleic acid from primary clinical samples is on-going, but the first report identifying *M. tuberculosis* from such as sample was published in 2014.[315]

Sequencing platforms are themselves evolving, with the latest Oxford Nanopore release promising to sequence a human genome in 15 minutes,[316] and presumably significantly less time for a mycobacterial genome ($3.3x10^9$ vs. $4.4x10^6$ base pairs). This would not only be faster than the Xpert MTB/RIF, which takes 2 hours,[195] but would produce all the required diagnostic information at once. Given that Oxford Nanopore's MinION platform is portable – the size of a smart phone – and that it can be powered by a laptop[317] that would not require a continuous, uninterrupted power supply and might even be powered by solar energy,[318] the prospect of a genuine field-based, random

access, point-of-care test that provides full diagnostic information within minutes is coming into view. The WHO, who backed the roll-out of Xpert MTB/RIF, have described that process as a ‘test case’ for the translation of new technology for patient benefit.[196] Although the Oxford Nanopore platforms still require some refinement before fulfilling their promise,[183,319] the success of these or similar platforms will be dictated as much by price as by their eventual utility. With a human genome sequence expected to initially cost $1,000, further reductions in cost are likely to lead to wide-spread adoption.[317] Pathogen sequences are massively smaller and the cost therefore likely to be less. At concessional cost, Xpert MTB/RIF cartridges are currently sold at around $10,[320] although the cumulative cost to high-burden countries remains high.[321] At a similar price point a MinION-like device would offer more information, faster, and might have a more realistic chance of actually being deployed as a point-of-care test.

The cost effectiveness of passive case finding, household based contact investigations, and active case finding have recently been assessed in Kampala, with household targeted contact investigations being identified as the most cost effective.[322] The high prior probability of finding additional active or latent cases within households containing patients with a diagnosis of tuberculosis raises the question of whether contact investigations targeted by WGS might have a similar prior probability of success, and hence similar cost effectiveness in low-income settings. This would be an important study for the future. However, if WGS is implemented widely as a diagnostic assay, the driving force is more likely to be the need for fast and accurate DST, and not outbreak investigation. The Xpert MTB/RIF was designed as a screening test

for MDR tuberculosis, facilitating point-of-care based decisions on whether to include isoniazid and rifampicin in treatment regimens or not. If a mutation in rpoB is detected, rifampicin resistance is assumed and isoniazid resistance inferred, but unlike WGS data, the result cannot help clinicians decide between alternative treatment regimes.

The results presented in chapter 5 demonstrate that WGS data can already predict DST for a wider range of drugs, and across a wider range of alleles relevant to those drugs, than either LPAs or Xpert MTB/RIF. Current understanding of the genetic causes of drug resistance is therefore already sufficiently great that in a hypothetical future where bedside WGS is available, the results could provide clinicians with significantly more data to base treatment decisions on than is currently provided by the Xpert MTB/RIF.

Where phenotyping is currently performed, a higher level of understanding of the genetic determinants of resistance is required before justifiably replacing it completely with WGS. The tuberculosis research community is working towards compiling a catalogue of genetic variation encoding all, or nearly all, phenotypic resistance to each anti-tuberculosis drug. The work outlined in chapter 5 presents one approach that involves targeting a small number of candidate genes. Although largely successful, it failed to identify any plausible explanatory variants for a number of resistant phenotypes, leaving open the possibility that they are encoded elsewhere in the genome. To find these, a strategic grant has been submitted (Derrick Crook, Principal Investigator) to the Wellcome Trust to facilitate the sequencing and analysis of up to 100,000 *M. tuberculosis* genomes from across the globe, with large scale GWAS and machine learning

techniques planned. Genotyping will thus be assessed for its capacity to predict phenotypes, although not as a measure of patient outcome for which additional studies are required. Chapter 5 describes a schema whereby phenotyping could be progressively phased out as the accuracy of genotyping increases. Eventually the hope is that all genetic variation can be tightly associated with MIC.

It is unlikely that phenotyping can be entirely dispensed with even in the most optimistic of scenarios because of the need to monitor for novel alleles and their implications. Once sufficient overall sensitivity and specificity for predicting phenotypes has been achieved, it may not be necessary to continue phenotyping all isolates with a novel mutation, but it may be worth phenotyping strains with novel mutations from patients suffering treatment failure. To maintain the requisite expertise in phenotyping after the throughput has fallen significantly, one could further centralise the process, making use of the WHO network of supra-national reference laboratories. These would then serve to monitor trends and respond to the emergence of new clones or resistance determinants, and have the responsibility of updating the existing catalogue of resistance determinants.

If the eradication of tuberculosis depends on the prevention of transmission, then rapid, accessible and complete molecular DST will contribute by enabling more precise tailoring of treatment to individual patients, thereby preventing onward transmission of all forms of tuberculosis, including, significantly, drug-resistant disease. If the data in chapter 4 and that from other studies showing how new migrants reactivate their latent tuberculosis within the first few years of

arrival in a new country are representative of the incubation period of tuberculosis,[263] then 5 to 10 years of an aggressive package of preventing transmission through early diagnosis, effective treatment, identifying outbreaks and addressing the indoor environments that facilitate transmission has a chance of successfully meeting WHO targets.

In this thesis I have argued that WGS technology can contribute to each of these strands of work. Disease control includes outbreak detection and intervention, as well as accurate diagnosis and effective treatment. In chapter 3 I outline a method for relating genome sequences to one another to identify and describe transmission events. In chapter 4 I apply these findings to understand the relative contributions of recent transmission and reactivated disease to overall incidence in a low burden western European setting, and suggest possible policy interventions. In chapter 5 I devise an algorithm to identify the molecular determinants of drug resistance that would underlie future WGS based DST, which in turn has the potential to contribute significantly towards the rapid and accurate diagnosis and treatment of both drug susceptible and drug resistant tuberculosis, thereby decreasing transmission.

These findings have been a significant factor influencing PHE to invest in a pilot study into the use of WGS as a diagnostic assay in the Midlands, and will form the basis of relevant analyses within that study. Over the course of 2015 PHE will be assessing the potential of MiSeq platforms to sequence mycobacterial cultures from early positive MGITs, as well as the potential of an Oxford based bioinformatics infrastructure to identify species, predict DST for members of the *M. tuberculosis complex*, and identify their nearest genetic neighbour from a

back catalogue of over 3,000 UK based sequences. If the approach can successfully obtain accreditation after the one year long pilot, the plan is to roll the technology out across England, replacing components of the current workflow such as MIRU-VNTR typing and the LPAs.

The knowledge and techniques developed in thesis will be built on during the pilot study, hopefully contributing towards a knowledge and experience base that can be shared with and applied to alternative settings in the future.

7 APPENDIX

7.1 S1

Uncharacterised alleles that only occur in resistant phenotypes

Alleles	Drug	Alleles	Drug
ahpC_54_insGT	INH	rpoB_389_insATGTCGT	RIF
ahpC_C-54T	INH	rpoB_A538V	RIF
ahpC_D73H	INH	rpoB_I480V	RIF
ahpC_E76K	INH	rpoB_P45S	RIF
ahpC_F108S	INH	rpoB_Q975H	RIF
ahpC_G-48A	INH	rpoB_R219C	RIF
ahpC_G-74A	INH	rpoB_S874F	RIF
ahpC_G32D	INH	rpoB_V168A	RIF
ahpC_T-42C	INH	rpoB_V496A	RIF
ahpC_W96C	INH	embA_68_delC	EMB
fabG1_T-8A	INH	embA_A813G	EMB
fabG1_T-8G	INH	embA_C-11A	EMB
inhA_I21V	INH	embA_C-8T	EMB
inhA_T162S	INH	embA_G759R	EMB
katG_24_insCA	INH	embB_G156C	EMB
katG_A-75G	INH	embB_M423T	EMB
katG_A109T	INH	embB_W1089R	EMB
katG_A244G	INH	embC_A387V	EMB
katG_C549S	INH	embR_458_delG	EMB
katG_D542E	INH	iniA_N50D	EMB
katG_G34A	INH	iniA_V544A	EMB
katG_L427F	INH	iniC_G194R	EMB
katG_Q471R	INH	iniC_G341V	EMB
katG_R78G	INH	gidB_504_delC	SM
katG_S211N	INH	tlyA_C-83T	SM
katG_T579A	INH	gyrA_D641E	CIP
katG_W149R	INH	gidB_L108P	AK
katG_Y155S	INH	gidB_L108P	CAP
ndh_D143A	INH	gidB_L108P	KAN
ndh_K208E	INH		
ndh_K32E	INH		

7.2 S2

Alleles previously identified in the literature as resistance determinants. The 'literature' was defined as any allele listed in the Dream TB database project (https://tbdreamdb.ki.se/Info/), any allele contained on one of the three line-probe assays (which also cover those identified by the Cepheid MTB/RIF Xpert). An additional manual search was also performed for any of the 120 resistance determinants that did not match any of these sources. Analysing the literature like this of course introduces bias, as I look harder for evidence of the resistance determining alleles in the literature than I do for the other alleles. However, that lineage-defining alleles have been characterised as resistance determinants in the published literature at all is evidence of noise, only more of which would be found by a further manual trawl for alleles. Each allele is counted for each of the drugs it is associated with (i.e. some alleles are counted more than once).

Allele	Drug	Characterisation
pncA_S104R	PZA	Resistance-determinant
gidB_R137W	SM	Resistance-determinant
katG_W328L	INH	Resistance-determinant
pncA_K96T	PZA	Resistance-determinant
rpoB_H445R	RIF	Resistance-determinant
rpsL_K43R	SM	Resistance-determinant
rpoB_S450F	RIF	Resistance-determinant
rpoB_D435V	RIF	Resistance-determinant
gyrA_D94A	OFX	Resistance-determinant
fabG1_T-8C	INH	Resistance-determinant
embB_M306I	EMB	Resistance-determinant
gidB_A200E	SM	Resistance-determinant
pncA_C138R	PZA	Resistance-determinant
rpoB_Q432K	RIF	Resistance-determinant
pncA_V180F	PZA	Resistance-determinant
embB_G406D	EMB	Resistance-determinant
rpoB_D435F	RIF	Resistance-determinant
pncA_W68C	PZA	Resistance-determinant

embB_D354A	EMB	Resistance-determinant
gyrA_S91P	MOX	Resistance-determinant
inhA_S94A	INH	Resistance-determinant
embB_Q497R	EMB	Resistance-determinant
pncA_V139L	PZA	Resistance-determinant
gyrA_D94G	CIP	Resistance-determinant
fabG1_C-15T	INH	Resistance-determinant
pncA_L4S	PZA	Resistance-determinant
rpoB_H445D	RIF	Resistance-determinant
katG_S315N	INH	Resistance-determinant
rpoB_S450W	RIF	Resistance-determinant
pncA_G162D	PZA	Resistance-determinant
pncA_D8N	PZA	Resistance-determinant
gyrA_D94N	MOX	Resistance-determinant
pncA_C14R	PZA	Resistance-determinant
rrs_A1401G	KAN	Resistance-determinant
pncA_A-11G	PZA	Resistance-determinant
rpoB_S431G	RIF	Resistance-determinant
rpsL_K88R	SM	Resistance-determinant
pncA_Q141*	PZA	Resistance-determinant
pncA_L172P	PZA	Resistance-determinant
gyrA_S91P	OFX	Resistance-determinant
pncA_G97D	PZA	Resistance-determinant
rpoB_L452P	RIF	Resistance-determinant
pncA_G132D	PZA	Resistance-determinant
fabG1_G-17T	INH	Resistance-determinant
rrs_A514C	SM	Resistance-determinant
rrs_C517T	SM	Resistance-determinant
katG_T180K	INH	Resistance-determinant
rrs_A1401G	CAP	Resistance-determinant
gidB_A134E	SM	Resistance-determinant
gyrA_D94N	OFX	Resistance-determinant
katG_S315T	INH	Resistance-determinant
pncA_H57D	PZA	Resistance-determinant
rpoB_I491F	RIF	Resistance-determinant
embB_M306V	EMB	Resistance-determinant
gyrA_D94A	MOX	Resistance-determinant
gidB_A80P	SM	Resistance-determinant
rrs_A1401G	AK	Resistance-determinant
rpoB_H445N	RIF	Resistance-determinant
pncA_D12A	PZA	Resistance-determinant
embB_G406S	EMB	Resistance-determinant
inhA_I194T	INH	Resistance-determinant
embB_G406A	EMB	Resistance-determinant
gyrA_A74S	CIP	Resistance-determinant
rpoB_S450L	RIF	Resistance-determinant
pncA_D8G	PZA	Resistance-determinant
gyrA_A90V	MOX	Resistance-determinant
rpoB_H445Y	RIF	Resistance-determinant

inhA_I21T	INH	Resistance-determinant
gyrA_S91P	CIP	Resistance-determinant
pncA_D136N	PZA	Resistance-determinant
pncA_L27P	PZA	Resistance-determinant
pncA_V125G	PZA	Resistance-determinant
eis_G-10A	KAN	Resistance-determinant
gyrA_D94G	OFX	Resistance-determinant
katG_W191R	INH	Resistance-determinant
pncA_Q10*	PZA	Resistance-determinant
gyrA_D94G	MOX	Resistance-determinant
embB_Q497K	EMB	Resistance-determinant
gyrA_A90V	OFX	Resistance-determinant
embR_C110Y	EMB	Lineage-defining
katG_R463L	INH	Lineage-defining
rmlD_S257P	EMB	Lineage-defining
embC_V981L	EMB	Lineage-defining
manB_D152N	EMB	Lineage-defining
gidB_L16R	SM	Lineage-defining
gyrA_S95T	OFX	Lineage-defining
embC_N394D	EMB	Lineage-defining
embC_R738Q	EMB	Lineage-defining
ndh_V18A	INH	Lineage-defining
embA_P913S	EMB	Lineage-defining
inhA_V78A	INH	Lineage-defining
gyrA_S95T	CIP	Lineage-defining
gyrA_T80A	MOX	Lineage-defining
gyrA_S95T	MOX	Lineage-defining
gyrA_T80A	OFX	Lineage-defining
gidB_E92D	SM	Lineage-defining
embB_E378A	EMB	Lineage-defining
embC_T270I	EMB	Lineage-defining
gyrA_T80A	CIP	Lineage-defining
iniC_P248A	EMB	Benign
ahpC_C-52T	INH	Benign
gyrA_D94H	MOX	Benign
rpoB_L430P	RIF	Benign
pncA_Y64D	PZA	Benign
pncA_V21G	PZA	Benign
iniA_S501W	EMB	Benign
pncA_L35R	PZA	Benign
ndh_R268H	INH	Benign
gyrA_D94H	CIP	Benign
pncA_A146T	PZA	Benign
embA_A201T	EMB	Benign
pncA_T47A	PZA	Benign
embB_M306L	EMB	Benign
gyrA_A90G	CIP	Benign

rpoB_D435Y	RIF	Benign
gyrA_A90G	OFX	Benign
embC_A307T	EMB	Benign
embB_F285L	EMB	Benign
pncA_V139A	PZA	Benign
rmlD_G-71T	EMB	Benign
katG_T394A	INH	Benign
embA_C-16T	EMB	Benign
gyrA_A90G	MOX	Benign
embB_L370R	EMB	Benign
embB_P430L	EMB	Benign
pncA_H71Y	PZA	Benign
pncA_T87M	PZA	Benign
gyrA_D94H	OFX	Benign
katG_M257I	INH	Benign
gidB_S70R	SM	Benign
embB_M306T	EMB	Uncharacterised
fabG1_T-8A	INH	Uncharacterised
rpsL_K88T	SM	Uncharacterised
gyrA_A90V	CIP	Uncharacterised
embA_C-12T	EMB	Uncharacterised
rpoB_I480V	RIF	Uncharacterised
ahpC_E76K	INH	Uncharacterised
fabG1_T-8G	INH	Uncharacterised
embA_C-11A	EMB	Uncharacterised
ahpC_G-74A	INH	Uncharacterised
inhA_I21V	INH	Uncharacterised
katG_D695A	INH	Uncharacterised
ahpC_G-48A	INH	Uncharacterised
ahpC_C-54T	INH	Uncharacterised
katG_Y155S	INH	Uncharacterised
ahpC_D73H	INH	Uncharacterised

7.3 S3

120 resistant determinants derived from the training set. Each allele can be counted more than once if it is characterised as a resistance determinant to more than one drug. The number of resistant and susceptible phenotypes with which each resistance determinant is associated is shown, and a pubmed identifier is given if the allele has been described as resistance determinants before.

Alleles	Drug	Resistant	Susceptible	Total	Pubmed link to reference
ahpC_C-57T	INH	1	0	1	
eis_G-10A	KAN	4	0	4	21300839[uid]
embB_D354A	EMB	1	1	2	20427375[uid]
embB_G406A	EMB	1	3	4	10639358[uid]
embB_G406D	EMB	2	2	4	10639358[uid]
embB_G406S	EMB	3	5	8	10639358[uid]
embB_H1002R	EMB	1	0	1	22646308[uid] **
embB_M306I	EMB	20	34	54	9257740[uid]
embB_M306V	EMB	18	3	21	9257740[uid]
embB_Q497K	EMB	1	2	3	10639358[uid]
embB_Q497R	EMB	8	4	12	10639358[uid]
fabG1_C-15T	INH	73	6	79	10428945[uid]
fabG1_G-17T	INH	5	1	6	14638486[uid]
fabG1_T-8C	INH	3	3	6	15793126[uid]
gidB_141_delC	SM	1	0	1	
gidB_202_delG	SM	4	7	11	
gidB_202_insGC	SM	1	1	2	
gidB_A134E	SM	1	0	1	17238915[uid]
gidB_A138T	SM	1	0	1	
gidB_A138V	SM	2	1	3	
gidB_A200E	SM	2	0	2	17238915[uid]
gidB_A80P	SM	1	0	1	24102832[uid]
gidB_G69D	SM	3	0	3	
gidB_H48N	SM	2	0	2	
gidB_L91P	SM	1	0	1	
gidB_P75L	SM	1	1	2	
gidB_R137W	SM	1	1	2	21444711[uid]
gidB_S70N	SM	1	0	1	
gidB_V65G	SM	1	0	1	
gidB_V88A	SM	2	0	2	
gyrA_A74S	CIP	2	0	2	17035499[uid]
gyrA_A90V	MOX	2	0	2	8031045[uid]

gyrA_A90V	OFX	2	0	2	8031045[uid]
gyrA_D94A	MOX	1	0	1	8031045[uid]
gyrA_D94A	OFX	1	0	1	8031045[uid]
gyrA_D94G	CIP	18	1	19	8031045[uid]
gyrA_D94G	MOX	9	4	13	8031045[uid]
gyrA_D94G	OFX	9	3	12	8031045[uid]
gyrA_D94N	MOX	1	0	1	8031045[uid]
gyrA_D94N	OFX	1	0	1	8031045[uid]
gyrA_S91P	CIP	2	2	4	8031045[uid]
gyrA_S91P	MOX	3	0	3	8031045[uid]
gyrA_S91P	OFX	3	0	3	8031045[uid]
inhA_I194T	INH	2	0	2	16495272[uid]
inhA_I21T	INH	2	0	2	14638486[uid]
inhA_S94A	INH	4	1	5	10815738[uid]
katG_1450_delC	INH	1	0	1	
katG_1910_delA	INH	1	0	1	
katG_471_delG	INH	2	0	2	
katG_L159P	INH	1	0	1	
katG_S315N	INH	2	1	3	9210694[uid]
katG_S315T	INH	195	1	196	8537659[uid]
katG_T180K	INH	1	0	1	16870753[uid]
katG_V633A	INH	1	1	2	
katG_W191R	INH	1	0	1	15793126[uid]
katG_W300C	INH	1	0	1	22646308[uid] **
katG_W328L	INH	1	0	1	9210694[uid]
katG_W90R	INH	1	0	1	
pncA_177_delC	PZA	1	0	1	
pncA_292_insAT	PZA	1	0	1	
pncA_409_delT	PZA	2	0	2	
pncA_491_insACC	PZA	1	0	1	
pncA_494_delC	PZA	1	0	1	
pncA_528_insGGCCGTCTGGC	PZA	1	0	1	
pncA_570_insCT	PZA	2	0	2	
pncA_A-11G	PZA	3	0	3	9056006[uid]
pncA_C138R	PZA	1	0	1	25336456[uid]
pncA_C14R	PZA	1	0	1	9055989[uid]
pncA_D12A	PZA	1	0	1	9055989[uid]
pncA_D136N	PZA	1	2	3	11641519[uid]
pncA_D49N	PZA	1	0	1	
pncA_D8G	PZA	1	0	1	11641519[uid]
pncA_D8N	PZA	1	0	1	25336456[uid]
pncA_G132D	PZA	1	0	1	9692180[uid]
pncA_G162D	PZA	1	0	1	17360809[uid]
pncA_G78C	PZA	1	0	1	
pncA_G97D	PZA	1	0	1	11083630[uid]
pncA_H57D	PZA	11	0	11	9056006[uid]
pncA_H57R	PZA	2	0	2	
pncA_K96T	PZA	1	0	1	9055989[uid]
pncA_L172P	PZA	1	0	1	9692180[uid]

pncA_L27P	PZA	1	0	1	17596354[uid]
pncA_L4S	PZA	4	0	4	11083630[uid]
pncA_Q10*	PZA	1	0	1	10390239[uid]
pncA_Q141*	PZA	1	0	1	11641519[uid]
pncA_S104R	PZA	2	0	2	9692180[uid]
pncA_T-12C	PZA	3	0	3	
pncA_V125G	PZA	4	0	4	18573039[uid]
pncA_V139L	PZA	1	0	1	11083630[uid]
pncA_V180F	PZA	1	0	1	11641519[uid]
pncA_V7L	PZA	3	0	3	
pncA_W68C	PZA	1	0	1	25336456[uid]
rpoB_1396_insATTC	RIF	2	0	2	
rpoB_1427_delTGGCCCC	RIF	1	0	1	
rpoB_D435F	RIF	1	0	1	16229229[uid]
rpoB_D435V	RIF	4	0	4	15184414[uid]
rpoB_H445D	RIF	6	0	6	14729930[uid]
rpoB_H445N	RIF	1	4	5	14729930[uid]
rpoB_H445R	RIF	3	0	3	9003625[uid]
rpoB_H445Y	RIF	12	0	12	8027320[uid]
rpoB_I491F	RIF	4	19	23	10565894[uid]
rpoB_L452P	RIF	4	4	8	10921994[uid]
rpoB_Q432K	RIF	1	0	1	8027320[uid]
rpoB_S431G	RIF	1	0	1	17360809[uid]
rpoB_S450F	RIF	4	0	4	15728936[uid]
rpoB_S450L	RIF	59	1	60	7946393[uid]
rpoB_S450W	RIF	3	1	4	7759399[uid]
rpoB_V170F	RIF	2	1	3	
rpoB_V262A	RIF	1	0	1	
rpoB_V359A	RIF	1	0	1	
rpsA_A440T	PZA	11	0	11	
rpsL_K43R	SM	24	0	24	7968530[uid]
rpsL_K88R	SM	11	0	11	7934937[uid]
rrs_A1401G	AK	5	0	5	8971706[uid]
rrs_A1401G	CAP	5	0	5	
rrs_A1401G	KAN	5	0	5	
rrs_A514C	SM	2	0	2	22943573[uid]
rrs_C513T	SM	2	0	2	
rrs_C517T	SM	2	0	2	15567277[uid]
tlyA_C-83T	CAP	1	0	1	

** publication based on subset of sequences included in this study.

7.4 S4

The performance of alleles characterised as resistance determinants in the training set when predicting phenotypes for the validation set.

		Phenotypically resistant		Phenotypically sensitive		
Drug	Variant	Resistance-	Resistance-	Resistance-	Resistance-	Link to
INH	ahpC_C-57T	1	0	0	0	N/A
INH	fabG1_C-15T	32	1	9	0	10428945[uid]
INH	fabG1_G-17T	2	0	0	0	14638486[uid]
INH	fabG1_T-8C	5	1	0	0	15793126[uid]
INH	inhA_I194T	3	1	1	0	16495272[uid]
INH	inhA_I21T	1	0	0	0	14638486[uid]
INH	inhA_S94A	0	0	2	0	10815738[uid]
INH	katG_S315N	6	0	0	0	9210694[uid]
INH	katG_S315T	271	5	8	0	8537659[uid]
INH	katG_W191R	1	0	0	0	15793126[uid]
RIF	rpoB_1396_insATTC	0	1	0	0	N/A
RIF	rpoB_D435F	1	0	0	0	16229229[uid]
RIF	rpoB_D435V	9	2	0	0	15184414[uid]
RIF	rpoB_H445D	8	1	0	0	14729930[uid]
RIF	rpoB_H445N	0	0	0	1	14729930[uid]
RIF	rpoB_H445R	2	1	0	0	9003625[uid]
RIF	rpoB_H445Y	10	5	1	0	8027320[uid]
RIF	rpoB_L452P	4	0	3	0	10921994[uid]
RIF	rpoB_Q432K	0	0	1	0	8027320[uid]
RIF	rpoB_S450L	224	7	3	0	15728936[uid]
RIF	rpoB_S450W	3	1	1	0	7759399[uid]
RIF	rpoB_V170F	1	1	0	0	N/A
EMB	embB_D354A	2	0	1	0	N/A
EMB	embB_G406A	3	0	2	0	10639358[uid]
EMB	embB_G406D	3	0	2	1	10639358[uid]
EMB	embB_G406S	1	0	0	0	10639358[uid]
EMB	embB_H1002R	2	0	1	0	22646308[uid]**
EMB	embB_M306I	38	1	27	2	9257740[uid]
EMB	embB_M306V	94	5	22	4	9257740[uid]
EMB	embB_Q497K	0	0	2	0	10639358[uid]
EMB	embB_Q497R	9	0	4	0	10639358[uid]
PZA	pncA_292_insAT	1	0	0	0	N/A
PZA	pncA_494_delC	1	0	0	0	N/A
PZA	pncA_A-11G	15	5	1	0	9056006[uid]
PZA	pncA_C138R	0	2	0	0	25336456[uid]
PZA	pncA_C14R	2	1	0	0	9055989[uid]

PZA	pncA_D136N	0	1	0	0	11641519[uid]
PZA	pncA_G97D	2	0	0	0	11083630[uid]
PZA	pncA_H57D	7	0	0	0	9056006[uid]
PZA	pncA_Q141*	1	0	0	0	11641519[uid]
PZA	pncA_T-12C	0	1	0	0	N/A
PZA	pncA_V125G	2	2	0	0	18573039[uid]
PZA	pncA_V139L	0	1	0	0	11083630[uid]
PZA	rpsA_A440T	7	0	1	0	N/A
SM	gidB_202_delG	4	0	2	0	N/A
SM	gidB_A134E	2	0	0	0	17238915[uid]
SM	gidB_V65G	1	1	0	0	N/A
SM	rpsL_K43R	229	3	7	1	7968530[uid]
SM	rpsL_K88R	16	1	0	0	7934937[uid]
SM	rrs_A514C	19	2	0	0	22943573[uid]
SM	rrs_C517T	6	1	2	0	15567277[uid]
CIP	.	.	.	.	.	
MOX	gyrA_D94A	1	0	0	0	8031045[uid]
MOX	gyrA_D94G	2	0	0	0	8031045[uid]
OFX	gyrA_A90V	0	2	0	0	8031045[uid]
OFX	gyrA_D94G	2	3	0	0	8031045[uid]
AK	rrs_A1401G	36	16	1	2	8971706[uid]
CAP	rrs_A1401G	31	13	6	4	N/A
KAN	rrs_A1401G	3	0	0	0	N/A

7.5 S5

Susceptible phenotypes predicted resistant in the validation set

Isolate	Drug	Resistance-determinant
4751-12	EMB	embB_D354A
687-05	EMB	embB_G406A
9703-04	EMB	embB_G406A
3004-06	EMB	embB_G406D
677-04	EMB	embB_G406D
1319-05	EMB	embB_G406D
4549-04	EMB	embB_H1002R
3672-04	EMB	embB_M306I
8121-04	EMB	embB_M306I
7950-05	EMB	embB_M306I
706-05	EMB	embB_M306I
3087-05	EMB	embB_M306I
9546-01	EMB	embB_M306I
688-05	EMB	embB_M306I
7426-01	EMB	embB_M306I
2991-06	EMB	embB_M306I
8660-04	EMB	embB_M306I
6519-01	EMB	embB_M306I
702-05	EMB	embB_M306I
685-05	EMB	embB_M306I
8662-04	EMB	embB_M306I
1603-06	EMB	embB_M306I
TRL0081319-S9	EMB	embB_M306I
8125-04	EMB	embB_M306I
8131-04	EMB	embB_M306I
10849-11	EMB	embB_M306I
TRL0080777-S13	EMB	embB_M306I
TRL0031029-S19	EMB	embB_M306I
8645-04	EMB	embB_M306I
RG00153143-S11	EMB	embB_M306I
5573-01	EMB	embB_M306I
TRL0039671-S15	EMB	embB_M306I
10564-01	EMB	embB_M306I
121-04	EMB	embB_M306I
TRL0078704-S32	EMB	embB_M306I
2383-05	EMB	embB_M306I

8648-04	EMB	embB_M306V
8671-04	EMB	embB_M306V
684-06	EMB	embB_M306V
4539-04	EMB	embB_M306V
8668-04	EMB	embB_M306V
8129-04	EMB	embB_M306V
116-04	EMB	embB_M306V
3082-05	EMB	embB_M306V
9689-04	EMB	embB_M306V
10535-06	EMB	embB_M306V
8644-04	EMB	embB_M306V
2126-06	EMB	embB_M306V
TRL0080791-S14	EMB	embB_M306V
8655-04	EMB	embB_M306V
8128-04	EMB	embB_M306V
TRL0080005-S15	EMB	embB_M306V
2393-05	EMB	embB_M306V
8092-01	EMB	embB_M306V
8650-04	EMB	embB_M306V
10468-07	EMB	embB_M306V
9690-04	EMB	embB_M306V
8661-04	EMB	embB_M306V
10536-05	EMB	embB_M306V
4546-04	EMB	embB_M306V
8647-04	EMB	embB_M306V
677-04	EMB	embB_M306V
---	---	---
6810-06	EMB	embB_Q497K
IF00091761-S29	EMB	embB_Q497K
---	---	---
701-05	EMB	embB_Q497R
9700-04	EMB	embB_Q497R
2839-11	EMB	embB_Q497R
784-02	EMB	embB_Q497R
---	---	---
TRL0036924-S16	INH	fabG1_C-15T
TRL0020863-S23	INH	fabG1_C-15T
1036-06	INH	fabG1_C-15T
10584-09	INH	fabG1_C-15T
9707-12	INH	fabG1_C-15T
646-10	INH	fabG1_C-15T
TRL0057764-S21	INH	fabG1_C-15T
TRL0024410-S8	INH	fabG1_C-15T
3362-12	INH	fabG1_C-15T
---	---	---
8844-10	SM	gidB_202_delG
4890-09	SM	gidB_202_delG
---	---	---
10584-09	INH	inhA_I194T

1956-05	INH	inhA_S94A
TRL0016331-S16	INH	inhA_S94A
1355-10	INH	katG_S315T
12397-12	INH	katG_S315T
735-10	INH	katG_S315T
TRL0022536-S28	INH	katG_S315T
1346-10	INH	katG_S315T
10468-07	INH	katG_S315T
4545-12	INH	katG_S315T
IK00117937-S5	INH	katG_S315T
3009-04	PZA	pncA_A-11G
TRL0029138-S10	RIF	rpoB_H445N
116-04	RIF	rpoB_H445Y
925-08	RIF	rpoB_L452P
TRL0029796-S8	RIF	rpoB_L452P
TRL0050689-S4	RIF	rpoB_L452P
IF00140841-S30	RIF	rpoB_Q432K
TRL0080930-S14	RIF	rpoB_S450L
10468-07	RIF	rpoB_S450L
TRL0077954-S26	RIF	rpoB_S450L
12355-12	RIF	rpoB_S450W
10468-07	SM	rpsL_K43R
4545-12	SM	rpsL_K43R
853-06	SM	rpsL_K43R
1563-10	SM	rpsL_K43R
4825-07	SM	rpsL_K43R
IF00140841-S30	SM	rpsL_K43R
3826-09	SM	rpsL_K43R
2280-07	SM	rpsL_K43R
116-04	AK	rrs_A1401G
3670-04	AK	rrs_A1401G
8132-04	AK	rrs_A1401G
10468-07	SM	rrs_C517T
TRL0083637-S3	SM	rrs_C517T

7.6 S6

100 iterations of the algorithm were each conducted after randomly assigning 1825 isolates to training set and 1826 to a validation set. Genotypic predictions for the phenotypes in each of the 100 validation sets are shown. As in Table 9, algorithmic characterisations are based on: R (resistance determining allele); Rx (resistance determinant as a mixed base call); S0 (zero alleles present); Ss (only benign alleles present); U (unclassified allele present in the absence of a resistance-determining allele).

Phenotypically Resistant Genotype					**Phenotypically Sensitive Genotype**				
R	**R_x**	**S_0**	**S_s**	**U**	**R**	**R_x**	**S_0**	**S_s**	**U**
786	30	53	27	98	143	9	6754	565	591
786	37	46	14	95	130	12	6842	504	565
877	38	52	14	106	147	9	6754	536	558
807	37	56	8	106	129	14	6789	531	540
783	42	54	18	92	125	38	6804	487	606
776	36	44	12	86	131	12	6780	559	594
878	44	50	13	97	135	7	6708	534	535
848	46	59	11	88	102	14	6744	546	593
817	32	56	17	85	124	10	6783	570	545
843	39	59	12	97	128	10	6837	569	503
850	40	43	13	92	130	12	6810	510	564
835	42	53	18	90	118	10	6746	562	574
796	36	50	13	100	115	10	6722	582	558
828	38	61	18	88	112	6	6792	552	571
884	42	58	15	104	129	9	6762	569	506
843	35	59	7	130	134	10	6733	528	628
882	39	57	16	106	107	10	6744	531	562
883	41	37	12	100	114	12	6791	556	535
820	42	58	13	93	139	11	6775	519	595
850	32	66	22	109	131	15	6746	590	532
829	47	65	18	93	124	15	6738	536	585
814	31	59	10	97	146	8	6763	566	568
812	38	48	12	92	135	7	6774	544	558
836	43	58	10	115	122	8	6743	514	601
843	33	50	18	87	144	30	6768	557	517
822	37	54	18	126	122	13	6753	605	467
836	37	60	14	102	133	8	6837	558	586

816	37	53	11	112	112	11	6788	511	617
846	38	52	19	98	112	12	6776	582	571
794	39	58	22	90	116	13	6722	513	622
827	33	59	16	88	124	12	6756	548	534
702	29	57	18	95	116	11	6872	542	578
859	36	65	20	79	122	10	6705	587	540
846	40	58	22	118	134	9	6764	526	573
815	43	39	25	97	107	11	6784	526	571
776	34	56	17	91	125	11	6826	541	583
865	40	58	15	94	133	9	6760	563	548
830	40	57	16	112	120	12	6782	518	548
805	38	58	12	90	130	9	6858	582	520
844	39	49	22	103	133	14	6721	549	556
851	41	52	22	96	131	10	6703	536	590
815	42	62	10	104	152	9	6797	504	582
851	50	63	28	102	133	10	6750	523	584
849	39	60	16	104	132	8	6787	536	559
928	40	45	16	99	139	8	6685	614	501
838	39	51	10	105	112	5	6767	519	581
829	38	61	18	85	120	11	6804	514	533
829	38	44	16	111	133	11	6714	613	568
842	37	56	13	104	121	6	6734	531	600
835	48	61	12	106	127	12	6721	558	567
811	34	49	19	105	112	9	6827	538	536
884	44	48	18	104	118	10	6669	560	615
827	35	58	14	79	137	12	6807	514	564
860	37	60	16	122	122	8	6774	565	535
807	41	65	13	97	146	12	6793	559	559
844	45	48	19	104	133	9	6775	551	482
839	48	49	15	107	132	11	6729	542	577
838	42	48	14	81	139	31	6754	508	565
836	47	42	9	95	136	14	6788	581	511
872	31	50	17	109	139	5	6779	552	524
830	40	47	17	108	141	6	6734	621	524
845	36	60	20	107	125	13	6753	596	527
844	35	59	21	87	119	11	6791	520	558
872	42	57	24	91	128	11	6758	542	557
875	35	60	14	88	128	11	6792	512	616
782	39	51	14	98	119	6	6742	524	563
781	33	65	17	109	123	8	6817	578	554
887	41	48	12	105	124	10	6728	551	539
823	32	56	12	99	113	13	6779	535	532
878	38	57	9	101	118	13	6766	539	535
835	32	45	13	94	115	7	6812	518	617
836	32	64	22	95	145	8	6767	585	540
825	43	60	11	81	140	14	6764	567	537

806	39	54	29	106	113	12	6763	594	551
844	32	48	11	90	136	9	6799	546	534
812	40	55	15	92	126	10	6729	524	617
770	37	52	18	96	131	8	6759	517	583
808	30	55	19	111	112	8	6792	568	530
953	46	57	19	88	134	10	6690	554	554
873	39	54	15	108	131	14	6770	547	513
847	43	51	22	81	105	8	6771	566	521
854	32	58	19	95	121	7	6737	507	592
818	36	54	9	103	130	10	6724	553	587
811	29	63	17	98	108	10	6788	605	474
852	40	57	16	105	117	14	6825	581	514
819	49	60	19	87	104	8	6761	595	530
832	43	54	17	107	127	7	6735	525	619
783	46	48	24	82	113	9	6874	513	583
845	32	53	24	112	121	12	6766	531	589
800	37	52	12	119	133	6	6786	553	561
860	35	50	17	84	136	8	6744	630	515
855	35	60	24	99	133	8	6735	570	502
832	38	50	24	103	127	10	6718	502	612
763	42	54	13	109	141	14	6748	549	545
803	41	54	22	105	124	11	6755	541	541
849	33	66	17	114	127	8	6719	500	606
789	43	55	21	102	123	12	6751	588	499
811	36	48	17	102	113	8	6752	527	622
837	36	59	14	96	126	10	6777	530	608
841	38	57	17	104	129	13	6769	534	534

7.7 S7

Resistance determinants derived from the combined training and validation sets (i.e. by applying the algorithm to all 3651 isolates).

Variant	Drug	Resistant	Susceptible
ahpC_C-57T	INH	2	0
ahpC_C-72T	INH	1	1
eis_G-10A	KAN	4	0
embA_C-12T	EMB	12	7
embA_C-16G	EMB	5	1
embA_C-16T	EMB	6	4
embB_D328Y	EMB	2	1
embB_D354A	EMB	3	2
embB_G406A	EMB	4	5
embB_G406D	EMB	5	5
embB_G406S	EMB	4	5
embB_M306I	EMB	59	63
embB_M306V	EMB	117	29
embB_N1033K	EMB	1	0
embB_Q497K	EMB	1	4
embB_Q497R	EMB	17	8
fabG1_C-15T	INH	106	15
fabG1_G-17T	INH	7	1
fabG1_T-8C	INH	9	3
gidB_141_delC	SM	1	0
gidB_202_delG	SM	8	9
gidB_202_insGC	SM	1	1
gidB_215_delC	SM	6	3
gidB_399_delAAACTCGGTG	CAP	1	0
gidB_399_delAAACTCGGTG	SM	1	0
gidB_451_delG	SM	4	4
gidB_451_insGC	SM	1	2
gidB_A134E	SM	3	0
gidB_A138T	SM	1	0
gidB_A138V	SM	2	1
gidB_A19P	SM	1	2
gidB_A200E	SM	2	0
gidB_A205E	SM	1	0
gidB_A80P	SM	1	0
gidB_C52F	SM	1	0
gidB_D85A	SM	3	0
gidB_E173*	SM	1	0
gidB_G117V	SM	2	0

gidB_G30D	SM	4	0
gidB_G34V	SM	2	0
gidB_G69D	SM	3	0
gidB_G73A	SM	1	0
gidB_H48N	SM	2	0
gidB_H48Q	SM	1	0
gidB_I11N	SM	1	0
gidB_I162S	SM	1	0
gidB_L26F	SM	1	0
gidB_L79S	SM	5	2
gidB_L79W	SM	1	0
gidB_L91P	SM	1	0
gidB_P75L	SM	1	1
gidB_P75R	SM	5	0
gidB_P93L	SM	1	0
gidB_Q125*	SM	1	0
gidB_R118L	SM	1	0
gidB_R118S	SM	3	1
gidB_R137P	SM	2	1
gidB_R137W	SM	1	1
gidB_R47W	SM	1	0
gidB_R64W	SM	1	0
gidB_R83P	SM	1	0
gidB_S136*	SM	1	1
gidB_S149R	SM	1	2
gidB_S70N	SM	1	0
gidB_V203L	SM	1	0
gidB_V41I	SM	1	1
gidB_V65G	SM	3	0
gidB_V88A	SM	2	0
gidB_Y195H	CAP	2	28
gyrA_A74S	CIP	2	0
gyrA_A90V	MOX	2	0
gyrA_A90V	OFX	4	0
gyrA_D94A	MOX	2	0
gyrA_D94A	OFX	1	0
gyrA_D94G	CIP	18	1
gyrA_D94G	MOX	11	4
gyrA_D94G	OFX	14	3
gyrA_D94H	MOX	1	1
gyrA_D94H	OFX	1	1
gyrA_D94N	MOX	1	0
gyrA_D94N	OFX	1	0
gyrA_S91P	CIP	2	2
gyrA_S91P	MOX	3	0
gyrA_S91P	OFX	3	0

inhA_I194T	INH	6	1
inhA_I21T	INH	3	0
inhA_S94A	INH	4	3
katG_1157_delT	INH	1	0
katG_121_insTA	INH	1	0
katG_1388_delA	INH	2	0
katG_1450_delC	INH	1	0
katG_1465_delC	INH	1	0
katG_1910_delA	INH	1	0
katG_471_delG	INH	2	0
katG_A109V	INH	1	0
katG_A614E	INH	1	0
katG_D142G	INH	1	0
katG_G125D	INH	1	0
katG_G182R	INH	1	0
katG_G297V	INH	1	0
katG_L141F	INH	3	0
katG_L159P	INH	1	0
katG_L627P	INH	1	0
katG_L704S	INH	1	0
katG_P232R	INH	1	0
katG_R104Q	INH	2	0
katG_S315I	INH	1	0
katG_S315N	INH	8	1
katG_S315T	INH	471	9
katG_S481L	INH	1	1
katG_S700P	INH	2	0
katG_T180K	INH	1	0
katG_V633A	INH	1	1
katG_W191G	INH	1	0
katG_W191R	INH	2	0
katG_W300C	INH	1	0
katG_W300S	INH	1	0
katG_W328L	INH	1	0
katG_W505*	INH	1	0
katG_W90R	INH	1	0
ndh_G225D	INH	1	0
pncA_174_delG	PZA	1	0
pncA_177_delC	PZA	1	0
pncA_285_insCT	PZA	1	0
pncA_292_insAT	PZA	2	0
pncA_409_delT	PZA	2	0
pncA_489_delC	PZA	9	0
pncA_491_insACC	PZA	1	0
pncA_494_delC	PZA	2	0
pncA_528_insGGCCGTCTGGC	PZA	1	0

pncA_556_insTG	PZA	1	0
pncA_563_insAC	PZA	1	0
pncA_570_insCT	PZA	2	0
pncA_592_delC	PZA	1	0
pncA_617_insTC	PZA	1	0
pncA_A-11G	PZA	23	1
pncA_C138R	PZA	3	0
pncA_C14R	PZA	4	0
pncA_D12A	PZA	1	0
pncA_D136N	PZA	2	2
pncA_D49N	PZA	1	0
pncA_D8G	PZA	1	0
pncA_D8N	PZA	1	0
pncA_F81V	PZA	1	1
pncA_G132D	PZA	1	0
pncA_G162D	PZA	1	0
pncA_G78C	PZA	1	0
pncA_G97C	PZA	1	0
pncA_G97D	PZA	3	0
pncA_G97R	PZA	1	0
pncA_H137R	PZA	1	2
pncA_H51Q	PZA	1	1
pncA_H57D	PZA	18	0
pncA_H57R	PZA	2	0
pncA_H71Q	PZA	1	0
pncA_H71Y	PZA	2	1
pncA_I133T	PZA	20	15
pncA_K48E	PZA	1	2
pncA_K96T	PZA	1	0
pncA_L151S	PZA	1	0
pncA_L159V	PZA	1	1
pncA_L172P	PZA	1	0
pncA_L27P	PZA	1	0
pncA_L4S	PZA	4	0
pncA_M175V	PZA	1	1
pncA_P54L	PZA	2	2
pncA_P54Q	PZA	1	1
pncA_Q10*	PZA	1	0
pncA_Q10P	PZA	38	3
pncA_Q141*	PZA	2	0
pncA_Q141P	PZA	5	1
pncA_S104G	PZA	1	1
pncA_S104R	PZA	2	0
pncA_S32I	PZA	1	0
pncA_T-12C	PZA	4	0
pncA_T114P	PZA	1	0

pncA_T135P	PZA	1	0
pncA_T47A	PZA	1	1
pncA_V125G	PZA	8	0
pncA_V139L	PZA	2	0
pncA_V180F	PZA	1	0
pncA_V180G	PZA	2	0
pncA_V21G	PZA	2	1
pncA_V7L	PZA	3	0
pncA_W68C	PZA	1	0
pncA_W68R	PZA	4	0
pncA_Y99*	PZA	1	0
rpoB_1377_delG	RIF	1	0
rpoB_1392_insGCCA	RIF	2	0
rpoB_1394_delC	RIF	1	0
rpoB_1396_insATTC	RIF	3	0
rpoB_1398_delT	RIF	1	0
rpoB_1427_delTGGCCCC	RIF	1	0
rpoB_D435F	RIF	2	0
rpoB_D435V	RIF	15	0
rpoB_G981D	RIF	1	0
rpoB_H445D	RIF	15	0
rpoB_H445L	RIF	2	0
rpoB_H445N	RIF	1	5
rpoB_H445R	RIF	6	0
rpoB_H445Y	RIF	27	1
rpoB_I491F	RIF	4	19
rpoB_L430P	RIF	3	4
rpoB_L452P	RIF	8	7
rpoB_M434I	RIF	1	0
rpoB_Q432K	RIF	1	1
rpoB_Q432L	RIF	1	0
rpoB_Q432P	RIF	2	0
rpoB_S441L	RIF	1	0
rpoB_S450F	RIF	4	0
rpoB_S450L	RIF	290	4
rpoB_S450W	RIF	7	2
rpoB_T676P	RIF	1	0
rpoB_V170F	RIF	4	1
rpoB_V359A	RIF	1	0
rpsA_E67D	PZA	1	0
rpsA_V260I	PZA	1	16
rpsL_K43R	SM	256	8
rpsL_K88R	SM	28	0
rrs_A1325C	SM	1	0
rrs_A1401G	AK	57	3
rrs_A1401G	CAP	49	10

rrs_A1401G	KAN	8	0
rrs_A514C	SM	23	0
rrs_A514T	SM	1	0
rrs_C1402T	AK	1	1
rrs_C1402T	CAP	2	1
rrs_C513T	SM	2	0
rrs_C517T	KAN	1	5
rrs_C517T	SM	9	2
rrs_C905A	SM	1	0
tlyA_C-83T	CAP	1	0

7.8 S8

Characterisation of alleles by the original training set and by all the 3651 samples combined.

Previously seen in the literature	Characterisation in the training-set	Characterisation in the final, combined set of 3651	Which isolates was the final characterisation based on?	Number of alleles
No	Allele not seen	Resistance-determinant	Validation-set	71
Yes	Allele not seen	Resistance-determinant	Validation-set	26
No	Allele not seen	Benign	Validation-set	872
Yes	Allele not seen	Benign	Validation-set	14
No	Allele not seen	Uncharacterised	Validation-set	528
Yes	Allele not seen	Uncharacterised	Validation-set	8
No	Resistance-determinant	Resistance-determinant	Training-set	30
Yes	Resistance-determinant	Resistance-determinant	Training-set	30
No	Resistance-determinant	Resistance-determinant	Combined sets	8
Yes	Resistance-determinant	Resistance-determinant	Combined sets	48
No	Resistance-determinant	Benign	Combined sets	2
No	Resistance-determinant	Uncharacterised	Training-set	1
Yes	Resistance-determinant	Uncharacterised	Training-set	1
No	Benign	Resistance-determinant	Combined sets	9
Yes	Benign	Resistance-determinant	Combined sets	7
No	Benign	Benign	Training-set	607
Yes	Benign	Benign	Training-set	13
No	Benign	Benign	Combined sets	113
Yes	Benign	Benign	Combined sets	11
No	Benign	Uncharacterised	Combined sets	12
No	Uncharacterised	Resistance-determinant	Combined sets	2
Yes	Uncharacterised	Resistance-determinant	Combined sets	1
No	Uncharacterised	Benign	Combined sets	2
Yes	Uncharacterised	Benign	Combined sets	1
No	Uncharacterised	Uncharacterised	Training-set	57
Yes	Uncharacterised	Uncharacterised	Training-set	9
No	Uncharacterised	Uncharacterised	Combined sets	24
Yes	Uncharacterised	Uncharacterised	Combined sets	5

8 REFERENCES

1 Evans AS. Causation and disease: the Henle-Koch postulates revisited. *The Yale journal of biology and medicine* 1976;**49**:175.

2 Koch R. Die Aetiologie der Tuberkulose. *Berliner Klinische Wochenschrift* 1882;:221–30.

3 Bowditch H. Is consumption ever contagious, or communicated by one person to another in any manner? *Boston Society for Medical Observation* 1864:1–14.

4 Koch R. The Nobel Lecture on how the fight against tuberculosis now stands. *The Lancet* 1906;**167**:1449–51.

5 Wilson LG. Commentary: Medicine, population, and tuberculosis. *International Journal of Epidemiology* 2005;**34**:521–4.

6 Newsholme A. An Inquiry into the Principal Causes of the Reduction in the Death-rate from Phthisis during the last Forty Years, with Special Reference to the Segregation of Phthisical Patients in General Institutions. *J Hyg (Lond)* 1906;**6**:304–84.

7 Comas I, Coscolla M, Luo T, *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* 2013;**45**:1176–82.

8 Phillips L. Infectious disease: TB's revenge. *Nature* 2013;**493**:14–6.

9 Donoghue HD, Spigelman M, Greenblatt CL, *et al.* Tuberculosis: from prehistory to Robert Koch, as revealed by ancient DNA. *Lancet Infect Dis* 2004;**4**:584–92.

10 Davies P, Grange JM. Factors affecting susceptibility and resistance to tuberculosis. *Thorax* 2001;**56**:ii23–9.

11 Mohamed J. Epidemics and public health in early colonial Somaliland. *Social Science & Medicine* 1999;**48**:507–21.

12 Daniel TM. The early history of tuberculosis in central East Africa: insights from the clinical records of the first twenty years of Mengo Hospital and review of relevant literature. *Int J Tuberc Lung Dis* 1998;**2**:784–90.

13 Fayrer J. Scrofula,Tuberculosis, and Phthisis in India. *BMJ* 1881;**1**:808–12.

14 Wilkinson E. Notes on the Prevalence of Tuberculosis in India. *Proc R Soc Med* 1914;**7**:195–226.

15 Arnold D. Imperial Medicine and Indigenous Societies. Manchester University Press, 1988.:5–6. http://books.google.co.uk/books?id=0Xi7AAAAIAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false (accessed 14 Oct2014).

16 Tilt EJ. Health in India for British Women, and on the Prevention of Disease in Tropical Climates. Churchill, 1875.:4–5and116.https://archive.org/stream/healthinindiafo00tiltgoog#page/n20/mode/2up (Accessed 14 October 2014).

17 Young DB, Perkins MD, Duncan K, *et al.* Confronting the scientific obstacles to global control of tuberculosis. *J Clin Invest* 2008;**118**:1255–65.

18 World Health Organization. Global tuberculosis report 2013. http://www.who.int/tb/publications/global_report/en/ (Accessed 11 April 2014)

19 Philips JA, Ernst JD. Tuberculosis pathogenesis and immunity. *Annu Rev Pathol* 2012;**7**:353–84.

20 Kaufmann SHE. How can immunology contribute to the control of tuberculosis? *Nat Rev Immunol* 2001;**1**:20–30.

21 Walzl G, Ronacher K, Hanekom W, *et al.* Immunological biomarkers of tuberculosis. *Nat Rev Immunol* 2011;**11**:343–54.

22 Mack U, Migliori GB, Sester M, *et al.* LTBI: latent tuberculosis infection or lasting immune responses to M. tuberculosis? A TBNET consensus statement. ERJ May 1, 2009 33:5; 956-973

23 Joshi R, Patil S, Kalantri S, *et al.* Prevalence of abnormal radiological findings in health care workers with latent tuberculosis infection and correlations with T cell immune response. *PLoS ONE* 2007;**2**:e805.

24 Behr MA, Waters WR. Is tuberculosis a lymphatic disease with a pulmonary portal? *Lancet Infect Dis* 2014;**14**:250–5.

25 3rd CEB, Boshoff HI, Dartois V, *et al.* The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nature Reviews Microbiology* 2009;**7**:845–55.

26 Ramakrishnan L. Revisiting the role of the granuloma in tuberculosis. *Nat Rev Immunol* 2012;**12**:352–66.

27 Keane J, Gershon S, Wise RP, *et al.* Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent. *N Engl J Med* 2001;**345**:1098–104.

28 Harris J, Keane J. How tumour necrosis factor blockers interfere with tuberculosis immunity. *Clinical & Experimental Immunology* 2010;**161**:1–9

29 Zumla A, Raviglione M, Hafner R, *et al.* Current Concepts: Tuberculosis. *N Engl J Med* 2013;**368**:745–55.

30 Hunter RL. Pathology of post primary tuberculosis of the lung: An illustrated critical review. *Tuberculosis* 2011;**91**:497–509.

31 Thwaites G. Neurological aspects of tropical disease: Tuberculous meningitis. *Journal of Neurology, Neurosurgery & Psychiatry* 2000;**68**:289–99.

32 Dass B, Puet TA, Watanakunakorn C. Tuberculosis of the spine (Pott's disease) presenting as “compression fractures”. *Spinal Cord* 2002;**40**:604–8.

33 Geppert EF. The Pathogenesis of Pulmonary and Miliary Tuberculosis. *Arch Intern Med* 1979;**139**:1381.

34 Kim JH, Langston AA, Gallis HA. Miliary tuberculosis: epidemiology, clinical manifestations, diagnosis, and outcome. *Review of Infectious Diseases* 1990;**12**:583–90.

35 Ray S, Kundu S, Sonthalia N, *et al.* Diagnosis and management of miliary tuberculosis: current state and future perspectives. *TCRM* 2013;9.

36 Comas I, Chakravartti J, Small PM, *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat Genet* 2010;**42**:498–503.

37 Ernst JD. The immunological life cycle of tuberculosis. *Nat Rev Immunol* 2012;**12**:581–91.

38 Gandhi NR, Moll A, Sturm AW, *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**:1575–80.

39 Geldmacher C, Ngwenyama N, Schuetz A, *et al.* Preferential infection and depletion of Mycobacterium tuberculosis-specific CD4 T cells after HIV-1 infection. *J Exp Med* 2010;**207**:2869–81.

40 Kwan CK, Ernst JD. HIV and tuberculosis: a deadly human syndemic. *Clinical Microbiology Reviews* 2011;**24**:351–76.

41 Roy A, Eisenhut M, Harris RJ, *et al.* Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis. *BMJ* 2014;**349**:g4643–3.

42 Kaufmann SHE. Novel tuberculosis vaccination strategies based on understanding the immune response. *J Intern Med* 2010;**267**:337–53.

43 Tameris MD, Hatherill M, Landry BS, *et al.* Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: a randomised, placebo-controlled phase 2b trial. *Lancet*

2013;**381**:1021–8.

44 Nayak S, Acharjya B. Mantoux test and its interpretation. *Indian Dermatol Online J* 2012;**3**:2.

45 Heaf F. The multiple-puncture tuberculin test. *The Lancet* 1951;**2**:151–3.

46 Carruthers KJ. Comparison of the Heaf (multiple puncture) and Mantoux tests using several tuberculins. *Tubercle* 1969;**50**:22–41.

47 Lalvani A. Diagnosing Tuberculosis Infection in the 21st Century: New Tools To Tackle an Old Enemy. *Chest* 2007;**131**:1898–906.

48 Lawn SD, Zumla AI. Tuberculosis. *The Lancet* 2011;**378**:57–72.

49 Mao TE, Okada K, Yamada N, *et al.* A cross-sectional study of tuberculosis prevalence in Cambodia between 2002 and 2011. Bulletin of the World Health Organization, Article ID: BLT.13.131581

50 Tiemersma EW, van der Werf MJ, Borgdorff MW, *et al.* Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: a systematic review. *PLoS ONE* 2011;**6**:e17601.

51 Marais BJ, Gie RP, Schaaf HS, *et al.* The natural history of childhood intra-thoracic tuberculosis: a critical review of literature from the pre-chemotherapy era [State of the Art]. *Int J Tuberc Lung Dis* 2004;**8**:392–402.

52 Rodrigo TT, Caylà JAJ, de Olalla PPG, *et al.* Characteristics of tuberculosis patients who generate secondary cases. *Int J Tuberc Lung Dis* 1997;**1**:352–7.

53 Riley RL, Mills CC, Nyka W, *et al. Aerial dissemination of pulmonary tuberculosis. A two-year study of contagion in a tuberculosis ward. 1959.* 1959.

54 Escombe AR, Moore DAJ, Gilman RH, *et al.* The infectiousness of tuberculosis patients coinfected with HIV. *Plos Med* 2008;**5**:e188.

55 Tuberculosis in the UK: 2012 report. Health Protection Agency. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/332560/TB_Annual_Report_2012.pdf (Accessed 25 April 2013)

56 Greenwood D, Slack RCB, Peutherer JF. Medical Microbiology. Churchill Livingstone: p.200.

57 Joyce Wang MAB. Building a better bacillus: the emergence of Mycobacterium tuberculosis. *Frontiers in Microbiology* 2014;**5**.

58 Euzeby J, editor. LPSN: List of prokaryotic names with standing in nomenclature. bacterio.net. http://www.bacterio.net/mycobacterium.html (Accessed 20 August 2014).

59 Tortoli E. Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. *Clinical Microbiology Reviews* 2003;**16**:319–54.

60 Smith NH, Gordon SV, la Rua-Domenech de R, *et al.* Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. *Nature Reviews Microbiology* 2006;**4**:670–81.

61 Ford CB, Lin PL, Chase MR, *et al.* Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat Genet* 2011;**43**:482–6.

62 Titford M. Progress in the Development of Microscopical Techniques for Diagnostic Pathology. *Journal of Histotechnology* 2009;**32**:9–19.

63 Drobniewski FA, Hoffner S, Rusch-Gerdes S, *et al.* Recommended standards for modern tuberculosis laboratory services in Europe. *European Respiratory Journal* 2006;**28**:903–9.

64 Steingart KR, Henry M, Ng V, *et al.* Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. *Lancet Infect Dis* 2006;**6**:570–81.

65 Lewis JJ, Chihota VN, van der Meulen M, *et al.* 'Proof-Of-Concept' Evaluation of an Automated Sputum Smear Microscopy System for Tuberculosis Diagnosis. *PLoS ONE* 2012;**7**:e50173.

66 Pallen MJ. Death knell for guineapig test. *Tubercle* 1972;**53**:31–4.

67 Mitchison DA. The Diagnosis and Therapy of Tuberculosis During the Past 100 Years. *Am J Respir Crit Care Med 171,2005: p.699-706*

68 Richter E, Rüsch-Gerdes S, Hillemann D. Drug-susceptibility testing in TB: current status and future prospects. *Expert Review of Respiratory Medicine* 2009;**3**:497–510.

69 Moore DAJ, Evans CAW, Gilman RH, *et al.* Microscopic-Observation Drug-Susceptibility Assay for the Diagnosis of TB. *N Engl J Med* 2006;**355**:1539–50.

70 Gale GL. Atypical Mycobacteria in a Tuberculosis Hospital. *Can Med Assoc J* 1976;**114**:612–4.

71 Babady NE, Wengenack NL. Clinical Laboratory Diagnostics for Mycobacterium tuberculosis. *in Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis*;p.8–11. http://cdn.intechopen.com/pdfs-wm/28531.pdf (Accessed 30 August 2014)

72 Weiszfeiler JG, Karasseva V, Karczag E. Mycobacterium simiae and related mycobacteria. *Rev Infect Dis* 1981;**3**:1040–5.

73 Roberts GD, Goodman NL, Heifets L, *et al.* Evaluation of the BACTEC radiometric method for recovery of mycobacteria and drug susceptibility testing of Mycobacterium tuberculosis from acid-fast smear-positive specimens. *Journal of Clinical Microbiology* 1983;**18**:689–96.

74 Hanna BA, Ebrahimzadeh A, Elliott LB, *et al.* Multicenter evaluation of the BACTEC MGIT 960 system for recovery of mycobacteria. *Journal of Clinical Microbiology* 1999;**37**:748–52.

75 Ghodbane R, Raoult D, Drancourt M. Dramatic reduction of culture time of Mycobacterium tuberculosis. *Sci Rep* 2014;**4**.

76 Canetti G, Hauduroy P, Grosset J, *et al.* Mycobacteria - Laboratory Methods for Testing Drug Sensitivity and Resistance. *Bull World Health Organ* 1963;**29**:565–&.

77 Drobniewski F, Rusch-Gerdes S, Hoffner S, *et al.* Antimicrobial susceptibility testing of Mycobacterium tuberculosis (EUCAST document E.DEF 8.1)--report of the Subcommittee on Antimicrobial Susceptibility Testing of Mycobacterium tuberculosis of the European Committee for Antimicrobial Susceptibility Testing (EUCAST) of the European Society of Clinical Microbiology and Infectious Diseases (ESCMID). *Clin Microbiol Infect* 2007;**13**:1144–56.

78 Böttger EC. The ins and outs of Mycobacterium tuberculosis drug susceptibility testing. *Clin Microbiol Infect* 2011;**17**:1128–34.

79 World Health Organization. *Noncommercial Culture and Drug-Susceptibility Testing Methods for Screening Patients at Risk for Multidrug-Resistant Tuberculosis: Policy Statement*. Geneva 2011. ISBN-13: 978-92-4-150162-0

80 Grandjean L, Moore DA. Tuberculosis in the developing world: recent advances in diagnosis with special consideration of extensively drug-resistant tuberculosis (XDR-TB). *Current opinion in infectious diseases* 2008;**21**:454.

81 Riva MA. From milk to rifampicin and back again: history of failures and successes in the treatment for tuberculosis. *J Antibiot* 2014; 1-5.

82 Warren P. The Evolution of the Sanatorium: The First Half-Century, 1854-1904. *Canadian Bulletin of Medical History* 2006;**23**:457–76.

83 Diacon A, von G-B, Donald PR. From Magic Mountain to Table Mountain. *Swiss Med Wkly* Published Online First: 22 August 2012.

84 Pope AS. The role of the sanatorium in tuberculosis control. *The Milbank Memorial Fund Quarterly* 1938: 327–37.

85 Hart P. Chemotherapy of Tuberculosis—Part I. *Br Med J* 1946;**2**:805–10.

86 Da Silva PEA, Palomino JC. Molecular basis and mechanisms of drug

resistance in Mycobacterium tuberculosis: classical and new drugs. J Antimicrob Chemother May 2011, 1-14.

87 Marshall G, Blacklock J, Cameron C, *et al.* Streptomycin treatment of pulmonary tuberculosis. A medical research council investigation. *Br Med J* 1948;**2**:769–82.

88 Marshall G, Blacklock J, Cameron C, *et al.* Streptomycin treatment of tuberculous meningitis. *The Lancet* 1948;**251**:582–96.

89 Marshall G, Cruickshank R, Daniels M, *et al.* Treatment of Pulmonary Tuberculosis with Streptomycin and Para-Amino-Salicylic Acid: A Medical Research Council Investigation. *Br Med J* 1950;**2**:1073.

90 Marshall G. PREVENTION of streptomycin resistance by combined chemotherapy; a Medical Research Council investigation. *Br Med J* 1952;**1**:1157–62.

91 Caution in the chemotherapy of tuberculosis. *JAMA* 1952;**149**:1224–5.

92 Marshall G, Crofton J, Cruickshank R, *et al.* Treatment of Pulmonary Tuberculosis with Isoniazid: A Medical Research Council Investigation. *Br Med J* 1952;**2**:735.

93 Marshall G, Crofton J, Cruickshank R, *et al.* Isoniazid in Treatment of Pulmonary Tuberculosis. Br Med J 1953: 521-536

94 Fox W, Wiener A, Mitchison DA, *et al.* The prevalence of drug-resistant tubercle bacilli in untreated patients with pulmonary tuberculosis; a national survey, 1955-56. *Tubercle* 1957;**38**:71–84.

95 Long-term chemotherapy in the treatment of chronic pulmonary tuberculosis with cavitation: A report to the Medical Research Council by their Tuberculosis Chemotherapy Trials Committee. *Tubercle* 1962;**43**:201–67.

96 Bignall JR, Rist N. An international investigation of the efficacy of chemotherapy in previously untreated patients with pulmonary tuberculosis. *Bull Int Union Tuberc* 1964;**2**:83–191.

97 Mitchison D, Davies G. The chemotherapy of tuberculosis: past, present and future. *Int J Tuberc Lung Dis* 2012;**16**:724–32.

98 Alahari A, Trivelli X, Guérardel Y, *et al.* Thiacetazone, an Antitubercular Drug that Inhibits Cyclopropanation of Cell Wall Mycolic Acids in Mycobacteria. *PLoS ONE* 2007;**2**:e1343.

99 Kinsley BJ. Comparative trial of isoniazid in combination with thiacetazone or a substituted diphenylthiourea (SU1906) or PAS in the treatment of acute pulmonary tuberculosis in east africans: A Co-operative Investigation in East African Hospitals and Laboratories with the Collaboration of the British Medical Research Council. *Tubercle*

1960;**41**:399–423.

100 Heffernan JF, Tall R. Isoniazid with thiacetazone (thioacetazone) in the treatment of pulmonary tuberculosis in East Africa-second report of fifth investigation. A co-operative study in East African hospitals, clinics and laboratories with the collaboration of the East African and British Medical Research Councils. *Tubercle* 1970;**51**:353–8.

101 Tuberculosis Chemotherapy Centre, Madras. *Tubercle* 1968;**49**:114–21.

102 Tuberculosis Chemotherapy Centre, Madras. *A Concurrent Comparison of Home and Sanatorium Treatment of Pulmonary Tuberculosis in South India*. Bull. Wld Hlth Org.1959;21,51-144

103 Donomae I, Yamamoto K. Clinical Evaluation of Ethambutol in Pulmonary Tuberculosis. Annals New York Academy of Sciences. 1966;:849–81.http://onlinelibrary.wiley.com/store/10.1111/j.1749-6632.1966.tb45528.x/asset/j.1749-6632.1966.tb45528.x.pdf?v=1&t=hzo9134z&s=bf08b5512fc9d084f7c021174bde21acb0e3f23b (Accessed 4 September 2014).

104 Co-operative controlled trial of a standard regimen of streptomycin, PAS and isoniazid and three alternative regimens of chemotherapy in Britain. A report from the British Medical Research Council. *Tubercle* 1973;**54**:99–129.

105 Third report. East African-British Medical Research Councils.Controlled clinical trial of four short-course (6-month) regimens of chemotherapy for treatment of pulmonary tuberculosis. *Lancet* 1974;**2**:237–40.

106 Second East African/British Medical Research Council Study. Controlled clinical trial of four short-course (6-month) regimens of chemotherapy for treatment of pulmonary tuberculosis. *The Lancet* 1974;**2**:1100–6.

107 Second report. Third East African/British Medical Research Council Study. Controlled clinical trial of four short-course regimens of chemotherapy for two durations in the treatment of pulmonary tuberculosis. *Tubercle* 1980;**61**:59–69.

108 Second report: the results up to 24 months. Hong Kong Chest Service/British Medical Research Council.Controlled trial of 4 three-times-weekly regimens and a daily regimen all given for 6 months for pulmonary tuberculosis. *Tubercle* 1982;**63**:89–98.

109 British Thoracic Society. A controlled trial of 6 months chemotherapy in pulmonary tuberculosis-Final report: results during 36 months after the end of chemotherapy and beyond. *Br J Dis Chest* (1984) 78, 330.

110 Chum HJ, Ilmolelian G, Rieder HL, *et al.* Impact of the change from an injectable to a fully oral regimen on patient adherence to ambulatory tuberculosis treatment in Dar es Salaam, Tanzania. *Tuber Lung Dis* 1995;**76**:286–9.

111 Rabarijaona L, Boisier P, Ratsirahonana O, *et al.* Replacement of streptomycin by ethambutol in the intensive phase of tuberculosis treatment: no effect on compliance. *Int J Tuberc Lung Dis* 1999;**3**:42–6.

112 World Health Organization. Treatment of tuberculosis: guidelines for national programmes, third edition; Revision (June 2004). who.int. 2004.http://www.who.int/tb/publications/tb_2003_313_chap4_rev.pdf (Accessed 9 September 2004).

113 Jindani A, Nunn PAJ, Enarson PDA. Two 8-month regimens of chemotherapy for treatment of newly diagnosed pulmonary tuberculosis: international multicentre randomised trial. *The Lancet* 2004;**364**:1244–51.

114 Gillespie SH, Crook AM, McHugh TD, *et al.* Four-Month Moxifloxacin-Based Regimens for Drug-Sensitive Tuberculosis. *N Engl J Med* Published Online First: 7 September 2014.

115 Merle CS, Fielding K, Sow OB, *et al.* A Four-Month Gatifloxacin-Containing Regimen for Treating Tuberculosis. *N Engl J Med* 2014;**371**:1588–98.

116 Kempker RR, Vashakidze S, Solomonia N, *et al.* Surgical treatment of drug-resistant tuberculosis. *Lancet Infect Dis* 2012;**12**:157–66.

117 Sakula A. Forlanini,Carlo Inventor of Artificial Pneumothorax for Treatment of Pulmonary Tuberculosis. *Thorax* 1983;**38**:326–32.

118 Gregory L Calligaro LMGSKD. The medical and surgical treatment of drug-resistant tuberculosis. *Journal of Thoracic Disease* 2014;**6**:186.

119 Nunn P, Reid A, De Cock KM. Tuberculosis and HIV Infection: The Global Setting. *J Inf Dis* 2007;196:S5-14.

120 Vynnycky EE, Fine PEP. Interpreting the decline in tuberculosis: the role of secular trends in effective contact. *International Journal of Epidemiology* 1999;**28**:327–34.

121 Murray C. World Tuberculosis Burden. *Lancet* 1990;**335**:1043–4.

122 Glynn JR. Resurgence of tuberculosis and the impact of HIV infection. *British Medical Bulletin* 1998,54 (No 3),579-593

123 Zumla A, Malon P, Henderson J, *et al.* Impact of HIV infection on tuberculosis. *Postgrad Med J* 2000;76:259-268

124 Raviglione MC, Rieder HL, Styblo K, *et al.* Tuberculosis trends in eastern Europe and the former USSR. whqlibdoc.who.int. http://whqlibdoc.who.int/hq/1994/WHO_TB_94.176.pdf?ua=1 (Accessed 19 August 2014).

125 Russell DG, Barry CE, Flynn JL. Tuberculosis: what we don't know can, and does, hurt us. *Science* 2010;**328**:852–6.

126 Glaziou P, Floyd K, Korenromp EL, *et al.* WHO | Lives saved by tuberculosis control and prospects for achieving the 2015 global target for reducing tuberculosis mortality. *WHO* 2011. http://www.who.int/bulletin/volumes/89/8/11-087510/en/ (Accessed 19 August 2014)

127 Kochi A. TB: A Global Emergency. whqlibdoc.who.int. http://whqlibdoc.who.int/hq/1994/WHO_TB_94.177.pdf?ua=1 (accessed 19 Aug2014).

128 World Health Organization. Tuberculosis Programme. *Framework for Effective Tuberculosis Control*. WHO/TB/94.179 http://whqlibdoc.who.int/hq/1994/WHO_TB_94.179.pdf (Accessed 3 October 2014)

129 Programme WGT. *An Expanded DOTS Framework for Effective Tuberculosis Control*. 2002. WHO/CDS/TB/2002.297 http://www.who.int/tb/publications/expanded_dots_framework/en/ (Accessed 3 October 2014)

130 United Nations Millennium Development Goals. http://www.un.org/millenniumgoals/ (Accessed 3 October 2014)

131 World Health Organization. *Global Tuberculosis Report 2014*. 2014. http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809_eng.pdf?ua=1 (Accessed 28 November 2014)

132 ECDC and WHO/Europe joint report on tuberculosis surveillance and monitoring in Europe. *Euro Surveill* 2014;**19**:49–9.

133 Public Health England. Tuberculosis in the UK: 2013 report. http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317139689583 (Accessed 13 October 2013).

134 Public Health England. Tuberculosis in the UK: 2014 report. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/360335/TB_Annual_report__4_0_300914.pdf (Accessed 9 October 2014).

135 Public Health England. Annual TB update 2014. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/329393/PHE_annual_tuberculosis_update_2014.pdf (Accessed 19 December 2014).

136 Burki T. Tackling tuberculosis in London's homeless population. *The Lancet* 2010;**376**:2055–6.

137 LOnnroth K, Jaramillo E, Williams BG, *et al.* Drivers of tuberculosis epidemics: The role of risk factors and social determinants. *Social Science & Medicine* 2009;**68**:2240–6.

138 Riley RL. Airborne Pulmonary Tuberculosis. *Bacteriological Reviews*

1961;**25**:243.

139 Morrison J, Pai M, Hopewell PC. Tuberculosis and latent tuberculosis infection in close contacts of people with pulmonary tuberculosis in low-income and middle-income countries: a systematic review and meta-analysis. *Lancet Infect Dis* 2008;**8**:359–68.

140 Ewer K, Deeks J, Alvarez L, *et al.* Comparison of T-cell-based assay with tuberculin skin test for diagnosis of Mycobacterium tuberculosis infection in a school tuberculosis outbreak. *The Lancet* 2005;**361**:1168–73.

141 Gardy JL, Johnston JC, Sui SJH, *et al.* Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *N Engl J Med* 2011;**364**:730–9.

142 Valway SE, Sanchez MP, Shinnick TF, *et al.* An outbreak involving extensive transmission of a virulent strain of Mycobacterium tuberculosis. *N Engl J Med* 1998;**338**:633–9.

143 Curtis ABA, Ridzon RR, Vogel RR, *et al.* Extensive transmission of Mycobacterium tuberculosis from a child. *N Engl J Med* 1999;**341**:1491–5.

144 Riley EC, Murphy G, Riley RL. Airborne spread of measles in a suburban elementary school. *Amercian Journal of Epidemiology* 107:5;421-432

145 Andrews JR, Morrow C, Wood R. Modeling the Role of Public Transportation in Sustaining Tuberculosis Transmission in South Africa. *American Journal of Epidemiology* 2013;**177**:556–61.

146 Horna-Campos OJ, Bedoya-Lama A, Romero-Sandoval NC, *et al.* Risk of tuberculosis in public transport sector workers, Lima, Peru. *Int J Tuberc Lung Dis* 2010;**14**:714–9.

147 Feske ML, Teeter LD, Musser JM, *et al.* Giving TB wheels: Public transportation as a risk factor for tuberculosis transmission. *Tuberculosis (Edinb)* 2011;**91 Suppl 1**:S16–23.

148 Wood R, Johnstone Robertson S, Uys P, *et al.* Tuberculosis transmission to young children in a South African community: modeling household and community infection risks. *Clin Infect Dis* 2010;**51**:401–8.

149 Johnstone Robertson S, Lawn SD, Welte A, *et al.* Tuberculosis in a South African prison - a transmission modelling analysis. *S Afr Med J* 2011;**101**:809–13.

150 Richardson ET, Morrow CD, Kalil DB, *et al.* Shared air: a renewed focus on ventilation for the prevention of tuberculosis transmission. *PLoS ONE* 2014;**9**:e96334.

151 Kenyon TA, Valway SE, Ihle WW, *et al.* Transmission of Multidrug-Resistant Mycobacterium tuberculosis during a Long Airplane Flight. *N*

Engl J Med 1996;**334**:933–8.

152 WHO. *Tuberculosis and Air Travel: Guidelines for Prevention and Control*. 3rd ed. Geneva: World Health Organization 2008. http://www.who.int/tb/publications/2008/WHO_HTM_TB_2008.399_eng.pdf (Accessed 6 October 2014)

153 National Collaborating Centre for Chronic Conditions (UK), Centre for Clinical Practice at NICE (UK). *Tuberculosis: Clinical Diagnosis and Management of Tuberculosis, and Measures for Its Prevention and Control*. London: National Institute for Health and Clinical Excellence (UK) 2011. http://www.nice.org.uk/guidance/cg117/resources/cg117-tuberculosis-full-guideline3 (Accessed 19 December 2014)

154 Eang MT, Satha P, Yadav RP, *et al.* Early detection of tuberculosis through community-based active case finding in Cambodia. *BMC Public Health* 2012;**12**:469–9.

155 Golub JE, Mohan CI, Comstock GW, *et al.* Active case finding of tuberculosis: historical perspective and future prospects. *Int J Tuberc Lung Dis* 2005;**9**:1183–203.

156 Bothamley GH, Ditiu L, Migliori GB, *et al.* Active case finding of tuberculosis in Europe: a Tuberculosis Network European Trials Group (TBNET) survey. *European Respiratory Journal* 2008;**32**:1023–30.

157 National Tuberculosis Controllers Association, Centers for Disease Control and Prevention (CDC). Guidelines for the investigation of contacts of persons with infectious tuberculosis. Recommendations from the National Tuberculosis Controllers Association and CDC. MMWR Recomm Rep. 2005;**54**:1–47.

158 Cook VJ, Shah L, Gardy J, *et al.* Recommendations on modern contact investigation methods for enhancing tuberculosis control [Review article]. *Int J Tuberc Lung Dis* 2011;**16**:297–305.

159 Tardin A, Dominicé Dao M, Ninet B, *et al.* Tuberculosis cluster in an immigrant community: case identification issues and a transcultural perspective. *Tropical Medicine & International Health* 2009;**14**:995–1002.

160 Asghar RJ, Patlan DE, Miner MC, *et al.* Limited utility of name-based tuberculosis contact investigations among persons using illicit drugs: results of an outbreak investigation. *J Urban Health* 2009;**86**:776–80. doi:10.1007/s11524-009-9378-z

161 Lobue P, Menzies D. Treatment of latent tuberculosis infection: An update. *Respirology* 2010;**15**:603–22.

162 Thompson NJ. Efficacy of various durations of isoniazid preventive therapy for tuberculosis: five years of follow-up in the IUAT trial. International Union Against Tuberculosis Committee on Prophylaxis. *Bull World Health Organ* 1982;**60**:555–64.

163 Snider DE. Preventive Therapy With Isoniazid. *JAMA* 1986;**255**:1579–83.

164 Churchyard GJ, Fielding KL, Lewis JJ, *et al.* A Trial of Mass Isoniazid Preventive Therapy for Tuberculosis Control. *N Engl J Med* 2014;**370**:301–10.

165 Pareek M, Watson JP, Ormerod LP, *et al.* Screening of immigrants in the UK for imported latent tuberculosis: a multicentre cohort study and cost-effectiveness analysis. *The Lancet* 2011;**11**:435–44.

166 Pareek M, Abubakar I, White PJ, *et al.* Tuberculosis screening of migrants to low-burden nations: insights from evaluation of UK practice. *European Respiratory Journal* 2011;**37**:1175–82.

167 Munsiff SS, Ahuja SD, King L, *et al.* Ensuring accountability: the contribution of the cohort review method to tuberculosis control in New York City. *Int J Tuberc Lung Dis* 2006;**10**:1133–9.

168 Anderson C, White J, Abubakar I, *et al.* Raising standards in UK TB control: introducing cohort review.

169 NHS Commissioning Board, Service specification No.2, Neonatal BCG immunisation programme November 2012. gov.uk. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213154/02-Neonatal-BCG-immunisation-programme.pdf (Accessed 9 October 2015).

170 McShane H. Tuberculosis vaccines: beyond bacille Calmette–Guérin. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2011;**366**:2782.

171 Abubakar I, Pimpin L, Ariti C, *et al.* Systematic review and meta-analysis of the current evidence on the duration of protection by bacillus Calmette-Guérin vaccination against tuberculosis. *Health Technol Assess* 2013;**17**:1–372–v–vi.

172 Dahm R. Friedrich Miescher and the discovery of DNA. *Dev Biol* 2005;**278**:274–88.

173 Watson JD, Crick FH. Molecular structure of nucleic acids. *Nature* 1953;**171**:737–8.

174 Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;**74**:5463–7.

175 Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci USA* 1977;**74**:560–4.

176 Staden R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* 1979;**6**:2601–10.

177 Messing J, Crea R, Seeburg PH. A system for shotgun DNA sequencing.

Nucleic Acids Res 1981;**9**:309–21.

178 Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* 1987;**155**:335–50.

179 Cole ST, Brosch R, Parkhill J, *et al.* Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* 1998;**393**:537–44.

180 Brenner S, Johnson M, Bridgham J, *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 2000;**18**:630–4.

181 Jarvie T. Next generation sequencing technologies. *Drug Discov Today Technol* 2005;**2**:255–60.

182 Check E. Illumina board rejects Roche offer. blogs.nature.com. http://blogs.nature.com/news/2012/02/illumina-board-rejects-roche-offer.html#wpn-more-15023 (Accessed 11 October 2014).

183 Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources* 2014;**14**:1097–102.

184 Illumina | Sequencing and array-based solutions for genetic research. illumina.com. http://www.illumina.com/ (Accessed 12 October 2014).

185 Illumina. Technology spotlight: Illumina Sequencing. res.illumina.com. http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf (Accessed 11 October 2014).

186 Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010;**95**:315–27.

187 Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Meth* 2009;**6**:S6–S12.

188 David J Edwards KEH. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation* 2013;**3**:2.

189 Didelot X, Bowden R, Wilson DJ, *et al.* Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet* 2012;**13**:601–12.

190 Underwood A, Green J. Call for a Quality Standard for Sequence-Based Assays in Clinical Microbiology: Necessity for Quality Assessment of Sequences Used in Microbial Identification and Typing. *Journal of Clinical Microbiology J Clin Micro* 2011,49(1)23-26.

191 Catanzaro A, Davidson B, Fujiwara P, *et al.* Rapid diagnostic tests for tuberculosis: what is the appropriate use? American Thoracic Society Workshop. *American Journal of Respiratory and Critical Care Medicine*

1997;**155**:1804–14.

192 Heym B, Honoré N, Schurra C, *et al.* Implications of multidrug resistance for the future of short-course chemotherapy of tuberculosis: a molecular study. *The Lancet* 1994;**344**:293–8.

193 Mazars E, Lesjean S, Banuls AL, *et al.* High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology. *Proc Natl Acad Sci USA* 2001;**98**:1901.

194 Richter E, Weizenegger M, Rusch-Gerdes S, *et al.* Evaluation of Genotype MTBC Assay for Differentiation of Clinical Mycobacterium tuberculosis Complex Isolates. *Journal of Clinical Microbiology* 2003;**41**:2672–5.

195 Boehme CC, Nabeta P, Hillemann D, *et al.* Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. *N Engl J Med* 2010;**363**:1005–15.

196 Weyer K, Mirzayev F, Migliori GB, *et al.* Rapid molecular TB diagnosis: evidence, policy making and global implementation of Xpert MTB/RIF. *Eur Respir J* 2013;**42**:252–71.

197 Zhang Y, Heym B, Allen B, *et al.* The catalase—peroxidase gene and isoniazid resistance of Mycobacterium tuberculosis. *Nature* 1992;**358**:591–3.

198 Haas WH, Schilke K, Brand J *et al*. Molecular analysis of katG gene mutations in strains of Mycobacterium tuberculosis complex from Africa. *Antimicrob Agents Chemother* 1997;**41**:1601.

199 Banerjee A, Dubnau E, Quemard A, *et al.* inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. *Science* 1994;**263**:227–30.

200 Telenti A, Imboden P, Marchesi F, *et al.* Detection of rifampicin-resistance mutations in Mycobacterium tuberculosis. *The Lancet* 1993;**341**:647–51. doi:10.1016/0140-6736(93)90417-F

201 Musser JM. Antimicrobial agent resistance in mycobacteria: molecular genetic insights. *Clinical Microbiology Reviews* 1995;**8**:496–514.

202 Sreevatsan S, Stockbauer KE, Pan X *et al*. Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations. *Antimicrob Agents Chemother* 1997;**41**:1677.

203 N Honoré STC. Streptomycin resistance in mycobacteria. *Antimicrob Agents Chemother* 1994;**38**:238.

204 Takiff HE, Salazar L, Guerrero C, *et al.* Cloning and nucleotide sequence of Mycobacterium tuberculosis gyrA and gyrB genes and detection of quinolone resistance mutations. *Antimicrob Agents Chemother*

1994;**38**:773–80.

205 Suzuki Y, Katsukawa C, Tamaru A, *et al.* Detection of Kanamycin-ResistantMycobacterium tuberculosis by Identifying Mutations in the 16S rRNA Gene. *J Clin Micro*,1998;36:5,1220-1225.

206 Cooksey RC, Morlock GP, Glickman S, *et al.* Evaluation of a line probe assay kit for characterization of rpoB mutations in rifampin-resistant Mycobacterium tuberculosis isolates from New York City. *J Clin Micro,* 1997;35:51281-1283

207 Hillemann D, Weizenegger M, Kubica T, *et al.* Use of the genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in Mycobacterium tuberculosis complex isolates. *Journal of Clinical Microbiology* 2005;**43**:3699–703.

208 Brossier F, Veziris N, Jarlier V, *et al.* Performance of MTBDR plus for detecting high/low levels of Mycobacterium tuberculosis resistance to isoniazid. *Int J Tuberc Lung Dis* 2009;**13**:260–5.

209 Brossier F, Veziris N, Aubry A, *et al.* Detection by GenoType MTBDRsl test of complex mechanisms of resistance to second-line drugs and ethambutol in multidrug-resistant Mycobacterium tuberculosis complex isolates. *Journal of Clinical Microbiology* 2010;**48**:1683–9.

210 Ritter C, Lucke K, Sirgel FA, *et al.* Evaluation of the AID TB resistance line probe assay for rapid detection of genetic alterations associated with drug resistance in Mycobacterium tuberculosis strains. *Journal of Clinical Microbiology* 2014;**52**:940–6.

211 Schürch AC, van Soolingen D. DNA fingerprinting of Mycobacterium tuberculosis: from phage typing to whole-genome sequencing. *Infect Genet Evol* 2012;**12**:602–9.

212 Oelemann MC, Diel R, Vatin V, *et al.* Assessment of an Optimized Mycobacterial Interspersed Repetitive- Unit-Variable-Number Tandem-Repeat Typing System Combined with Spoligotyping for Population-Based Molecular Epidemiology Studies of Tuberculosis. *Journal of Clinical Microbiology* 2007;**45**:691–7.

213 Allix-Beguec C, Harmsen D, Weniger T, *et al.* Evaluation and Strategy for Use of MIRU-VNTRplus, a Multifunctional Database for Online Analysis of Genotyping Data and Phylogenetic Identification of Mycobacterium tuberculosis Complex Isolates. *Journal of Clinical Microbiology* 2008;**46**:2692–9.

214 Allix-Beguec C, Fauville-Dufaux M, Supply P. Three-Year Population-Based Evaluation of Standardized Mycobacterial Interspersed Repetitive-Unit-Variable-Number Tandem-Repeat Typing of Mycobacterium tuberculosis. *Journal of Clinical Microbiology* 2008;**46**:1398–406.

215 Hawkey PM, Smith EG, Evans JT, *et al.* Mycobacterial Interspersed

Repetitive Unit Typing of Mycobacterium tuberculosis Compared to IS6110-Based Restriction Fragment Length Polymorphism Analysis for Investigation of Apparently Clustered Cases of Tuberculosis. *Journal of Clinical Microbiology* 2003;**41**:3514–20.

216 Abubakar I, Lipman M, Anderson C, *et al.* Tuberculosis in the UK--time to regain control. *BMJ* 2011;**343**:d4281–1.

217 Li J, Driver CR, Munsiff SS, *et al.* Finding contacts of homeless tuberculosis patients in New York City. *Int J Tuberc Lung Dis* 2003;**7**:S397–404.

218 Schurch AC, Kremer K, Daviena O, *et al.* High-Resolution Typing by Integration of Genome Sequencing Data in a Large Tuberculosis Cluster. *Journal of Clinical Microbiology* 2010;**48**:3403–6.

219 Schürch AC, Kremer K, Kiers A, *et al.* The tempo and mode of molecular evolution of Mycobacterium tuberculosis at patient-to-patient scale. *Infection, Genetic and Evolution* 2010;**10**:108–14.

220 Casali N, Nikolayevskyy V, Balabanova Y, *et al.* Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Research* 2012;**22**:735–45.

221 Evans JT. Global Origin of Mycobacterium tuberculosis in the Midlands, UK. *Emerg Infect Dis* 2010;**16**.

222 Gibson A, Brown T, Baker L, *et al.* Can 15-Locus Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Analysis Provide Insight into the Evolution of Mycobacterium tuberculosis? *Applied and Environmental Microbiology* 2005;**71**:8207–13.

223 Didelot X, Eyre DW, Cule M *et al*. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. *Genome Biology* 2012;**13**:R118.

224 Evans JT, Serafino Wani RL, Anderson L, *et al.* A Geographically-Restricted but Prevalent Mycobacterium tuberculosis Strain Identified in the West Midlands Region of the UK between 1995 and 2008. *PLoS ONE* 2011;**6**:e17930.

225 van Soolingen D, Hermans PW, de Haas PE, *et al.* 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. *Journal of Clinical Microbiology* 1991;**29**:2578–86.

226 Dragan AI, Casas-Finet JR, Bishop ES, *et al.* Characterization of PicoGreen Interaction with dsDNA and the Origin of Its Fluorescence Enhancement upon Binding. *Biophysical Journal* 2010;**99**:3010–9.

227 Gerton Lunter MG. Stampy: A statistical algorithm for sensitive and fast

mapping of Illumina sequence reads. *Genome Research* 2011;**21**:936–9.

228 Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

229 Eyre DW, Golubchik T, Gordon NC, *et al.* A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. *BMJ Open* 2012;**2**:e001124.

230 Young BCB, Golubchik TT, Batty EME, *et al.* Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. *Proc Natl Acad Sci USA* 2012;**109**:4550–5.

231 Rodrigo AG, Felsenstein J. Coalescent appraoches to HIV population genetics. In: Crandall KA, ed. *Evolution of HIV*. Baltimore, Md.: : Johns Hopkins University Press, 1999 233–72.

232 Guindon S, Dufayard J-F, Lefort V, *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol.* 59(3):307-321, 2010

233 Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol Biol Evol* 2010;**27**:221–4.

234 Liu X, Gutacker MM, Musser JM, *et al.* Evidence for Recombination in Mycobacterium tuberculosis. *Journal of Bacteriology* 2006;**188**:8169–77.

235 Hirsh AE, Tsolaki AG, DeRiemer K, *et al.* Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proc Natl Acad Sci USA* 2004;**101**:4871–6.

236 Bryant JM, Schurch AC, van Deutekom H. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC infectious Diseases* 2013,13:110;3-12

237 Pérez-Lago L, Comas I, Navarro Y, *et al.* Whole Genome Sequencing Analysis of Intrapatient Microevolution in Mycobacterium tuberculosis: Potential Impact on the Inference of Tuberculosis Transmission. *J Inf Dis* 2014 Jan1;209(1):98-108

238 Bryant JM, Harris SR, Parkhill J, *et al.* Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. *The Lancet Respiratory Medicine* 2013;**1**:786–92.

239 Guerra-Assunção JA, Houben RMGJ, Crampin AC, *et al.* Recurrence due to Relapse or Reinfection With Mycobacterium tuberculosis: A Whole-Genome Sequencing Approach in a Large, Population-Based Cohort With a High HIV Infection Prevalence and Active Follow-up. *J Infect Dis* 2015, 211(7):1154-1163

240 Colijn C, Gardy J. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health* 2014(1):96-108

241 Kline SE, Hedemark LL, Davies SF. Outbreak of tuberculosis among regular patrons of a neighborhood bar. *N Engl J Med* 1995;**333**:222–7.

242 Kato-Maeda M, Ho C, Passarelli B, *et al.* Use of Whole Genome Sequencing to Determine the Microevolution of Mycobacterium tuberculosis during an Outbreak. *PLoS ONE* 2013;**8**:e58235.

243 Mallard K, McNerney R, Crampin AC, *et al.* Molecular Detection of Mixed Infections of Mycobacterium tuberculosis Strains in Sputum Samples from Patients in Karonga District, Malawi. *Journal of Clinical Microbiology* 2010;**48**:4512–8.

244 Worby CJ, Lipsitch M, Hanage WP. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput Biol* 2014;**10**:e1003549–9.

245 Kohl TA, Diel R, Harmsen D, *et al.* Whole-Genome-Based Mycobacterium tuberculosis Surveillance: a Standardized, Portable, and Expandable Approach. *J Clin Micro* 2014, 52(7),2479-2486

246 Stucki D, Ballif M, Bodmer T, *et al.* Tracking a tuberculosis outbreak over 21 years: strain-specific single nucleotide polymorphism-typing combined with targeted whole genome sequencing. *J Infect Dis 2014* Oct 30 [Epub ahead of print]

247 Evans JT, Smith EG, Banerjee A, *et al.* Cluster of human tuberculosis caused by Mycobacterium bovis: evidence for person-to-person transmission in the UK. *Lancet* 2007;**369**:1270–6.

248 Niemann S, Köser CU, Gagneux S, *et al.* Genomic Diversity among Drug Sensitive and Multidrug Resistant Isolates of Mycobacterium tuberculosis with Identical DNA Fingerprints. *PLoS ONE* 2009;**4**:e7407.

249 Al-Hajoj SAM, Akkerman O, Parwati I, *et al.* Microevolution of Mycobacterium tuberculosis in a tuberculosis patient. *Journal of Clinical Microbiology* 2010;**48**:3813–6.

250 Roetzer A, Diel R, Kohl TA, *et al.* Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *Plos Med* 2013;**10**:e1001387.

251 Middelkoop K, Mathema B, Myer L, *et al.* Transmission of Tuberculosis in a South African Community With a High Prevalence of HIV Infection. *J of Infect DIs* 2015;211:53-61

252 Roetzer A, Schuback S, Diel R, *et al.* Evaluation of Mycobacterium tuberculosis typing methods in a 4-year study in Schleswig-Holstein, Northern Germany. *Journal of Clinical Microbiology* 2011;**49**:4173–8.

253 Walker TM, Ip CL, Harrell RH, *et al.* Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis* 2013;**13**:137–46.

254 Kruijshaar ME, Abubakar I, Dedicoat M, *et al.* Evidence for a national problem: continued rise in tuberculosis case numbers in urban areas outside London. *Thorax* 2012;**67**:275–7.

255 Health Protection Agency. TB Strain Typing Cluster Investigation Handbook for Health Protection Units, 2nd Edition, September 2011. http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317131018354 (Accessed 2 December 2013).

256 World Health Organization. *Global Tuberculosis Report 2012*. http://apps.who.int/iris/bitstream/10665/75938/1/9789241564502_eng.pdf (Accessed 25 April 2013)

257 Erkens C, Slump E, Kamphorst M, *et al.* Coverage and yield of entry and follow-up screening for tuberculosis among new immigrants. *European Respiratory Journal* 2008;**32**:153–61.

258 Svensson E, Millet J, Lindqvist A, *et al.* Impact of immigration on tuberculosis epidemiology in a low-incidence country. *Clinical Microbiology and Infection* 2010;**17**:881–7.

259 Garzelli C, Rindi L. Molecular epidemiological approaches to study the epidemiology of tuberculosis in low-incidence settings receiving immigrants. *Infect Genet Evol* 2012;**12**:610–8.

260 Dahle UR, Eldholm V, Winje BA, *et al.* Impact of Immigration on the Molecular Epidemiology of Mycobacterium tuberculosis in a Low-Incidence Country. *Am J resp Crit Care Med*, 2007;176,930-935.

261 Borgdorff MW, Nagelkerke N, van Soolingen D, *et al.* Analysis of Tuberculosis Transmission between Nationalities in the Netherlands in the Period 1993–1995 Using DNA Fingerprinting. *Am J of Epidem* 1998;147(2),187-195.

262 Small PM, Hopewell PC, Singh SP, *et al.* The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med* 1994;**330**:1703–9.

263 Borgdorff MW, Sebek M, Geskus RB, *et al.* The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. *International Journal of Epidemiology* 2011;**40**:964–70.

264 Hargreaves S, Carballo M, Friedland JS. Screening migrants for tuberculosis: where next? *Lancet Infect Dis* 2009;**9**:139–40.

265 Ormerod LP. Further evidence supporting programmatic screening for, and treatment of latent TB Infection (LTBI) in new entrants to the UK from high TB prevalence countries, *Thorax* 2013,68 (3):201.

266 UK pre-entry tuberculosis screening brief report 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/328468/TB_preentry_screening_brief_report_2013.pdf (Accessed 18 December 2014).

267 Metcalf EP, Davies JC, Wood F, *et al.* Unwrapping the diagnosis of tuberculosis in primary care: a qualitative study. *Br J Gen Pract* 2007;**57**:116–22.

268 Story A, Aldridge RW, Abubakar I, *et al.* Active case finding for pulmonary tuberculosis using mobile digital chest radiography: an observational study. *Int J Tuberc Lung Dis* 2012;**16**:1461–7.

269 Martínez-Lirola M, Alonso-Rodriguez N, Sánchez ML, *et al.* Advanced Survey of Tuberculosis Transmission in a Complex Socioepidemiologic Scenario with a High Proportion of Cases in Immigrants. *Clin Infect Dis* 2008;**47**:8–14.

270 Genomics England: 100K Genome Project. genomicsengland.co.uk. http://www.genomicsengland.co.uk/100k-genome-project/ (Accessed 7 November 2013).

271 Casali N, Nikolayevskyy V, Balabanova Y, *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 2014;**46**:279–86.

272 Feng Y, Liu S, Wang Q, *et al.* Rapid Diagnosis of Drug Resistance to Fluoroquinolones, Amikacin, Capreomycin, Kanamycin and Ethambutol Using Genotype MTBDRsl Assay: A Meta-Analysis. *PLoS ONE* 2013;**8**:e55292.

273 Drobniewski F, Nikolayevskyy V, Maxeiner H, *et al.* Rapid diagnostics of tuberculosis and drug resistance in the industrialized world: clinical and public health benefits and barriers to implementation. *BMC Medicine* 2013;**11**:190.

274 Daum LT, Rodriguez JD, Worthy SA, *et al.* Next-generation ion torrent sequencing of drug resistance mutations in Mycobacterium tuberculosis strains. *Journal of Clinical Microbiology* 2012;**50**:3831–7.

275 Köser CU, Bryant JM, Becq J, *et al.* Whole-Genome Sequencing for Rapid Susceptibility Testing of M. tuberculosis. *N Engl J Med* 2013;**369**:290–2.

276 Clark TG, Mallard K, Coll F, *et al.* Elucidating Emergence and Transmission of Multidrug-Resistant Tuberculosis in Treatment Experienced Patients by Whole Genome Sequencing. *PLoS ONE* 2013;8(12)e83012

277 Farhat MR, Shapiro BJ, Kieser KJ, *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet* 2013;45(10):1183-9

278 Feuerriegel S, Oberhauser B, George A, *et al.* Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting. *BMC Microbiol* 2012;**12**:90–0.

279 Iqbal Z, Caccamo M, Turner I, *et al.* De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* 2012;**44**:226–32.

280 Stamatakis A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* 2014;30(9):1312-3

281 Pupko T, Pe I, Shamir R, *et al.* A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Mol Biol Evol* 17(6):890-896

282 Maynard Smith J, Smith NH. Detecting recombination from gene trees. *Mol Biol Evol* 1998;**15**:590–9.

283 Chewapreecha C, Marttinen P, Croucher NJ, *et al.* Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet* 2014;**10**:e1004547.

284 Sreevatsan S, Pan X, Zhang Y, *et al.* Mutations associated with pyrazinamide resistance in pncA of Mycobacterium tuberculosis complex organisms. *Antimicrob Agents Chemother* 1997;**41**:636–40.

285 Molina-Moya B, Lacoma A, Prat C, *et al.* AID TB resistance line probe assay for rapid detection of resistant Mycobacterium tuberculosis in clinical samples. *Journal of Infection* Published Online First: October 2014.

286 Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput Biol* 2015;**11**:e1004041. d

287 Zhang Z, Wang Y, Pang Y, *et al.* Ethambutol Resistance as Determined by Broth Dilution Method Correlates Better than Sequencing Results with embB Mutations in Multidrug-Resistant Mycobacterium tuberculosis Isolates. *J Clin Micro* 2014;52(2),638-641

288 Ocheretina O, Escuyer VE, Mabou M-M, *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. *PLoS ONE* 2014;**9**:e90569.

289 Rodwell TC, Valafar F, Douglas J, *et al.* Predicting Extensively Drug-Resistant Mycobacterium tuberculosis Phenotypes with Genetic Mutations. *J Clin Micro* 2014;52(3),781-789

290 Lin SYG, Rodwell TC, Victor TC, *et al.* Pyrosequencing for Rapid Detection of Extensively Drug-Resistant Mycobacterium tuberculosis in Clinical Isolates and Clinical Specimens. *Journal of Clinical Microbiology* 2014;**52**:475–82.

291 Comas I, Homolka S, Niemann S, *et al.* Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. *PLoS ONE* 2009;**4**:e7815.

292 *TB Drug Resistance Mutation Database*. http://www.tbdreamdb.com (Accessed 12 September 2013).

293 Safi H, Lingaraju S, Amin A, *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl. *Nat Genet* 2013;**45**:1190–7.

294 Köser C, Feuerriegel S, Summers DK, *et al*. Importance of the Genetic Diversity within the Mycobacterium tuberculosis Complex for the Development of Novel Antibiotics and Diagnostic Tests of Drug Resistance. *Antimicrob Agents Chemother* 2012;**56**:6080.

295 Scorpio A, Lindholm-Levy P, Heifets L, *et al.* Characterization of pncA mutations in pyrazinamide-resistant Mycobacterium tuberculosis. *Antimicrob Agents Chemother* 1997;**41**:540–3.

296 Miotto P, Cabibbe AM, Feuerriegel S, *et al.* Mycobacterium tuberculosis Pyrazinamide Resistance Determinants: a Multicenter Study. *MBio* 2014:5(5)e01819-14.

297 Thumamo BP, Asuquo AE, Abia-Bassey LN, *et al.* Molecular epidemiology and genetic diversity of Mycobacterium tuberculosis complex in the Cross River State, Nigeria. *Infection, Genetics and Evolution* 2012;**12**:671–7.

298 Horne D, Pinto LM, Arentz M, *et al*. Diagnostic Accuracy and Reproducibility of WHO-Endorsed Phenotypic Drug Susceptibility Testing Methods for First-Line and Second-Line Antituberculosis Drugs. *Journal of Clinical Microbiology* 2013;**51**:393–401.

299 Zhang Y, Wade MM, Scorpio A, *et al.* Mode of action of pyrazinamide: disruption of Mycobacterium tuberculosis membrane transport and energetics by pyrazinoic acid. *J Antimicrob Chemother* 2003;**52**:790–5.

300 Loman NJ, Misra RV, Dallman TJ, *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature Publishing Group* 2012;**30**:434–9.

301 Koller D, Friedman N. Probabilistic Graphical Models. mitpress.mit.edu. http://mitpress.mit.edu/books/probabilistic-graphical-models Accessed 11 November 2014).

302 Public Health England. Collaborative Tuberculosis Strategy for England 2015 to 2020. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/396263/CollaborativeTBStrategyEngland_FINAL.pdf (Accessed 5 February 2015)

303 The End TB Strategy. who.int. http://who.int/tb/post2015_TBstrategy.pdf?ua=1 (Accessed 3 January 2015).

304 Frieden TR, Fujiwara PI, WASHKO RM, *et al.* Tuberculosis in New-York-City - Turning the Tide. *N Engl J Med* 1995;**333**:229–33.

305 Paolo WF, Nosanchuk JD. Tuberculosis in New York city: recent lessons and a look ahead. *Lancet Infect Dis* 2004;**4**:287–93.

306 Johnstone-Robertson SP, Mark D, Morrow C, *et al.* Social Mixing Patterns Within a South African Township Community: Implications for Respiratory Disease Transmission and Control. *American Journal of Epidemiology* 2011;**174**:1246–55.

307 Boehme CC, Nicol MP, Nabeta P, *et al.* Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. *Lancet* 2011;**377**:1495–505.

308 Theron G, Zijenah L, Chanda D, *et al.* Feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing for tuberculosis in primary-care settings in Africa: a multicentre, randomised, controlled trial. *The Lancet* 2014;**383**:424–35.

309 CROI 2014: GeneXpert real world study yields mixed results, highlights SA health system challenges | Science Speaks: HIV & TB News. sciencespeaksblog.org. http://sciencespeaksblog.org/2014/03/05/croi-2014-genexpert-real-world-study-yields-mixed-results-highlights-sa-health-system-challenges/ (Accessed 8 February 2015).

310 Qin ZZ, Pai M, Van Gemert W, *et al.* How is Xpert MTB/RIF being implemented in 22 high tuberculosis burden countries? *Eur Respir J* 2015;**45**:549–54.

311 Hillemann D, Rüsch-Gerdes S, Richter E. Evaluation of the GenoType MTBDRplus Assay for Rifampin and Isoniazid Susceptibility Testing of Mycobacterium tuberculosis Strains and Clinical Specimens. *J Clin Micro* 2007;45(8),2635-2640.

312 Mironova SS, Pimkina EE, Kontsevaya II, *et al.* Performance of the GenoType® MTBDRPlus assay in routine settings: a multicenter study. *Eur J Clin Microbiol Infect Dis* 2012;**31**:1381–7.

313 Köser CU, Holden MTG, Ellington MJ, *et al.* Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak. *N Engl J Med* 2012;**366**:2267–75.

314 Votintseva AA, Pankhurst LJ, Anson LW, *et al.* Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *J Clin Micro 2015,* published online 28th January

315 Doughty EL, Sergeant MJ, Adetifa I, *et al.* 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.

316 Press releases - Oxford Nanopore Technologies. nanoporetech.com. https://www.nanoporetech.com/news/press-releases/view/39 (Accessed 10 February 2015).

317 Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology* 2012;**30(4)**:295–6.

318 New-generation solar panels far cheaper, more efficient: scientists. *http://www.reuters.com/article/2015/01/27/us-global-renewables-solar-idUSKBN0L020720150127* (Accessed 10 February 2015)

319 Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience* 2014;**3**:22.

320 FIND - Price for Xpert® MTB/RIF and FIND country list. http://www.finddiagnostics.org/about/what_we_do/successes/find-negotiated-prices/xpert_mtb_rif.html (Accessed 10 February 2015)

321 Wejse C. Point-of-care diagnostics for tuberculosis elimination? *The Lancet* 2014;**383**:388–90.

322 Sekandi JN, Dobbin K, Oloya J, *et al.* PLOS ONE: Cost-Effectiveness Analysis of Community Active Case Finding and Household Contact Investigation for Tuberculosis Case Detection in Urban Africa.**10**:e0117009.