Investigating the genomic basis of antimicrobial resistance in *Mycobacterium tuberculosis* (*Mtb*) using genome-wide methodologies

Yaa Emily Adoma Oppong

Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy
University of London
September 2019

Department of Infection Biology

Faculty of Infectious and Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by BBSRC

Research group affiliation(s): Martin Hibberd
Taane Clark

DECLARATION OF OWN WORK

All students are required to complete the following declaration when submitting their thesis.

Please note: Assessment misconduct includes any activity that compromises the integrity of your research or assessment of is will be considered under the Assessment Irregularity Policy. This includes plagiarism, cheating and failure to follow correct progression and examination procedures.

Please see the following documents for further guidance:

- Research Degrees Handbook
- Assessment Irregularities Policy

Supervisors should be consulted if there are any doubts about what is permissible.

1. STUDENT DETAILS

Student ID Number	LSH1500052	**Title**	Ms.
First Name(s)	Yaa Emily Adoma		
Surname/Family Name	Oppong		
Programme of Study	PhD		
LSHTM Email (if this is no longer active, please provide an alternative)	yaa.oppong@lshtm.ac.uk		

2. TITLE OF THESIS

Title of Thesis	Investigating the genomic basis of antimicrobial resistance in *Mycobacterium tuberculosis* (*Mtb*) using genome-wide methodologies

3. DECLARATION

I have read and understood the LSHTM's definition of plagiarism and cheating. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people.

I have read and understood the LSHTM's definition and policy on the use of third parties (either paid or unpaid) who have contributed to the preparation of this thesis by providing copy editing and, or, proof reading services. I declare that no changes to the intellectual content or substance of this thesis were made as a result of this advice, and, that I have fully acknowledged all such contributions.

I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law or infringe any third party's copyright or other intellectual property right.

Student Signature	
Date	27/09/19

ABSTRACT OF THESES

1. STUDENT DETAILS

Student ID Number	LSH1500052	**Title**	Ms.
First Name(s)	Yaa Emily Adoma		
Surname/Family Name	Oppong		
Programme of Study	PhD		
LSHTM Email (if this is no longer active, please provide an alternative)	yaa.oppong@lshtm.ac.uk		

2. TITLE OF THESIS

Title of Thesis	Investigating the genomic basis of antimicrobial resistance in *Mycobacterium tuberculosis* (Mtb) using genome-wide methodologies

3. NOTES FOR CANDIDATES

- Type your abstract on the page two of this document
- Use single-space typing
- **Limit your abstract to one side of the sheet**
- Submit your Abstract to the Assessments team in the Registry: https://www.lshtm.ac.uk/study/studentservices/registry-services
- This abstract will be forwarded to the LSHTM Library, which will send this sheet to the British Library and to ASLIB (Association of Special Libraries and Information Bureau) for publication in Index to Theses.

4. ABSTRACT

Characterizing the drug resistance mutations that have evolved in *Mycobacterium tuberculosis* (*Mtb*), has important implications for control of tuberculosis (TB) disease, through more accurate and timely use of therapy. Whole genome sequencing of *Mtb* can assist this characterization by providing insights into loci and specific mutations underlying drug resistance and the transmission success that enables their spread.

We hypothesised that genetic variation outside of known resistance-conferring mutations might give additional information concerning drug resistance and fitness. Firstly, we explored the effect of lineage on the identification of drug resistance associations, applying novel lineage level genome-wide association study (GWAS) and convergence-based (PhyC) methods to drug resistance phenotypes of a global dataset of *Mtb* lineages 2 and 4. We identified known drug resistance variants and novel associations, uniquely identifying associations for lineage-specific GWAS analyses and reporting 17 novel associations between antimicrobial resistance phenotypes and *Mtb* genomic variants, demonstrating the utility of lineage-specific GWAS.

To further examine the genomic basis of extensively drug resistant (XDR)-TB*,* we next applied the GWAS and PhyC techniques to a global dataset of 18,255 *Mtb* isolates. Through GWAS we identified 20 loci in novel associations within highly drug-resistant *Mtb* strains*.* Cluster-based GWAS and a lack of overlap with associations identified through convergent-evolution-based analyses confirmed that many such associations have been driven by transmission in outbreaks of XDR-TB.

We then investigated the feasibility of applying a learning classifier system to this dataset to predict rifampicin resistance and discover candidate loci for novel involvement, finally enabling a sensitivity of 93.7% and a specificity of 94.8% of rifampicin resistance prediction.

Finally, we applied this methodology to the XDR phenotype in lineages 2 and 4 of a global dataset (n=13,270), achieving high accuracy of prediction and identifying a number of candidate loci for involvement in XDR, including candidates for epistasis.

Acknowledgements

I wish to thank;

My supervisors Martin Hibberd and Taane Clark.

The members of the group including Jody, Ernest, Matt R, Ben, Neneh, Matt H, Pepita, Dan, Amy, Gary and Anna.

Colleagues who have helped me throughout my studies; Francesc Coll, Stéphane Hué and Sonal Shah.

BBSRC and the LiDo team for continued support.

My mother, Annette.

My father, Kofi.

My brother, Kwabena.

My Nana, Christine.

Leah, Elena, Elisa, Sarah and Hannah.

Table of Contents

Abbreviation List

GWAS Genome-wide association analysis

HLA Human leukocyte antigen

Indel Insertion/deletion

LCS Learning classifier system

MIC Minimum Inhibitory Concentration

Mtb *Mycobacterium tuberculosis*

MDR Multidrug resistant

NGS Next Generation Sequencing

PCA Principal component analysis

SNP Single Nucleotide Polymorphism

SVM Support vector machine

WGS Whole genome sequencing

XDR Extensively drug resistant

Chapter 1:
Introduction

Tuberculosis Disease

Global Tuberculosis Disease Burden

With an estimated 10.0 million people developing tuberculosis (TB) in 2017 and an estimated 1.6 million deaths [1], the global burden of TB is overwhelming. Worldwide, it is estimated that between 1.7 billion individuals are infected with the causative agent, *Mycobacterium tuberculosis* (*Mtb*), with 5-10% of those infected developing active disease [1]. There is great variation in the distribution of tuberculosis disease globally. Of those who developed TB in 2017, two thirds were in eight countries; India, China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh and South Africa [1]. Most high income countries had less than 10 new cases per 100,000 population, whilst there were more than 500 cases per 100,000 population in countries such as Mozambique, the Philippines and South Africa [1].

Global Diversity of *Mtb*

Mtb is a member of a larger group of related species, known as the *Mtb* complex. Today, *Mtb* has seven lineages, defined on the basis of molecular typing, which are endemic in different locations around the globe, with some persisting in geographical regions (lineages 5 and 6 in West Africa) and others across continents (lineage 2- East Asian/Beijing strains or lineage 4 – Euro-American strains) leading to the hypothesis that the strain-types are specifically adapted to people of different genetic backgrounds [2]. These lineages may vary in propensity to transmit, virulence,

site of infection and ultimately propensity to cause disease [3–5] but results are inconsistent and there is considerable inter-strain variation within lineages [6, 7]. Recent research into lineage 4 alludes to this variation, suggesting different evolutionary strategies are employed by different sublineages [8]. A set of single nucleotide polymorphisms (SNPs) has been identified that can be used to barcode sub-lineages [9], leading to informatic tools that position sequenced samples within a global phylogeny [10].

This global distribution of modern *Mtb* lineages is thought to be explained by their distribution through human migration and evolution resulting from changes in selection pressure as human populations underwent increases in population density or neutral evolution [5, 11–14].

Tuberculosis Disease Aetiology

Under the classical model of TB, upon infection by *Mtb* there are three possible outcomes; clearance, latent disease or active disease. Transmission occurs during active disease due to the release of *Mtb* bacteria from the lungs in aerosol form. Inhalation of the *Mtb* bacterium can then result in macrophage infection within the lungs, invoking an immune response; macrophages likely phagocytose the bacteria, which are then resistant to macrophage killing mechanisms.

The primary site of interaction between host and *Mtb* is the granuloma; which forms as a result of host immune response. Latent disease occurs when the granuloma successfully contains the *Mtb* infection. This is an active and complex process and lack

of containment allows dissemination of *Mtb* and progression to active TB disease [15–17]. This granuloma diversity may impact in disease progression and outcomes; there is evidence that there is a spectrum of latent disease [18] and that reactivation risk is granuloma specific, with the potential of *Mtb* dissemination from only one or a few granuloma to cause active disease [19, 20].

A number of reasons have been put forward to explain host heterogeneity in response to *Mtb* infection including; environmental host factors such as nutrition or HIV+ status, resulting in immunosuppression, *Mtb* genetics and host genetics. Indeed, a number of studies have found associations between human ethnicity and *Mtb* populations [2, 21], and additionally human ethnicity and *M. africanum* populations [22]. Such findings have led to suggestions of population specific adaptations in *Mtb* and host-pathogen coevolution [5]. Furthermore, ethnicity has been implicated in clinical TB phenotype [23].

Vaccines

Attenuated *Mycobacterium bovis* strain bacillus Callmette Guerin (BCG) has been used as a vaccine against TB since 1921. BCG shows variable efficacy [24], with factors such as previous exposure to non-tuberculous mycobacteria, human genetic variation and genetic variation of the vaccine strain itself potentially implicated [25]. There remains a need for a new vaccine with increased efficacy [26].

Treatment and the Evolution of Resistance

Regardless of the initial emergence of *Mtb*, the widespread use and misuse of chemical antimicrobials since the 1960s likely represents a major new selection pressure governing *Mtb* evolution. Indeed, the emergence of drug resistance in *Mtb* is threatening disease control efforts. The evolution of drug resistance has occurred, despite clonal reproduction and a lack of lateral gene transfer in *Mtb*. *Mtb* resistance has developed to all anti-*Mtb* drugs, usually relatively shortly after their introduction and now isolates occur with multiple different drug resistances.

Drug-resistant TB is phenotypically categorised as singly resistant to any anti-Mtb drug, multi-drug resistant (MDR), resistance to two first–line treatments, rifampicin and isoniazid; extensively drug-resistant (XDR), defined as MDR alongside resistance to fluoroquinolones and at least one second-line injectable; or totally drug-resistant (TDR or XXDR). In 2017, MDR TB amounted to 3.5% of new TB cases globally and 8.5% of these were XDR [1]. Treatment success rates for MDR and XDR TB are only 55% and 34%, respectively [1].

Two new drugs, bedaquiline and delaminid, have recently been introduced, but there remains a need for further development of new drugs and drug regimens, alongside increased drug susceptibility testing, better diagnosis and easier access to continued treatment [1].

Diagnosis and Surveillance

Currently culture-based assays are the standard for diagnosis and drug susceptibility testing of clinical *Mtb* isolates. Due to the slow growth of *Mtb*, this process is time consuming. For simplicity and speed, determining drug susceptibility is routinely done using agreed standardised cut-offs, resulting in the binary resistant or susceptible phenotype. However, these thresholds are subject to change over time and do not reflect the full diversity of the drug resistant phenotype. Additionally, it is possible to measure the minimum inhibitory concentration (MIC) required to prevent growth of an *Mtb* isolate, generating a more accurate continuous variable. Not only would this help to improve appropriate drug use for patients, but this data would greatly increase the information on determining the genetics of the drug resistance.

Rapid molecular tests offer an interesting alternative to standard drug susceptibility testing, such as XPERT MTB/RIF, which is PCR-based and is able to diagnose tuberculosis alongside resistance to rifampicin, as well as new methods showing promise as point of care tests [27].

Genomics of Drug Resistance

De novo emergence of drug resistance has been observed, with the presence of multiple unfixed drug-resistance mutations and selective sweeps in *Mtb* populations within patients [28–30]. Additionally, transmission of resistant strains is frequently observed [31, 32]. Indeed, many mutations associated with antimicrobial resistance have been identified [33], some have been associated with no fitness cost and others

with additional compensatory mutations that may increase fitness and enable transmission [34]; this area requires further investigation. Such mutations include both point mutations, for example, single nucleotide polymorphisms (SNPs) such as in *rpoB* [35] and structural variants such as the *dfrA-thyA* double deletion linked to para-aminosalicylic acid resistance [36]. Genes involved in resistance to some drugs are well known; for example, mutations for rifampicin (in *rpoB* and *rpoC*) and isoniazid (in *katG*) are well characterised [33]. However, the mechanisms for ethambutol (*embB*), pyrazinimde (*pncA*) and second line drugs are not fully known. As whole genome sequencing (WGS) is applied to *Mtb* more routinely [37] , association approaches using genomic variation have the potential to provide new insights into these resistance mechanisms. Compensatory mutations such as those in *rpoA* and *rpoC,* associated with the *rpoB* rifampicin resistance mutations, have been associated with transmission of drug resistant strains [38].

Whole Genome Sequencing

WGS is increasingly being applied to *Mtb,* either after culturing or directly from sputum[39–42]. A number of tools have been developed to predict drug susceptibility from genome sequence [43–50].

Additionally, WGS has important implications for TB surveillance; it allows the phylogenetic reconstruction of relationships between strains and thus the inference of transmission events [51–54]. Such inference of transmission events can be used to inform public health strategy. Furthermore, WGS can enable the monitoring of drug

resistance evolution at a genomic level, for example determining the relative importance of within patient evolution versus transmission of drug resistance [28–32].

Genome-Wide Discovery Methods

Genome Wide Association

The genome-wide association study (GWAS) approach has been applied as a method to discover genomic variants involved in drug resistance phenotypes.

This approach seeks to identify genomic variants associated with phenotypes of interest, classically through case-control designs, in which the genomic data of a group with the phenotype of interest is compared to that of a group without the phenotype of interest, and statistical association is assessed. Such methods are firmly established in the field of human genetics research.

However, there are important differences between humans and bacterial pathogens in relation to GWAS methodology that must be considered. Haploid organisms like bacteria often form highly structured populations resulting from transmission and clonal reproduction. They may even group into distinct lineages, which may have important biological differences encoded in their genome. Thus, there is a need to deal with population structure, such that relatedness is accounted for, minimising spurious associations as a result of common genetic background, whilst maximising sensitivity to detect biologically relevant effects. Furthermore, for many bacterial species, there are added complications, such as lateral gene transfer, which

can lead to variable gene content between isolates; there may be differences in the presence or absence of entire genes [55].

With the development of linear mixed model approaches that seek to account for relatedness [56] and increasing availability of bacterial whole genome sequences, there is new potential in the application of GWAS to understand evolutionary dynamics in bacterial pathogens [57].

It is interesting to note that human GWAS typically rely on typed SNPs, and exploit patterns of known linkage disequilibrium, the patterns by which variants in close proximity are commonly co-inherited, within human populations, to link SNPs to potential causal genomic variants. However, WGS data negates the need for characterised linkage disequilibrium patterns and allows the possibility of detecting causal variants directly.

The GWAS approach, widely used in human genetics, is increasingly being applied to pathogen research and shows great promise [58]. It allows the identification of variants across the genome, associated with specific phenotypes, and has been used in humans, for example, to identify variants in the class II human leukocyte antigens (HLA) region associated with susceptibility to TB infection [59]. In order to prevent spurious associations, pathogen GWAS face the need to deal with the much higher levels of population structure seen in bacteria compared to humans, whilst maximising sensitivity [56, 60]. This is especially prescient for *Mtb* due to its clonality.

GWAS has been demonstrated as a useful methodology for investigating drug resistance in *Mtb,* identifying genes known to be involved in drug resistance phenotypes [61–64].

Convergence-Based Methods

PhyC is a methodology that seeks to identify genomic signatures of selection resulting from convergent evolution. This occurs when a mutation independently becomes fixed multiple times. To detect such events, ancestral reconstruction is performed for each variant, using parsimony to infer at which node in the phylogenetic tree the mutation event likely occurred. The variant frequency in branches with the phenotype in question can then be compared to the variant frequency in branches without the phenotype and tested for statistical difference [65].

These methods have yet to be universally applied across pathogen species. As a clonal bacterial pathogen with no evidence of lateral gene transfer, *Mtb*, the causal agent of TB, may prove to be a useful organism on which to develop such methods. This is especially prescient in relation to drug resistance phenotypes, as the use of anti-microbial therapy represents a strong selective force on *Mtb*.

Convergence-based methods have been used to identify resistance mutations in *Mtb* [65, 66]. Such methods seek to identify convergent evolution in phenotypically resistant strains. This occurs when mutations in the same gene or nucleotide position repeatedly and independently become fixed, thus signalling positive selection for a particular phenotype.

Previous applications of convergence-based methods in *Mtb* have resulted in the identification of a number of new targets of independent selection in relation to drug resistance, including a mutation in *ponA1* [65] and *prpR* [66]. It was shown through functional genetic analyses that strains carrying the *ponA1* mutation showed a survival advantage during in vitro growth in the presence of rifampicin [65]; whilst common variants in *prpR* were found to confer multidrug tolerance [66].

Furthermore, both association and convergent evolution analyses have been successfully combined together, resulting in the finding that loss of function mutations in *ald* confer resistance to D-cycloserine in *Mtb* [67]. Although there have been notable difficulties in disentangling individual drug resistance-conferring variants from drug resistance phenotypes that are not directly related, due to co-occurring drug resistance phenotypes as a result of combined drug therapy, and there is a need for further methods development [61, 67].

Application of Machine Learning

Broadly, there has been interest in the application of machine learning methods to predict drug resistance in *Mtb.* A number of methods have been employed, demonstrating their potential utility in predicting resistance phenotypes from *Mtb* genome sequence, including support vector machines (SVM), k-nearest neighbour clustering, random forests and neural networks [68–76]. Such applications include approaches that did not require the need for mapping next generation sequencing (NGS) reads to a reference sequence [68]. Unlike GWAS and phyC, which detect

additive effects, such approaches might offer capability to detect epistasis [68]. However, explainability of some machine learning models can be limited, and due to high computational costs, such analyses are often restricted to specific loci already known to be involved in resistance.

Learning classifier systems (LCS) may offer an interesting approach to disentangling epistatic interactions in relation to drug resistance in *Mtb*. LCS broadly work through the creation of populations of rules; collectively these rules form the model. Each rule predicts phenotype based on the state of one or more attributes. During prediction, rules across the whole population 'vote' for a specific phenotype. Rule populations are formed through supervised learning; a genetic algorithm is employed in which rules can reproduce, recombine and mutate as a function of their prediction accuracy, with preference for more general rules to prevent overfitting [77]. Thus, LCS can be considered to employ an evolutionary approach to learning.

In this way, LCS may cope with epistasis- where multiple attributes contribute to a phenotype, as well as heterogeneity within the genomics of a specific phenotype- where multiple different genomic mechanisms can underpin the phenotype. Further to this, inspection of rules within the rule population may provide insight into the biological mechanisms involved in a given phenotype.

Data

The data used throughout this work comprises of WGS data coupled with drug resistance phenotype data aggregated from multiple studies (see Chapters 2-5). Clinical

isolates from individual patients underwent WGS and drug susceptibility testing; isolates were cultured and susceptibility was tested using phenotypic testing protocols recognized by WHO [63,78]. This resulted in a binary phenotype for resistance versus susceptibility. Each isolate was not necessarily tested for susceptibility to each drug; where isolates were found to be susceptible to first-line treatments, they often did not undergo additional drug susceptibility testing for second-line treatments.

Project Overview

This work is structured in four parts, as explained below.

Title	Published
Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage- specific associations with drug resistance	2019
Genome-wide analyses identify novel associations with Extensively Drug Resistant tuberculosis	Under Review
Genome-wide machine learning classifier applied to *Mycobacterium tuberculosis* as a novel approach to unravel genomic complexity associated with drug resistance	In prep.
Genome-wide Learning Classifier System applied to Extensively Drug Resistant *Mycobacterium tuberculosis* discovers novel resistance mechanisms	In prep.

References

1. World Health Organisation. Global Tuberculosis Report 2018. Geneva; 2018.

2. Reed MB, Pichler VK, Mcintosh F, Mattia A, Fallow A, Masala S, et al. Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. J Clin Microbiol. 2009;47:1119–28.

3. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. Clin Infect Dis. 2012;54:211–9.

4. Krishnan N, Malaga W, Constant P, Caws M, Thi Hoang Chau T, Salmons J, et al. *Mycobacterium tuberculosis* lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. PLoS One. 2011;6;e23870.

5. Gagneux S. Host-pathogen coevolution in human tuberculosis. Philos Trans R Soc Lond B Biol Sci. 2012;367:850–9. doi:10.1098/rstb.2011.0316.

6. Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. PLoS Pathog. 2011;7; e1001307.

7. Mathema B, Kurepina N, Yang G, Shashkina E, Manca C, Mehaffy C, et al. Epidemiologic consequences of microvariation in *Mycobacterium tuberculosis*. J Infect Dis. 2012;205:964–74.

8. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted

sublineages. Nat Genet. 2016; doi:10.1038/ng.3704.

9. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun. 2014;5:4812. doi:10.1038/ncomms5812.

10. Benavente ED, Coll F, Furnham N, McNerney R, Glynn JR, Campino S, et al. PhyTB: Phylogenetic tree visualisation and sample positioning for *M. tuberculosis*. BMC Bioinformatics. 2015;16:155. doi:10.1186/s12859-015-0603-3.

11. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature. 2014; doi:10.1038/nature13591.

12. Kay GL, Sergeant MJ, Zhou Z, Chan JZ-M, Millard A, Quick J, et al. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nat Commun. 2015;6:6717. doi:10.1038/ncomms7717.

13. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45:1176–82. doi:10.1038/ng.2744.

14. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, et al. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog. 2008;4;e1000160.

15. Queval CJ, Brosch R, Simeone R. The macrophage: A disputed fortress in the battle against Mycobacterium tuberculosis. Front Microbiol. 2017;8;1–11.

16. de Martino M, Lodi L, Galli L, Chiappini E. Immune Response to *Mycobacterium*

tuberculosis: A Narrative Review. Front Pediatr. 2019;7;1–8.

17. BoseDasgupta S, Pieters J. Striking the right balance determines TB or not TB. Front Immunol. 2014;5 ;1–9.

18. Barry CE, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, et al. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. Nat Rev Microbiol. 2009;7:845–55. doi:10.1038/nrmicro2236.

19. Lin PL, Ford CB, Coleman MT, Myers AJ, Gawande R, Ioerger T, et al. Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. Nat Med. 2013;20:75–9. doi:10.1038/nm.3412.

20. Cadena AM, Fortune SM, Flynn JL. Heterogeneity in tuberculosis. Nat Rev Immunol. 2017;1–12. doi:10.1038/nri.2017.69.

21. Gagneux S, Deriemer K, Van T, Kato-maeda M, Jong BC De. Variable host – pathogen compatibility in *Mycobacterium tuberculosis*. 2006;103:2869–73.

22. Asante-Poku A, Yeboah-Manu D, Otchere ID, Aboagye SY, Stucki D, Hattendorf J, et al. Mycobacterium africanum is associated with patient ethnicity in Ghana. PLoS Negl Trop Dis. 2015;9;e3370.

23. Pareek M, Evans J, Innes J, Smith G, Hingley-Wilson S, Lougheed KE, et al. Ethnicity and mycobacterial lineage as determinants of tuberculosis disease phenotype. Thorax. 2012;221–9.

24. Colditz GA, Brewer TF, Berkey CS, Burdick E, Fineberg H V, Mosteller F. Vaccine in the prevention of tuberculosis efficacy of BCG. JAMA. 1994;271:698–702.

25. Abdallah AM, Hill-Cawthorne GA, Otto TD, Coll F, Guerra-Assunção JA, Gao G, et al.

Genomic expression catalogue of a global collection of BCG vaccine strains show evidence for highly diverged metabolic and cell-wall adaptations. Scientific Reports. 2015;5;15443. doi:10.1038/srep15443.

26. World Health Organisation. Global Tuberculosis Report 2015. 2015.

27. Xie YL, Chakravorty S, Armstrong DT, Hall SL, Via LE, Song T, et al. Evaluation of a rapid molecular drug-susceptibility test for tuberculosis. N Engl J Med. 2017;377;1043–54.

28. Mariam SH, Werngren J, Aronsson J, Hoffner S, Andersson DI. Dynamics of antibiotic resistant *Mycobacterium tuberculosis* during long-term infection and antibiotic treatment. PLoS One. 2011;6:e21147. doi:10.1371/journal.pone.0021147.

29. Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, et al. Dynamic Population Changes in Mycobacterium tuberculosis During Acquisition and Fixation of Drug Resistance in Patients. J Infect Dis. 2012;206:1724–33. doi:10.1093/infdis/jis601.

30. Fortune SM. The surprising diversity of *Mycobacterium tuberculosis*: Change you can believe in. J Infect Dis. 2012;206:1642–4.

31. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One. 2013;8:1–12.

32. Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, et al. Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. Nat Commun. 2015;6;7119. doi:10.1038/ncomms8119.

33. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G,

et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7:51. doi:10.1186/s13073-015-0164-0.

34. Müller B, Borrell S, Rose G, Gagneux S. The heterogeneous evolution of multidrug-resistant *Mycobacterium tuberculosis*. Trends Genet. 2013;29:160–9.

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

36. Moradigaravand D, Grandjean L, Martinez E, Li H, Zheng J, Coronel J, et al. *DfrA-thyA* double deletion in *para* -aminosalicylic acid resistant *Mycobacterium tuberculosis* Beijing strains. Antimicrob Agents Chemother. 2016; March:AAC.00253-16. doi:10.1128/AAC.00253-16.

37. Galagan JE. Genomic insights into tuberculosis. Nat Rev Genet. 2014;15:307–20. doi:10.1038/nrg3664.

38. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet. 2014;46:279–86. doi:10.1038/ng.2878.

39. Nguyen TNA, Berre VA Le, Bañuls AL, Nguyen TVA. Molecular diagnosis of drug-resistant tuberculosis; A literature review. Front Microbiol. 2019;10;1–12.

40. Lowenthal P, Lin S-YG, Desmond E, Shah N, Flood J, Barry PM. Evaluation of the impact of a sequencing assay for detection of drug resistance on the clinical management of tuberculosis. Clin Infect Dis. 2019;69:668–75.

41. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, et al. Direct whole-

genome sequencing of sputum accurately identifies drug-Resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. J Clin Microbiol. 2018;56:1–11. doi:10.1128/JCM.00666-18.

42. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. J Clin Microbiol. 2015;53:2230–7.

43. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun. 2015;6;018564.

44. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: Targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics. 2014;15:1–12.

45. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: A web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. J Clin Microbiol. 2015;53:1908–14.

46. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med. 2019;11:41. doi:10.1186/s13073-019-0650-x.

47. Mahé P, El Azami M, Barlas P, Tournoud M. A large scale evaluation of TBProfiler and Mykrobe for antibiotic resistance prediction in *Mycobacterium tuberculosis* . PeerJ. 2019;7:e6857.

48. van Beek J, Haanperä M, Smit PW, Mentula S, Soini H. Evaluation of whole genome

sequencing and software tools for drug susceptibility testing of *Mycobacterium tuberculosis*. Clin Microbiol Infect. 2019;25:82–6.

49. Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, et al. TGS-TB: Total genotyping solution for *Mycobacterium tuberculosis* using short-read whole-genome sequencing. PLoS One. 2015;10:1–12.

50. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. N Engl J Med. 2018;379:1403–15. doi:10.1056/NEJMoa1800474.

51. Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, et al. Tracing *Mycobacterium tuberculosis* transmission by whole genome sequencing in a high incidence setting: A retrospective population-based study in East Greenland. Sci Rep. 2016;6;1–8. doi:10.1038/srep33180.

52. Packer S, Green C, Brooks-Pollock E, Chaintarli K, Harrison S, Beck CR. Social network analysis and whole genome sequencing in a cohort study to investigate TB transmission in an educational setting. BMC Infect Dis. 2019;19:1–8.

53. Lee RS, Behr MA. The implications of whole-genome sequencing in the control of tuberculosis. Ther Adv Infect Dis. 2016;3:47–62.

54. Lee RS, Pai M. Real-time sequencing of *Mycobacterium tuberculosis*: Are we there yet? J Clin Microbiol. 2017;55:1249–54. doi:10.1128/JCM.00358-17.

55. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2016. doi:10.1038/nrg.2016.132.

56. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in

bacterial association studies. 2015. http://arxiv.org/abs/1510.06863.

57. Falush D. Bacterial genomics: Microbial GWAS coming of age. Nat Microbiol. 2016;1:16059. doi:10.1038/nmicrobiol.2016.59.

58. Cain AK, Lees JA. Using genomics to combat infectious diseases on a global scale. Genome Biol. 2015;16:250. doi:10.1186/s13059-015-0822-y.

59. Sveinbjornsson G, Gudbjartsson DF, Halldorsson B V, Kristinsson KG, Gottfredsson M, Barrett JC, et al. HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. Nat Genet. 2016;48:318–22. doi:10.1038/ng.3498.

60. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178:1709–23. doi:10.1534/genetics.107.080101.

61. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 2013;45:1255–60. doi:10.1038/ng.2735.

62. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE V., Furnham N, et al. *Mycobacterium tuberculosis* whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med. 2016;14:31. doi:10.1186/s12916-016-0575-9.

63. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. Nat Genet. 2018;50:307–16. doi:10.1038/s41588-017-0029-0.

64. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, et al. GWAS for

quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nat Commun. 2019;10;2128. doi:10.1038/s41467-019-10110-6.

65. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet. 2013;45:1183–9. doi:10.1038/ng.2747.

66. Hicks ND, Yang J, Zhang X, Zhao B, Grad YH, Liu L, et al. Clinically prevalent mutations in *Mycobacterium tuberculosis* alter propionate metabolism and mediate multidrug tolerance. Nat Microbiol. 2018;3:1032–42. doi:10.1038/s41564-018-0218-3.

67. Desjardins CA, Cohen KA, Munsamy V, Abeel T, Maharaj K, Walker BJ, et al. Genomic and functional analyses of *Mycobacterium tuberculosis* strains implicate ald in D-cycloserine resistance. Nat Genet. 2016;48 October 2015:1–9. doi:10.1038/ng.3548.

68. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, et al. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. Nat Commun. 2018;9:4306. doi:10.1038/s41467-018-06634-y.

69. Yang Y, Walker TM, Iqbal Z, Walker AS, Daniel J, Peto TEA, et al. DeepAMR for predicting co-occurrent resistance of *Mycobacterium tuberculosis*. 2018;:1–8.

70. Chen ML, Doddi A, Royer J, Freschi L, Ezewudo M, Kohane IS, et al. Deep learning predicts tuberculosis drug resistance status from whole-genome sequencing data. 2018. bioRxiv. http://dx.doi.org/10.1101/275628.

71. Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, Peto TEA, et al.

Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics. 2018;35;2276-2282. doi:10.1093/bioinformatics/bty949.

72. Sergeev RS, Kavaliou I, Sataneuski U, Gabrielian A, Rosenthal A, Tartakovsky M, et al. Genome-wide analysis of MDR and XDR tuberculosis from Belarus: Machine-learning approach. IEEE/ACM Trans Comput Biol Bioinforma. 2019;16:1–1. doi:10.1109/TCBB.2017.2720669.

73. Chowdhury AS, Khaledian E, Broschat SL, Science C, States U. Capreomycin resistance prediction in two species of *Mycobacterium* using a stacked ensemble method. Journal of Applied Microbiology. 2019;127;1656-1664.

74. Huang H, Ding N, Yang T, Li C, Jia X, Wang G, et al. Cross-sectional whole-genome sequencing and epidemiological study of multidrug-resistant *Mycobacterium tuberculosis* in China. Clin Infect Dis. 2019;69:405–13.

75. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial resistance prediction in PATRIC and RAST. Sci Rep. 2016;6:1–12. doi:10.1038/srep27930.

76. Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, et al. DeepAMR for predicting co-occurrent resistance of *Mycobacterium tuberculosis* . Bioinformatics. 2019;35;3240-324.

77. Urbanowicz RJ, Moore JH. Learning Classifier Systems: A Complete Introduction, Review, and Roadmap. J. Artif. Evol. Appl. 2009;2009:1–25.

78. World Health Organization. Guidelines for surveillance of drug resistance in tuberculosis. WHO Geneva/IUATLD Paris. International Union Against Tuberculosis and

Lung Disease. [Internet]. 2009. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/9562116

Chapter 2:

Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage- specific associations with drug resistance

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1500052	**Title**	Ms
First Name(s)	Yaa Emily Adoma		
Surname/Family Name	Oppong		
Thesis Title	Investigating the genomic basis of antimicrobial resistance in *Mycobacterium tuberculosis* (*Mtb*) using genome-wide methodologies		
Primary Supervisor	Martin Hibberd		

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

<u>SECTION B – Paper already published</u>

Where was the work published?	BMC Genomics
When was the work published?	March 2019
If the work was published prior to registration for your research	NA

degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

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

SECTION C – Prepared for publication, but not yet published

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

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	**TC and MH conceived and directed the project. JPh generated the sequencing dataset. YO performed bioinformatic and statistical analyses under the supervision of TC and MH, and wrote the first draft of the manuscript. JPe, DM, AM, IP and MV contributed protocols and data. All authors commented and edited on various versions of the draft manuscript. All authors compiled and approved the final manuscript.**

SECTION E

Student Signature	
Date	

Supervisor Signature	
Date	

Oppong et al. BMC Genomics (2019) 20:252
https://doi.org/10.1186/s12864-019-5615-3

BMC Genomics

RESEARCH ARTICLE Open Access

Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance

Yaa E. A. Oppong[1*], Jody Phelan[1], João Perdigão[2], Diana Machado[3], Anabela Miranda[4], Isabel Portugal[2], Miguel Viveiros[3], Taane G. Clark[1,5†] and Martin L. Hibberd[1†]

Abstract

Background: Continuing evolution of *the Mycobacterium tuberculosis (Mtb)* complex genomes associated with resistance to anti-tuberculosis drugs is threatening tuberculosis disease control efforts. Both multi- and extensively drug resistant *Mtb* (MDR and XDR, respectively) are increasing in prevalence, but the full set of *Mtb* genes involved are not known. There is a need for increased sensitivity of genome-wide approaches in order to elucidate the genetic basis of anti-microbial drug resistance and gain a more detailed understanding of *Mtb* genome evolution in a context of widespread antimicrobial therapy. Population structure within the *Mtb* complex, due to clonal expansion, lack of lateral gene transfer and low levels of recombination between lineages, may be reducing statistical power to detect drug resistance associated variants.

Results: To investigate the effect of lineage-specific effects on the identification of drug resistance associations, we applied genome-wide association study (GWAS) and convergence-based (PhyC) methods to multiple drug resistance phenotypes of a global dataset of *Mtb* lineages 2 and 4, using both lineage-wise and combined approaches. We identify both well-established drug resistance variants and novel associations; uniquely identifying associations for both lineage-specific and -combined GWAS analyses. We report 17 potential novel associations between antimicrobial resistance phenotypes and *Mtb* genomic variants.

Conclusions: For GWAS, both lineage-specific and -combined analyses are useful, whereas PhyC may perform better in contexts of greater diversity. Unique associations with XDR in lineage-specific analyses provide evidence of diverging evolutionary trajectories between lineages 2 and 4 in response to antimicrobial drug therapy.

Keywords: Drug resistance, Evolution, Mutations, *Mycobacterium tuberculosis*, Tuberculosis

Background

Despite clonal expansion and a lack of lateral gene transfer in *Mycobacterium tuberculosis* (*Mtb*), the evolution of drug resistance is threatening tuberculosis disease (TB) control efforts. Resistance to all anti-*Mtb* drugs has been observed, usually evolving relatively shortly after their introduction. Drug-resistant TB is phenotypically categorised as multi-drug resistant (MDR) when resistant to two first–line drugs, rifampicin and isoniazid; extensively drug-resistant (XDR) occurs when MDR *Mtb* have additional resistance to fluoroquinolones and at least one second-line injectable. Only 50% of patients receiving treatment for MDR TB, globally, were successfully treated in 2014 [1].

De novo emergence of drug resistance has been observed, with the presence of multiple unfixed drug-resistance mutations and selective sweeps in *Mtb* populations within patients [2–4]. Additionally, transmission of resistant strains is frequently observed [5, 6]. Indeed, many mutations associated with antimicrobial resistance have been identified [7], some have been associated with no fitness cost and others with additional compensatory mutations that may increase fitness and enable transmission [8]. These polymorphisms include both point mutations, for example,

* Correspondence: yaa.oppong@lshtm.ac.uk
Taane G Clark and Martin L. Hibberd are joint authors
[1]Pathogen Molecular Biology Department, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Full list of author information is available at the end of the article

single nucleotide polymorphisms (SNPs) such as in *rpoB* [9] and structural variants such as the *dfrA-thyA* double deletion linked to para-aminosalicylic acid resistance [10]. Genes involved in resistance to some drugs are well known; for example, mutations for rifampicin (in *rpoB* and *rpoC*) and isoniazid (in *katG*) are well characterised [7]. However, the mechanisms for ethambutol (*embB*), pyrazinamide (*pncA*) and second line drug resistance are not fully known. As whole genome sequencing of *Mtb* becomes more routinely applied [11], association approaches using genomic variation have the potential to provide new insights into these resistance mechanisms. Compensatory mutations such as those in *rpoA* and *rpoC*, associated with the *rpoB* rifampicin resistance mutations, have been associated with transmission of drug resistant strains [12]. Furthermore, as patients receive a cocktail of anti-*Mtb* drugs, multiple concomitant resistance can arise naturally, and this complicates the analysis of phenotype-genotype relationships [13].

The genome-wide association study (GWAS) approach has been widely used in human genetics; for example, to identify variants in the class II human leukocyte antigens (HLA) region associated with susceptibility to TB infection [14]. However, it is increasingly being applied to pathogen research and shows great promise [13, 15, 16]. It allows the identification of variants across the genome, associated with specific phenotypes. In order to prevent spurious associations, pathogen GWASs face the need to deal with the much higher levels of population structure seen in bacteria compared to humans, whilst maximising sensitivity [17, 18]. This is especially important for *Mtb* due to its clonality. This clonality is consistent with a phylogenetic tree structure and thus has led to the application of convergence-based methods, which have identified resistance mutations in *Mtb* [13, 19]. Such methods seek to identify convergent evolution in genetically diverse strains with similar resistance phenotypes. This happens when mutations in the same gene or nucleotide position occur repeatedly and independently become fixed, thus signaling their positive selection for a particular phenotype.

However, there remain questions as to the importance of historic genetic background variation in the evolution of drug resistance, such as between *Mtb* lineages, which have not been systematically explored [20]. The *Mtb* complex is categorised into seven lineages, defined on the basis of molecular typing, which are endemic in different locations around the globe. These lineages are known to have other distinctive features, with some persisting in geographical regions (lineages 5 and 6 in West Africa) and others spreading across continents (lineage 2- East Asian and lineage 4 – Euro-American strains). This observation has led to the hypothesis that the strain-types are specifically adapted to people of different genetic backgrounds [21]. These lineages may vary in their propensity to transmit, their virulence, site of infection and ultimately propensity to cause disease [22–24], but results are inconsistent and there is considerable inter-strain variation within lineages [25, 26]. Recent research into lineage 4 alludes to this variation, suggesting different evolutionary strategies are employed by different sublineages [27]. A set of single nucleotide polymorphisms (SNPs) has been identified that can be used to barcode sub-lineages [28], leading to informatic tools that position sequenced samples within a global phylogeny [29]. Thus, lineage-based genetic differences may also be important in resistance adaptations to anti-*Mtb* drug exposure.

The current study applies lineage-specific and lineage-combined GWAS, alongside convergence-based PhyC methods, to gain insight into lineage-specific drug resistance evolution. We focus on the modern lineage 2 and lineage 4 isolates, which are known to be drug resistant globally, and use a large dataset involving *Mtb* isolate sequences from more than 12 countries ($n > 4400$).

Results

Genomic variants and population structure

High quality SNP and insertion and deletion (indel) variants were characterised in relation to the H37Rv reference genome, from raw sequence data from a convenience sample of existing data for isolates in lineages 2 ($n = 702$) and 4 ($n = 3706$). These isolates are within a global drug resistance data set [13], which has been further complemented by additional phenotypic data (see Methods). After removing variants that are monomorphic within each dataset, the final lineage-combined dataset consisted of 157,726 SNPs, 5998 deletions and 2926 insertions across the 4408 isolates (see Additional file 1). The median number of SNPs per sample in the lineage 2 dataset, after removing monomorphic variants, was 332 (range: 189–386) and in lineage 4 was 724 (range: 10–870) (significant difference between lineages with Wilcoxon test p-value < minimum calculable (2.2×10^{-16})). Lineage 4 contains the H37Rv reference strain, but also has increased strain-type diversity [13, 28]. The median number of indels per sample in lineage 2 was 31 (range: 7–42) and in lineage 4 was 40 (range: 2–61) (significant difference between lineages Wilcoxon test: p-value < minimum calculable (2.2×10^{-16})) (see Additional file 1). The majority of variants were rare, with 75% of them found to have a non-reference variant frequency (defined as the number of isolates with a non-reference allele at a specific variant position divided by the total number of isolates with a non-missing allele at this position) of less than 0.0028 and 0.00054 in lineages 2 and 4, respectively (see Additional file 1 and Additional file 2). A principal component analysis (PCA) using the variants revealed the expected clustering by lineage and greater diversity within lineage 4 (see Additional file 3). Within lineage 2, the first 10 principal components account for 71.9% of the variation

(see Additional file 3 and Additional file 4) and the mean pairwise variant distance was 1074 (range: 0–6270) (see Additional file 3). Within lineage 4, the first 10 principal components account for 88.9% of the variation (see Additional file 3 and Additional file 4) and the mean pairwise variant distance was 1458 (range: 0–11,780) (see Additional file 3). There are 567 isolates with < 10 variants different from at least one other isolate, indicative of potential transmission events, which can confound an association analysis. A phylogenetic tree constructed using the variants mimicked the relationships observed in the PCA, with isolates clustering by sublineage on both (see Additional file 3 and Fig. 1).

Drug resistance phenotypes

Overall, analyses were conducted for 17 drug resistance phenotypes, including for 12 individual drugs and 5 composite phenotypes. The 12 individual drug resistance phenotypes with frequency of resistance ranging from 3.3% (MOX in lineage 4) to 43.0% (STM in lineage 2), and the composite phenotypes of MDR (lineage 2 35.7%; lineage 4 9.5%) and XDR (lineage 2 9.9%; lineage 4 1.2%). The combined second-line drug resistance phenotypes for resistance to any fluoroquinones (FQ) and resistance to any aminoglycosides (AG) were also considered (see Additional file 5). The completeness of drug-resistance phenotype data is variable. Rifampicin was the most tested for (tested for in 92.0% of isolates); while ciprofloxacin was the least (tested for in 4.2% of isolates) (see Additional file 6). Furthermore, there is evidence of multiple concomitant resistance with 44.1% of MDR isolates also resistant to ethambutol.

Convergence-based analyses, variant-based GWAS and locus-based identified known resistance conferring variants

We performed convergence-based analyses (PhyC), GWAS across loci (locus-based) and GWAS on individual variants (variant-based). Each were conducted in a lineage-specific and lineage-combined manner. Due to the close relatedness between some samples, for the GWAS analyses, we applied specialized regression models with random effects that have been implemented in a human setting to handle “cryptic relatedness” [13] (see Methods).

In total, PhyC analysis of the combined lineages identified 53 variants in 20 different loci, with individual lineage analyses identifying a subset of these loci (see Table 1, Additional file 7). Eleven of these loci were not identified by GWAS techniques, including eight loci with known involvement in antimicrobial resistance; *thyX-hsdS.1* (para-aminosalicylic acid), *rpoC* (rifampicin), *pncA-Rv2044c* (pyrazinamide), *eis-Rv2417c* (aminoglycosides), *folC* (para-aminosalicylic acid), *fabG1* (isoniazid), *oxyR’-ahpC* (isoniazid) and *gyrB* (fluoroquinolones) (see Table 1, Additional file 8).

Locus-based GWAS identified 23 different loci (see Table 2, Fig. 2, Additional file 7). Fourteen such loci were identified by locus-based GWAS exclusively; of these 14 loci, *gid* is known to be involved in streptomycin resistance and *inhA* is known to be involved in isoniazid and ethionamide resistance [30, 31] (see Additional file 8). Variant-based GWAS identified eleven variants in nine different loci. No known associations were identified by variant-based GWAS exclusively; however, three novel associations were identified (*RV0197, recF, argJ*) (see Table 3, Additional file 8). Three loci were identified by locus-based GWAS and PhyC but not variant-based GWAS: *pncA* (pyrazinamide), *embC-embA* and *embB* (ethambutol) (see Fig. 3a and b, Additional file 8).

Effects of lineage-specific analysis on identifying known resistance associated variants

Lineage 2 specific

Overall, for locus-based GWAS analyses across the 16 phenotypes, two loci were identified exclusively to lineage 2 specific analyses; *rrs (KAN; p*-value = 1.40×10^{-22}) and *Rv3128c-Rv3129* (MDR; p-value = 7.4×10^{-22}) (see Fig. 2a). For locus-based GWAS, *pncA* was found in association with XDR exclusively, however for lineage 4 *pncA* was found in association with PZA exclusively; greater variation was found in the *pncA* locus for lineage 2 (see Fig. 3c and d). For the variant-based GWAS analyses there were no lineage 2 exclusive associations. Furthermore, no lineage 2 exclusive associations were identified by PhyC analyses.

Lineage 4 specific

Overall, for the locus-based GWAS analyses, seven loci were identified exclusively by lineage 4 specific analyses (*inhA*, *fadB4-Rv3142c*, *tuf*, *cut5b-Rv3725*, *Rv3007c*, *Rv2668*, *moeX*) (see Fig. 2b). All of which were found in significant association with the XDR phenotype. For locus-based GWAS, *gid* was identified in association with streptomycin by lineage 4 specific analyses and – combined analyses but not lineage 2 specific analyses; there is greater variation within the *gid* locus for lineage 4 (see Fig. 3e and f). The variant-based GWAS analyses identified no lineage 4 exclusive analyses. Moreover, no lineage 4 exclusive associations were identified by PhyC analyses.

Lineages 2 and 4 combined

Four loci were solely identified through combined lineage PhyC analyses; *Rv3115-moeB2* (MDR, STM; min. *p*-value = 6.7×10^{-4}), *eis-Rv2417c* (STM; min. p-value = 1.4×10^{-05}), *whib6-Rv3863* (EMB; p-value = 9×10^{-4}) and *oxyR’-ahpC* (INH, PZA; *p*-values = 6.8×10^{-4}, 9×10^{-4}, respectively) (see Table 1). For each loci identified by PhyC, there were consistently the same number or more

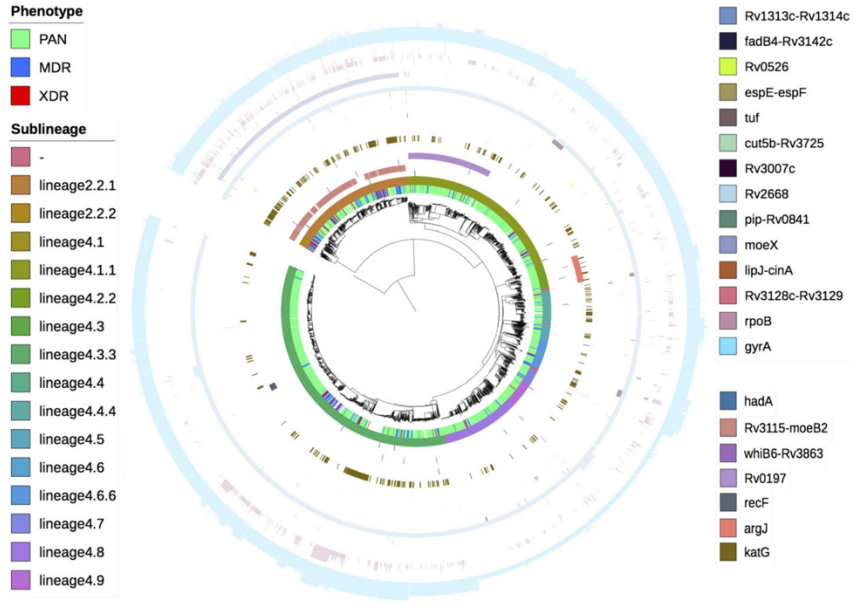


Fig. 1 Phylogenetic tree of all samples: coloured by phenotype, sublineage and genotype for novel variants alongside *katG*, *rpoB* and *gyrA*. From inner to outer, each track is coloured by; drug-resistance phenotype, sublineage, variant genotype for; *hadA* (732110), Rv3115-moeB2 (3482717), *whiB6-Rv3863* (4338594), *Rv0197* (232574), *recF* (4047), *argJ* (1867614), *katG* (2155168); locus genotype for; *Rv1313c-Rv1314c*, *fadB4-Rv3142c*, *Rv0526*, *espE-espF*, *tuf*, *cut5b-Rv3725*, *Rv3007c*, *Rv2668*, *pip-Rv0841*, *moeX*, *lipJ-cinA*, *Rv3128c-Rv3129*, *rpoB*, *gyrA*. Variant genotype is coloured in samples where a non-reference variant is present with respect to H37rv reference and variant position is given in brackets. Locus genotype is coloured in samples with one or more non-synonymous or intergenic variants at each locus with respect to H37rv reference, relative height of the bar reflects number of variants at each loci

associations identified by the -combined versus the lineage-specific analyses (see Table 1).

For locus-based GWAS, four loci were identified in association with XDR by the combined lineages 2 and 4 analyses exclusively; *Rv0526* (p-value = 8.70×10^{-37}; thioredoxin protein) and three intergenic regions; *espE-espF* (p-value = 5.70×10^{-31}), *pip-Rv0841* (p-value = 8.60×10^{-23}) and *lipJ-cinA* (p-value = 6.20×10^{-22}) (see Table 2, Fig. 2c).

For variant-based GWAS, one locus was identified by the combined lineages 2 and 4 analyses exclusively; *argJ*, in association with XDR (p-value = 6.9×10^{-26}) (see Table 3).

Novel resistance-associated variants identified

Across all analyses, we report 17 potentially novel associations between antimicrobial resistance and genomic variants in *Mtb;* 7 such associations were identified exclusively by lineage-specific analyses (see Tables 1, 2, 3). Twelve were identified by locus-based GWAS, three were identified by variant-based GWAS and two were identified by PhyC. All novel associations identified by GWAS were found in association with the XDR phenotype. There was no overlap in novel associations identified between methods.

Two potentially novel associations were identified by PhyC; *hadA* (lineage 4, 2 + 4; INH, MDR and STM; 1.1×10^{-4} < p-values< 4×10^{-4}) and *Rv3115-moeB2* (lineages 2 + 4; MDR; STM, min. p-value = 6.7×10^{-4}) (see Table 1, Fig. 1). The *Rv3115-moeB2* variant displays a different pattern of variation within lineage 2 than within lineage 4 (see Fig. 1).

Twelve potentially novel associations were identified by locus-based GWAS (see Table 1). Six loci were identified exclusively in lineage 4 all in association with XDR; *fadB4-Rv3142c* (p-value = 4.6×10^{-38}), *tuf* (p-value = 1.5×10^{-29}), *Rv3007c* (p-value = 7.8×10^{-24}), *cut5b-Rv3725* (p-value = 5.1×10^{-27}), *Rv2668* (p-value = 1.3×10^{-23}) and *moeX* (p-value = 5.5×10^{-22}). *Rv1313c-Rv1314c* was identified by both lineage 4 and lineage-combined analyses in association with XDR (min. p-value = 1.4×10^{-54}). Four loci were identified exclusively by lineage-combined analyses in association with XDR; *Rv0526* (p-value = 8.7×10^{-37}), *espE-espF* (p-value = 5.7×10^{-31}), *pip-Rv0841* (p-value = 8.6×10^{-23}) and *lipJ-cinA* (p-value = 6.2×10^{-22}). *Rv3128c-Rv3129* was identified exclusively by the lineage 2 analysis in association with MDR (p-value = 7.4×10^{-22}) (see Table 2, Fig. 1).

Three potentially novel associations were identified by variant-based GWAS, all in association with XDR; in the *Rv0197* locus (lineage 4, 2 + 4; min. p-value = 9.5×10^{-62}), in the *recF* locus (lineage 4, 2 + 4; min. p-value = 1.2×10^{-52}, respectively) and the *argJ* locus (lineages 2 + 4; p-value = 6.9×10^{-26}) (see Table 3, Fig. 1).

Table 1 Significant associations between genomic variants and drug resistance phenotypes identified by PhyC

Locus	Known Phenotype Association [13]	Lineage	Observed Phenotype Association (position; *p*-value)	Total
rpoB	RMP	2	INH(761,155;1.2×10^{-06}, 761,139;5.6×10^{-06}), MDR(761,155;1.5×10^{-12}, 761,139;1.4×10^{-07}, 761,140;5.5×10^{-04}), RMP(761,155;2.4×10^{-11}, 761,139;6.7×10^{-09}, 761,110;4.6×10^{-04}, 761,140;9.3×10^{-04}), XDR(761,155;2.6×10^{-04}, 761,139;7.5×10^{-04})	11
rpoB	RMP	4	AG(761,155;2.0×10^{-5}), EMB(761,155;2.2×10^{-14}), INH(761,155;3.4×10^{-32}, 761,139;6.2×10^{-11}, 761,110;7.6×10^{-08}, 761,140;3.3×10^{-07}), MDR(761,155;2.1×10^{-45}, 761,139;2.8×10^{-20}, 761,140;1.3×10^{-09}, 761,110;1×10^{-08}, 759,939;2.8×10^{-04}), PZA(761,155;1.5×10^{-12}, 761,110;2.2×10^{-06}), RMP(761,155;5.7×10^{-54}, 761,139;9.7×10^{-27}, 761,110;2.6×10^{-12}, 761,140;3.4×10^{-11}, 761,998;1.1×10^{-05}, 761,109;7.3×10^{-05}, 759,939;4.9×10^{-04}), STRx10P(761,155;2.1×10^{-16}, 761,110;7.5×10^{-04}), XDR(761,155;7.1×10^{-14}, 761,110;2.7×10^{-05})	23
rpoB	RMP	2 + 4	AG(761,155;2.3×10^{-5}), EMB(761,155;3.8×10^{-18}, 761,140;6.3×10^{-06}, 761,110;1.1×10^{-05}, 761,139;2×10^{-05}), INH(761,155;2.3×10^{-42}, 761,139;3.5×10^{-19}, 761,140;2.3×10^{-11}, 761,110;1.1×10^{-09}, 761,161;1.5×10^{-06}), MDR(761,155;7.8×10^{-63}, 761,139;1.7×10^{-31}, 761,140;4.8×10^{-15}, 761,110;1.3×10^{-13}, 761,161;5.8×10^{-05}, 761,095;1.8×10^{-04}, 759,939;2.6×10^{-04}, 761,109;2.6×10^{-04}), PZA(761,155;5.6×10^{-16}, 761,110;2×10^{-07}, 761,139;5.8×10^{-04}), RMP(761,155;3.5×10^{-70}, 761,139;3.2×10^{-38}, 761,110;2.4×10^{-17}, 761,140;2.2×10^{-15}, 761,161;4.4×10^{-07}, 761,109;3.5×10^{-05}, 761,998;1×10^{-04}, 761,095;3.1×10^{-04}, 759,939;4.7×10^{-04}, 760,314;4.7×10^{-04}), STM(761,155;8.4×10^{-21}, 761,110;2.1×10^{-07}, 761,139;2.6×10^{-07}, 761,140;8.5×10^{-05}, 761,161;1.8×10^{-04}), XDR(761,155;2.2×10^{-18}, 761,110;2.6×10^{-08}, 761,139;9.7×10^{-08}, 761,161;1.9×10^{-06}, 761,109;6.2×10^{-05})	40
embB	EMB	2	EMB(4,247,429;3×10^{-07}, 4,247,431;1.8×10^{-04}), INH(4,247,429;9.3×10^{-10}, 4,247,431;3.1 × 10–05), MDR(4,247,429;2.5×10^{-08}, 4,247,431;1.1×10^{-04}), RMP(4,247,429;7.6×10^{-09}, 4,247,431;1.3×10^{-04}, 4,247,730;8.7×10^{-04}), STM(4,247,429;1×10^{-05}, 4,247,431;1.1×10^{-04}), XDR(4,247,429;4.3×10^{-06}, 4,247,431;1.2×10^{-04}, 4,247,730;1.2×10^{-04})	14
embB	EMB	4	AG(4,247,431;7.1×10^{-4}), EMB(4,247,431;3×10^{-11}, 4,247,729;3×10^{-08}, 4,247,730;1×10^{-07}, 4,248,003;1×10^{-07}, 4,247,429;1.5×10^{-06}, 4,247,574;8.1×10^{-04}), INH(4,247,431;6.6×10^{-19}, 4,247,730;1.4×10^{-09}, 4,247,429;1.7×10^{-09}, 4,247,729;6.3×10^{-06}, 4,247,574;1.8×10^{-05}, 4,248,003;2.8×10^{-05}), MDR(4,247,431;2.3×10^{-21}, 4,247,429;1.5×10^{-09}, 4,247,730;8.8×10^{-08}, 4,247,574;6×10^{-07}, 4,247,729;4.2×10^{-06}, 4,248,003;6×10^{-04}), PZA(4,247,431;1.2×10^{-04}, 4,247,730;2.2×10^{-04}, 4,248,003;5.3×10^{-04}), RMP(4,247,431;2.2×10^{-18}, 4,247,429;1.5×10^{-10}, 4,247,730;1.4×10^{-08}, 4,247,574;6.6×10^{-05}, 4,247,729;1.3×10^{-04}, 4,248,003;1.5×10^{-04}), STM(4,247,431;1.5×10^{-08}, 4,247,729;3.6×10^{-06}, 4,247,574;7.2 × 10–04), XDR(4,247,429;1.3×10^{-04})	31
embB	EMB	2 + 4	AG(4,247,431;3.5×10^{-4}), EMB(4,247,429;9.2×10^{-21}, 4,247,431;1.5×10^{-16}, 4,247,729;2.1×10^{-09}, 4,247,730;6.4×10^{-08}, 4,248,003;2.8×10^{-07}, 4,247,574;1.4×10^{-04}, 4,249,518;1.9×10^{-04}), FQ(4,247,730;9.5×10^{-07}), INH(4,247,429;2.7×10^{-27}, 4,247,431;1×10^{-25}, 4,247,730;8.4×10^{-14}, 4,248,003;1.2×10^{-08}, 4,247,729;1.4×10^{-07}, 4,247,574;1.7×10^{-07}), MDR(4,247,431;3.2×10^{-26}, 4,247,429;6.1×10^{-26}, 4,247,730;2×10^{-12}, 4,247,574;2.5×10^{-09}, 4,247,729;1.3×10^{-07}, 4,248,003;1.5×10^{-07}), PZA(4,247,730;6.3×10^{-08}, 4,247,431;2×10^{-05}, 4,247,429;2.9×10^{-04}, 4,248,003;4.6×10^{-04}), RMP(4,247,429;4.1×10^{-29}, 4,247,431;4.8×10^{-24}, 4,247,730;3.1×10^{-13}, 4,248,003;3.5×10^{-07}, 4,247,574;4.7×10^{-07}, 4,247,729;2.5×10^{-06}, 4,247,469;4.7×10^{-04}), STRx10P(4,247,431;2.2×10^{-14}, 4,247,429;2.9×10^{-13}, 4,247,729;1.4×10^{-05}, 4,248,003;2.6×10^{-05}, 4,247,730;5.5×10^{-05}, 4,247,574;6.9×10^{-05}), XDR(4,247,429;4.4×10^{-13}, 4,247,431;8.9×10^{-10}, 4,247,730;2.6×10^{-08})	41
katG	INH	2	INH(2,155,168;2.7×10^{-07}), MDR(2,155,168;4.5×10^{-08}), RMP(2,155,168;5.7×10^{-04}), STM(2,155,168;8.3×10^{-04}), XDR(2,155,168;4.1×10^{-09})	5
katG	INH	4	EMB(2,155,168;1.5×10^{-07}), INH(2,155,168;2×10^{-63}, 2,155,167;8.5×10^{-05}), MDR(2,155,168;3×10^{-58}, 2,155,167;2.8×10^{-04}), PZA(2,155,168;1.5×10^{-09}), RMP(2,155,168;2.9×10^{-29}), STRx10P(2,155,168;2.8×10^{-11}), XDR(2,155,168;1.8×10^{-14})	9
katG	INH	2 + 4	EMB(2,155,168;4.8×10^{-11}), INH(2,155,168;7.1×10^{-72}, 2,155,167;1.1×10^{-04}), MDR(2,155,168;3.3×10^{-68}, 2,155,167;2.6×10^{-04}), PZA(2,155,168;1.7×10^{-11}), RMP(2,155,168;2.5×10^{-36}), STRx10P(2,155,168;3.9×10^{-18}), XDR(2,155,168;3.5×10^{-28})	9
rpsL	STM	2	INH(781,687;5.9×10^{-05}), MDR(781,687;5.3×10^{-05}), RMP(781,687;4.8×10^{-04}), STM(781,687;4.1×10^{-08})	4
rpsL	STM	4	AG(781,687;3.8×10^{-4}), INH(781,687;4.3×10^{-15}), MDR(781,687;3.9×10^{-12}), PZA(781,687;6.1×10^{-06}), RMP(781,687;8.3×10^{-10}), STM(781,687;9.6×10^{-14}, 781,822;2.3×10^{-04})	6
rpsL	STM	2 + 4	AG(781,687;3.8×10^{-5}), EMB(781,687;3.5×10^{-05}), FQ(781,687;8.3×10^{-05}), INH(781,687;2.3×10^{-26}, 781,822;6.4×10^{-05}), MDR(781,687;2.3×10^{-25}, 781,822;4.1×10^{-06}), PZA(781,687;1.5×10^{-08}), RMP(781,687;4.8×10^{-22}, 781,822;8.6×10^{-06}), STM(781,687;3.4×10^{-30}, 781,822;2.6×10^{-07}), XDR(781,687;4.3×10^{-09})	13

Table 1 Significant associations between genomic variants and drug resistance phenotypes identified by PhyC *(Continued)*

Locus	Known Phenotype Association [13]	Lineage	Observed Phenotype Association (position; *p*-value)	Total
Rv1482c-fabG1	INH, ETH	2	INH(1,673,425;9×10^{-06}), MDR(1,673,425;5.7×10^{-05})	2
Rv1482c-fabG1	INH, ETH	4	INH(1,673,425;2.2×10^{-20}), MDR(1,673,425;2×10^{-07}), XDR(1,673,425;3.9×10^{-05})	3
Rv1482c-fabG1	INH, ETH	2 + 4	EMB(1,673,432;5.4×10^{-05}), ETH(1,673,425;7.6×10^{-04}), FQ(1,673,432;9×10^{-04}), INH(1,673,425;6.4×10^{-27}, 1,673,432;8.3×10^{-07}), MDR(1,673,425;4.4×10^{-14}, 1,673,432;6.4×10^{-05}), RMP(1,673,432;7.9×10^{-06}, 1,673,425;3×10^{-05}), STEP(1,673,432;8.5×10^{-05}, 1,673,425;3.6×10^{-04}), XDR(1,673,425;8.7×10^{-06}, 1,673,432;1.2×10^{-05})	14
gyrA	FQ	2	EMB(7582;1.3×10^{-05}), ETH(7582;9.5×10^{-04}), FQ(7582;2.3×10^{-09}, 7570;7.6 × 10–06, 7581;4.1×10^{-04}), INH(7582;9.9×10^{-06}), MDR(7582;7.9×10^{-05}), OFL(7582;1.4×10^{-06}, 7570;8.5×10^{-04}, 7581;8.5×10^{-04}), RMP(7582;2.6×10^{-06}, 7570;1.2×10^{-04}, 7581;9.3×10^{-04}), STM(7582;6.5×10^{-04}), XDRvMDR(7570;9.6×10^{-04}), XDR(7570;5.7×10^{-07}, 7582;7.5×10^{-07}, 7581;7.5×10^{-04})	21
gyrA	FQ	4	EMB(7570;1.2×10^{-05}), FQ(7570;1.9×10^{-08}, 7582;3×10^{-06}, 7581;2.1×10^{-05}), INH(7570;3.2×10^{-10}, 7581;9.2×10^{-05}, 7582;1.1×10^{-04}), KAN(7570;1.5×10^{-04}), MDR(7570;1×10^{-08}, 7582;4.2×10^{-06}, 7581;5×10^{-05}), OFL(7570;2.5×10^{-04}, 7582;5.6×10^{-04}), PZA(7570;1×10^{-05}, 7581;1.3×10^{-04}), RMP(7570;3.4×10^{-11}, 7582;5×10^{-08}, 7581;5.1×10^{-06}), XDR(7570;3.3×10^{-10}, 7582;2.7×10^{-05}, 7572;3.6×10^{-04})	24
gyrA	FQ	2 + 4	AMK(7570;6.5×10^{-04}), CAP(7570;9.9×10^{-04}), EMB(7582;4.2×10^{-13}, 7570;1.1×10^{-08}, 7572;3.7×10^{-04}, 7581;4.8×10^{-04}), ETH(7582;1.3×10^{-04}), FQ(7582;3.6×10^{-18}, 7570;4.1×10^{-14}, 7581;2.4×10^{-09}, 7572;6.1×10^{-04}), INH(7582;4.5×10^{-15}, 7570;4.2×10^{-14}, 7581;1.2×10^{-08}, 7572;8.3×10^{-06}), KAN(7570;5.7×10^{-06}, 7572;6.7×10^{-05}), MDR(7582;2.2×10^{-15}, 7570;1.8×10^{-11}, 7581;6.4×10^{-07}), OFL(7582;1.3 × 10–10, 7570;2 × 10–07, 7581;5.8×10^{-06}), PZA(7581;1.2×10^{-07}, 7570;2.4×10^{-07}, 7572;9.7×10^{-05}, 7582;4.6×10^{-04}), RMP(7582;2.3×10^{-20}, 7570;2.5×10^{-17}, 7581;1.8 × 10–10, 7572;1×10^{-06}), STM(7582;1.9×10^{-10}, 7570;1.5×10^{-06}, 7581;1.8×10^{-04}), XDRvMDR(7570;4.2×10^{-05}, 7582;5.7×10^{-04}), XDR(7570;3.4×10^{-19}, 7582;3×10^{-16}, 7572;9.7×10^{-08}, 7581;2.9×10^{-07})	44
rrs	STM, AG	2	AMK(1,473,246;1.8×10^{-04}), CAP(1,473,246;5×10^{-08}), INH(1,473,246;5×10^{-06}), KAN(1,473,246;1.3×10^{-11}), RMP(1,473,246;8.9×10^{-06}), STM(1,473,246;4.1×10^{-04}), XDRvMDR(1,473,246;3.6×10^{-05}), XDR(1,473,246;7.4×10^{-11})	8
rrs	STM, AG	4	AG(1,473,246;2.6×10^{-7}), AMK(1,473,246;4.6×10^{-06}), CAP(1,473,246;2×10^{-06}), CIP(1,473,246;9.4×10^{-04}), EMB(1,473,246;7.1×10^{-07}), FQ(1,473,246;2.5×10^{-04}), INH(1,473,246;3.2×10^{-10}), KAN(1,473,246;3.3×10^{-10}), MDR(1,473,246;4.6×10^{-06}), PZA(1,473,246;9.4×10^{-10}), RMP(1,473,246;1.9×10^{-16}), STM(1,473,246;3.1×10^{-05}, 1,472,359;2.3×10^{-04}), XDRvMDR(1,473,246;1.9×10^{-05}), XDR(1,473,246;1.6×10^{-13})	15
rrs	STM, AG	2 + 4	AG(1,473,246;7.5×10^{-5}), AMK(1,473,246;3.9×10^{-11}), CAP(1,473,246;7.2×10^{-14}), CIP(1,473,246;6.5×10^{-04}), EMB(1,473,246;2.5×10^{-11}), FQ(1,473,246;3.5×10^{-07}), INH(1,473,246;3.6×10^{-20}, 1,472,359;1.2×10^{-05}), KAN(1,473,246;7.9×10^{-22}), MDR(1,473,246;1.8×10^{-11}, 1,472,359;4.4×10^{-04}), PZA(1,473,246;2.6×10^{-10}), RMP(1,473,246;7.3×10^{-26}), STM(1,473,246;1.3×10^{-11}, 1,472,359;1.5×10^{-08}), XDRvMDR(1,473,246;2.1×10^{-09}), XDR(1,473,246;7.9×10^{-29}, 1,472,359;1.5×10^{-04})	18
thyX-hsdS.1	PAS	2	XDR(3,067,961;7.5×10^{-04})	1
thyX-hsdS.1	PAS	4	INH(3,067,961;4.9×10^{-04}), STM(3,067,961;3.2×10^{-04})	2
thyX-hsdS.1	PAS	2 + 4	EMB(3,067,961;1×10^{-05}), INH(3,067,961;1.4×10^{-07}), MDR(3,067,961;6.4×10^{-05}), RMP(3,067,961;9.4×10^{-05}), STM(3,067,961;2.3×10^{-07}), XDR(3,067,961;1.2×10^{-05})	6
rpoC	RMP	4	EMB(764,817;2×10^{-04}), MDR(764,817;6×10^{-07}, 764,840;2.8×10^{-04}), PZA(764,817;8.3×10^{-08}), RMP(764,817;3.6×10^{-08}, 764,840;4.9×10^{-04}, 767,123;4.9×10^{-04}), STM(764,817;3.4×10^{-04})	8
rpoC	RMP	2 + 4	EMB(764,817;1.9×10^{-04}), INH(764,817;1.1×10^{-05}, 764,840;1.1×10^{-04}), MDR(764,817;4.9×10^{-10}, 764,840;9.7×10^{-06}), PZA(764,817;2×10^{-07}), RMP(764,817;1.4×10^{-09}, 764,840;2.2×10^{-05}, 764,363;4.7×10^{-04}, 767,123;4.7×10^{-04}), STM(764,817;3.8×10^{-06})	11
embC-embA	EMB	2	EMB(4,243,217;1.7×10^{-04})	1
embC-embA	EMB	4	EMB(4,243,221;3.7×10^{-04}, 4,243,190;3.8 × 10–04), INH(4,243,221;8.5×10^{-05}), MDR(4,243,217;4.6×10^{-06}, 4,243,221;4.6×10^{-06}, 4,243,190;3.6×10^{-05}), RMP(4,243,221;1.1×10^{-05}, 4,243,190;4.9×10^{-04})	8
embC-embA	EMB	2 + 4	EMB(4,243,217;1.4×10^{-07}, 4,243,190;3×10^{-07}, 4,243,221;1.3×10^{-06}), INH(4,243,217;4×10^{-08}, 4,243,221;2.3×10^{-06}, 4,243,190;3×10^{-05}), MDR(4,243,217;1.2×10^{-09}, 4,243,221;6.9×10^{-08}, 4,243,190;1.9×10^{-06}), RMP(4,243,221;2.2×10^{-07}, 4,243,217;6.1×10^{-07}, 4,243,190;4.8×10^{-06}), STM(4,243,217;8.6×10^{-04})	13
hadA	Novel	4	INH(732,110;4×10^{-04}), MDR(732,110;2.8×10^{-04})	2

Table 1 Significant associations between genomic variants and drug resistance phenotypes identified by PhyC *(Continued)*

Locus	Known Phenotype Association [13]	Lineage	Observed Phenotype Association (position; *p*-value)	Total
hadA	Novel	2 + 4	INH(732,110;1.1×10^{-04}), MDR(732,110;2.6×10^{-04}), STM(732,110;4×10^{-04})	3
pncA	PZA	4	EMB(2,288,868;3.8×10^{-04}), MDR(2,288,764;2.8×10^{-04}), RMP(2,288,764;4.9×10^{-04})	3
pncA	PZA	2 + 4	EMB(2,288,820;1.9×10^{-04}, 2,289,103;1.9×10^{-04}), MDR(2,289,207;2.6×10^{-04}), PZA(2,289,207;9.7×10^{-05}), RMP(2,288,778;4.7×10^{-04}, 2,288,820;4.7×10^{-04})	6
pncA-Rv2044c	PZA	4	RMP(2,289,252;4.9×10^{-04}), XDR(2,289,252;3.6×10^{-04})	2
pncA-Rv2044c	PZA	2 + 4	INH(2,289,252;1.1×10^{-04}), MDR(2,289,252;5×10^{-05}), PZA(2,289,252;2×10^{-07}), RMP(2,289,252;4.8×10^{-06}), XDR(2,289,252;6.2×10^{-05})	5
Rv3115-moeB2	Novel	2 + 4	MDR(3,482,717;6.7×10^{-04}), STM(3,482,717;6.7×10^{-04})	2
eis-Rv2417c	AG	2 + 4	EMB(2,715,342;1.6×10^{-05}), FQ(2,715,342;1.7×10^{-04}), INH(2,715,342;1.1×10^{-04}), KAN(2,715,342;5.4×10^{-04}), RMP(2,715,342;2.2×10^{-05}), STM(2,715,342;1.4×10^{-05})	7
folC	PAS	4	EMB(2,747,471;3.8×10^{-04})	1
folC	PAS	2 + 4	EMB(2,747,471;3.7×10^{-04}), INH(2,747,471;1.1×10^{-04}), STM(2,747,471;4×10^{-04})	3
whiB6-Rv3863	Putative STM or ETH	2 + 4	EMB(4,338,594;9×10^{-04})	1
fabG1	INH [53]	4	INH(1,674,048;6.3×10^{-06})	1
fabG1	INH [53]	2 + 4	INH(1,674,048;5.5×10^{-06})	1
oxyR'-ahpC	INH	2 + 4	INH(2,726,141;6.8×10^{-04}), PZA(2,726,141;9×10^{-04})	2
gyrB	FQ	4	RMP(6620;4.9×10^{-04})	1
gyrB	FQ	2 + 4	RMP(6620;4.7×10^{-04})	1

(*p-values* < 1E-3) Drug resistance phenotype abbreviations are as given in methods. 'Total' refers to the total number of significantly associated variants for the locus and lineage in question. AMK = Amikacin-resistance, AG = Aminoglycoside-resistance, CAP = Capreomycin-resistance, CIP = Ciprofloxacin-resistance, EMB = Ethambutol-resistance, ETH = Ethionamide-resistance, FQ = Fluoroquinolone-resistance, INH = Isoniazid-resistance, KAN = Kanamycin-resistance, MDR = Multidrug-resistant, OFL = Ofloxacin-resistance, PAN = pan-susceptible (no known drug-resistance), PZA = Pyrazinamide-resistance, RMP = Rifampicin-resistance, STM = Streptomycin-resistance, XDR = Extensively drug-resistant

Discussion

Our results highlight that lineage specific analyses are able to provide new insights into genetic associations with drug resistance phenotypes, despite a smaller sample size than a pan-lineage approach. Lineage specific associations were found within lineage 2, such as the novel association between *Rv3128c-Rv3129* and MDR. We also identified lineage-specific novel associations within lineage 4, such as the association between *fadB4-Rv3142c* and XDR. This indicates biological differences between these lineages with respect to drug resistance and perhaps in evolutionary trajectory. Novel associations specific to combined analyses indicate convergent evolution between lineages 2 and 4 at the same loci, with variant frequency too low for lineage-specific analyses to detect, that would most likely be detected in larger scale combined analyses (as previously described[13]). Lineage-specific GWAS is complementary to lineage-combined approaches, with their application in tandem potentially improving the power to detect *Mtb* genomic variants evolving under differing evolutionary dynamics.

Overall, despite conservative significance thresholds based on permutation, 17 potential novel associations were identified between antimicrobial resistance and *Mtb* loci and thus warrant experimental validation. For GWAS, 15 novel associations were identified, one in relation to the MDR phenotype and 14 in relation to the XDR phenotype; 7 were lineage specific. This might suggest an evolutionary shift amongst XDR strains. It may be feasible to consider XDR as a highly complex phenotype encompassing transmissibility [32]; unless evolution of XDR from pan-susceptible strains frequently happens within one patient, it is likely that XDR strains have gone through numerous cycles of active disease, transmission and treatment within recent history. The fact that many of these associations are lineage specific lends weight to such a hypothesis, suggesting differing evolutionary trajectories between lineages 2 and 4. Genetic drift might contribute to such divergence; there are numerous bottlenecks during the natural infectious cycle for *Mtb*, driven by host immune system, anti-TB drug therapy and transmission [33].

Some of the novel associated variants may be involved directly in drug resistance such as *hadA*, whose gene product, similar to InhA, is involved in fatty acid synthesis type II (FAS-II)) and thus may be involved in isoniazid resistance [34, 35]. One of the novel associated loci, *Rv0197,* identified here by variant-based GWAS in association with XDR, was previously identified through PhyC in association with a transmissibility phenotype [36]. *EspE* was identified by this previous analysis also [36], and it remains possible that the *espE-espF* intergenic region, identified here by locus-based GWAS in

Table 2 Significant associations between loci and drug resistance phenotypes identified by locus-based GWAS

Locus	Known Phenotype Association [13]	Lineage	Observed Phenotype Association	P-value	PhyC
rpoB	RMP	2	XDR, RMP, MDR	1.9×10^{-72}, 5.2×10^{-58}, 1.4×10^{-44}	11
rpoB	RMP	4	RMP, MDR, PZA, XDR, EMB	2×10^{-94}, 1.1×10^{-35}, 1.1×10^{-33}, 2.5×10^{-30}, 6.4×10^{-23}	23
rpoB	RMP	2 + 4	RMP, MDR, XDR, PZA	1.6×10^{-126}, 4.1×10^{-77}, 4.1×10^{-66}, 2.1×10^{-24}	40
gyrA	FQ	2	MOX, FQ, XDR, OFL	5.2×10^{-117}, 2.5×10^{-45}, 4.3×10^{-23}, 5.1×10^{-22}	21
gyrA	FQ	4	FQ, CIP, XDR	4.1×10^{-38}, 5.4×10^{-36}, 6.5×10^{-27}	24
gyrA	FQ	2 + 4	FQ, OFL, CIP	1.2×10^{-63}, 1.7×10^{-32}, 2.2×10^{-26}	44
pncA	PZA	2	XDR	1.50×10^{-25}	0
pncA	PZA	4	PZA	4.50×10^{-103}	3
pncA	PZA	2 + 4	PZA, EMB, XDR	5.3×10^{-101}, 1.2×10^{-29}, 6.7×10^{-22}	6
embC-embA	EMB	4	XDR, EMB	6.1×10^{-69}, 2.8×10^{-26}	8
embC-embA	EMB	2 + 4	XDR, EMB	3.3×10^{-37}, 1.4×10^{-23}	13
katG	INH	2	INH, MDR	3.6×10^{-34}, 7.3×10^{-24}	5
katG	INH	4	INH	1.20×10^{-44}	9
katG	INH	2 + 4	INH, MDR	1.5×10^{-61}, 1.5×10^{-36}	9
embB	EMB	2	EMB	7.20×10^{-27}	14
embB	EMB	4	EMB	1.80×10^{-56}	31
embB	EMB	2 + 4	EMB	3.30×10^{-55}	41
gid	STM	4	STM	7.40×10^{-55}	0
gid	STM	2 + 4	STM	1.30×10^{-53}	0
Rv1313c-Rv1314c		4	XDR	1.40×10^{-54}	0
Rv1313c-Rv1314c		2 + 4	XDR	3.30×10^{-32}	0
rpsL	STM	2	STM	1.90×10^{-38}	4
rpsL	STM	4	STM	5.60×10^{-26}	6
rpsL	STM	2 + 4	STM	6.00×10^{-41}	13
fadB4-Rv3142c		4	XDR	4.60×10^{-38}	0
Rv0526		2 + 4	XDR	8.70×10^{-37}	0
Rv1482c-fabG1	INH, ETH	4	INH	1.70×10^{-34}	3
Rv1482c-fabG1	INH, ETH	2 + 4	INH	3.30×10^{-30}	14
espE-espF		2 + 4	XDR	5.70×10^{-31}	0
tuf		4	XDR	1.50×10^{-29}	0
inhA	INH, ETH	4	XDR	2.40×10^{-28}	0
cut5b-Rv3725		4	XDR	5.10×10^{-27}	0
Rv3007c		4	XDR	7.80×10^{-24}	0
Rv2668		4	XDR	1.30×10^{-23}	0
pip-Rv0841		2 + 4	XDR	8.60×10^{-23}	0
rrs	STM, AG	2	KAN	1.40×10^{-22}	8
moeX		4	XDR	5.50×10^{-22}	0
lipJ-cinA		2 + 4	XDR	6.20×10^{-22}	0
Rv3128c-Rv3129		2	MDR	7.40×10^{-22}	0

(P-values <1E-21) Novel associations are given in bold. 'PhyC' column refers to the number of associations identified through PhyC analysis for the locus and lineage in question. AMK = Amikacin-resistance, AG = Aminoglycoside-resistance, CAP = Capreomycin-resistance, CIP = Ciprofloxacin-resistance, EMB = Ethambutol-resistance, ETH = Ethionamide-resistance, FQ = Fluoroquinolone-resistance, INH = Isoniazid-resistance, KAN = Kanamycin-resistance, MDR = Multidrug-resistant, OFL = Ofloxacin-resistance, PAN = pan-susceptible (no known drug-resistance), PZA = Pyrazinamide-resistance, RMP = Rifampicin-resistance, STM = Streptomycin-resistance, XDR = Extensively drug-resistant

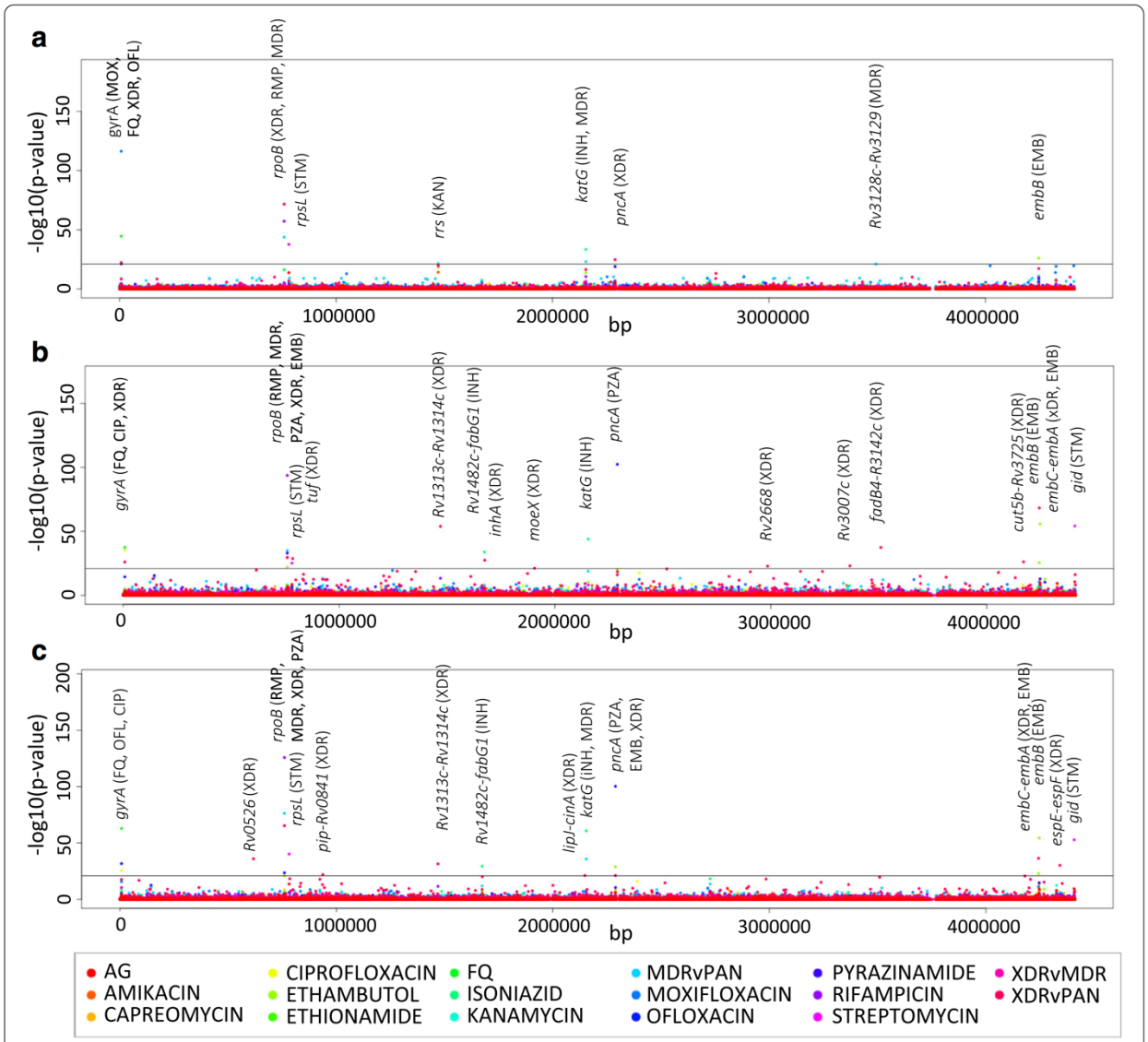


Fig. 2 Locus-based GWAS results **a** Manhattan plot for locus-based GWAS for lineage 2. **b** Manhattan plot for locus-based GWAS for lineage 4 **c** Manhattan plot for locus-based GWAS for lineages 2 and 4 combined. *P-value* threshold <1E-21. Phenotypes with which loci were found to be significantly associated are given in brackets next to locus name. AMK = Amikacin-resistance, AG = Aminoglycoside-resistance, CAP = Capreomycin-resistance, CIP = Ciprofloxacin-resistance, EMB = Ethambutol-resistance, ETH = Ethionamide-resistance, FQ = Fluoroquinolone-resistance, INH = Isoniazid-resistance, KAN = Kanamycin-resistance, MDR = Multidrug-resistant, OFL = Ofloxacin-resistance, PAN = pan-susceptible (no known drug-resistance), PZA = Pyrazinamide-resistance, RMP = Rifampicin-resistance, STM = Streptomycin-resistance, XDR = Extensively drug-resistant

association with XDR, may be related by regulation to *espE*. Additionally, both *espE-espF* and *whiB6-Rv3863* have been linked to Esx-1 which has been implicated in virulence regulation. The *WhiB6-Rv3863* intergenic region, which was also identified through previous PhyC analyses including our dataset [13], may additionally be linked to the DosR regulon. This regulon is composed of 48 co-regulated genes and is considered essential for persistence of latent *Mtb* [37–40]. Interestingly, the *whiB6-Rv3863* variant identified shows a markedly different distribution between lineages 2 and 4, showing greater frequency in lineage 2 (see Fig. 1).

Apart from *Rv0197*, a further two variant-based GWAS SNPs were identified (*recF* and *argJ*), however both are synonymous variants. These may be examples of background variants 'hitchhiking' alongside causal variants, or may play a biological role. Notably, a number of identified loci are potentially involved in molybdenum cofactor biosynthesis; *Rv3115-moeB2, moeX* [41], and *Rv0197 (mycobrowser: Gene Ontology: molybdenum ion*

Table 3 Significant associations between genomic variants and drug resistance phenotypes identified by variant-based GWAS

Variant Locus	Variant Position	Type	Known Phenotype Association [13]	Lineage	Observed Phenotype Association (*p-value*)	PhyC
rrs	1,473,246	inter	STM, AG	2	CAP(2×10^{-31}), KAN(1.1×10^{-37})	8
rrs	1,473,246	inter	STM, AG	4	KAN(6.7×10^{-31})	15
rrs	1,473,246	inter	STM, AG	2 + 4	AMK(2.4×10^{-39}), CAP(3.9×10^{-48}), KAN(6.5×10^{-69}), XDRvMDR(5.3×10^{-27})	18
katG	2,155,168	NS	INH	2	XDR(2.1×10^{-42})	5
katG	2,155,168	NS	INH	4	INH(6.1×10^{-65}), MDR(6×10^{-45}), XDR(1.5×10^{-29})	9
katG	2,155,168	NS	INH	2 + 4	INH($4.4 \times 10–^{56}$), MDR(7.4×10^{-25})	9
Rv0197	232,574	NS	Novel	4	XDR(9.5×10^{-62})	0
Rv0197	232,574	NS	Novel	2 + 4	XDR(232,574;3.8×10^{-51})	0
rpoB	761,155	NS	RMP	2	XDR(3.5×10^{-25})	4
rpoB	761,155	NS	RMP	4	MDR(1.2×10^{-27}), PZA(1.9×10^{-28}), RMP(2.6×10^{-42}, 7.1×10^{-31}, 761,139;3.4×10^{-23}), XDR(3.8×10^{-57})	7
rpoB	761,139	NS	RMP	4	RMP(3.4×10^{-23})	3
rpoB	761,155	NS	RMP	2 + 4	MDR(5×10^{-23}), PZA(6×10^{-26}), RMP(2×10^{-38}), XDR(1.3×10^{-27})	7
rpoB	761,139	NS	RMP	2 + 4	PZA(4×10^{-23}), RMP(2.2×10^{-29}),	7
recF	4047	S	Novel	4	XDR(1.2×10^{-52})	0
recF	4047	S	Novel	2 + 4	XDR(8.6×10^{-41})	0
Rv1482c-fabG1	1,673,425	inter	INH, ETH	4	INH(1.1×10^{-36})	3
Rv1482c-fabG1	1,673,425	inter	INH, ETH	2 + 4	INH(1.1×10^{-35})	14
rpsL	781,687	NS	STM	2	STM(3×10^{-27})	4
rpsL	781,687	NS	STM	2 + 4	STM(6.3×10^{-28})	6
argJ	1,867,614	S	Novel	2 + 4	XDR(6.9×10^{-26})	0
gyrA	7570	NS	FQ	4	XDR(8.6×10^{-23})	24
gyrA	7582	NS	FQ	2 + 4	CIP(1.3×10^{-24}), FQ(4.6×10^{-22})	44

(*p-values* < 1E-22) NS = non-synonymous, S = synonymous, inter = intergenic region. Novel associations are given in bold. 'PhyC' column refers to the number of associations identified through PhyC analysis for the locus and lineage in question; AMK = Amikacin-resistance, AG = Aminoglycoside-resistance, CAP = Capreomycin-resistance, CIP = Ciprofloxacin-resistance, EMB = Ethambutol-resistance, ETH = Ethionamide-resistance, FQ = Fluoroquinolone-resistance, INH = Isoniazid-resistance, KAN = Kanamycin-resistance, MDR = Multidrug-resistant, OFL = Ofloxacin-resistance, PAN = pan-susceptible (no known drug-resistance), PZA = Pyrazinamide-resistance, RMP = Rifampicin-resistance, STM = Streptomycin-resistance, XDR = Extensively drug-resistant

binding) (Mycobrowser). Molybdenum cofactor is found in molybdenum enzymes, which are responsible for a number of functions from dormancy regulation to energy source metabolism [41, 42]. Interestingly, these three loci were each identified by a different analyses type; variant-based GWAS, locus-based GWAS and PhyC, respectively. Functional studies may be useful in providing further insight into the role of variants identified here.

Recognizing that drug resistance phenotypes may be subtly different, depending on the genetic background of the strain, could be important and might relate directly to drug resistance, or to fitness more broadly, such as through increased virulence and transmission. With the recognition of XDR transmission [36, 43], our study suggests that further critical information on lineage and transmission clustering (obtained from the genome sequence) would also be important to determine the full impact of specific mutations, that might lead to further phenotypic descriptions related to transmission, virulence and degree of drug resistance.

The results show the differing evolutionary insights offered by locus- and variant-based GWAS, and convergence-based methodologies. Both variant-based and locus-based GWAS led to unique loci being identified. The *rrs* locus was found in lineage 2 only locus-based GWAS analyses, but for both variant-based GWAS and PhyC analyses, *rrs* was identified in both lineage-specific and lineage-combined analyses. Neutral variation within the *rrs* gene may be diluting the signal from causal drug resistance variants in the lineage 4 locus-based GWAS analysis.

inhA was not identified by variant-based GWAS or PhyC, only lineage 4 specific locus-based GWAS. A sub-type of the Portuguese Lisboa (lineage 4) strain is known to have *inhA* markers involved in isoniazid resistance [44], and a different mechanism to other lineages. Whilst *inhA* was not identified by lineage-combined

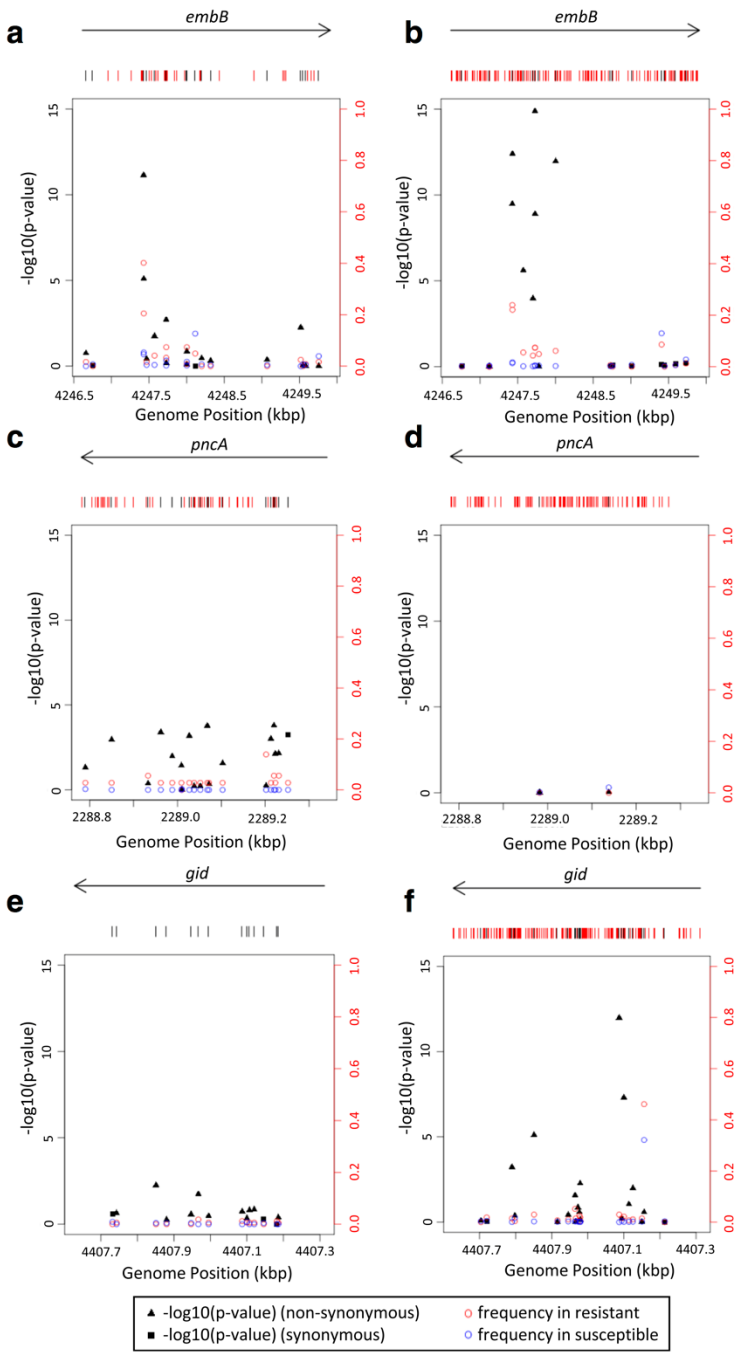


Fig. 3 Locus maps showing variant-based GWAS results (on the left y-axis) and variant frequency (on the right y-axis): **a** lineage 2 ethambutol analysis for *embB;* **b** lineage 4 ethambutol analysis for *embB;* **c** lineage2 XDR analysis for *pncA;* **d** lineage 4 XDR analysis for *pncA;* **e** lineage 2 streptomycin analysis for *gid;* **f** lineage 4 streptomycin analysis for *gid*

GWAS, it is notable that *Rv1482c-fabG1* and *katG* were; both these loci also play a role in isoniazid resistance, suggesting different mechanisms of resistance to these drugs between lineage 2 and lineage 4.

In cases where drug resistance is driven by rare variants and genetic heterogeneity exists within a single gene, such as in *pncA*, where multiple alleles can cause pyrazinamide resistance, locus-based analyses may be more powerful. Indeed, *pncA* was identified here by locus-based GWAS but not variant-based GWAS. Convergence-based PhyC analysis seems to have greater sensitivity in combined-lineage analyses. Unlike GWAS, the

success of PhyC in detecting antimicrobial resistance associated variants is determined by the magnitude of convergent evolution within the *Mtb* population under question [19]. Indeed, there were important differences between the GWAS and PhyC results outlined here. These differences might provide insight into the relative importance of within patient evolution of antimicrobial resistance versus transmission of antimicrobial resistant strains. In instances where a mutation is highly transmissible and consequently increases in frequency with only one or few mutation events, it might be expected that GWAS would be a more powerful analytical tool, due to the lack of convergent-evolution.

It is notable that lineage 2 had a smaller sample size than the lineage 4 dataset, this may contribute to the greater sensitivity in lineage 4 specific analyses. In order to assess the extent to which the lower significance levels in the lineage 2 GWAS were as a result of smaller sample size in comparison to lineage, it would be interesting to repeat the GWAS analyses with a larger and perhaps more geographically spread lineage 2 dataset. Additionally, statistical power is potentially limited in the current analyses by low resolution phenotypic data, with not all drugs tested on all samples, primarily due to second line drugs only being tested where there is multidrug resistance. For example, for lineage 2 there were only 8 resistant and 120 susceptible isolates for moxifloxacin. Despite this, the most significant gene-based GWAS result for lineage 2 was for *gyrA*, identified in relation to moxifloxacin resistance, showing the sensitivity of the method. Nevertheless, to identify variants with smaller effect sizes, increased phenotypic resolution may prove useful. Further work could explore the use of minimum inhibitory concentration values, where available, being incorporated into resistance phenotypes.

Conclusions

In summary, GWAS and PhyC are sensitive, robust and complementary methodologies in examining evolution of antimicrobial resistance in *Mtb*. Within GWAS analyses, locus-based and variant-based approaches are both useful and complementary, as are lineage-combined and lineage-specific analyses. These different methodological approaches can be used to detect different evolutionary dynamics and thus their similarities and differences are informative. Evidence presented here suggests the importance of lineage- specific paths of evolution towards drug resistance in *Mtb*. It will be interesting to see how methodologies outlined here might apply to other *Mtb* lineages and other pathogen species in an anti-microbial resistance context, or indeed in relation to other phenotypes of interest such as transmissibility.

Methods

Isolates, phenotypic methods, sequencing and variant calling

The raw sequence data used here ($n = 4408$) form part of a subset of a larger dataset ($n = 6465$), which represents multiple populations from different geographic areas (see Additional file 9), and is described elsewhere [13]. In particular, only lineages 2 ($n = 702$) and 4 ($n = 3706$) from the larger dataset are used, with additional phenotypic data for the samples collected in Portugal. Drug resistance phenotypes were available for amikacin, capreomycin, ciprofloxacin, ethambutol, ethionamide, isoniazid, kanamycin, moxifloxacin, ofloxacin, pyrazinamide, rifampicin, streptomycin, resistance to any fluoroquinolone; levofloxacin, moxifloxacin, ciprofloxacin or ofloxacin (FQ), resistance to any of the aminoglycosides; kanamycin, amikacin, or streptomycin (AG), combined isoniazid and rifampicin resistance, but not XDR (MDR), MDR plus resistance to a fluoroquinolone (ciprofloxacin, levofloxacin, moxifloxacin) and to a second line injectable (amikacin, kanamycin, capreomycin) (XDR), and pan-susceptible, susceptibility to rifampicin and isoniazid plus no other known resistance (PAN). Isoniazid, rifampicin, ethambutol, streptomycin and pyrazinamide are first-line drugs. Amikacin, capreomycin, ofloxacin, para-aminosalicylic acid, moxifloxacin and cycloserine are second-line drugs. Samples found to be MDR, underwent testing for second-line drugs. Para-aminosalycylic acid, levofloxacin, rifabutin and cycloserine resistance phenotypes were excluded from analyses due to lack of data. Where present, levofloxacin data was used in defining the aggregate phenotypes of FQ; however, there was not enough levofloxacin phenotypic data to use in individual drug-resistance analyses.

All samples underwent Illumina sequencing generating paired-end reads of at least 50 bp with at least 50-fold average genome coverage. The raw sequence data were aligned to the H37Rv reference genome (Genbank accession number: NC_000962.3) using the BWA mem algorithm [45]. The SAMtools/BCFtools [46] and GATK [47] software was used to call SNPs and small insertions or deletions (indels) using default options. The overlapping set of variants from the two algorithms was retained for further analysis. Alleles were additionally called across the whole genome (including SNP sites) using a coverage-based approach [16, 28]. A missing call was assigned if the total depth of coverage at a site did not reach a minimum of 20 reads or none of the four nucleotides accounted for at least 75% of the total coverage. The final dataset consisted of 157,726 SNPs, 2926 insertions and 5998 deletions across the 4408 isolates. Monomorphic variants within each of the three datasets ('lineage 4-specific', 'lineage 2-specific' and 'lineages 2 and 4 combined') were removed.

Phylogenetic tree and PhyC

Sublineage was assigned based on SNPs (see Additional file 10). PCA was conducted on the pairwise variant distance matrix for lineages separately and combined. A maximum likelihood phylogenetic tree was constructed for the 157,726 SNP sites present in lineages 2 and 4 isolates using ExaML [48] using the standard model and rooted with *M. canettii* as the outgroup. The ITOL v3 tool was used for visualisation [49]. PhyC [19] analysis was performed using an in-house pipeline as described by Phelan et al. (2016) [16] . A significance cut-off of $< 10^{-3}$ was applied, and this threshold was established based on permutation analysis.

Association analyses

Genome wide association study (GWAS) analyses were performed using GEMMA software [50]. The general parameters were; default missingness (< 0.05) and a minor allele frequency cut-off of 0.001. Kinship matrices were used to account for relatedness. Analyses were performed based on SNPs and short indels (range: 1 to 70 bp) ("variant-based"); and mutations aggregated over coding and intergenic loci ("locus-based"). For coding loci, only non-synonymous variants were aggregated. A linear mixed model was used for both types of analysis, and a likelihood ratio test was used to assess statistical significant of the variants and loci. Each analysis considered a different drug susceptibility phenotype, namely: amikacin resistant (AMK) vs. non-amikacin resistant, AG resistant vs. non-AG resistant, capreomycin resistant (CAP) vs. non-capreomycin resistant, ciprofloxacin resistant (CIP) vs. non-ciprofloxacin resistance, ethambutol resistant (EMB) vs. non-ethambutol resistant, ethionamide resistant (ETH) vs. non-ethionamide resistant, isoniazid resistant (INH) vs. non-isoniazid resistant, kanamycin resistant (KAN) vs. non-kanamycin resistant, moxifloxacin resistant (MOX) vs. non-moxifloxacin resistant, ofloxacin resistant vs. non-ofloxacin resistant (OFL), pyrazinamide resistant (PZA) vs. non-pyrazinamide resistant, rifampicin resistant (RMP) vs. non-rifampicin resistant, streptomycin (STM) vs. non-streptomycin resistant, FQ vs. non-FQ, MDR vs. PAN ("MDR"), XDR vs. PAN ("XDR") and XDR vs. MDR ("XDRvMDR"). Analyses were performed with lineage 4 only ($n = 3706$), lineage 2 only ($n = 701$, after removing 1 outlier identified by PCA) and lineages 2 and 4 combined. Analyses were repeated accounting for different numbers of principal components, from 0 to 5, to assess the effects on significance. A significance threshold of $< 10^{-21}$ based on permutation.

All statistical analyses, including PCA, were performed in R software (r-project.org) and its qqman package [51] was used to construct Manhattan plots and quantile-quantile (qq)-plots. Pairwise variant distance between isolates was calculated in R [52], using absolute distance between isolates including all variants for lineage 2 and lineage 4.

Additional files

Additional file 1: Variant Summary Tables, Summary tables of variants called in comparison to the H37rv reference, with monomorphic variants removed for each dataset. **a** Total numbers of variants by lineage; **b** Number of variants per sample; **c** Non-reference variant frequency summary; variants called in comparison to the H37rv reference. (PPTX 39 kb)

Additional file 2: Non-reference variant frequency histogram, A histogram showing log10(frequency + 1) of non-reference alleles compared to the H37rv reference for **a** lineage 2 and **b** lineage4. (PPTX 69 kb)

Additional file 3: Population diversity within investigated strains, **a** Principal component 1 (PC1) by principal component 2 (PC2) for lineage 2, The first 10 principal components account for 71.9% of the variation in lineage 2; **b** Distance plot for lineage 2 showing pairwise number of variant differences between samples; **c** Principal component 1 (PC1) by principal component 2 (PC2) for lineage 4, the first 10 principal components account for 88.9% of the variation in lineage 4. **d** Distance plot for lineage 2 showing pairwise number of variant differences between samples. (PPTX 5650 kb)

Additional file 4: Scree plots for the principal component analyses, Scree plots showing the proportion of variation accounted for by the first ten principal components, calculated for the pairwise distances within **a** lineage 4 and **b** lineage 2. (PPTX 142 kb)

Additional file 5: Drug-resistance phenotype frequency table, Drug-resistance phenotype frequency table by lineage. 'Totals' shows the number and percentage of each lineage with a known drug-resistance phenotype. (PPTX 45 kb)

Additional file 6: Cross-resistance phenotype table, Cross-Resistance Table upper diagonal shows proportion of samples phenotyped for both vertical and horizontal phenotype, that test positive for vertical phenotype. Diagonal shows number of samples with each phenotype. Lower diagonal shows number of samples with phenotype for both horizontal and vertical phenotype. (PPTX 45 kb)

Additional file 7: Variant Position Table, Table detailing variants at all positions with at least one non-synonymous variant found to be significantly associated with a phenotype in any of the variant-based analyses. (PPTX 52 kb)

Additional file 8: Locus Comparison Table, Locus comparison table showing which analyses and in which lineage each loci was identified. An 'x' indicates a locus which was not identified by the method of analysis in question. Loci without a known association with the phenotype are highlighted in bold. (PPTX 44 kb)

Additional file 9: Study frequency table, Study frequency table, showing numbers and percentage of strains from each study by lineage. (PPTX 40 kb)

Additional file 10: Sublineage frequency table, Numbers and percentage by lineage assigned to each sublineage. (PPTX 36 kb)

Abbreviations

AG: Aminoglycoside-resistance; AMK: Amikacin-resistance; CAP: Capreomycin-resistance; CIP: Ciprofloxacin-resistance; EMB: Ethambutol-resistance; ETH: Ethionamide-resistance; FQ: Fluoroquinolone-resistance; GWAS: Genome-wide Association Study; INH: Isoniazid-resistance; KAN: Kanamycin-resistance; MDR: Multidrug-resistant; MOX: Moxifloxacin-resistance; OFL: Ofloxacin-resistance; PAN: Pan-susceptible; no known drug-resistance; PAS: Para-Aminosalicylic Acid-resistance; PCA: Principal Component Analysis; PZA: Pyrazinamide-resistance; RMP: Rifampicin-resistance; STM: Streptomycin-resistance; XDR: Extensively drug-resistant

Acknowledgments

The authors wish to thank the study participants.

Funding

YO and JPh are funded by a BBSRC PhD studentship (Grant no. BB/J014567/1). TC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). The MRC eMedLab computing resource was used for bioinformatics and statistical analysis. Further funds were received from Fundação para a Ciência

e a Tecnologia, Portugal, through the grants UID/Multi/04413/2013 (DM and MV). The funding bodies played no role in design of the study, collection, analysis and interpretation of data or in writing the manuscript.

Availability of data and materials
The analysis was performed on raw sequencing data available from the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena) under the following study accession numbers; PRJEB10385, ERP006619, ERP002611, ERP000192, SRP018402 and ERP008770, as utilized in Coll et al. 2017 [13].

Authors' contributions
TC and MH conceived and directed the project. JPh generated the sequencing dataset. YO performed bioinformatic and statistical analyses under the supervision of TC and MH, and wrote the first draft of the manuscript. JPe, DM, AM, IP and MV contributed protocols and data. All authors commented and edited on various versions of the draft manuscript. All authors compiled and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Publisher's Note

Author details
[1]Pathogen Molecular Biology Department, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]iMed.ULisboa – Research Institute for Medicines, Faculdade de Farmácia, Universidade de Lisboa, Lisbon, Portugal. [3]Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, UNL, Lisbon, Portugal. [4]National Mycobacterium Reference Laboratory, Porto, Portugal. [5]Faculty of Epidemiology and Population Health, LSHTM, London, UK.

Received: 11 April 2018 Accepted: 15 March 2019
Published online: 29 March 2019

References
1. World Health Organisation. Global Tuberculosis Report; 2015. p. 2015.
2. Mariam SH, Werngren J, Aronsson J, Hoffner S, Andersson DI. Dynamics of Antibiotic Resistant *Mycobacterium tuberculosis* during Long-Term Infection and Antibiotic Treatment. PLoS One. 2011;6:e21147. https://dx.plos.org/10.1371/journal.pone.0021147.
3. Fortune SM. The surprising diversity of mycobacterium tuberculosis: change you can believe in. J Infect Dis. 2012;206:1642–4.
4. Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, et al. Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients. J Infect Dis. 2012;206:1724–33 Available from: http://jid.oxfordjournals.org/lookup/doi/10.1093/infdis/jis601.
5. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One. 2013;8:1–12.
6. Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, et al. Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. Nat. Commun. 2015;6:7119 Available from: http://www.nature.com/ncomms/2015/150511/ncomms8119/full/ncomms8119.html.
7. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7:51 Available from: http://genomemedicine.com/content/7/1/51.
8. Müller B, Borrell S, Rose G, Gagneux S. The heterogeneous evolution of multidrug-resistant Mycobacterium tuberculosis. Trends Genet. 2013;29:160–9.
9. Telenti A, Imboden P, Marchesi F, Matter L, Schopfer K, Bodmer T, et al. Detection of rifampicin-resistance mutations in Mycobacterium tuberculosis. Lancet. 1993;341:647–51 Available from: http://linkinghub.elsevier.com/retrieve/pii/014067369390417F.
10. Moradigaravand D, Grandjean L, Martinez E, Li H, Zheng J, Coronel J, et al. *DfrA-thyA* double deletion in *para* -aminosalicylic acid resistant *Mycobacterium tuberculosis* Beijing strains. Antimicrob. Agents Chemother. 2016:AAC.00253, 16 Available from: http://aac.asm.org/lookup/doi/10.1128/AAC.00253-16.
11. Galagan JE. Genomic insights into tuberculosis. Nat. Rev. genet. 2014;15:307–20 Available from: http://www.ncbi.nlm.nih.gov/pubmed/24662221.
12. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat. Genet. 2014;46:279–86 Available from: https://doi.org/10.1038/ng.2878.
13. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018;50:307–16 Available from: http://www.nature.com/articles/s41588-017-0029-0.
14. Sveinbjornsson G, Gudbjartsson DF, Halldorsson B V, Kristinsson KG, Gottfredsson M, Barrett JC, et al. HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. Nat. Genet. 2016;48:318–322. Available from: http://www.nature.com/doifinder/10.1038/ng.3498.
15. Cain AK, Lees JA. Using genomics to combat infectious diseases on a global scale. Genome Biol; 2015;16:250. Available from: http://genomebiology.com/2015/16/1/250
16. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE V., Furnham N, et al. *Mycobacterium tuberculosis* whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med; 2016;14:31. Available from: http://www.biomedcentral.com/1741-7015/14/31
17. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178:1709–23 Available from: http://www.genetics.org/cgi/doi/10.1534/genetics.107.080101.
18. Earle SG, Wu C, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat. Microbiol. 2016;1:16041. Available from: https://doi.org/10.1038/nmicrobiol.2016.41.
19. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat. Genet. 2013;45:1183–9. Available from: http://www.nature.com/articles/ng.2747.
20. Desjardins CA, Cohen KA, Munsamy V, Abeel T, Maharaj K, Walker BJ, et al. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. Nat. Genet. 2016;48:1–9 Available from: http://www.nature.com/doifinder/10.1038/ng.3548.
21. Reed MB, Pichler VK, Mcintosh F, Mattia A, Fallow A, Masala S, et al. Major mycobacterium tuberculosis lineages associate with patient country of origin. J Clin Microbiol. 2009;47:1119–28.
22. Gagneux S. Host-pathogen coevolution in human tuberculosis. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 2012;367:850–9 Available from: http://rstb.royalsocietypublishing.org/content/367/1590/850.short.
23. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship between mycobacterium tuberculosis phylogenetic lineage and clinical site of tuberculosis. Clin Infect Dis. 2012;54:211–9.
24. Krishnan N, Malaga W, Constant P, Caws M, Thi Hoang Chau T, Salmons J, et al. *Mycobacterium tuberculosis* lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. PLoS One. 2011;6.
25. Portevin D, Gagneux S, Comas I, Young D. Human macrophage responses to clinical isolates from the Mycobacterium tuberculosis complex discriminate between ancient and modern lineages. PLoS Pathog. 2011;7.
26. Mathema B, Kurepina N, Yang G, Shashkina E, Manca C, Mehaffy C, et al. Epidemiologic consequences of microvariation in Mycobacterium tuberculosis. J Infect Dis. 2012;205:964–74.
27. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. Nat. Genet. 2016; Available from: http://www.nature.com/doifinder/10.1038/ng.3704.
28. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat. Commun. 2014;5:4812. Available from: http://www.nature.com/articles/ncomms5812.

29. Benavente ED, Coll F, Furnham N, McNerney R, Glynn JR, Campino S, et al. PhyTB: Phylogenetic tree visualisation and sample positioning for M. tuberculosis. BMC Bioinformatics. 2015;16:155. Available from: http://www.biomedcentral.com/1471-2105/16/155.
30. Perdigão J, Macedo R, Machado D, Silva C, Jordão L, Couto I, et al. GidB mutation as a phylogenetic marker for Q1 cluster Mycobacterium tuberculosis isolates and intermediate-level streptomycin resistance determinant in Lisbon, Portugal. Clin. Microbiol. Infect. 2014;20.
31. Machado D, Perdigão J, Ramos J, Couto I, Portugal I, Ritter C, et al. High-level resistance to isoniazid and ethionamide in multidrug-resistant Mycobacterium tuberculosis of the Lisboa family is associated with inhA double mutations. J Antimicrob Chemother. 2013;68:1728–32.
32. Guerra-Assunção J, Crampin A, Houben R, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife. 2015;4:1–17 Available from: http://elifesciences.org/lookup/doi/10.7554/eLife.05166.
33. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High Functional Diversity in *Mycobacterium tuberculosis* Driven by Genetic Drift and Human Demography. Blaser MJ, editor. PLoS Biol. 2008;6:e311. Available from: https://dx.plos.org/10.1371/journal.pbio.0060311.
34. Dong Y, Qiu X, Shaw N, Xu Y, Sun Y, Li X, et al. Molecular basis for the inhibition of β-hydroxyacyl-ACP dehydratase HadAB complex from mycobacterium tuberculosis by flavonoid inhibitors. Protein Cell Higher Education Press. 2015;6:504–17.
35. Gannoun-Zaki L, Alibaud L, Kremer L. Point mutations within the fatty acid synthase type II dehydratase components HadA or HadC contribute to isoxyl resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2013;57:629–32.
36. Nebenzahl-Guimaraes H, Van Laarhoven A, Farhat MR, Koeken VACM, Mandemakers JJ, Zomer A, et al. Transmissible mycobacterium tuberculosis strains share genetic markers and immune phenotypes. Am J Respir Crit Care Med. 2017;195:1519–27.
37. Chen Z, Hu Y, Cumming BM, Lu P, Feng L, Deng J, et al. Mycobacterial WhiB6 Differentially Regulates ESX-1 and the Dos Regulon to Modulate Granuloma Formation and Virulence in Zebrafish. Cell Rep. The Author(s); 2016;16:2512–2524. Available from: https://doi.org/10.1016/j.celrep.2016.07.080
38. Solans L, Aguiló N, Samper S, Pawlik A, Frigui W, Martín C, et al. A specific polymorphism in Mycobacterium tuberculosis H37Rv causes differential ESAT-6 expression and identifies whiB6 as a novel ESX-1 component. Infect Immun. 2014;82:3446–56.
39. Chen T, He L, Deng W, Xie J. The Mycobacterium DosR regulon structure and diversity revealed by comparative genomic analysis. J Cell Biochem. 2013;6:1–6.
40. Domenech P, Zou J, Averback A, Syed N, Curtis D, Donato S, et al. Unique regulation of the DosR regulon in the Beijing lineage of Mycobacterium tuberculosis. J Bacteriol. 2017;199:1–19.
41. Shi T, Xie J. Molybdenum enzymes and molybdenum cofactor in mycobacteria. J Cell Biochem. 2011;112:2721–8.
42. Levillain F, Poquet Y, Mallet L, Mazères S, Marceau M, Brosch R, et al. Horizontal acquisition of a hypoxiaresponsive molybdenum cofactor biosynthesis pathway contributed to *Mycobacterium tuberculosis* pathoadaptation. PLOS Pathog. 2017;13:e1006752. Available from: https://dx.plos.org/10.1371/journal.ppat.1006752.
43. Shah NS, Auld SC, Brust JCM, Mathema B, Ismail N, Moodley P, et al. Transmission of extensively drug-resistant tuberculosis in South Africa. N Engl J Med [Internet]. 2017;376:243–53 Available from: https://doi.org/10.1056/NEJMoa1604544.
44. Perdigão J, Silva H, Machado D, Macedo R, Maltez F, Silva C, et al. Unraveling *Mycobacterium tuberculosis* genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. BMC Genomics. 2014;15:991 Available from: http://www.biomedcentral.com/1471-2164/15/991.
45. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843–51.
46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009;25:2078–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19505943.
47. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. 2011;43:491–8.
48. Kozlov AM, Aberer AJ, Stamatakis A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics. 2015;31:2577–9.
49. Letunic I, Bork P. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23:127–8.
50. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. Nature Publishing Group; 2012;44:821–824. Available from: https://doi.org/10.1038/ng.2310.
51. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J. Open Source Softw. 2018;3:731. Available from: http://joss.theoj.org/papers/10.21105/joss.00731.
52. Core R. Team. R: a language and environment for statistical computing. [internet]. Vienna, Austria: R Foundation for Statistical. Computing. 2015; Available from: https://www.r-project.org/.
53. Torres JN, Paul LV, Rodwell TC, Victor TC, Amallraja AM, Elghraoui A, et al. Novel katG mutations causing isoniazid resistance in clinical M. tuberculosis isolates. Emerg. Microbes Infect. 2015;4:e42 Available from: http://www.nature.com/doifinder/10.1038/emi.2015.42.

Additional File 1

a Total numbers of variants by lineage; variants called in comparison to the H37Rv reference, with monomorphic variants removed for each dataset.

Lineage	SNP	Deletion	Insertion
2	27,254	1,334	635
4	132,074	4,880	2,403
2+4	157,726	5,998	2,926

b Number of variants per sample;; variants called in comparison to the

Lineage	Type	Mean	Median	Min.	Max.
2	indel	30.7	31	7	42
4	indel	39.2	40	2	61
2+4	indel	47.4	42	2	103
2	SNP	327.6	332	189	386
4	SNP	689.5	724	10	870
2+4	SNP	777.2	733	10	1300

c Non-reference variant frequency summary; variants called in comparison to the H37rv reference; variants called in comparison to the H37rv reference, with monomorphic variants removed for each dataset.

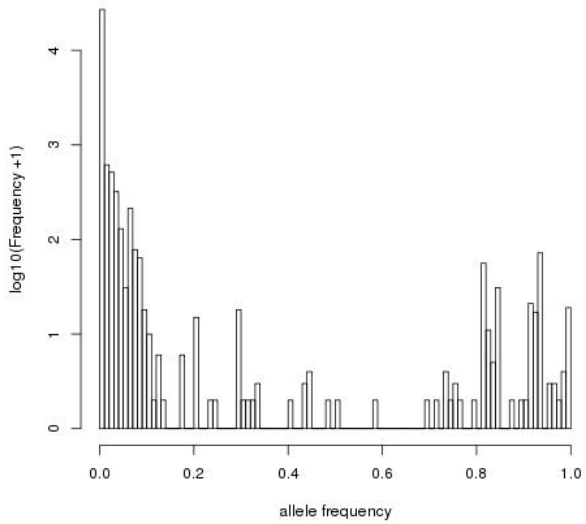
lineage	Mean	Min.	1st Quartile	Median	3rd Quartile	Max.
2	0.0122	0.0014	0.0014	0.0014	0.0028	0.9929
4	0.0052	0.0003	0.0003	0.0003	0.0005	0.9987
2+4	0.0049	0.0002	0.0002	0.0002	0.0005	0.9989

Additional File 2

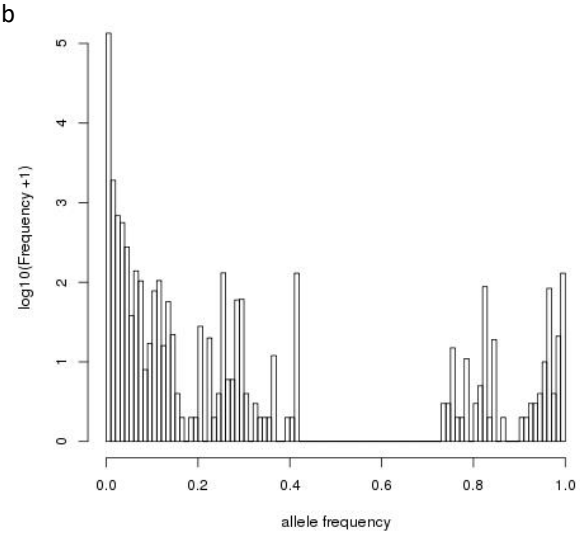
Non-reference variant frequency histogram

Histograms showing log10(frequency+1) of non-reference alleles compared to the H37rv reference for **a** lineage 2 and **b** lineage4**.**

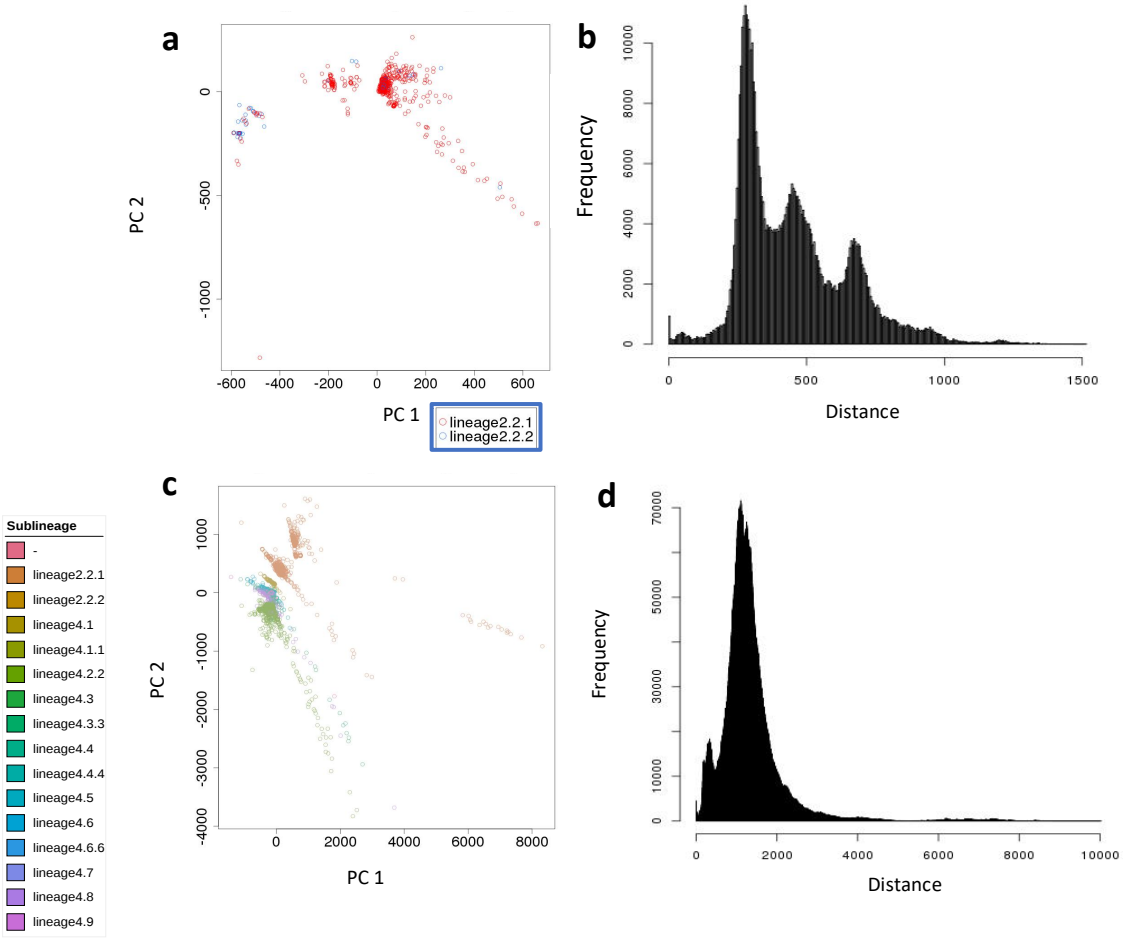
a



b



Additional File 3

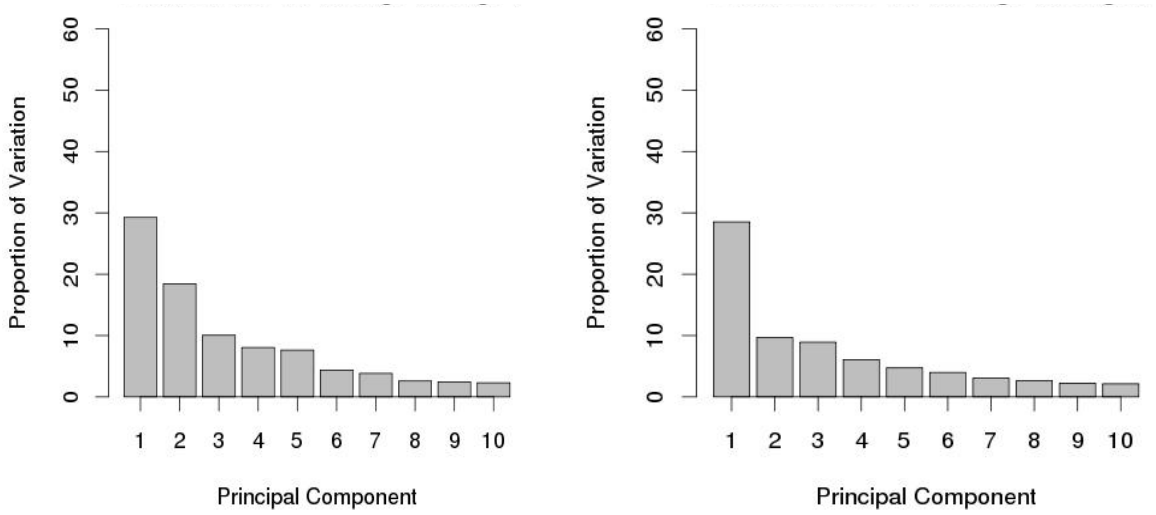


Population diversity within investigated strains.

a Principal component 1 (PC1) by principal component 2 (PC2) for lineage 2, The first 10 principal components account for 71.9% of the variation in lineage 2; **b** Distance plot for lineage 2 showing pairwise number of variant differences between samples; **c** Principal component 1 (PC1) by principal component 2 (PC2) for lineage 4, the first 10 principal components account for 88.9% of the variation in lineage 4. **d** Distance plot for lineage 2 showing pairwise number of variant differences between samples.

Additional File 4

Scree plots showing the proportion of variation accounted for by the first ten principal components, calculated for the pairwise distances within **a** lineage 4 and **b** lineage 2.



Additional File 5

Drug-resistance phenotype frequency table by lineage; 'Total' shows the number and percentage by lineage of known drug-resistance phenotypes.

Drug	Lineage	Suseptible		Resistant		Total	
MDR	2	328	64.3%	182	35.7%	510	72.6%
MDR	4	2704	90.5%	285	9.5%	2989	80.7%
XDR	2	328	90.1%	36	9.9%	364	51.9%
XDR	4	2704	98.8%	34	1.2%	2738	73.9%
XDRvMDR	2	182	83.5%	36	16.5%	218	31.1%
XDRvMDR	4	285	89.3%	34	10.7%	319	8.6%
FQ	2	235	75.1%	78	24.9%	313	44.6%
FQ	4	692	87.8%	96	12.2%	788	21.3%
AG	2	58	75.3%	19	24.7%	77	11.0%
AG	4	261	75.0%	87	25.0%	348	9.4%
RMP	2	364	59.2%	251	40.8%	615	87.6%
RMP	4	2974	86.4%	467	13.6%	3441	92.8%
INH	2	354	57.9%	257	42.1%	611	87.0%
INH	4	2770	82.5%	586	17.5%	3356	90.6%
EMB	2	411	77.1%	122	22.9%	533	75.9%
EMB	4	2361	93.5%	163	6.5%	2524	68.1%
PZA	2	281	86.7%	43	13.3%	324	46.2%
PZA	4	2019	95.6%	92	4.4%	2111	57.0%
STREP	2	223	57.0%	168	43.0%	391	55.7%
STREP	4	1499	84.7%	271	15.3%	1770	47.8%
CIP	2	22	91.7%	2	8.3%	24	3.4%
CIP	4	140	87.5%	20	12.5%	160	4.3%
MOX	2	120	93.8%	8	6.3%	128	18.2%
MOX	4	320	96.7%	11	3.3%	331	8.9%
OFL	2	109	61.2%	69	38.8%	178	25.4%
OFL	4	282	79.2%	74	20.8%	356	9.6%
KAN	2	178	86.0%	29	14.0%	207	29.5%
KAN	4	379	92.4%	31	7.6%	410	11.1%
AMK	2	121	79.6%	31	20.4%	152	21.7%
AMK	4	510	94.4%	30	5.6%	540	14.6%
CAP	2	218	88.6%	28	11.4%	246	35.0%
CAP	4	531	92.7%	42	7.3%	573	15.5%
ETH	2	130	76.0%	41	24.0%	171	24.4%
ETH	4	240	86.6%	37	13.4%	277	7.5%

Additional File 6

	RMP	INH	CIP	MOX	OFL	AMK	CAP	KAN	STREP	AMB	PZA	ETH
RMP	**4056**	0.14	0.07	0.03	0.26	0.09	0.08	0.09	0.15	0.08	0.04	0.17
INH	3936	**3967**	0.08	0.03	0.34	0.10	0.10	0.09	0.17	0.09	0.05	0.24
CIP	183	179	**184**	0.15	NA	0.11	0.04	0.21	0.09	0.08	0.08	0.45
MOX	440	459	27	**459**	0.12	0.03	0.02	0.02	0.04	0.02	0.02	0.10
OFL	529	414	0	49	**534**	0.09	0.08	0.10	0.23	0.20	0.23	0.11
AMK	682	572	35	350	358	**692**	0.06	0.05	0.06	0.06	0.05	0.07
CAP	813	702	89	351	429	604	**819**	0.06	0.07	0.07	0.06	0.09
KAN	611	617	38	326	305	438	597	**617**	0.07	0.07	0.05	0.10
STREP	2146	2160	80	347	344	488	688	602	**2161**	0.08	0.05	0.22
EMB	3033	3051	180	457	410	568	702	616	2155	**3057**	0.03	0.13
PZA	2429	2428	147	433	81	391	414	361	1645	2410	**2435**	0.30
ETH	443	328	11	49	417	259	338	216	257	326	92	**448**
Proportion of samples phenotyped	0.92	0.90	0.04	0.10	0.12	0.16	0.19	0.14	0.49	0.69	0.55	0.10

Cross-Resistance Table upper triangle shows proportion of samples with a known phenotype for both vertical and horizontal phenotype, that test positive for vertical phenotype. Diagonal (in bold) shows number of samples with a known phenotype for each phenotype. Lower triangle shows number of samples with a known phenotype for both horizontal and vertical phenotype.

Additional File 7

Position	Type	Reference	Alternative	Variant
6620	SNP	G	C,A	461D>H, 461D>N
7570	SNP	C	G,T	90A>G, 90A>V
7572	SNP	T	C	91S>P
7581	SNP	G	C,A,T	94D>P, 94D>F, 94D>S
7572	SNP	T	C	91S>P
232574	SNP	G	T	115G>V
732110	SNP	T	A,G	61C>S, 61C>G
759939	SNP	C	G,T	45P>A, 45P>S
760314	SNP	G	C,T	170V>L, 170V>F
761095	SNP	T	C,G	430L>P, 430L>R
761109	SNP	G	T	435D>Y
761110	SNP	A	C,G,T	435D>A, 435D>G, 435D>V
761139	SNP	C	A,G,T	445H>N, 445H>D, 445H>Y
761140	SNP	A	C,G,T	445H>P, 445H>R, 445H>L
761155	SNP	C	G,T	450S>W, 450S>L
761161	SNP	T	C	452L>P
761998	SNP	T	C	731L>P
764363	SNP	G	C,A,T	332G>R, 332G>S, 332G>C
764817	SNP	T	C,G	483V>A, 483V>G
764840	SNP	A	G	491I>V
767123	SNP	G	A,T	1252V>M, 1252V>L
781687	SNP	A	G	43K>R
781822	SNP	A	C,G,T	88K>T, 88K>R, 88K>M
2155167	SNP	G	T,A	315S>R, 315S
2155168	SNP	C	A,G,T	315S>I, 315S>K, 315S>T, 315S>I
2288764	SNP	T	C,G	160T>A, 160T>P
2288778	SNP	A	C,G	155V>G, 155V>A
2288820	SNP	T	G	141Q>P
2288868	SNP	A	C,T	125V>G, 125V>D
2289103	SNP	T	C,G	47T>A, 47T>P
2289207	SNP	T	C,G	12D>G, 12D>A
2747471	SNP	A	C,G	43I>S, 43I>T
4247429	SNP	A	C,G	306M>L, 306M>V
4247431	SNP	G	C,A,T	306M>I, 306M>I, 306M>I
4247469	SNP	A	C,G	319Y>S, 319Y>C
4247574	SNP	A	C	354D>A
4247729	SNP	G	A,T	406G>T, 406G>Y
4247730	SNP	G	C,A	406G>A, 406G>D
4248003	SNP	A	C,G	497Q>P, 497Q>R
4249518	SNP	A	G	1002H>R
761139	Indel	CACA	C	445HK>Q
2288778	Indel	A	AC	155VLVDLTAGVSADTTVAALEEMRTASVELVCSS*>155GAGGPDS GCVGRYHRRRAGGDAHRQRRVGLQLL
2289207	Indel	T	TC	12DFCEGGSLAVTGGAALARAISDYLAEAADYHHVVATKDFHIDPGD HFSGTPDYSSSWPPHCVSGTPGADFHPSLDTSAIEAVFYKGAYTGAYS GFEGVDENGTPLLNWLRQRGVDEVDVVGIATDHCVRQTAEDAVRN GLATRVLVDLTAGVSADTTVAALEEMRTASVELVCSS*>12GLLRGW LAGGNRWRRAGPRHQRLPGRSGGLPSRRGNQGLPHRPG*

Variant Position Table detailing variants at all positions with at least one non-synonymous variant found to be significantly associated with a phenotype in any of the variant-based analyses.

Additional File 8

Locus Comparison Table showing which analyses and in which lineage each loci was identified. A 'x' indicates a locus which was not identified by the method of analysis in question. Loci without a known association with the phenotype are highlighted in bold. There were 9 loci identified by Coll et al. (2018) using the wider non-lineage specific dataset that were not identified here[13].

Locus	Known Phenotype Association[13]	PhyC	Locus-based GWAS	Variant-based GWAS
argJ	Novel	x	x	2+4
cut5b-Rv3725	Novel	x	4	x
eis-Rv2417c	AG	2+4	x	x
embB	EMB	2, 4, 2+4	2, 4, 2+4	x
embC-embA	EMB	2, 4, 2+4	4, 2+4	x
espE-espF	Novel	x	2+4	x
fabG1	INH	4, 2+4	x	x
fadB4-Rv3142c	Novel	x	4	x
folC	PAS	4, 2+4	x	x
gid	STM	x	4, 2+4	x
gyrA	FQ	2, 4, 2+4	2, 4, 2+4	4, 2+4
gyrB	FQ	4, 2+4	x	x
hadA	Novel	4, 2+4	x	x
inhA	INH, ETH	x	4	x
katG	INH	2, 4, 2+4	2, 4, 2+4	2, 4, 2+4
lipJ-cinA	Novel	x	2+4	x
moeX	Novel	x	4	x
oxyR'-ahpC	INH	2+4	x	x
pip-Rv0841	Novel	x	2+4	x
pncA	PZA	4, 2+4	2, 4, 2+4	x
pncA-Rv2044c	PZA	4, 2+4	x	x
recF	Novel		4, 2+4	4, 2+4
rpoB	RMP	2, 4, 2+4	2, 4, 2+4	2, 4, 2+4
rpoC	RMP	4, 2+4	x	x
rpsL	STM	2, 4, 2+4	2, 4, 2+4	2, 2+4
rrs	STM, AG	2, 4, 2+4	2	2, 4, 2+4
Rv0197	Novel	x	x	4, 2+4
Rv0526	Novel	x	2+4	x
Rv1313c-Rv1314c	Novel	x	4, 2+4	x
Rv1482c-fabG1	INH, ETH	2, 4, 2+4	4, 2+4	4, 2+4
Rv2668	Novel	x	4	x
Rv3007c	Novel	x	4	x
Rv3115-moeB2	Novel	2+4	x	x
Rv3128c-Rv3129	Novel	x	2	x
thyX-hsdS.1	PAS	2, 4, 2+4	x	x
tuf	Novel	x	4	x
whiB6-Rv3863	Putative STM or ETH	2+4	x	x

Additional File 9

Study frequency table, showing numbers and percentage of strains from each study by lineage.

Study	Number in Lineage 2	Percentage of Lineage 2	Number in Lineage 4	Percentage of Lineage 4	Total	Percentage of Total
Brazil	0	0.0	92	2.5	92	2.1
China	122	17.4	37	1.0	159	3.6
South Africa	127	18.1	320	8.6	447	10.1
Columbia	0	0.0	15	0.4	15	0.3
Mixed**	38	5.4	50	1.4	88	2.0
Germany	0	0.0	20	0.5	20	0.5
Pakistan	0	0.0	4	0.1	4	0.1
Malawi	71	10.1	1116	30.1	1187	26.9
Lisbon	5	0.7	73	2.0	78	1.8
Peru	6	0.9	72	1.9	78	1.8
Porto	15	2.1	116	3.1	131	3.0
Russia	2	0.3	0	0.0	2	0.1
Vietnam	19	2.7	8	0.2	27	0.6
UK	263	37.5	1697	45.8	1960	44.5
WHO*	34	4.8	86	2.3	120	2.7

Additional File 10

Sublineage frequency table; numbers and percentage by lineage assigned to each sublineage.

Sublineage	Frequency	Percentage
2.2.1	655	93.3
2.2.2	47	6.7
	702	100.0
4.3.3	1,540	41.6
4.1.1	913	24.6
4.8	421	11.4
4.4.4	181	4.9
4.9	141	3.8
4.6.6	117	3.2
4.2.2	116	3.1
4.5	89	2.4
4.7	77	2.1
4.6	37	1.0
4	34	0.9
4.1	30	0.8
4.3	6	0.2
4.4	4	0.1
	3,706	100.0

Chapter 3:

Genome-wide analyses identify novel associations with Extensively Drug Resistant tuberculosis

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1500052	**Title**	Ms
First Name(s)	Yaa Emily Adoma		
Surname/Family Name	Oppong		
Thesis Title	Investigating the genomic basis of antimicrobial resistance in *Mycobacterium tuberculosis* (*Mtb*) using genome-wide methodologies		
Primary Supervisor	Martin Hibberd		

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

<u>SECTION B – Paper already published</u>

Where was the work published?	
When was the work published?	
If the work was published prior to registration for your research	

degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	

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

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	American Journal of Respiratory and Critical Care Medicine
Please list the paper's authors in the intended authorship order:	Yaa E A Oppong, Jody E Phelan, Martin L Hibberd, Taane G Clark
Stage of publication	Undergoing Revisions

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	**YO performed bioinformatic and statistical analyses under the supervision of TC and MH, and wrote the first draft of the manuscript.**

SECTION E

Student Signature	
Date	

Supervisor Signature	
Date	

Genome-wide analyses identify novel associations with Extensively Drug Resistant tuberculosis

Yaa E A Oppong[1, §], Jody E Phelan[1], Martin L Hibberd[1,*], Taane G Clark[1,2,*]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

[2] Faculty of Epidemiology and Population Health, LSHTM, London, United Kingdom

* Joint authors

§ Corresponding author:

Yaa E. A. Oppong

Pathogen Molecular Biology Department

Faculty of Infectious and Tropical Diseases

London School of Hygiene and Tropical Medicine

Keppel Street, London WC1E 7HT

Tel. +44 (0) 20 7636 8636

e-mail: yaa.oppong@lshtm.ac.uk

Author Contributions:

Funding*:* YO is funded by a BBSRC PhD studentship (Grant no. BB/J014567/1). JP Is supported by a Newton Institutional Links Grant (British Council. 261868591). TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). These funding bodies did not have a role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Running head: Genomic analysis of Extensively Drug Resistant tuberculosis

Descriptor Number: 11.2- Epidemiology of Tuberculosis

Total Word Count: Main body- (max. 3500) excluding the abstract, references, legends.

Abstract- (max 250).

This article has an online data supplement, which is accessible from this issue's table of content online at **www.atsjournals.org**

ABSTRACT

Rationale: The transmission of extensively drug resistant *Mycobacterium tuberculosis* (XDR-TB) is threatening the control of tuberculosis disease. Whole genome sequencing of XDR-TB and non-XDR-TB strains can provide much needed insights into the loci and specific mutations underlying drug resistance and transmission.

Objectives: To examine the genomic basis of XDR-TB.

Methods: We characterised 613,821 single nucleotide polymorphisms (SNPs) from whole genome sequencing data for a global dataset of 18,255 *M. tuberculosis* isolates representing the four main *M. tuberculosis* lineages. The SNPs were applied in genome-wide association study (GWAS) and convergent evolution analysis approaches to identify genetic markers of XDR-TB.

Measurements and Main Results: Through GWAS we identify 20 loci in novel associations within highly drug-resistant *M. tuberculosis* strains, including *Rv2000* and *espA-ephA,* that may enhance transmissibility. Cluster-based GWAS and a lack of overlap with associations identified through convergent-evolution-based analyses confirmed that many of the novel associations have been driven by transmission in outbreaks of highly-resistant *M. tuberculosis*.

Conclusions: Our XDR-TB genomic analysis revealed direct resistance-conferring mutations, as well as markers of high transmissibility.

ABSTRACT WORD COUNT: 175

KEYWORDS: *Mycobacterium tuberculosis*, antimicrobial resistance, transmission

INTRODUCTION

Tuberculosis disease (TB), caused by *Mycobacterium tuberculosis* (*Mtb*), is an important global public health issue, with 10 million new cases, and 1.6 million deaths in 2017 (1). The evolution of antimicrobial resistance in *Mtb* poses a serious threat to global TB control efforts. Resistant *Mtb* can be categorised as multi-drug resistant (MDR-TB), defined as resistance to two first–line drugs, rifampicin and isoniazid, or extensively-drug resistant (XDR-TB), defined as MDR-TB *Mtb* that is additionally resistant to fluoroquinolones and at least one second-line injectable. About 8.5% of MDR-TB cases were XDR-TB in 2017. At least 123 countries have had at least one XDR-TB case. Transmission of XDR-TB *Mtb* has been observed in the community, including in TB endemic South Africa (2). There is evidence of a genomic basis of transmissibility (3,4) (Sobkowiak et al., under review), though such complex phenotypes have been less studied than drug resistance outcomes. Genome wide association studies (GWAS) and convergent evolution-based approaches are proving useful in characterising the genetic basis of drug resistance in *Mtb* (5–7), with evidence of differing resistance genomics between the seven lineages into which members of the *Mtb* complex are categorised (8,9). They have also identified compensatory mutations, defined as mutations that restore fitness when costly resistance conferring variants are present in the genome, which have been suggested as drivers of transmission (10,11). The importance of these compensatory mutations in the *Mtb* genome in contributing to XDR-TB remains to be established (12,13).

An improved understanding of XDR-TB evolution, and in particular transmission, is critical for TB disease control. With the increasing availability of whole genome sequence (WGS) data for highly resistant XDR-TB isolates, it has become possible to take a more detailed look at the related *Mtb* population genomics. Here we perform GWAS analysis on a global dataset of >18,000 lineage 1, 2 ,3 and 4 *Mtb* isolates to identify genomic variants associated with XDR-TB.

METHODS

Genomic Data and processing

Mtb WGS raw data was available from across 18,255 isolates (see **Supplementary Table E1** for ENA Project accession numbers), including those described in recent studies (14,15). Variants were called from the raw WGS data in relation to the H37rv reference genome using in-house pipelines, as described in (5). In brief, sequencing reads were trimmed using trimmomatic (16) to remove low quality sequences, and then mapped against the H37Rv reference genome (AL123456) using BWA (17) (v0.7.17). SNPs were called using SAMtools/BCFtools (v1.8) (18) in regions where at least 10 reads were present. SNPs were converted into a FASTA format alignment, which was used by ExaML software (19) to reconstruct the phylogeny for each lineage and combined. Drug resistance profiles and lineages were predicted *in-silico* using TBProfiler (v2.0) software (15). Lineages 1, 2, 3, and 4 were used in downstream analyses and isolates assigned to multiple lineages were thought to be mixed infections and were removed.

Phenotypic Data

Resistance phenotypes were generated through laboratory-based drug susceptibility testing, as described in (5). Resistance phenotypes associated with the WGS data were collated (see **Supplementary Table E2**), providing a binary drug resistance phenotype for 15 potential drugs; amikacin, capreomycin, ciprofloxacin, cycloserine, ethambutol, ethionamide, isoniazid, kanamycin, moxifloxacin, ofloxacin, para aminosalicylic acid (PAS), pyrazinamide, rifabutin, rifampicin and streptomycin. Phenotypic data was not complete across all drugs, as resistance to first line treatments leads to second-line assessment (see **Supplementary Table E2**). Rifampicin and isoniazid drug susceptibility testing data was the most complete (>99%), and ciprofloxacin, PAS and cycloserine the least complete (<3%). Five composite "resistance" phenotypes were inferred from the phenotypic data: (i) aminoglycosides (amikacin, kanamycin or streptomycin), (ii) fluoroquinolones (ofloxacin, moxifloxacin or ciprofloxacin), (iii) MDR-TB (isoniazid and rifampicin, but not XDR-TB), (iv) XDR-TB (MDR-TB + resistance to fluoroquinolones or second line injectables (amikacin, kanamycin or capreomycin)) and (v) pan susceptible (susceptible to rifampicin and isoniazid and with no other known resistance). Low frequency phenotypes in each lineage were removed from downstream analyses as such; for each lineage, any total phenotype frequency of <100 or resistance frequency of <25 isolates (**Supplementary Table E3**).

Association Analyses

GWAS analyses were conducted using a statistical mixed model implemented in GEMMA software (20) for lineages 1, 2, 3 and 4 separately and combined. Models

contained individual SNPs ("variant-based") or aggregated mutations in gene and intergenic regions ("locus-based"). They also incorporated a kinship matrix ("random effect") to account for bacterial genetic relatedness, as described in (9). A conservative p-value cut-off of $1x10^{-21}$ (1E-21) was applied, as described in (9). Convergent-based phyC analyses were conducted on each SNP across each lineage separately and combined (6). For both GWAS and phyC analyses, the phenotypes of XDR-TB and MDR-TB were compared to the pan-susceptible phenotype. XDR-TB was additionally compared to the MDR-TB phenotype. For variant GWAS a minor allele frequency cut-off of > 0.001 (0.1%) was applied. Additionally, GWAS was conducted on isolates belonging to XDR-TB transmission clusters (determined by pairwise SNP distance <10) compared to all other (XDR-TB) isolates. PE/PPE genes were removed from analyses due to their repetitive nature and consequent difficulties in mapping these regions (21).

Functional Classification

Functional classification was conducted for loci identified to be in novel association with resistance phenotypes, using the STRING database (22). For intergenic loci; the functions of both flanking genes were included. For each novel loci, genes linked by neighbourhood, co-occurrence, co-expression, experiments, databases, text mining or homology were identified using STRING (22), to look for known resistance-conferring genes.

RESULTS

Lineages, drug resistance and transmission

In total 18,255 *Mtb* isolates were included in analyses. The majority were in lineage 4 (n=8,892; 47.3%), followed by lineage 2 (n=4,761; 25.3%), lineage 3 (n=2,860; 15.2%) and lineage 1 (n=2,270; 12.1%). There was a total of 10,924 (58.2%) pan susceptible, 3,804 (20.3%) MDR-TB and 351 (1.9%) XDR-TB (**Supplementary Table E3**). There were 6 (1.7%), 163 (46.4%), 28 (8.0%), and 154 (44.0%) XDR-TB isolates in lineages 1, 2, 3 and 4, respectively. The majority of pan susceptible isolates were in lineage 4 (n=5,838, 53.4%) (see **Supplementary Table E3**). Lineage 1 had the greatest diversity, as measured by the number of SNPs difference (mean=896.3, median=983.0), whereas lineage 2 had the lowest (mean=322.0, median=244.9) (see **Supplementary Table E4, Supplementary Figure E1**).

By looking at the XDR-TB isolates that were <10 SNPs different from each other, we determined potential transmission networks (lineage (no. samples in networks): 1 (0), 2 (67), 3 (7), and 4 (95)) (see **Supplementary Figure E2, Supplementary Figure E3, Supplementary Figure E4, Supplementary Figure E5, Supplementary Figure E6, Supplementary Figure E7**). The most numerous sub-lineage with transmission events being 4.3.3 (Euro-American (LAM), n=54), followed by 2.2.2 (East-Asian (Beijing), n=45) (see **Supplementary Table E4**). Of the XDR-TB isolates that were <10 SNPs different to at least one other isolate, 107 were from South Africa, 22 from Belarus, 21 from Portugal, 7 from Argentina, 7 from Pakistan, 3 from Brazil and 2 from China (see **Supplementary Table E4**).

GWAS analyses reveal known and novel loci associated with MDR-TB and XDR-TB

A GWAS association approach revealed eleven loci known to be involved with drug resistance phenotypes, including MDR-TB and XDR-TB. These loci were *embB*, *embC-embA*, *gyrA*, *gyrB*, *inhA, katG*, *oxyR'-ahpC*, *pncA*, *rpoB*, *Rv1482c-fabG1* and *ubiA* (see **Supplementary Tables E5** and **E6**). An analysis of the same phenotypes using phyC detected eleven loci: *eis-Rv2417c*, ***embB***, ***embC-embA***, ***gyrA***, ***inhA***, ***katG***, ***rpoB***, *rpsL, rrs,* and ***Rv1482c-fabG1*** (see **Supplementary Table E7,** overlaps with GWAS indicated in bold).

Across all GWAS analyses there were 31 novel associations relating to 20 loci (see **Tables 1** and **2**). There were 21 novel associations with XDR-TB compared to pan susceptibility, involving 15 loci (*aroG*, *cydB-cydA*, *echA2-mazF1, pks6*, *PPE13-Rv0879c*, *recF*, *Rv0197*, *Rv0530A-Rv0531*, *Rv1373, Rv1616*, *Rv1924c-fadD31*, *Rv3235*, *Rv3238c-Rv3239c*, *Rv3554-Rv3555c*, *Rv3755c-proZ*); seven novel associations with XDR-TB compared to MDR-TB, involving three loci (*espA-ephA*, *Rv0571c-Rv0572c*, *Rv2000*); and three novel associations with fluoroquinolones, involving three loci (*folD-relJ*, *Rv0530A-Rv0531*, *ligC-Rv3732*). No novel associations were identified with individual drug resistance phenotypes in either the locus- or SNP-based GWAS, except ciprofloxacin for lineage-combined, which this was excluded due to lineage-specific phenotype frequency threshold (see **Tables 1 and 2**).

Distinguishing Lineage-combined from lineage specific effects

For lineage-combined locus-based GWAS analyses, one novel association was identified between *Rv3755c-proZ* and XDR-TB compared to pan susceptible (p-value= 4.15E-59)

(see **Table 1**). For SNP-based GWAS analyses, seven novel associations were identified for seven loci. Three loci ((*Rv1616*, *pks6*, *Rv1373* and *Rv0197*) in association with XDR-TB compared to pan susceptible (p-values< 2.78E-23). A further four loci (*Rv2000*, *Rv0571c-Rv0572c* and *espA-ephA*) in association with XDR-TB compared to MDR-TB (p-values< 2.11E-22) (see **Table 2**). Additionally, we identified one significant novel association with ciprofloxacin involving *Rv2128* (p-value= 4.4E-22). The only novel association identified solely by lineage-combined GWAS analyses was *pks6* (see **Table 2**). PhyC based analysis identified 20 non-synonymous SNPs in novel association with resistance phenotypes in 14 loci; *Rv0336*, *Rv1765c*, *Rv3611*, *Rv0797, Rv1150, Rv2015c, Rv1588c, pks12*, *Rv0515*, *Rv0094c*, *Rv2186c*, *Rv2512c*, *Rv1042c, Rv3115, Rv3193c* (P-values< 10E-5, see **Supplementary Table E8**).

Lineage-specific analysis

No novel associations were found for any of the lineage 1 specific GWAS analyses (see **Tables 1 and 2**; **Supplementary table E5**). Seven associations with known resistance loci were identified (locus-based: *rpoB* , *Rv1482c- fabG1*, *katG*, *pncA* and *embB*; variant-based: *katG*, *embB*, *rpoB*, rrs, *Rv1482c-fabG1* and *inhA*). PhyC analyses found two novel associations (*Rv1204c* and *Rv2186c)*, both in association with ethambutol (both Fisher-test p-value= 2.57E-06) (see **Supplementary Table E8**).

For lineage 2, specific locus-based GWAS analyses, novel associations were identified for three loci (*Rv1924c-fadD31*, *aroG* and *Rv3235*) in association with XDR-TB compared to pan susceptible (p-values< 7.76E-25, respectively) (see **Table 1**). For SNP-based GWAS analyses, *Rv1373* was identified in association with XDR-TB compared to

pan susceptible (p-value= 2.62E-31) (**Table 2**). Application of phyC did not detect lineage 2 specific associations (see **Supplementary Table E8**).

For lineage 3 specific locus-based GWAS analyses, five novel associations were identified for four loci; *Rv0530A-Rv0531* and *Rv3554-Rv3555c* in association with XDR-TB compared to pan susceptible (p-values< 1.11E-22) and *folD-relJ*, *Rv0530A-Rv0531* and *ligC-Rv3732* in association with fluoroquinolones (p-values < 8.96E-23) (see **Table 1**). For SNP-based GWAS analyses, one novel association was identified; *PPE13-Rv0879c* identified in association with XDR-TB compared to pan susceptible (p-value= 1.04E-24) (see **Table 2**). There were no associations identified by lineage 3 specific phyC (see **Supplementary Table E8**).

For lineage 4 specific locus-based GWAS analyses, four novel associations were identified for four loci; *Rv3755c-proZ*, *Rv3238c-Rv3239c* and *cydB-cydA*, in association with XDR-TB compared to pan susceptible (p-values < 2.24E-24) and *Rv0571c-Rv0572c* in association with XDR-TB compared to MDR-TB (p-value= 5.93E-31) (see **Table 1**). For SNP-based GWAS analyses, seven novel associations were identified across seven loci. In particular, four loci (*recF*, *Rv1616*, *echA2-mazF1* and *Rv0197)* were revealed in association with XDR-TB compared to pan susceptible (p-values < 6.02E-22). Whilst, a further three loci (*Rv2000*, *Rv0571c-Rv0572c* and *espA-ephA*) were in association with XDR-TB compared to MDR-TB (p-values< 2.23E-29) (see **Table 2**). The *Rv0571c-Rv0572c* was identified by locus-based and SNP-based GWAS analyses in association analysis with XDR-TB compared to MDR-TB. The application of PhyC revealed no novel

associations, but found known associations of *rpoB* with MDR-TB and rifampicin resistance (fisher-test p-value< 1.72E-07) (see **Supplementary Table E8**).

Functional Characterisation

Analysis of the functional annotation of the locus-based GWAS hits revealed that LigC, of *ligC-Rv3732*, is a DNA ligase involved in DNA repair, and Rv3239c*,* of *Rv3238c-Rv3239c,* is similar to antibiotic resistance, and efflux proteins (see **Table 1**). Annotation of the locus-based GWAS hits with associated proteins from STRING revealed, *folD* of *folD-relJ* is associated with *thyX* and *thyA*, *Rv0571c* of *Rv0571c-Rv0572c* is associated with *hspX*, and *Rv3554-Rv3555c* is associated with *fas* and *echA20* (see **Supplementary Table E9**). Similarly, analysis of the SNP-based GWAS hits revealed that *recF* is involved in replication and repair and *espA-ephA* is associated with ESX-1 secretion. Annotation of the SNP-based GWAS hits with associated proteins from STRING revealed several pathways: (i) *recF* is associated with *gyrA* and *gyrB,* (ii) *Rv2000* is associated with *ubiA* and *ephA* (see **Supplementary Figure E6**), (iii) *Rv0571c-Rv0572c* Is associated with *hspX,* and (iv) *pks6* is associated with *fas* (see **Supplementary Table E9**).

Functional annotation of the novel associations identified by phyC revealed that, amongst other functions, four loci are transposases or putative transposases (*Rv2512c, Rv1042c, Rv3115*, *Rv0797*) and three loci are members of the 13E11 repeat family (*Rv0336, Rv0515, Rv0094c*) (see **Supplementary Table E8**).

Transmission Cluster GWAS

The locus-based transmission cluster GWAS comparing XDR-TB transmission clusters to all isolates revealed an association with *Rv0571c-Rv0572c* (p-value= 1.45E-34) (See **Supplementary Table E10**). The similar SNP-based analyses revealed associations in two known drug resistant loci: *rpoB* (p-value=3.08E-284) and *ubiA* (p-value=1.9E-129) (see **Supplementary Table E10**). *rpoB* was previously identified by locus-based GWAS, SNP-based GWAS and phyC (see **Supplementary Tables E5, E6,** and **E7**). *ubiA* was previously identified by SNP-based GWAS and phyC, but not locus-based GWAS (see **Supplementary Tables E5, E6, E7**). Additionally, the SNP-based transmission cluster GWAS, which compared XDR-TB transmission clusters to all isolates, revealed associations with three loci identified by the non-cluster-based methods. In particular, these were *Rv2000* (p-value= 2.09E-310), Rv0571c-*Rv0572c* (p-value= 1.33E-201) and *espA-ephA* (p-value= 4.46E-173) (see **Supplementary Table E10**). The SNP-based transmission cluster GWAS comparing *Mtb* isolates in XDR-TB transmission clusters to all other isolates, revealed associations with nine loci identified by the non-cluster-based methods. In particular, these were *Rv1061* (p-value= 8.57E-95), *mce2B* (p-value= 2.32E-81), *iniA* (p-value= 1.88E-30), *Rv2425c* (p-value= 1.88E-30)*, Rv1144-mmpL13a* (p-value= 1.07E-27), *Rv3471c* (p-value= 1.07E-27), *atsD* (p-value= 1.31E-24), *secD* (p-value= 8.76E-24), and *Rv2499c* (p-value= 2.29E-23) (see **Supplementary Table E10**). A locus-based GWAS of XDR-TB transmission clusters compared to non-clustering XDR-TB also identified *Rv0571c-Rv0572c* (p-value=2.15E-21). SNP-based GWAS of cluster XDR compared to non-cluster XDR also identified *rpoB* (p-value= 6.78E-27), *Rv2000* (p-

value= 6.78E-27), *espA-ephA* (p-value= 4.87E-23) and *Rv0571c-Rv0572c* (p-value= 9.5E-23) (see **Supplementary Table E10**).

Rv2000, which had the most significant association from the SNP-based transmission cluster GWAS comparing XDR-TB transmission clusters to all isolates (see **Supplementary Table E10**), was found in 48 clustering isolates and seven isolates that did not belong to an XDR transmission cluster, all of which belonged to sub-lineage 4.3.3. Of the non-XDR-cluster isolates with the *Rv2000* mutation, one was XDR-TB, two were MDR-TB and none were pan-susceptible (see **Figure 1).** All *Rv2000* mutant isolates were from South Africa. A STRING pathway analysis linked *Rv2000* to *ephA*, *Rv2001*, *fabG3*, *ubiA*, *yidC*, *ctpB*, *fhaA*, *atsD*, *Rv0493c* and *mmpL13a* (see **Supplementary Figure E7**).

DISCUSSION

The application of GWAS to this global *Mtb* dataset has revealed a number of lineage specific associations between genomic variants and XDR-TB. This is one of the largest MDR-/XDR-TB genomic studies, and one reason for the numerous novel associations with XDR-TB could be increased statistical power. However, many of these variants could have become frequent due to transmission in XDR-TB outbreak settings. Lack of overlap between novel GWAS findings and novel phyC findings further supports this conclusion; suggesting higher relative importance of transmission versus convergent evolution in the development of these GWAS identified variants. Indeed, if there is a genomic contribution to transmissibility, as has been suggested (3), transmissibility and

drug resistance might be presumed to coevolve; more virulent, less latent strains, would likely experience more drug therapy and increased selection for resistance. A genetic component that increases the rate of transmission of MDR-TB or XDR-TB strains could account for these observations, meaning that these strains could be exposed to more drug therapy and thus more selection pressure to develop both drug resistance and transmission success. Occurrence of outbreaks are likely multifactorial, caused by factors including behaviour, environment and genomics. Thus, it is unclear the relative importance of selection versus genetic drift in increasing frequencies of the associated variants identified amongst XDR-TB outbreaks; nevertheless, our work shows these variants are compatible with XDR-TB transmission and present in clinically important outbreak strains and thus warrant further study.

Some of the variants identified may play a role in directly conferring resistance whilst others may play a role in compensating phenotypes with reduced fitness as a result of costly resistance mutations. Interestingly, a number of novel associated variants identified have been linked to known resistance conferring mutations, such as *folD-relJ* (linked to *thyX; thyA*), *recF* (linked to *gyrA* and *gyrB*) and *Rv2000* (linked to *ubiA*). The *Rv0197* locus has previously been implicated in transmissibility (3), here we find it to be in association with XDR-TB for lineage 4 and lineages-combined GWAS analyses.

The *Rv2000* locus may be of particular interest, considering its highly significant association with XDR-TB transmission. Rv2000 has an unknown function (22), but interestingly has been linked to ephA including through potential gene fusion, coexpression of putative homologs in other species and interaction of putative

homologs in other species (22). The intergenic region *espA-ephA* was detected by our analyses. EphA is an epoxide hydrolase with suggested involvement in detoxification of extraneous host- cell epoxides (22). There is potential for these loci to be implicated in the transmissible XDR-TB phenotype and thus they merit further exploration. Additionally, *Rv2000* has been linked in the literature to *ubiA*, which is itself linked to ethambutol resistance, and *mmpl13a*, a probable conserved transmembrane transport protein; the intergenic region *Rv1144-mmpL13a* was identified by cluster variant GWAS analysis (23,24).

Overall, most novel associations identified by phyC were identified through lineage-combined rather than lineage specific analyses. This is concordant with the idea that, in contrast to GWAS, phyC gains power from more diverse datasets with a greater magnitude of convergent evolution and thus may be particularly suited to identifying independent mutation events leading to drug resistance .

In conclusion, through lineage specific GWAS, we have identified genetic associations with XDR-TB that are cluster specific and not directly associated with drug resistance. These loci are consistent with expansion of XDR-TB clones through transmission chains and may be compensatory or transmission associated. Further work into the biological impact of these variants would provide a deeper understanding into the potential biological mechanisms of their involvement with XDR-TB. The identification of genetically distinct XDR-TB transmission chains may have important implications for TB control.

ACKNOWLEDGEMENTS

Sequence analysis was performed on the MRC (HDR) UK funded eMedlab computing resource.

REFERENCES

1. World Health Organisation. Global Tuberculosis Report 2018. Geneva; 2018.
2. Dheda K, Esmail A, Mcnerney R, Theron G, Streicher EM, van Helden P, et al. Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. Lancet Respir Med [Internet]. 2017;2600(16):1–13. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2213260016304337%0Ahttp://dx.doi.org/10.1016/
3. Nebenzahl-Guimaraes H, van Laarhoven A, Farhat MR, Koeken VACM, Mandemakers JJ, Zomer A, et al. Transmissible *Mycobacterium tuberculosis* strains share genetic markers and immune phenotypes. Am J Respir Crit Care Med [Internet]. 2017 Jun;195(11):1519–27. Available from: http://www.atsjournals.org/doi/10.1164/rccm.201605-1042OC
4. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet [Internet]. 2018 Jun 21;50(6):849–56. Available from: http://dx.doi.org/10.1038/s41588-018-0117-9
5. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. Nat Genet [Internet]. 2018;50(2):307–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29358649

6. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet [Internet]. 2013 Oct 1;45(10):1183–9. Available from: http://www.nature.com/articles/ng.2747

7. Hicks ND, Yang J, Zhang X, Zhao B, Grad YH, Liu L, et al. Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. Nat Microbiol [Internet]. 2018;3(9):1032–42. Available from: http://dx.doi.org/10.1038/s41564-018-0218-3

8. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun [Internet]. 2014 Dec 1;5(1):4812. Available from: http://www.nature.com/articles/ncomms5812

9. Oppong YEA, Phelan J, Perdigão J, Machado D, Miranda A, Viveiros M, et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics [Internet]. 2019;1–15. Available from: http://link.springer.com/article/10.1186/s12864-019-5615-3?utm_source=researcher_app&utm_medium=referral&utm_campaign=MKEF_USG_Researcher_inbound

10. Meftahi N, Namouchi A, Mhenni B, Brandis G, Hughes D, Mardassi H. Evidence for the critical role of a secondary site rpoB mutation in the compensatory evolution and successful transmission of an MDR tuberculosis outbreak strain. J

Antimicrob Chemother. 2016;71(2):324–32.

11. Merker M, Barbier M, Cox H, Rasigade J-P, Feuerriegel S, Kohl TA, et al. Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia. Elife [Internet]. 2018;7:1–31. Available from: https://elifesciences.org/articles/38200

12. Nguyen QH, Contamin L, Nguyen TVA, Bañuls A-L. Insights into the processes that drive the evolution of drug resistance in *Mycobacterium tuberculosis*. Evol Appl [Internet]. 2018;(September 2017):1498–511. Available from: http://doi.wiley.com/10.1111/eva.12654

13. Müller B, Borrell S, Rose G, Gagneux S. The heterogeneous evolution of multidrug-resistant Mycobacterium tuberculosis. Trends Genet. 2013;29(3):160–9.

14. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018;50(2).

15. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med [Internet]. 2019;11(1):41. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0650-x

16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina

sequence data. Bioinformatics [Internet]. 2014 Aug 1;30(15):2114–20. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170

17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics [Internet]. 2009 Jul 15;25(14):1754–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19451168

18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 Aug 15;25(16):2078–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19505943

19. Kozlov AM, Aberer AJ, Stamatakis A. ExaML version 3: A tool for phylogenomic analyses on supercomputers. Bioinformatics. 2015;31(15):2577–9.

20. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet [Internet]. 2012;44(7):821–4. Available from: http://dx.doi.org/10.1038/ng.2310%5Cnhttp://www.nature.com/ng/journal/v44/n7/abs/ng.2310.html

21. Phelan J, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, et al. Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. In Preperation [Internet]. 2015;1–12. Available from: http://dx.doi.org/10.1186/s12864-016-2467-y

22. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of

life. Nucleic Acids Res [Internet]. 2015 Jan 28;43(D1):D447–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25352553%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4383874

23. Motiwala AS, Dai Y, Jones-López EC, Hwang S, Lee JS, Cho SN, et al. Mutations in Extensively Drug-Resistant Mycobacterium tuberculosis That Do Not Code for Known Drug-Resistance Mechanisms . J Infect Dis. 2010;201(6):881–8.

24. Loman N, Pallen M. XDR-TB genome sequencing: a glimpse of the microbiology of the future. Future Microbiol [Internet]. 2008 Apr;3(2):111–3. Available from: https://www.futuremedicine.com/doi/10.2217/17460913.3.2.111

25. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med [Internet]. 2018 Oct 11;379(15):1403–15. Available from: http://www.nejm.org/doi/10.1056/NEJMoa1800474

FIGURE LEGENDS

Figure 1- Maximum likelihood phylogenetic tree based on SNPs for lineage 4 showing cluster GWAS novel variants and transmission chain.

TABLES

Table 1 showing all novel associations identified by locus-based GWAS across all lineages and resistance phenotypes.

Locus	Comparison	Lineage	P-value	Function
Rv3755c-proZ	XDR-TB vs. Pan Susc.	All	4.15E-59	Conserved protein
Rv0530A-Rv0531	XDR-TB vs. Pan Susc.	3	2.50E-56	Conserved protein
Rv3755c-proZ	XDR-TB vs. Pan Susc.	4	2.49E-48	Conserved protein
folD-relJ	Fluoroquinolones	3	3.96E-44	Bifunctional protein FolD
Rv0530A-Rv0531	Fluoroquinolones	3	1.01E-41	Conserved protein
Rv3238c-Rv3239c	XDR-TB vs. Pan Susc.	4	5.77E-39	Probable conserved integral membrane protein
aroG	XDR-TB vs. Pan Susc.	2	2.62E-31	Phospho-2-dehydro-3-deoxyheptonate aldolase
Rv1924c-fadD31	XDR-TB vs. Pan Susc.	2	2.62E-31	Uncharacterized protein
Rv0571c -Rv0572c	XDR-TB vs MDR-TB	4	5.93E-31	Putative phosphoribosyl transferase
Rv3235	XDR-TB vs. Pan Susc.	2	7.76E-25	Hypothetical unknown ala-, arg-, pro-rich protein
cydB-cydA	XDR-TB vs. Pan Susc.	4	2.24E-24	Probable cydB, cytochrome D ubiquinol oxidase subunit II, integral membrane protein
ligC-Rv3732	Fluoroquinolones	3	8.96E-23	DNA ligase C
FdxB-Rv3555c	XDR-TB vs. Pan Susc.	3	1.11E-22	Possible electron transfer protein

XDR-TB = extensively drug resistant TB; MDR-TB = multi-drug resistant TB

Table 2 showing all novel associations identified by SNP-based GWAS across all lineages and resistance phenotypes, where the SNPs have a minor allele frequency of at least 0.1%.

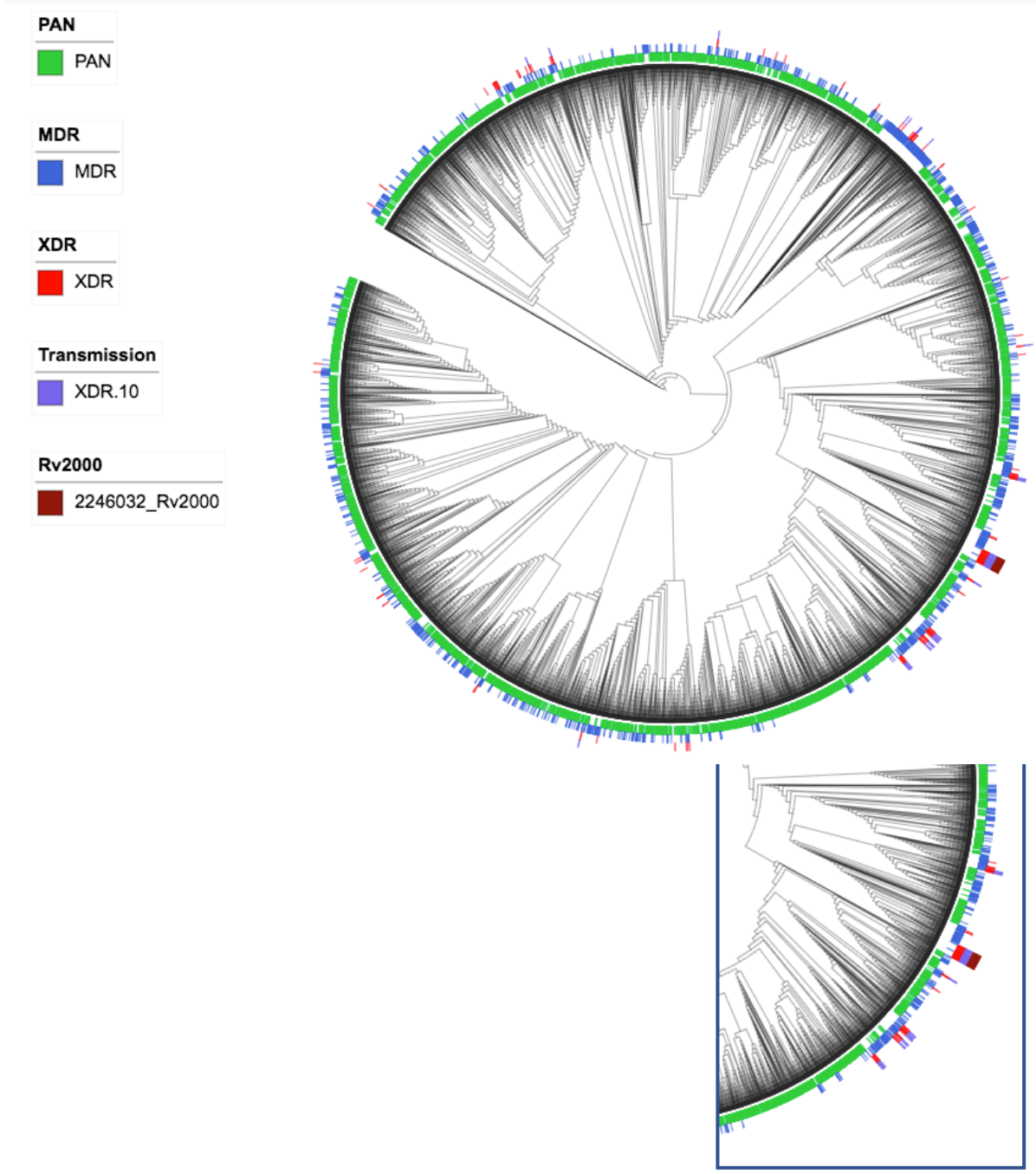
Locus	Genome Position (bp)	Comparison	Lineage	P-value	Odds ratio	Function
recF	4047	XDR-TB vs. Pan Susc.,	4	5.82E-69,	1.37,	DNA replication and repair protein
Rv1616	1815877	XDR-TB vs. Pan Susc., XDR-TB vs. Pan Susc.	All, 4	4.97E-52, 4.66E-43	0.46, 0.55	Conserved membrane protein
Rv2000	2246032	XDR-TB vs MDR-TB, XDR-TB vs MDR-TB	4, All	5.07E-38, 2.11E-29	375.51, 309.22	Unknown protein
Rv0571c-Rv0572c	664929	XDR-TB vs MDR-TB, XDR-TB vs MDR-TB	4, All	2.10E-34, 2.41E-27	703.67, 585.35	Putative phosphoribosyl transferase; Uncharacterized protein
echA2-mazF1	546914	XDR-TB vs. Pan Susc.	4	6.13E-33	0.71	Enoyl-CoA hydratase; Probable endoribonuclease
Rv1373	1546703	XDR-TB vs. Pan Susc., XDR-TB vs. Pan Susc.	2, All	2.62E-31, 1.15E-23	0.06, 1.40	Glycolipid sulfotransferase
espA-ephA	4056430	XDR-TB vs MDR-TB, XDR-TB vs MDR-TB	4, All	2.23E-29, 2.11E-22	150.30, 123.75	ESX-1 secretion-associated protein; Epoxide hydrolase
pks6	486978	XDR-TB vs. Pan Susc.	All	1.95E-25	0.84	Probable pks6,membrane-bound polyketide synthase
PPE13-	978350	XDR-TB vs. Pan	3	1.04E-	2.57	Uncharacterized

Rv0879c		Susc.		24		PPE family protein; Possible conserved transmembrane protein
Rv0197	232574	XDR-TB vs. Pan Susc., XDR-TB vs. Pan Susc.	All, 4	2.78E-23, 6.02E-22	0.90, 1.10	probable molybdopterin oxidoreductase

XDR = extensively drug resistant TB; MDR-TB = multi-drug resistant TB

FIGURES

Figure 1- Maximum likelihood phylogenetic tree based on SNPs for lineage 4 showing distribution of *Rv2000* variant and transmission chain



Genome-wide analyses identify novel associations with Extensively Drug Resistant tuberculosis

Yaa E A Oppong, Jody E Phelan, Martin L Hibberd*, Taane G Clark*

ONLINE DATA SUPPLEMENT

Supplementary Table E1- ENA project codes of isolates used in the study

Project	Drug-resistant	MDR-TB	Susceptible	XDR-TB	Total
cryptic_nejm_2018(25)	754	2391	5124	0	8269
PRJEB10385	210	303	106	95	780
PRJEB11653	77	35	14	0	126
PRJEB14199	34	14	0	77	125
PRJEB15857	8	20	20	0	48
PRJEB2138	4	13	8	11	38
PRJEB2221	19	6	333	0	358
PRJEB2358	30	2	290	0	322
PRJEB2424	2	40	3	0	51
PRJEB2777	0	0	93	0	93
PRJEB2794	79	7	1198	0	1284
PRJEB5162	14	2	176	0	192
PRJEB6276	0	0	3	0	3
PRJEB6945	0	0	55	0	55
PRJEB7056	177	42	890	0	1186
PRJEB7281	16	41	38	2	110
PRJEB7669	3	234	0	11	248
PRJEB7727	12	5	13	0	59
PRJEB9680	71	249	717	2	1039
PRJNA183624	43	140	85	67	335
PRJNA187550	0	94	44	23	161
PRJNA200335	6	47	25	58	136
PRJNA235852	35	20	157	0	212
PRJNA282721	283	91	1485	5	1864
PRJNA355614	0	0	0	0	1633
PRJNA376471	0	8	12	0	20
PRJNA49659	0	0	35	0	36
Total	**1877**	**3804**	**10924**	**351**	**18783**

MDR-TB = multi-drug-resistant TB; XDR-TB = extensively-drug-resistant TB

Supplementary Table E2- Number of tested isolates for each drug

Drug	No. Tested	% Tested	Susceptible	Resistant	Resistant %
Rifampicin	16296	86.8%	11879	4417	27.1%
Isoniazid	16211	86.3%	11053	5158	31.8%
Ethambutol	14655	78.0%	12108	2547	17.4%
Pyrazinamide	11808	62.9%	10016	1792	15.2%
Streptomycin	5161	27.5%	3833	1328	25.7%
Ofloxacin	1982	10.6%	1477	505	25.5%
Kanamycin	1830	9.7%	1194	636	34.8%
Capreomycin	1721	9.2%	1333	388	22.5%
Amikacin	1427	7.6%	1093	334	23.4%
Ethionamide	938	5.0%	609	329	35.1%
Moxifloxacin	870	4.6%	767	103	11.8%
PAS	406	2.2%	363	43	10.6%
Cycloserine	390	2.1%	286	104	26.7%
Ciprofloxacin	394	2.1%	331	63	16.0%
Fluoroquinolones	-	-	10924	574	-
MDR-TB	-	-	10924	3707	-
XDR-TB	-	-	10924	342	-

PAS = Para-aminosalisylic acid; MDR-TB = multi-drug-resistant TB; XDR-TB = extensively-drug-resistant TB

Supplementary Table E3- Distribution of drug resistance types by lineage

Lineage	Drug-resistant	MDR-TB	Susceptible	XDR-TB	Total
1	207	147	1477	6	2270 (12.1%)
2	481	1591	1423	163	4761 (25.3%)
3	214	424	2186	28	2860 (15.2%)
4	975	1642	5838	154	8892 (47.3%)
Total	1877 (10.0%)	3804 (20.3%)	10924 (58.2%)	351 (1.9%)	18783 (100.0%)

MDR-TB = multi-drug-resistant TB; XDR-TB = extensively-drug-resistant TB

Supplementary Table E4- XDR-TB isolates of pairwise distance <10 SNPs cluster summary by lineage

Sub-lineage	Country	Frequency	Total
2.2.1	Belarus	14	22
	China	2	
	South Africa	6	
2.2.2	South	45	45
3	Pakistan	2	2
3.1	Pakistan	2	2
3.1.2	Pakistan	3	3
4.1.1.3	South Africa	4	4
4.1.2.1	Argentina	7	9
	Belarus	2	
4.3.2.1	South Africa	2	2
4.3.3	Belarus	6	54
	South Africa	48	
4.3.4.1	Brazil	3	3
4.3.4.2	Portugal	21	21
4.4.1.1	South Africa	2	2

Supplementary table E5- Locus-based GWAS hits involving known resistance loci

Locus	P-value	Drug	Lineage
pncA	0	Pyrazinamide	All
rpoB	9.80E-278	Rifampicin	All
katG	2.31E-273	Isoniazid	All
pncA	4.13E-255	Pyrazinamide	4
katG	4.76E-219	MDR-TB vs. PAN	All
rpoB	4.19E-215	MDR-TB vs. PAN	All
rpoB	2.97E-178	Rifampicin	4
katG	2.02E-156	Isoniazid	4
rpoB	1.20E-150	Rifampicin	1
rpoB	2.74E-142	Rifampicin	3
rpoB	4.18E-130	FQ	All
oxyR'-ahpC	1.92E-129	Isoniazid	All
Rv3795	3.50E-126	Ethambutol	All
Rv3795	4.58E-119	Ethambutol	4
katG	2.92E-116	MDR-TB vs. PAN	4
rpoB	4.26E-116	MDR-TB vs. PAN	4
oxyR'-ahpC	2.13E-105	MDR-TB vs. PAN	All
rpsL	1.74E-103	Streptomycin	2
rpoB	4.80E-95	MDR-TB vs. PAN	3
rpoB	1.22E-93	Rifampicin	2
rpsL	8.28E-93	Streptomycin	All
gyrA	4.89E-92	Ofloxacin	4
katG	2.61E-90	Isoniazid	2
rpoB	5.23E-89	MDR-TB vs. PAN	1
rpoB	4.20E-85	FQ	2
Rv3919c	2.23E-84	AG	4
Rv3795	1.16E-79	Ethambutol	3
pncA	3.08E-77	Pyrazinamide	3
katG	3.29E-76	MDR-TB vs. PAN	2
rpoB	9.97E-74	AG	All
Rv3795	1.22E-73	Pyrazinamide	4
Rv3919c	1.66E-73	AG	All
rpoB	1.40E-67	FQ	4
katG	6.10E-65	Isoniazid	3
Rv1482c-fabG1	2.78E-61	Isoniazid	1
rpoB	8.80E-61	XDR-TB vs. PAN	All
pncA	1.45E-59	XDR-TB vs. PAN	All
rpoB	3.46E-58	MDR-TB vs. PAN	2

katG	1.29E-57	AG	All
Rv3795	4.19E-57	Pyrazinamide	All
rpsL	4.87E-57	AG	2
rpoB	8.19E-55	Pyrazinamide	4
Rv3795	2.19E-50	Ethambutol	2
Rv1482c-fabG1	4.76E-49	Isoniazid	3
katG	6.07E-49	MDR-TB vs. PAN	3
pncA	7.42E-48	Ethambutol	All
rpoB	1.64E-47	AG	3
oxyR'-ahpC	9.07E-47	Isoniazid	4
Rv3919c	1.27E-46	Streptomycin	All
katG	1.56E-46	AG	4
katG	2.39E-46	FQ	All
pncA	3.48E-46	FQ	All
katG	3.88E-45	Isoniazid	1
eis-Rv2417c	1.99E-44	FQ	3
oxyR'-ahpC	3.75E-44	Isoniazid	2
pncA	7.35E-43	XDR-TB vs. PAN	2
rpoB	1.05E-42	AG	4
katG	2.66E-42	AG	3
rpsL	5.86E-42	AG	All
embB	1.29E-40	Ethambutol	1
oxyR'-ahpC	1.63E-40	MDR-TB vs. PAN	4
Rv1482c-fabG1	2.86E-40	MDR-TB vs. PAN	1
embC-embA	4.23E-40	Ethambutol	4
gyrA	6.09E-40	Ciprofloxacin	All
rpoB	9.64E-40	FQ	3
oxyR'-ahpC	1.73E-39	AG	All
pncA	2.18E-39	FQ	2
pncA	2.69E-39	Pyrazinamide	2
Rv3795	3.88E-39	Pyrazinamide	3
rpoB	5.01E-39	Ethambutol	All
Rv1482c-fabG1	7.57E-39	Isoniazid	All
pncA	3.31E-38	Ethambutol	4
embC-embA	6.73E-38	Ethambutol	All
rpoB	1.48E-37	XDR-TB vs. PAN	2
rpoB	1.97E-37	Ethambutol	1
rpoB	9.87E-37	Isoniazid	All
oxyR'-ahpC	1.22E-35	MDR-TB vs. PAN	2
katG	1.60E-35	Rifampicin	All
rpoB	8.11E-35	Pyrazinamide	All
Rv1482c- fabG1	9.44E-35	Isoniazid	4

rpoB	1.84E-34	XDR-TB vs. PAN	4
Rv3919c	8.20E-34	Streptomycin	4
pncA	1.07E-32	XDR-TB vs. PAN	4
rpoB	2.08E-32	Ethambutol	4
rpoB	3.00E-30	Ethambutol	3
Rv3806c	1.02E-29	Ethambutol	All
rpoB	1.50E-29	AG	1
katG	2.51E-29	FQ	4
rpsL	6.53E-29	Streptomycin	4
gyrA	4.28E-28	Ciprofloxacin	3
katG	4.45E-28	AG	1
oxyR'-ahpC	5.77E-28	FQ	All
rpoB	1.42E-26	AG	2
embC-embA	9.67E-26	XDR-TB vs. PAN	All
rpoB	9.79E-26	XDR-TB vs. PAN	3
katG	1.00E-25	MDR-TB vs. PAN	1
Rv3795	1.01E-25	FQ	All
pncA	1.13E-25	Pyrazinamide	1
Rv3806c	6.80E-25	Ethambutol	4
katG	8.82E-24	FQ	2
katG	9.40E-23	AG	2
gyrA	9.94E-23	Moxifloxacin	All
gyrA	5.92E-22	XDR-TB vs. MDR-TB	4

AG = Aminoglycosides; MDR-TB = multi-drug resistant TB; FQ = Fluoroquinolones; PAN = Pan susceptible; XDR-TB = extensively drug resistant TB

Supplementary Table E6- SNP-based GWAS hits involving known resistance loci

Position	P-value	Locus	Drug	Lineage
2155168	8.01E-203	*katG*	AG	3
761110	9.06E-164	*rpoB*	FQ	3
761155	7.14E-159	*rpoB*	XDR-TB vs. PAN	3
761155	8.35E-157	*rpoB*	XDR-TB vs. PAN	All
1473246	5.13E-126	*rrs*	Amikacin	All
761155	5.45E-125	*rpoB*	XDR-TB vs. PAN	4
761155	2.12E-120	*rpoB*	FQ	3
2155168	6.06E-109	*katG*	Isoniazid	All
761155	4.05E-101	*rpoB*	FQ	All
2155168	1.04E-99	*katG*	AG	4
2155168	1.60E-96	*katG*	Isoniazid	4
761110	1.40E-94	*rpoB*	XDR-TB vs. PAN	4
1473246	1.05E-93	*rrs*	Amikacin	4
2155168	3.80E-91	*katG*	FQ	4
2155168	1.73E-90	*katG*	Isoniazid	1
761110	2.66E-86	*rpoB*	XDR-TB vs. PAN	All
4247429	2.44E-85	*embB*	Ethambutol	1
2155168	1.06E-84	*katG*	AG	1
1473246	4.09E-84	*rrs*	Capreomycin	All
2155168	1.38E-82	*katG*	AG	All
2155168	1.29E-81	*katG*	MDR-TB vs. PAN	1
761155	1.68E-79	*rpoB*	Rifampicin	1
2155168	4.32E-79	*katG*	FQ	All
761155	2.28E-78	*rpoB*	Rifampicin	All
761155	2.73E-78	*rpoB*	FQ	4
760314	2.08E-73	*rpoB*	Rifampicin	All
760314	7.78E-72	*rpoB*	MDR-TB vs. PAN	All
1472362	2.18E-71	*rrs*	AG	1
761155	1.31E-70	*rpoB*	Rifampicin	4
1473246	2.42E-70	*rrs*	XDR-TB vs. PAN	3
2155168	1.40E-68	*katG*	XDR-TB vs. PAN	3
1473246	9.54E-68	*rrs*	Capreomycin	4
761155	2.16E-66	*rpoB*	AG	3
1473246	2.85E-66	*rrs*	Kanamycin	All
1673425	1.01E-64	*Rv1482c-fabG1*	Isoniazid	1
2155168	3.17E-64	*katG*	Isoniazid	3
761139	1.62E-60	*rpoB*	Rifampicin	All
761110	2.30E-60	*rpoB*	FQ	All

761110	1.05E-58	*rpoB*	Rifampicin	All
761155	3.06E-55	*rpoB*	AG	1
781687	5.21E-55	*rpsL*	Streptomycin	All
761139	3.37E-54	*rpoB*	Rifampicin	4
2155168	9.94E-54	*katG*	XDR-TB vs. PAN	All
2155168	2.09E-53	*katG*	XDR-TB vs. PAN	4
2155168	3.19E-53	*katG*	Isoniazid	2
1673425	5.05E-53	*Rv1482c-fabG1*	Isoniazid	All
1473246	1.16E-52	*rrs*	Kanamycin	4
761110	1.55E-51	*rpoB*	Rifampicin	4
2155168	3.86E-51	*katG*	FQ	3
781687	1.25E-50	*rpsL*	Streptomycin	2
2155168	5.75E-50	*katG*	MDR-TB vs. PAN	All
1673425	9.56E-50	*Rv1482c-fabG1*	Isoniazid	4
1673425	2.11E-49	*Rv1482c-fabG1*	MDR-TB vs. PAN	1
761155	4.81E-48	*rpoB*	MDR-TB vs. PAN	4
761155	7.37E-48	*rpoB*	Rifampicin	3
2155168	2.52E-47	*katG*	MDR-TB vs. PAN	3
2155168	5.66E-46	*katG*	MDR-TB vs. PAN	4
761155	3.01E-45	*rpoB*	MDR-TB vs. PAN	All
7582	1.86E-43	*gyrA*	Ciprofloxacin	All
7582	2.39E-42	*gyrA*	Ofloxacin	4
4247429	5.23E-42	*embB*	Ethambutol	All
7582	6.54E-42	*gyrA*	Ofloxacin	All
760314	1.87E-40	*rpoB*	Rifampicin	4
761139	5.33E-40	*rpoB*	Rifampicin	1
1472362	8.46E-40	*rrs*	Streptomycin	1
761110	1.80E-37	*rpoB*	MDR-TB vs. PAN	All
1673425	4.65E-37	*Rv1482c-fabG1*	Isoniazid	3
4247429	5.65E-37	*embB*	AG	1
761155	1.11E-36	*rpoB*	XDR-TB vs. PAN	2
1473246	6.28E-36	*rrs*	Pyrazinamide	3
761140	1.07E-34	*rpoB*	Rifampicin	All
763123	1.18E-34	*rpoB*	XDR-TB vs. MDR-TB	4
761155	1.25E-34	*rpoB*	FQ	2
4247429	5.17E-34	*embB*	Ethambutol	3
4247429	7.48E-34	*embB*	Ethambutol	4
761110	1.56E-33	*rpoB*	MDR-TB vs. PAN	4
4247431	3.75E-33	*embB*	Ethambutol	3
761155	4.01E-33	*rpoB*	MDR-TB vs. PAN	1
760314	2.44E-32	*rpoB*	MDR-TB vs. PAN	4
761139	1.04E-31	*rpoB*	MDR-TB vs. PAN	All

781822	2.41E-31	*rpsL*	Streptomycin	All
761139	3.60E-31	*rpoB*	MDR-TB vs. PAN	4
7570	9.79E-31	*gyrA*	Ofloxacin	All
1473246	1.52E-30	*rrs*	AG	3
1473246	1.52E-30	*rrs*	FQ	3
1472362	2.07E-30	*rrs*	Streptomycin	All
761110	2.93E-30	*rpoB*	FQ	4
4269271	4.08E-30	*ubiA*	XDR-TB vs. MDR-TB	4
761109	1.12E-29	*rpoB*	XDR-TB vs. PAN	All
761155	2.44E-29	*rpoB*	Rifampicin	2
761277	4.94E-29	*rpoB*	FQ	All
1473246	4.99E-29	*rrs*	XDR-TB vs. PAN	All
7582	8.16E-29	*gyrA*	XDR-TB vs. PAN	3
761140	3.12E-28	*rpoB*	Rifampicin	3
1473246	4.73E-28	*rrs*	XDR-TB vs. MDR-TB	3
761155	6.12E-28	*rpoB*	MDR-TB vs. PAN	3
2155168	9.08E-28	*katG*	MDR-TB vs. PAN	2
4249583	1.62E-27	*embB*	XDR-TB vs. PAN	3
1472362	1.72E-27	*rrs*	AG	All
2289213	1.91E-27	*pncA*	Pyrazinamide	4
761139	5.96E-27	*rpoB*	Rifampicin	2
2289213	7.81E-27	*pncA*	Pyrazinamide	All
761155	2.32E-26	*rpoB*	AG	4
763123	2.32E-26	*rpoB*	XDR-TB vs. MDR-TB	All
761155	3.24E-26	*rpoB*	AG	All
761110	3.63E-26	*rpoB*	Rifampicin	3
6750	5.62E-26	*gyrB*	XDR-TB vs. MDR-TB	4
1674481	7.65E-26	*inhA*	XDR-TB vs. PAN	4
7570	8.49E-26	*gyrA*	Ofloxacin	4
7582	8.61E-26	*gyrA*	Moxifloxacin	All
761109	1.06E-25	*rpoB*	FQ	3
761155	1.41E-25	*rpoB*	Ethambutol	1
2155168	1.55E-25	*katG*	XDR-TB vs. PAN	2
1674481	2.70E-25	*inhA*	XDR-TB vs. PAN	All
781687	3.09E-25	*rpsL*	AG	2
1472359	7.24E-25	*rrs*	AG	4
761140	1.02E-24	*rpoB*	MDR-TB vs. PAN	All
1472359	2.82E-24	*rrs*	AG	1
4247429	4.37E-24	*embB*	XDR-TB vs. PAN	All
781822	5.29E-24	*rpsL*	Streptomycin	2
4247431	7.44E-24	*embB*	Ethambutol	All
761109	9.52E-24	*rpoB*	XDR-TB vs. PAN	3

7582	1.21E-23	*gyrA*	Moxifloxacin	4
761161	1.81E-23	*rpoB*	Rifampicin	All
7581	7.49E-23	*gyrA*	Ofloxacin	All
1674782	1.75E-22	*inhA*	AG	1
781687	3.28E-22	*rpsL*	Streptomycin	4
761161	8.10E-22	*rpoB*	MDR-TB vs. PAN	All
761161	1.00E-21	*rpoB*	MDR-TB vs. PAN	4

AG = Aminoglycosides; MDR-TB =multi-drug resistant TB; FQ = Fluoroquinolones; PAN = Pan susceptible; XDR-TB = extensively drug resistant TB

Supplementary Table E7- Phyc hits involving known resistance loci ($P<10^{-5}$)

Lineage	Phenotype	Position	Fisher Test P-value	Locus
1	Isoniazid	2155168	5.63E-62	*katG*
1	Isoniazid	1673425	1.50E-40	*Rv1482c-fabG1*
1	Rifampicin	761155	2.97E-33	*rpoB*
All	Rifampicin	761155	4.85E-30	*rpoB*
1	MDR-TB vs. PAN	761155	6.46E-30	*rpoB*
All	MDR-TB vs. PAN	761155	7.61E-30	*rpoB*
1	MDR-TB vs. PAN	2155168	1.62E-29	*katG*
1	MDR-TB vs. PAN	1673425	1.05E-23	*Rv1482c-fabG1*
All	MDR-TB vs. PAN	781687	2.08E-22	*rpsL*
All	Rifampicin	4247431	2.19E-21	*embB*
All	Rifampicin	781687	1.93E-20	*rpsL*
All	Rifampicin	4247429	2.06E-19	*embB*
1	Isoniazid	761155	2.41E-19	*rpoB*
1	Ethambutol	4247429	8.14E-19	*embB*
All	MDR-TB vs. PAN	4247431	1.67E-17	*embB*
All	MDR-TB vs. PAN	1673425	3.11E-17	*Rv1482c-fabG1*
1	MDR-TB vs. PAN	4247429	3.44E-17	*embB*
All	MDR-TB vs. PAN	4247429	3.94E-17	*embB*
All	MDR-TB vs. PAN	761139	1.17E-16	*rpoB*
All	Ethambutol	4247429	3.90E-16	*embB*
1	Rifampicin	761139	5.69E-16	*rpoB*
All	Rifampicin	761139	8.64E-16	*rpoB*
All	MDR-TB vs. PAN	2155168	1.34E-15	*katG*
All	Rifampicin	7582	3.45E-15	*gyrA*
1	Rifampicin	2155168	4.59E-15	*katG*
1	AG	2155168	6.68E-15	*katG*
All	Ethambutol	761155	7.57E-15	*rpoB*
1	Ethambutol	761155	7.58E-15	*rpoB*
1	Rifampicin	4247429	7.69E-15	*embB*
1	MDR-TB vs. PAN	761139	7.70E-15	*rpoB*
All	MDR-TB vs. PAN	7582	1.21E-14	*gyrA*
1	Rifampicin	1673425	1.95E-13	*Rv1482c-fabG1*
All	AG	781687	2.91E-13	*rpsL*
1	Isoniazid	4247429	3.25E-12	*embB*
All	Rifampicin	2155168	5.12E-12	*katG*
All	Ethambutol	781687	7.64E-12	*rpsL*
All	AG	761155	8.00E-12	*rpoB*

All	Rifampicin	1673425	1.48E-11	*Rv1482c-fabG1*
All	Pyrazinamide	781687	2.05E-11	*rpsL*
All	Rifampicin	1473246	3.27E-11	*rrs*
All	Streptomycin	4247429	1.78E-10	*embB*
1	MDR-TB vs. PAN	761140	3.89E-10	*rpoB*
1	Isoniazid	761139	7.78E-10	*rpoB*
1	AG	1472362	1.11E-09	*rrs*
All	Streptomycin	781687	1.13E-09	*rpsL*
All	MDR-TB vs. PAN	1473246	1.13E-09	*rrs*
All	Ethambutol	4247431	1.17E-09	*embB*
1	AG	761155	1.22E-09	*rpoB*
All	AG	2155168	1.98E-09	*katG*
All	Rifampicin	7570	2.20E-09	*gyrA*
All	Ethambutol	2155168	2.82E-09	*katG*
1	Ethambutol	2155168	2.90E-09	*katG*
All	Pyrazinamide	761155	3.31E-09	*rpoB*
1	Isoniazid	761140	3.40E-09	*rpoB*
1	Isoniazid	1472362	3.40E-09	*rrs*
1	Rifampicin	761140	3.51E-09	*rpoB*
All	FQ	1473246	4.51E-09	*rrs*
All	AG	4247429	6.23E-09	*embB*
All	Ethambutol	761139	9.24E-09	*rpoB*
1	Streptomycin	1472362	2.28E-08	*rrs*
All	AG	7570	2.53E-08	*gyrA*
All	Rifampicin	4247730	3.00E-08	*embB*
All	AG	1473246	4.05E-08	*rrs*
All	Streptomycin	7582	5.64E-08	*gyrA*
All	Streptomycin	761155	6.03E-08	*rpoB*
All	FQ	7570	7.40E-08	*gyrA*
All	Streptomycin	2155168	8.62E-08	*katG*
1	MDR-TB vs. PAN	1472362	8.75E-08	*rrs*
4	MDR-TB vs. PAN	761155	9.17E-08	*rpoB*
1	Ethambutol	4243217	1.03E-07	*embC-embA*
1	Ethambutol	761139	1.10E-07	*rpoB*
1	FQ	2155168	1.14E-07	*katG*
All	Rifampicin	761140	1.18E-07	*rpoB*
All	MDR-TB vs. PAN	7570	1.32E-07	*gyrA*
All	Rifampicin	2715342	1.63E-07	*eis-Rv2417c*
1	FQ	1473246	1.67E-07	*rrs*
4	Rifampicin	761155	1.72E-07	*rpoB*
All	Amikacin	1473246	2.45E-07	*rrs*
All	Ethambutol	1473246	2.57E-07	*rrs*

All	Amikacin	781687	2.77E-07	*rpsL*
All	Pyrazinamide	1473246	3.04E-07	*rrs*
All	AG	761139	3.16E-07	*rpoB*
All	Pyrazinamide	4247429	4.31E-07	*embB*
All	Ofloxacin	7570	4.91E-07	*gyrA*
1	XDR-TB vs. PAN	761155	5.05E-07	*rpoB*
All	MDR-TB vs. PAN	2715342	5.20E-07	*eis-Rv2417c*
All	Rifampicin	1472359	5.49E-07	*rrs*
1	MDR-TB vs. PAN	4247431	5.78E-07	*embB*
All	Ethambutol	7570	5.82E-07	*gyrA*
1	FQ	4247429	6.65E-07	*embB*
4	Rifampicin	4247431	7.37E-07	*embB*
All	Streptomycin	761139	7.93E-07	*rpoB*
All	Streptomycin	7570	8.00E-07	*gyrA*
All	XDR-TB vs. PAN	1473246	9.43E-07	*rrs*
All	MDR-TB vs. PAN	4247730	1.08E-06	*embB*
All	ethambutol	7582	1.25E-06	*gyrA*
1	MDR-TB vs. PAN	4247730	1.31E-06	*embB*
1	MDR-TB vs. PAN	4243217	1.31E-06	*embC-embA*
1	MDR-TB vs. PAN	1472359	1.31E-06	*rrs*
1	Rifampicin	4247730	1.37E-06	*embB*
1	Rifampicin	4243217	1.37E-06	*embC-embA*
All	Ethambutol	761140	1.42E-06	*rpoB*
1	FQ	761155	1.66E-06	*rpoB*
All	Rifampicin	761161	1.72E-06	*rpoB*
All	MDR-TB vs. PAN	761140	2.00E-06	*rpoB*
All	MDR-TB vs. PAN	1472359	2.00E-06	*rrs*
All	AG	7582	2.27E-06	*gyrA*
1	Rifampicin	4247431	2.28E-06	*embB*
1	Ethambutol	4269293	2.57E-06	*ubiA*
1	Ethambutol	1673425	3.23E-06	*Rv1482c-fabG1*
All	FQ	2155168	3.92E-06	*katG*
All	Rifampicin	1673432	3.98E-06	*Rv1482c-fabG1*
All	MDR-TB vs. PAN	1673432	4.23E-06	*Rv1482c-fabG1*
1	AG	1673425	4.33E-06	*Rv1482c-fabG1*
All	AG	1673425	4.57E-06	*Rv1482c-fabG1*
All	MDR-TB vs. PAN	1674263	4.65E-06	*inhA*
1	Ethambutol	761140	5.22E-06	*rpoB*
1	MDR-TB vs. PAN	4248003	7.44E-06	*embB*
1	Pyrazinamide	761155	7.44E-06	*rpoB*
1	Rifampicin	4248003	7.73E-06	*embB*
1	Rifampicin	1472359	7.73E-06	*rrs*

1	Streptomycin	781822	8.03E-06	*rpsL*
1	Isoniazid	4247431	8.26E-06	*embB*

AG = Aminoglycosides; MDR-TB =multi-drug resistant TB; FQ = Fluoroquinolones; PAN = Pan susceptible

Supplementary Table E8- Novel non-synonymous PhyC hits ($P<10^{-5}$); for all phenotypes and lineages; specific and combined

Gene	SNP Position	Phenotype	Min. Fisher test P	Lineage	Function
Rv0336	401318	Rifampicin, MDR-TB vs. PAN, AG, Ethambutol, Pyrazinamide, FQ, Streptomycin, XDR-TB vs. PAN, Amikacin, Capreomycin, Moxifloxacin	1.50E-51	All	Conserved 13E12 repeat family protein
Rv1765c	1998063	Rifampicin, MDR-TB vs. PAN, Ethambutol, Amikacin, Pyrazinamide, Capreomycin, Streptomycin, AG, Rifabutin	4.25E-19	All	Uncharacterized protein
Rv1765c	1998063	MDR-TB vs. PAN	2.91E-16	All	Uncharacterized protein
Rv3611	4053161	Rifampicin, MDR-TB vs. PAN, AG, FQ	2.71E-15	All	Hypothetical unknown arg-, pro-rich protein.
Rv0797	891220	MDR-TB vs. PAN, Rifampicin, Ethambutol, AG	9.33E-13	All	Putative transposase for IS1547
Rv1150	1278278	Rifampicin, MDR-TB vs. PAN, Rifabutin, Pyrazinamide, Ethambutol, AG, Amikacin	1.10E-12	All	Possible fragment of transposase (pseudogene).
Rv2015c	2262620	Amikacin, Rifampicin, Capreomycin, MDR-TB vs. PAN, Ethambutol	4.00E-10	All	Uncharacterized protein

Rv0797	890549	Rifampicin, MDRvPAN	1.10E-09	All	Putative transposase for IS1547
Rv3611	4053050	Rifampicin, MDR-TB vs. PAN, Streptomycin	1.54E-09	All	Hypothetical unknown arg-, pro-rich protein.
Rv1588c	1789446	Ciprofloxacin	1.68E-08	All	Uncharacterized protein
pks12	2295685	Rifampicin, MDR-TB vs. PAN, Ethambutol	2.54E-07	All	Polyketide synthase
Rv0515	607677	MDR-TB vs. PAN, AG, Streptomycin	1.08E-06	All	Conserved 13E12 repeat family protein
Rv0094c	103756	AG	2.20E-06	All	Member of 13E12 repeat family
Rv2186c	2447616	Ethambutol	2.57E-06	1	Uncharacterized protein
Rv2512c	2829656	Pyrazinamide	3.69E-06	All	Transposase for insertion sequence element IS1081
Rv1042c	1165114	MDR-TB vs. PAN, Rifampicin	5.19E-06	All	Probable is like-2 transposase
Rv3115	3482432	Ethionamide	6.16E-06	All	Probable transposase
Rv3193c	3560945	XDR-TB vs. PAN	8.80E-06	All	Probable conserved transmembrane protein

AG = Aminoglycosides; MDR-TB =multi-drug resistant TB; FQ = Fluoroquinolones; PAN =

Pan susceptible; XDR-TB = extensively drug resistant TB

Supplementary Table E9- novel hits with their STRING associations

Locus	Rv	Analysis	STRING Associations
recF	*Rv0003*	variant	*dnaA dnaN gyrA* gyrB* recA recO recR recX rnpA Rv0004*
Rv0197	*Rv0197*	variant	*lppM narX Rv0196 Rv0218 Rv0236.1 Rv2172c Rv2813 Rv3403c Rv3549c Rv3777*
iniA	*Rv0342*	cluster variant	*iniB iniC Rv0049 Rv0110 Rv0312 Rv0339c Rv0340 Rv2264c Rv3863 whiB6*
pks6	*Rv0405*	variant	*fadD30 fas mbtA mbtB mbtD mbtE mbtH nrp pptT tesA*
Rv0571c-Rv0572c	*Rv0571c*	locus; variant; cluster variant; cluster locus	*rip3 ctpF hspX Rv0572c rtcB hrp1 Rv0569 Rv1734c Rv0574c vapB32*
Rv0571c-Rv0572c	*Rv0572c*	locus; variant; cluster variant; cluster locus	*Rv1734c Rv3127 Rv0571c Rv3126c rip3 Rv3129 rtcB Rv1812c Rv2003c vapB22*
Rv0530A-Rv0531	*Rv0530A*	locus; variant; cluster variant; cluster locus	*Rv0048c Rv0157A Rv0381c Rv0531 Rv0680c Rv0686 Rv2799 Rv3493c Rv3572 TB22.2*
Rv0530A-Rv0531	*Rv0531*	locus; variant; cluster variant; cluster locus	*aftA aftC Rv0466 Rv0955 Rv1476 Rv1610 Rv3035 Rv3635 Rv3668c Rv3802c*
mce2B	*Rv0590*	cluster variant	*lprL mce2A mce2C mce2D mce2F Rv0590A yrbE2A yrbE2B yrbE4A yrbE4B*
atsD	*Rv0663*	cluster variant	*atsA atsB rnz Rv0296c Rv0712 Rv2000 Rv3077 Rv3406 Rv3762c Rv3796*
Rv1061	*Rv1061*	cluster variant	*fadD14 Rv0426c Rv1059 Rv1060 Rv1062*
Rv1144-	*Rv1144*	cluster	*bioF1 echA10 hycD mcr mmpL13a purH Rv2635*

mmpL13a		variant	*Rv2636 Rv3836 Rv3837c*
Rv1144-mmpL13a	*Rv1145*	cluster variant	*efpA mmpL1 mmpL13a mmr Rv0590A Rv0849 Rv1144 Rv1877 Rv2000 Rv3836*
Rv1373	*Rv1373*	variant	*PE_PGRS2 PE_PGRS23 PE_PGRS48 pks18 PPE39 PPE40 PPE53 PPE64 Rv1371 Rv1676*
Rv1616	*Rv1616*	variant	*crcB1 pykA Rv1517 Rv1610 Rv1615Rv1619 Rv1861 Rv2203 Rv2433c tesB1*
cydB-cydA	*Rv1622c*	locus	*atpH ctaC ctaE cydA cydC cydD narG qcrA qcrB qcrC*
cydB-cydA	*Rv1623c*	locus	*atpA ctaC cydA cydB cydC cydD qcrA qcrB qcrC Rv1488 Rv1579c*
Rv1924c-fadD31	*Rv1924c*	locus	*fadD31*
Rv1924c-fadD31	*Rv1925*	locus	*fadD22 fadD9 mbtE mbtF nrp pks13 Rv0068 Rv0149 Rv0149 Rv1924c tesA*
Rv2000	*Rv2000*	variant; cluster variant	*atsD ctpB ephA fabG3 fhaA mmpL13a Rv0493c Rv2001 ubiA* yidC*
aroG	*Rv2178c*	locus	*aroB aroD aroE aroK pheA pheA Rv0948c Rv1885c Rv2179c Rv2180c trpA*
Rv2425c	*Rv2425c*	cluster variant	*kgd mobA oxyR proA Rv0370c Rv2424c Rv2426c Rv2456c*
Rv2499c	*Rv2499c*	cluster variant	*accA1 accD1 citE fadD35 fadE19 Rv0575c Rv2506 scoA scoB yrbE3B*
secD	*Rv2587c*	cluster variant	*ffh ftsY Rv2585c secA1 secE1 secF secG secY yajC yidC*
Rv3235	*Rv3235*	locus	*fadD16 Rv0428c Rv1045 Rv1125 Rv1682 Rv2712c Rv3231c Rv3231c Rv3289c Rv3415c tgs3*
Rv3238c-Rv3239c	*Rv3238c*	locus	*cbs cysA3 metB metC metH metZ mmuM sahH sseA sseB*
Rv3238c-Rv3239c	*Rv3239c*	locus	*gltB lipJ ribG Rv0104 Rv1364c Rv1937 Rv2209 Rv2434c Rv3238c secA1*
folD-relJ	*Rv3356c*	locus	*fmt gcvT glyA1 glyA2 guaA purH purN purU thyA* thyX**
folD-relJ	*Rv3357*	locus	*ccdA higB3 relB relE relF relG relK Rv3359 vapB2 vapC2*
Rv3471c	*Rv3471c*	cluster variant	*gmdA ilvB2 mhpE mrsA Rv0521 Rv1508A Rv2305 Rv3468c Rv3472 udgA*
Rv3554-Rv3555c	*Rv3554*	locus	*echA20 fadB3 fadH fas fdxD fprA ompA Rv3551 Rv3552 Rv3553*
Rv3554-Rv3555c	*Rv3555c*	locus	*fadA6 Rv0083 Rv0336 Rv0393 Rv0515 Rv0756c Rv1765c Rv2100 Rv2642 Rv3776*
espA-	*Rv3616c*	variant;	*eccCa1 eccCb1 espB espB espC espD espR esxA*

ephA		cluster variant	*esxB phoP Rv3612c*
espA-ephA	*Rv3617*	variant; cluster variant	*echA1 ephG espA folE lpqG Rv1056 Rv2000 Rv3612c Rv3613c Rv3618*
ligC-Rv3732	*Rv3731*	locus	*dnaN ligA mku polA recA recC recG Rv0269c Rv2090 Rv3730c*
ligC-Rv3732	*Rv3732*	locus	*csm4 csm6 ligC lppH lprF mmpS1 PPE46 Rv1682 Rv1754c Rv3096*
Rv3755c-proZ	*Rv3755c*	locus	*mmpR5 proV proW proX proZ Rv0513 Rv2739c Rv3479 Rv3489 Rv3760*
Rv3755c-proZ	*Rv3756c*	locus	*cobL hsaF lpqZ proV proW proX Rv0191 Rv1667c Rv3008 Rv3755c*

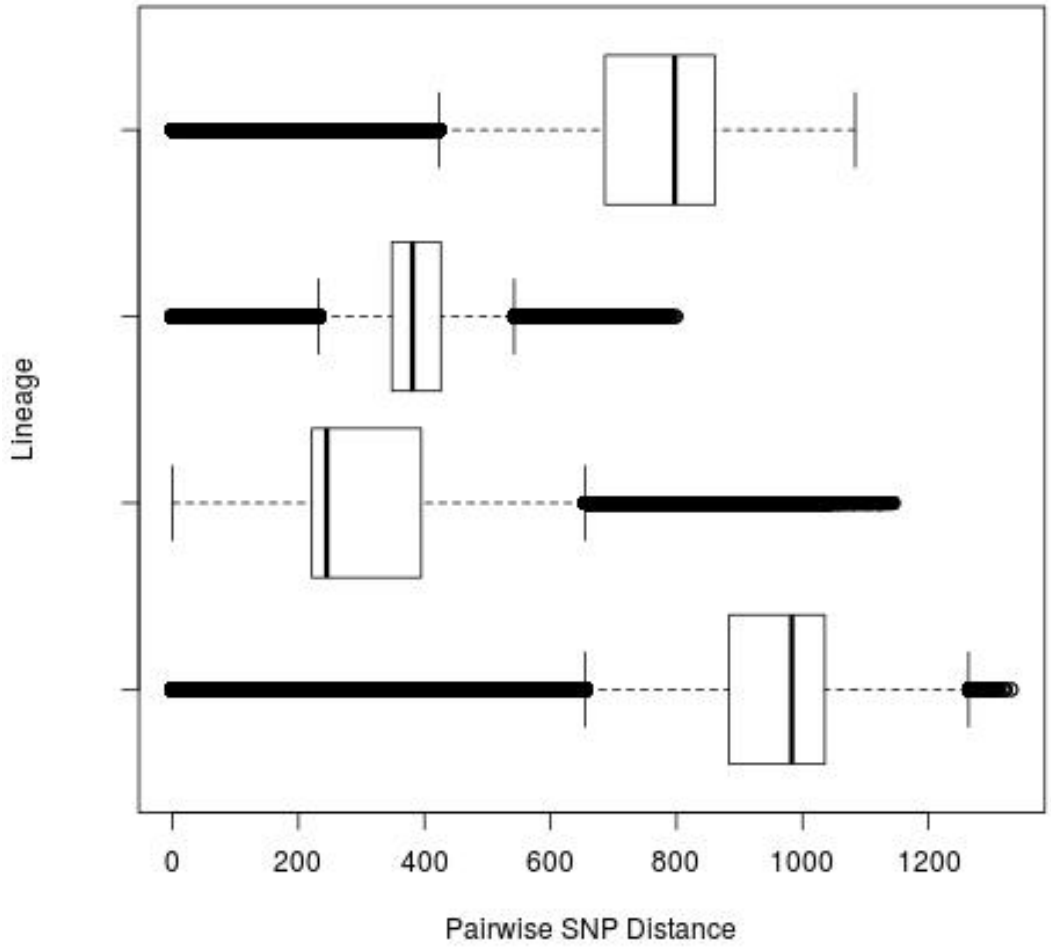
* Loci known to be involved in resistance

Supplementary Table E10- SNP-based cluster GWAS comparing the transmission clusters groups versus others

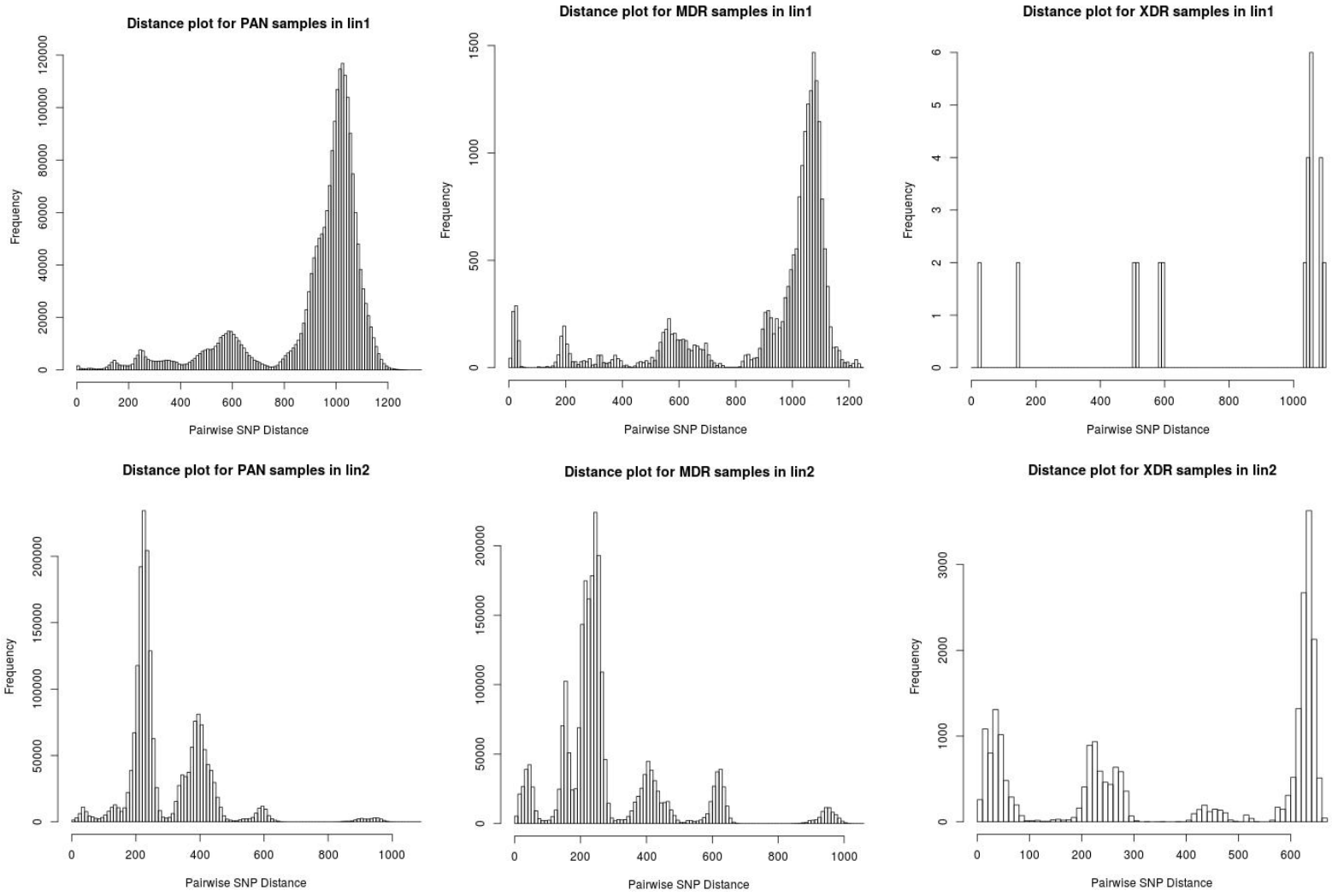
Position	P-value	Locus	Links to candidates*
2246032	2.09e-310	*Rv2000*	*ubiA*
763123	3.08E-284	*rpoB*	
664929	1.33E-201	*Rv0571c-Rv0572c****	
4056430	4.46E-173	*espA-ephA*	*Rv2000*
4269271	1.90E-129	*ubiA*	*Rv2000*
1184080	8.57E-95	*Rv1061*	
688228	2.32E-81	*mce2B*	
410962	1.88E-30	*iniA*	
2722670	1.88E-30	*Rv2425c*	
1272321	1.07E-27	*Rv1144-mmpL13a*	
3889150	1.07E-27	*Rv3471c*	
763123**	6.78E-27	*rpoB*	
2246032**	6.78E-27	*Rv2000*	*ubiA*
756757	1.31E-24	*atsD*	*Rv2000*
2915226	8.76E-24	*secD*	
2813575	2.29E-23	*Rv2499c*	
4056430**	4.87E-23	*espA-ephA*	*Rv2000*
664929**	9.50E-23	*Rv0571c-Rv0572c*	

* in STRING database; ** comparison group is the non-transmitted XDR-TB; *** also found in a locus based analysis (P=1.45E-34, vs. all; P=2.15E-21, vs. non-transmitted XDR-TB)

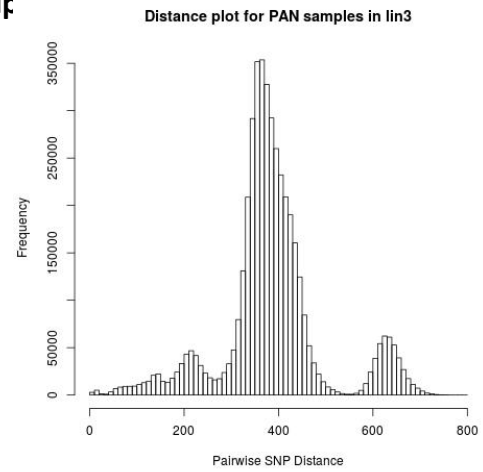
Supplementary Figure E1- Boxplot showing pairwise SNP distance summary by lineage

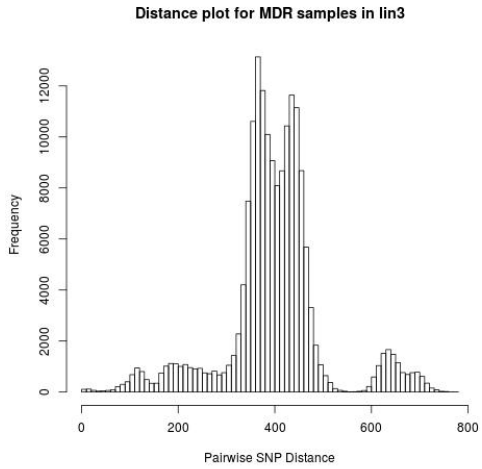


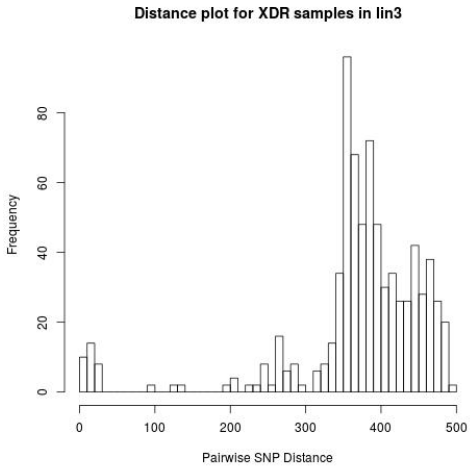
Supplementary Figure E2- within MDR distance plots, within XDR distance plots

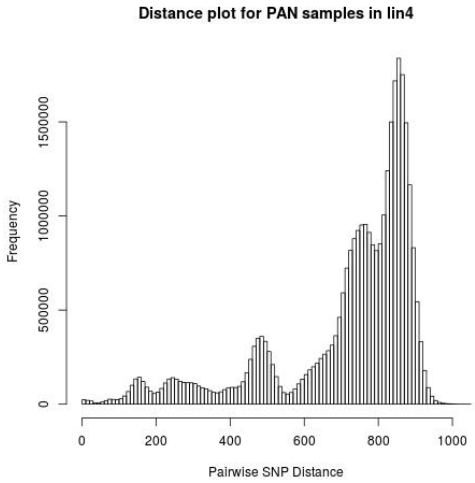


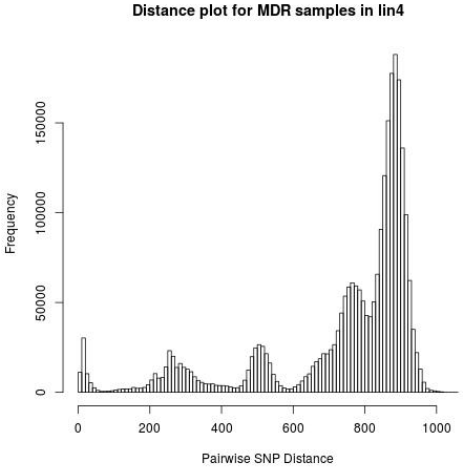
Sup

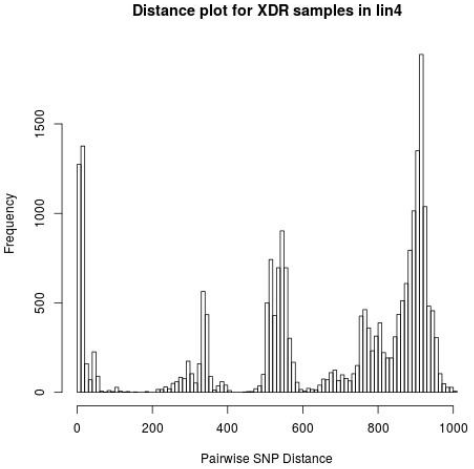

Distance plot for PAN samples in lin3
Frequency
0
50000
150000
250000
350000
0
200
400
600
800
Pairwise SNP Distance


Distance plot for MDR samples in lin3
Frequency
0
2000
4000
6000
8000
10000
12000
0
200
400
600
800
Pairwise SNP Distance

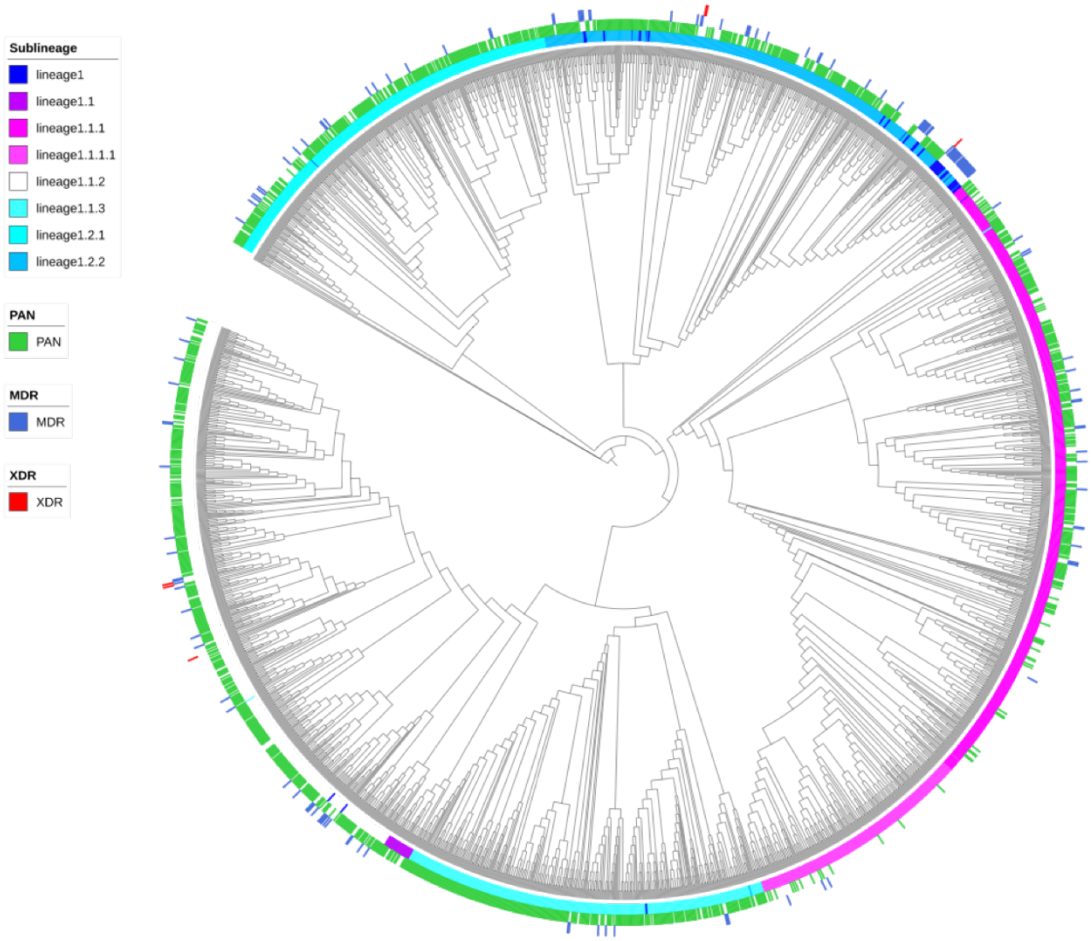

Distance plot for XDR samples in lin3
Frequency
0
20
40
60
80
0
100
200
300
400
500
Pairwise SNP Distance


Distance plot for PAN samples in lin4
Frequency
0
500000
1000000
1500000
0
200
400
600
800
1000
Pairwise SNP Distance

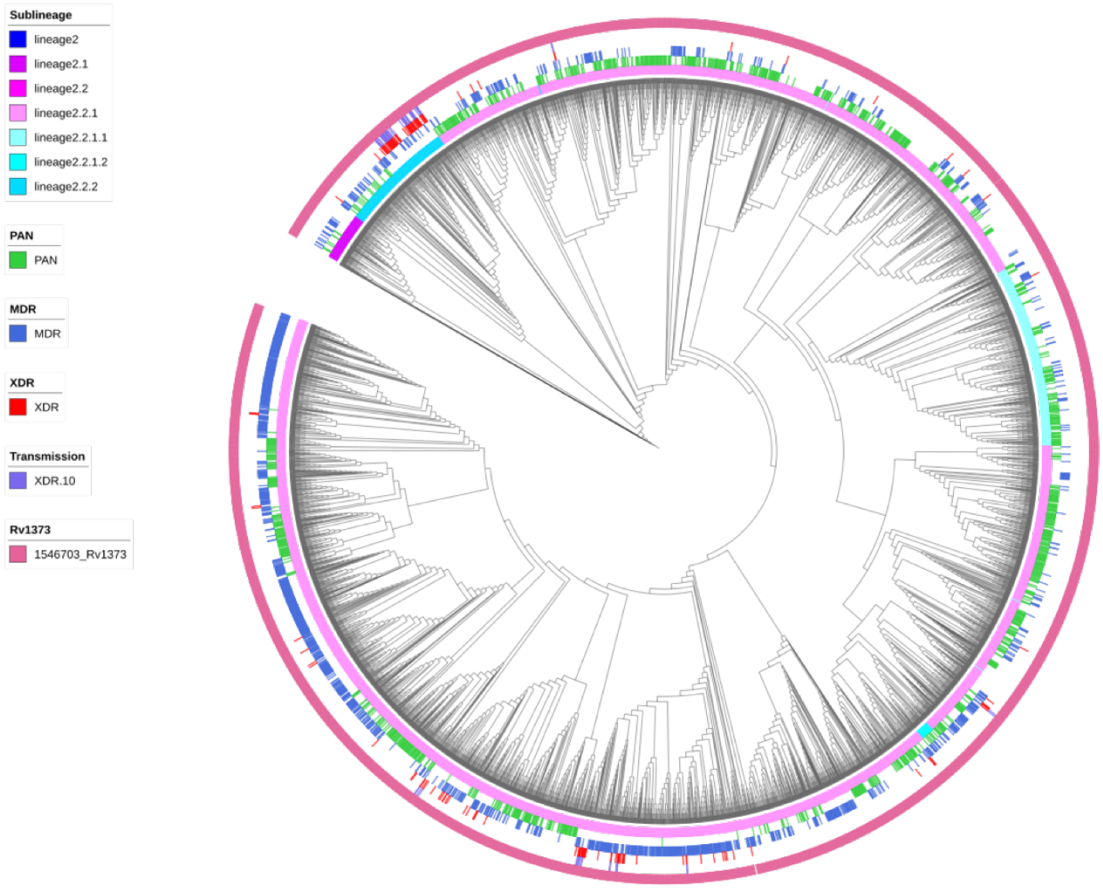

Distance plot for MDR samples in lin4
Frequency
0
50000
100000
150000
0
200
400
600
800
1000
Pairwise SNP Distance


Distance plot for XDR samples in lin4
Frequency
0
500
1000
1500
0
200
400
600
800
1000
Pairwise SNP Distance

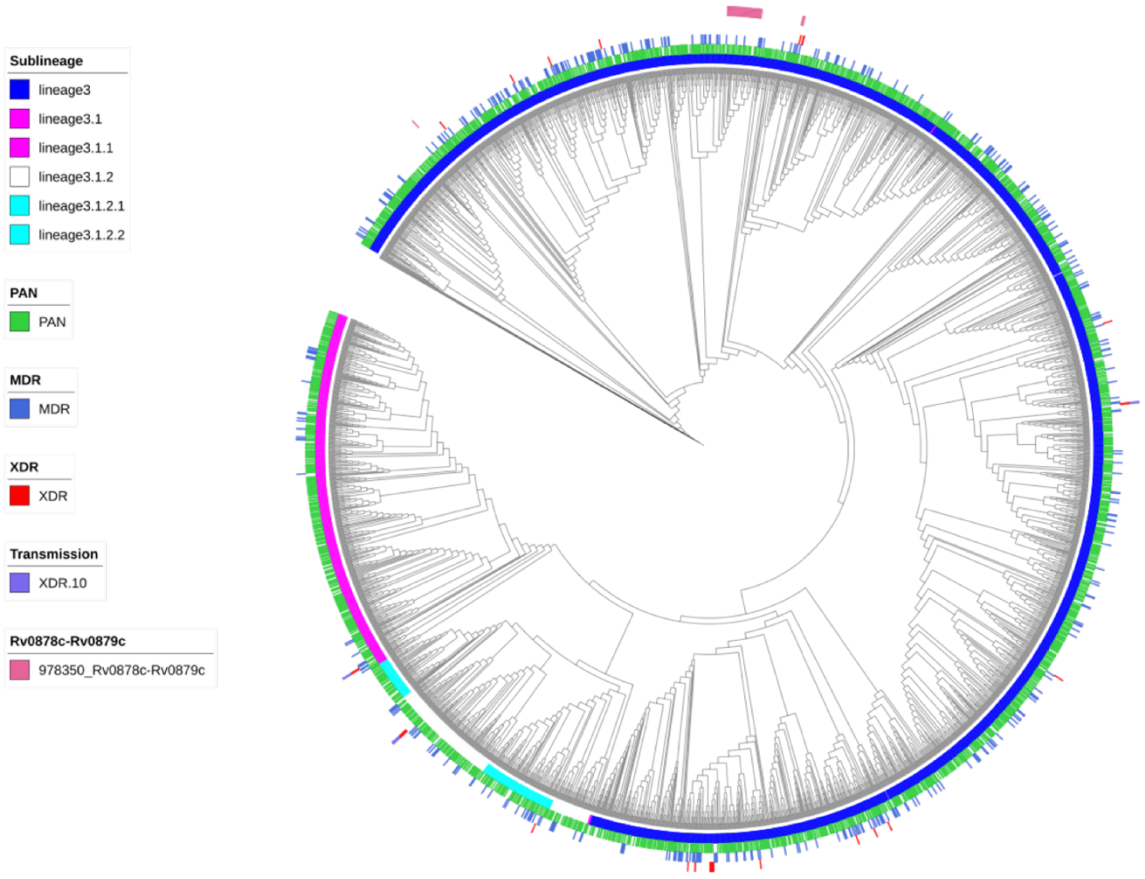
Supplementary Figure E3- lineage 1 maximum likelihood tree with novel associations



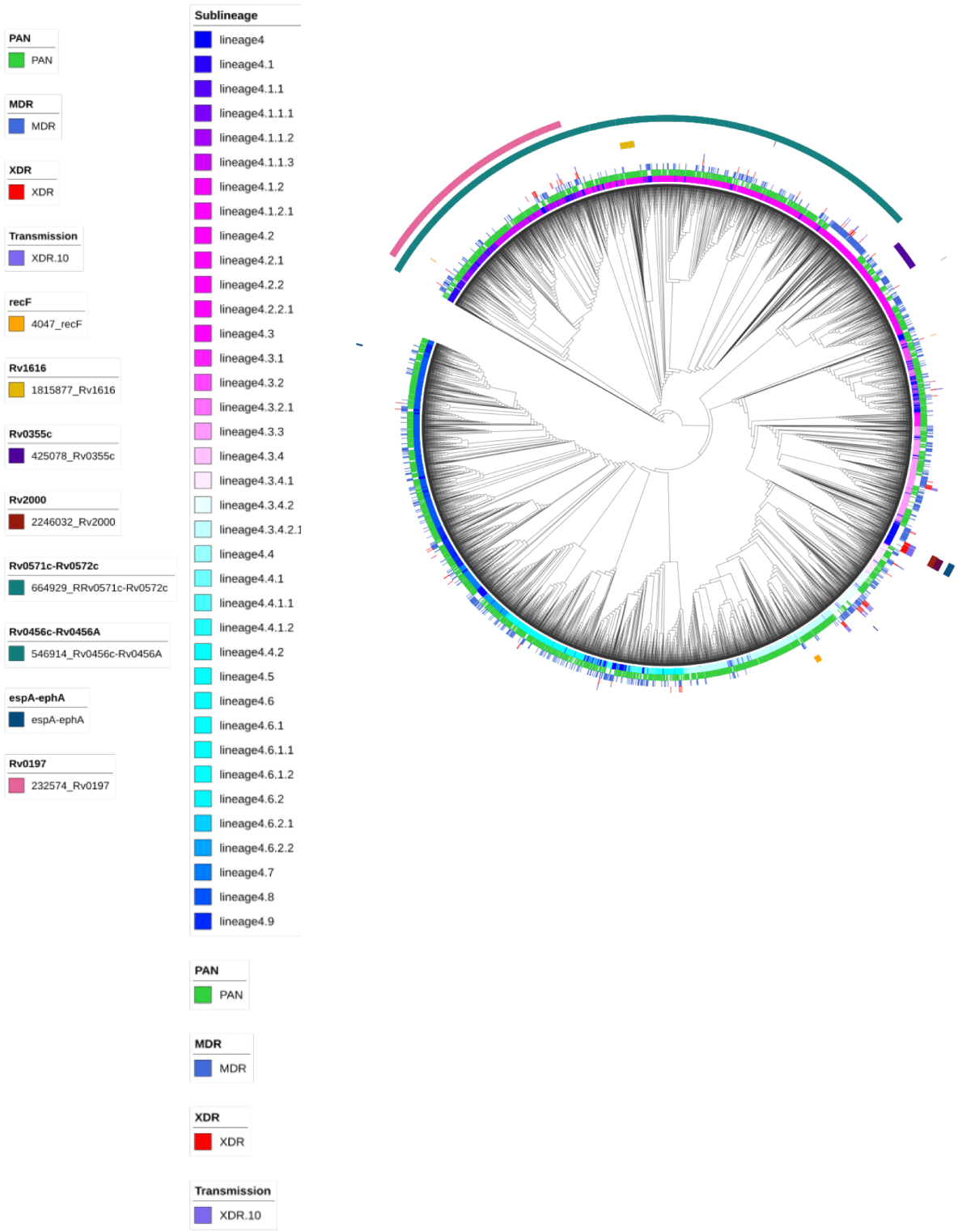
Supplementary Figure E4- lineage 2 maximum likelihood tree with novel associations



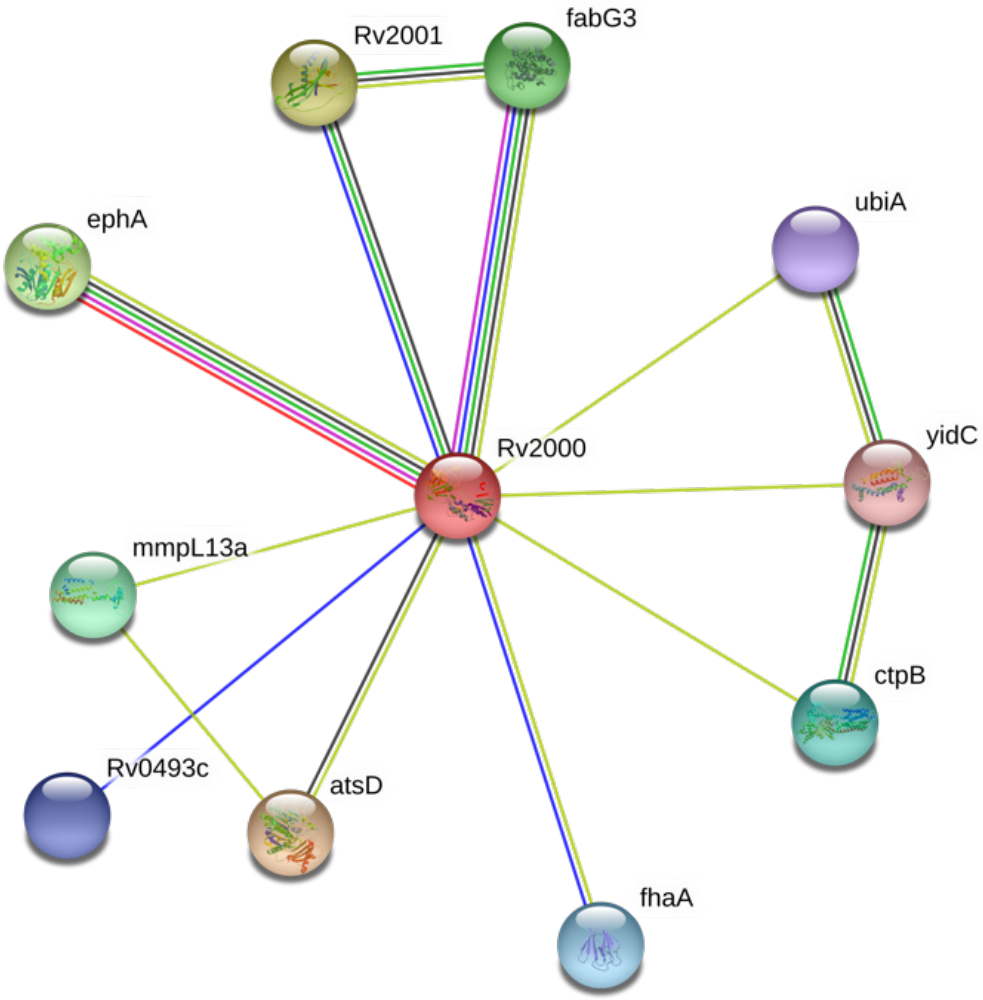
Supplementary Figure E5- lineage 3 maximum likelihood tree with novel associations



Supplementary Figure E6- Lineage 4 maximum likelihood tree with novel associations



Supplementary Figure E7- *Rv2000* string associations



Chapter 4:

Genome-wide machine learning classifier applied to *Mycobacterium tuberculosis* as a novel approach to unravel genomic complexity associated with drug resistance

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1500052	**Title**	Ms
First Name(s)	Yaa Emily Adoma		
Surname/Family Name	Oppong		
Thesis Title	Investigating the genomic basis of antimicrobial resistance in *Mycobacterium tuberculosis* (*Mtb*) using genome-wide methodologies		
Primary Supervisor	Martin Hibberd		

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

<u>SECTION B – Paper already published</u>

Where was the work published?	
When was the work published?	
If the work was published prior to registration for your research	

degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	

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

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Bioinformatics
Please list the paper's authors in the intended authorship order:	Yaa E A Oppong, Jody E Phelan, Martin L Hibberd, Taane G Clark
Stage of publication	Not yet submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	**YO conceived the project and performed bioinformatic and statistical analyses under the supervision of TC and MH, and wrote the first draft of the manuscript. JPh generated the sequencing dataset.**

SECTION E

Student Signature	
Date	

Supervisor Signature	
Date	

Genome-wide machine learning classifier applied to *Mycobacterium tuberculosis* as a novel approach to unravel genomic complexity associated with drug resistance

Yaa E A Oppong[1, §], Jody E Phelan[1], Martin L Hibberd[1,*], Taane G Clark[1,2,*]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

[2] Faculty of Epidemiology and Population Health, LSHTM, London, United Kingdom

* Joint authors

[§] Corresponding author

Yaa E A Oppong

Department of Infection Biology

Faculty of Infectious and Tropical Diseases

London School of Hygiene and Tropical Medicine, London, United Kingdom

yaa.oppong@lshtm.ac.uk

ABSTRACT

Motivation

The identification of genomic variants in whole genome sequencing data associated with *Mycobacterium tuberculosis* (*Mtb*) drug-resistance rely on genome-wide association study (GWAS) or convergent evolution approaches. However, these methods are potentially limited in their ability to detect multiple interacting loci contributing towards drug resistance. We investigate the use of a (machine) learning classifier system (LCS) to gain insights into within genome interactions, with the aim of developing an approach capable of disentangling complex drug resistance genomics, for use in predicting phenotypes and discovering novel loci involved in resistance.

Results

The LCS algorithm parameters and performance were benchmarked on the rifampicin anti-TB drug for which we know the underlying genes (e.g. *rpoB*) involved in resistance. We achieved a sensitivity of 93.7% and a specificity of 94.8% in predicting rifampicin resistance, and the model included *rpoB* and *rpoC,* known to be epistaticly involved in drug resistance. We then applied LCS to isoniazid resistance and multi-drug (rifampicin and isoniazid) resistance, achieving sensitivity of 86.8% and 93.5% and specificity of 94.2% and 97.6%, respectively. Additionally, we identified candidate loci for novel involvement in resistance such as *celA1.* Our work demonstrates the potential of machine learning approaches to disentangle complex genomic interactions and provide

mechanistic inferences in relation to drug resistance in *Mtb,* without the requirement for prior knowledge.

INTRODUCTION

The evolution of drug-resistant *Mycobacterium tuberculosis (Mtb)* is threatening the control of tuberculosis disease, with multi-drug resistant (MDR-TB) forms making up 82% of the 558,000 new cases in 2017 with resistance to rifampicin – the most effective first-line drug [1]. Large whole-genome sequence (WGS) datasets are increasingly becoming available for *Mtb,* coupled with drug resistance phenotype data, enabling the possibility of precision public health and diagnostics. This advance would require a deeper understanding of the genomic basis of drug resistance in *Mtb,* in all its complexity.

Drug resistance in *Mtb* can be due to mutations (including single nucleotide polymorphisms (SNPs), insertions and deletions (indels)) in genes coding for drug-targets or -converting enzymes. However, the full repertoire of resistance variants across the more than sixteen anti-TB drugs is unknown, and other mechanisms that are compensatory or related to efflux pumps have been described [2]. Current methods employed to gain insight into *Mtb* genomics and drug resistance include genome-wide association studies (GWAS) using linear mixed models [2,3] and convergent evolution analysis [4]. Recently, machine learning approaches such as random forest [5], support vector machines [6,7] and deep learning [8,9] have been applied and show promise. However, such methods are often limited in their ability to detect complex genetic interactions, known as epistasis, especially those involving more than two loci, in a mechanistically explainable way and often require prior knowledge.

Learning classifier systems (LCS) offer an alternative approach with scope to gain insight into epistatic interactions that contribute towards drug resistance, in a genome-wide fashion, without the use of prior knowledge. They work by constructing 'rules' involving combinations of variants that are predictive of a phenotype, through evolutionary optimization techniques. Unlike some other machine learning methods, LCS allows ease of interpretation of predictive loci and therefore underlying biological mechanisms. The availability of computationally efficient LCS implementation libraries allows for the approach to be applied on large datasets. Here, we apply LCS to an *Mtb* global dataset consisting of >18k isolates with whole genome sequences (>600k SNPs; >6000 genes/intergenic regions) and drug susceptibility testing-based phenotypes for first-line rifampicin and isoniazid drugs [10]. We optimize the parameters and assess the performance of LCS on rifampicin, which has known resistance gene (*rpoB*, and compensatory *rpoC*) mechanisms. We then apply the algorithms to establish mutation 'rules' for the prediction of isoniazid resistance and MDR-TB, and assess its feasibility as a tool for drug resistance phenotype prediction and discovery of drug resistance.

METHODS

Whole Genome Sequence Data and Processing

A global dataset of whole genome sequence data for 18,255 *Mtb* isolates from lineages 1, 2, 3 and 4 was collated, alongside phenotypic drug susceptibility testing data for rifampicin (n=16,296; 27.1% resistance), and isoniazid (n=16,211; 31.8% resistance), allowing MDR-TB (n=14,322; 25.9% resistant) to be determined (see **Supplementary Table 1**). 613,821 SNPs and their genotypes were called after aligning the sequence

data to the H37rv reference. SNP data was also collapsed by its gene or intergenic region, where a binary 'locus-type' was assigned; any loci containing one or more non-synonymous SNPs were coded as '1', any loci that did not contain one or more non-synonymous SNPs were coded as '0', producing a dataset of 6,879 loci.

Learning Classifier System

For all applications, data was randomly divided into a training set (80%) and test set (20%), and LCS was performed using ExSTraCS (version 2.0) [10] for 30,000,000 iterations. Two key parameters were varied in order to allow optimization: (i) maximum rule population size ("N"; 1000 to 5000, by 1000) and (ii) accuracy threshold applied to the subsumption mechanism ("*acc_sub"*; 0.90, 0.95, 0.99), which determines the minimum accuracy of a classifier in order for it to potentially subsume another classifer (one classifier subsumes another if it has greater or equal accuracy and is more general) [11]. The maximum rule population size should be large enough to fully explore the search space and allow useful classifiers to be kept in the rule population, whilst it not too large so as to unnecessarily increase the run time of the algorithm and keep classifiers in the population that are not useful [11]. Lower subsumption mechanism (sub_acc) values may perform better in noisier problem domains [11]. All other parameters were set as default [10], including the probability of a model crossover (default=0.8) and mutation (default=0.04) operation [11]. Within the final rule populations, the frequencies of individual loci, their co-occurring pairs and whole rules were quantified. Additionally, rule lengths (the number of loci in each rule) were inspected. Functions for each locus were taken from STRING [12] and

Mycobrowser [13]. Frequency refers to total number across the total rule population, macro frequency refers to the total number across unique rules. The final models were compared to the TB-Profiler mutation panel [14], and the predictive performance was estimated (and compared to TB-Profiler) using sensitivity and specificity metrics, assuming the drug susceptibility testing result as the gold standard.

RESULTS

Parameter Exploration and Application to Rifampicin

The predictive performance (sensitivity, specificity) of the LCS for rifampicin resistance was insensitive to the population sizes considered (see **Table 1**, **Supplementary Figure 1**). Similarly, the sub_acc parameter was tested, and the predictive performance was similar across the range of values but marginally higher for the default value of 0.99 (**Table 1, Supplementary Figure 2**). Therefore, the parameter settings for population size (N=2,000) and sub_acc (0.99) [11] (see **Table 1**) were applied in subsequent analyses.

The final LCS analysis for rifampicin contained 971 different rules, with the majority containing the *rpoB* gene (frequency = 850) (**Table 2, Figure 1**), a locus known to be causally involved in rifampicin resistance. The median number of loci in the rules was 4 (range: 1 to 14) (**Supplementary Table 2, Figure 1**). The most numerous rule in the population was *rpoB* by itself (frequency = 14), including the reference-and mutation-types to predict rifampicin susceptibility and resistance, respectively (**Supplementary Table 3**). *rpoB* was most commonly found in association with *katG* (n=182, isoniazid) (**Supplementary Table 4, Figure 2**), directly as a result of some patients having

additional resistance to isoniazid, and therefore MDR-TB. *rpoC* was also detected (frequency = 159, rifampicin), and is a known rifampicin resistance related compensatory locus (see **Figure 2**, **Supplementary Table 4**). Six other loci known to be involved in resistance to other drugs were identified: *pncA* (freq=588; pyrazinamide), *katG* (freq.= 478, isoniazid), *Rv1482c-fabG1* (freq.=461, isoniazid), *embB* (freq.=454, ethambutol), *rrs* (freq=314, streptomycin), and *rpsL* (freq.=132, streptomycin) (see **Table 2**). These were detected through co-occurring resistance mutations arising from TB patients receiving multiple drugs. Forty loci not previously established as being involved in drug resistance were identified, including *Rv2395* (frequency=305), *Rv2230c* (frequency=253), *pbpB* (frequency=233), *uspC* (frequency=225), *celA1* (frequency=209) and *guaB2* (frequency=196) (see **Table 2**). Non-reference *Rv2395, pbpB* and *uspC* and reference *guaB2* was found in rules predictive of rifampicin susceptibility. Whilst, reference *Rv2230c* and *celA1* was found in rules predictive of rifampicin resistance. The final model achieved a sensitivity in predicting rifampicin resistance of 93.7% and specificity of 95.8% (**Table 1**). This is comparable to current methods (see **Table 1**) using databases of known resistance variants [5–9,14].

Application to Isoniazid and MDR-TB

Applying the LCS to isoniazid resistance led to a final rule set consisting of 49 loci. The *Rv1482c-fabG1* (frequency = 836) operon, known to be related to isoniazid resistance, was the locus with the highest frequency in the rule population (**Supplementary Table**

5, Supplementary Table 6, Supplementary Figure 3). In total, 7 loci known to be involved in drug resistance were identified, many of which signal cross-resistance with other drugs, similar to the rifampicin analysis; *Rv1482c-fabG1* (frequency=836; isoniazid), *rpoB* (frequency=782; rifampicin), *rpsL* (frequency=667; streptomycin), *embB* (frequency=573; ethambutol), *katG* (frequency=467; isoniazid), *gid* (frequency=183; streptomycin) and *rrs* (frequency=110; strepromycin) (see **Supplementary Table 5**). Forty-two loci that have not previously been established as being involved in drug resistance were present in the rule population. The most frequent of which were *desA3-Rv3230c* (frequency=385), *cyp135B1-Rv0569* (frequency=278), *mmpS2* (frequency=235), *Rv3403c* (frequency=233) and *espF* (frequency=218) (see **Supplementary Table 5**). Reference form *desA3-Rv3230c* was found in rules predictive of isoniazid resistance. Whilst, reference *espF, cyp135B1-Rv0569, mmpS2* and *Rv3403c* in rules predictive of isoniazid susceptibility (see **Supplementary Table 6**). The most frequently co-occurring pair of loci was *rpoB* and *Rv1482c-fabG1* (frequency= 308) (see **Supplementary Table 7**). The most numerous rule was non-reference *rpoB* with non-reference *embB* in association with isoniazid resistance (see **Supplementary Table 6**). The median rule length was 4 loci (range: 1 to 13) (see **Supplementary Table 2**). The LCS achieved sensitivity of 86.8% and specificity of 94.2% in predicting isoniazid resistance (see **Table 1**).

Applying LCS to the MDR-TB compared to pan-susceptibility detected fifty-two different loci across 683 different rules. The most frequent individual locus was *rpoB* (frequency=488) (see **Supplementary Table 8, Supplementary Table 9, Supplementary**

Figure 4). Nine loci known to be involved in drug resistance were identified; *rpoB* (frequency=488; rifampicin), *embB* (frequency=434; ethambutol), *katG* (frequency=420; isoniazid), *pncA* (frequency=315; pyrazinamide), *rpsL* (frequency=309; streptomycin), *Rv1482c-fabG1* (frequency=288; isoniazid), *gid* (frequency=159; streptomycin), *inhA* (frequency=67, isoniazid) and *rpoA* (frequency=2; rifampicin) (see **Supplementary Table 8**). Forty-three loci that are not established drug resistance involved loci were identified. The most frequent of which were *PPE5* (frequency=197), *Rv3903c* (frequency=164), *Rv0075* (frequency=156), *Rv2512c* (frequency=153), *Rv3249c* (frequency=129) (see **Supplementary Table 8**). Non-reference *PPE5*, *Rv3903c, Rv0075, Rv3249c* and reference *Rv2512c* were found in rules predictive of pan-susceptibility (see **Supplementary Table 9**). The most frequently co-occurring pair of loci was *katG* and *embB* (frequency= 308) (see **Supplementary Table 10**). The most numerous rule was non-reference *rpoB* in prediction of MDR-TB (see **Supplementary Table 9**). The median rule length was 3 (range 1 to 10) (see **Supplementary Table 2**). The final model achieved prediction sensitivity of 93.5% and specificity of 97.6% (see **Table 1**).

DISCUSSION

We have applied LCS to a dataset of lineages 1, 2, 3 and 4 *Mtb* in prediction of rifampicin resistance, isoniazid resistance and MDR, achieving sensitivity and specificity comparable to current methods and discovering a number of candidate loci for novel involvement in these phenotypes.

One limitation of this method is that by collapsing loci into reference or non-reference on the basis of containing one or more non-synonymous mutation, within loci complexities may be masked, potentially reducing resolution. Nevertheless, application of LCS to this *Mtb* dataset has provided prediction accuracy for rifampicin resistance comparable to current prediction methods [14], without the need for any prior knowledge of resistance genomics. The simultaneous consideration of multiple loci, within rules, may be partially responsible for this predictive power. It is interesting to note we found shorter rule lengths for MDR than the other phenotypes, yet more individual loci overall. Perhaps, this points towards the high predictive power of multiple resistance loci across the rule population without the need to be present together in long rules. One situation in which longer rules might reduce predictive power is where genomes are highly variable, in terms of the combinations of loci in which resistance mutations are present; in other words, high genetic heterogeneity of MDR due to independent evolution.

Eleven loci known to be involved in drug resistance were identified across analyses by LCS, including *rpoB* which was the most frequent locus in the rule population in prediction of rifampicin resistance, and is known to be commonly responsible. Additionally, *rpoC* was identified by this analysis. It has been proposed that *rpoC* is in epistasis with *rpoB,* and it helps restore fitness when costly rifampicin-resistance conferring *rpoB* mutations are present [15,16].

Additionally, 125 loci were identified as novel candidates for involvement in resistance, and thus interesting subjects for further mechanistic exploration. One such locus is *sigA*

(*rpoD*), an RNA polymerase sigma factor [17]. It was identified in the rifampicin resistance analysis, always in reference form and always in prediction of rifampicin susceptibility. It is potentially an alternative mechanism directly conferring rifampicin resistance, or compensatory mutation in a similar fashion to *rpoC*. *celA1* was identified as predictive of rifampicin resistance when in reference form. *celA1* is a cellulase that has been shown to cause the breakdown of biofilms in *M. smegmatis*, where biofilms have been shown to be induced by exposure to rifampicin in *M. smegmatis* [18]. This observation might allude to a role of biofilm formation in drug resistance in *Mtb*. If reference *celA1* is more effective at biofilm breakdown than non-reference, the association between *celA1* and *rpoB* might suggest that biofilms are not required to provide rifampicin resistance when *rpoB* is present. Conversely, if reference *celA1* is less effective than non-reference, it could suggest that biofilm formation is a useful step in the evolution of *rpoB* in response to treatment with rifampicin.

pbpB is also an interesting candidate for future work. It has been shown to form part of a ternary septation complex involved in septum synthesis. In *M. smegmatis,* it is upregulated in starved cells as they transition into a non-replicative state [19], and thus may play a role in phenotypic drug resistance. Interestingly, the isocitrate lyase *aceA* was also shown to be upregulated under starvation conditions in *M. smegmatis* [19], the MDR-TB analysis described here identified *aceAb-PPE3*; perhaps this intergenic region could play a role in the regulation of *aceAb*- a putative isocitrate lyase subunit B in *Mtb* [12]. Furthermore, *uspC* is an amino-sugar transporter that allows the

optimisation of scarce nutrient resources in *Mtb* [20], and was identified here by LCS to be predictive of rifampicin resistance.

Four loci identified by the isoniazid-resistance LCS, *espF*, *eccA1*, *espR*, and *eccD1*, were all linked to the ESX-1 secretion system. ESX-1 is a secretion system involved in virulence, thus it may be important to investigate how this might relate to drug resistance. A number of loci identified mention tRNA in their function; *thrT-metT* (both tRNA anticodons) and *Rv2630-Rv2631*- found by the MDR-TB analysis, *tyrT-Rv0567* (both tRNA anticodons), *mpt53-Rv2879c* and *fusA1*- found by the isoniazid-resistance analysis and *rplS* and *leuU-parE2* (*leuU* is a tRNA anticodon) (see **Table 5**, **Table 4**, **Table 2**).

There is the additional possibility that variants in some loci identified here may not play a role in drug resistance phenotypes, but may instead be markers of population structure. However, as our analyses includes only non-synonymous variants, they may have important functional implications for resistant strains, regardless. For example, further work may be warranted regarding *guaB2*, identified in the rifampicin resistance analysis, and the impact such findings have on its potential as a new drug target, as has been suggested [21]. Similarly, *esxV* has been considered a potential vaccine candidate [22], it may be useful to assess how a potential link with rifampicin resistance would affect this.

Our results differ from previous applications of machine learning methods to identify epistasis [6]; perhaps, in part, because our approach does not rely on prior knowledge and is genome-wide; we did not restrict analyses to specific loci.

In conclusion, LCS is a feasible approach to both resistance prediction and the disentangling of complex drug resistance genomics in *Mtb*, detecting known epistatic variants and providing novel candidates for further investigation, with scope to explore other phenotypes in *Mtb* and beyond.

REFERENCES

1. World Health Organisation. Global Tuberculosis Report 2018. Geneva; 2018.

2. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat. Genet. 2018;50.

3. Oppong YEA, Phelan J, Perdigão J, Machado D, Miranda A, Viveiros M, et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics [Internet]. BMC Genomics; 2019;1–15. Available from: http://link.springer.com/article/10.1186/s12864-019-5615-3?utm_source=researcher_app&utm_medium=referral&utm_campaign=MKEF_USG_Researcher_inbound

4. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat. Genet. [Internet]. Nature Publishing Group; 2013;45:1183–9. Available from: http://www.nature.com/articles/ng.2747

5. Huang H, Ding N, Yang T, Li C, Jia X, Wang G, et al. Cross-sectional Whole-genome Sequencing and Epidemiological Study of Multidrug-resistant Mycobacterium tuberculosis in China. Clin. Infect. Dis. 2019;69:405–13.

6. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, et al. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nat. Commun. [Internet]. Springer US; 2018;9:4306.

Available from: http://dx.doi.org/10.1038/s41467-018-06634-y

7. Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics. 2018;1–7.

8. Yang Y, Walker TM, Iqbal Z, Walker AS, Daniel J, Peto TEA, et al. DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis. 2018;1–8.

9. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. EBioMedicine [Internet]. The Authors; 2019;43:356–69. Available from: https://doi.org/10.1016/j.ebiom.2019.04.016

10. Urbanowicz RJ, Moore JH. ExSTraCS 2.0: description and evaluation of a scalable learning classifier system. Evol. Intell. [Internet]. 2015;8:89–116. Available from: http://link.springer.com/10.1007/s12065-015-0128-8

11. Urbanowicz RJ, Moore JH. ExSTraCS User's Guide Version 2.0.2 Beta. 2014. p. 0–43.

12. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. [Internet]. 2015;43:D447–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25352553%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4383874

13. Kapopoulou A, Lew JM, Cole ST. The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis [Internet].

Elsevier Ltd; 2011;91:8–13. Available from: http://dx.doi.org/10.1016/j.tube.2010.09.006

14. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med. [Internet]. Genome Medicine; 2019;11:41. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0650-x

15. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat. Genet. [Internet]. Nature Publishing Group; 2014;46:279–86. Available from: http://dx.doi.org/10.1038/ng.2878

16. Xu Z, Zhou A, Wu J, Zhou A, Li J, Zhang S, et al. Transcriptional approach for decoding the mechanism of rpoC compensatory mutations for the fitness cost in rifampicin-resistant Mycobacterium tuberculosis. Front. Microbiol. 2018;9:1–12.

17. Hurst-Hess K, Biswas R, Yang Y, Rudra P, Lasek-Nesselquist E, Ghosh P. Mycobacterial SigA and SigB cotranscribe essential housekeeping genes during exponential growth. MBio. 2019;10:1–17.

18. Van Wyk N, Navarro D, Blaise M, Berrin JG, Henrissat B, Drancourt M, et al. Characterization of a mycobacterial cellulase and its impact on biofilm- and drug-induced cellulose production. Glycobiology. 2017;27:392–9.

19. Wu ML, Gengenbacher M, Chung JCS, Chen SL, Mollenkopf HJ, Kaufmann SHE, et al. Developmental transcriptome of resting cell formation in Mycobacterium smegmatis.

BMC Genomics. 2016;17:1–9.

20. Fullam E, Prokes I, Fütterer K, Besra GS. Structural and functional analysis of the solute-binding protein UspC from Mycobacterium tuberculosis that is specific for amino sugars. Open Biol. 2016;6.

21. Singh V, Pacitto A, Donini S, Ferraris DM, Boros S, Illyés E, et al. Synthesis and Structure–Activity relationship of 1-(5-isoquinolinesulfonyl)piperazine analogues as inhibitors of Mycobacterium tuberculosis IMPDH. Eur. J. Med. Chem. [Internet]. 2019;174:309–29. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0223523419303332

22. Mansury D, Ghazvini K, Jamehdar SA, Badiee A, Tafaghodi M, Nikpoor AR, et al. Increasing cellular immune response in liposomal formulations of DOTAP encapsulated by fusion protein Hspx, PPE44, And Esxv, as a potential tuberculosis vaccine candidate. Reports Biochem. Mol. Biol. 2019;7:156–66.

23. http://tbdr.lshtm.ac.uk/download.html

TABLES, FIGURES AND ADDITIONAL FILES

Table 1- Prediction accuracies across parameters

Table 2- Loci Frequency Table for Rifampicin Npop=2000

Supplementary Table 1- Study Accession Numbers

Supplementary Table 2- Rule Length Summary

Supplementary Table (see **[23]**)**-** Complete Rule Population Spreadsheet for rifampicin

Supplementary Table 4 (see **[23]**)**-** Co-occurrence Table for rifampicin Npop=2000

Supplementary Table 5 (see **[23]**)**-** Loci Frequency Table for Isoniazid Npop=2000

Supplementary Table 6 (see **[23]**)**-** Complete Rule Population Spreadsheet for isoniazid

Supplementary Table 7 (see **[23]**)**-** Co-occurrence Table for isoniazid Npop=2000

Supplementary Table 8 (see **[23]**)**-** Loci Frequency Table for MDR-TB Npop=2000

Supplementary Table 9 (see **[23]**)**-** Complete Rule Population Spreadsheet for MDR-TB

Supplementary Table 10 (see **[23]**)**-** Co-occurrence Table for MDR-TB Npop=2000

Figure 1- Rule plot for rifampicin

Figure 2- Co-occurrence heatmap for rifampicin

Figure 3- Co-occurrence heatmap for isoniaizid

Figure 4- Co-occurrence heatmap for MDR-TB

Supplementary Figure 1- Accuracy Plots for all parameters

Supplementary Figure 2- Numerosity plots for all parameters- rifampicin

Supplementary Figure 3- Rule plot for isoniazid

Supplementary Figure 4- Rule plot for MDR-TB

Figure 1

Rule plot for rifampicin: clockwise showing from top left; Frequency of locus against locus number; Rule index (the highest index is the most frequent rule) against rule length; Rule index (the highest index is the most frequent rule) against locus number- red indicates resistance, blue indicates susceptibility, hollow points represent reference, filled points represent non-reference; Rule index against frequency (also referred to as 'numerosity').

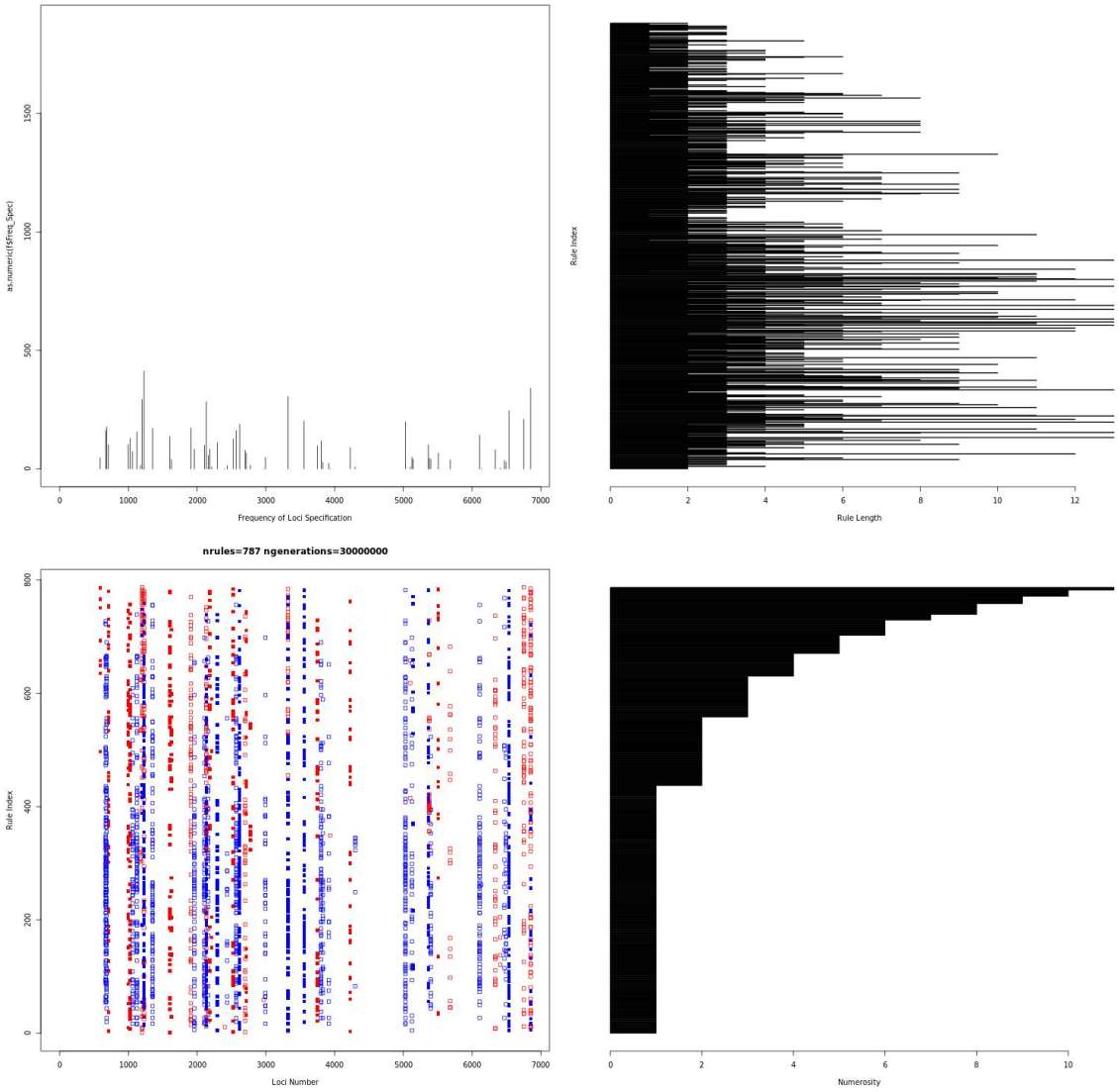


Figure 2

Co-occurrence heat map for rifampicin: Heat map showing co-occurring pairs of loci found in rules predictive rifampicin.

Figure 3

Co-occurrence heat map for isoniazid: Heat map showing co-occurring pairs of loci found in rules predictive isoniazid.

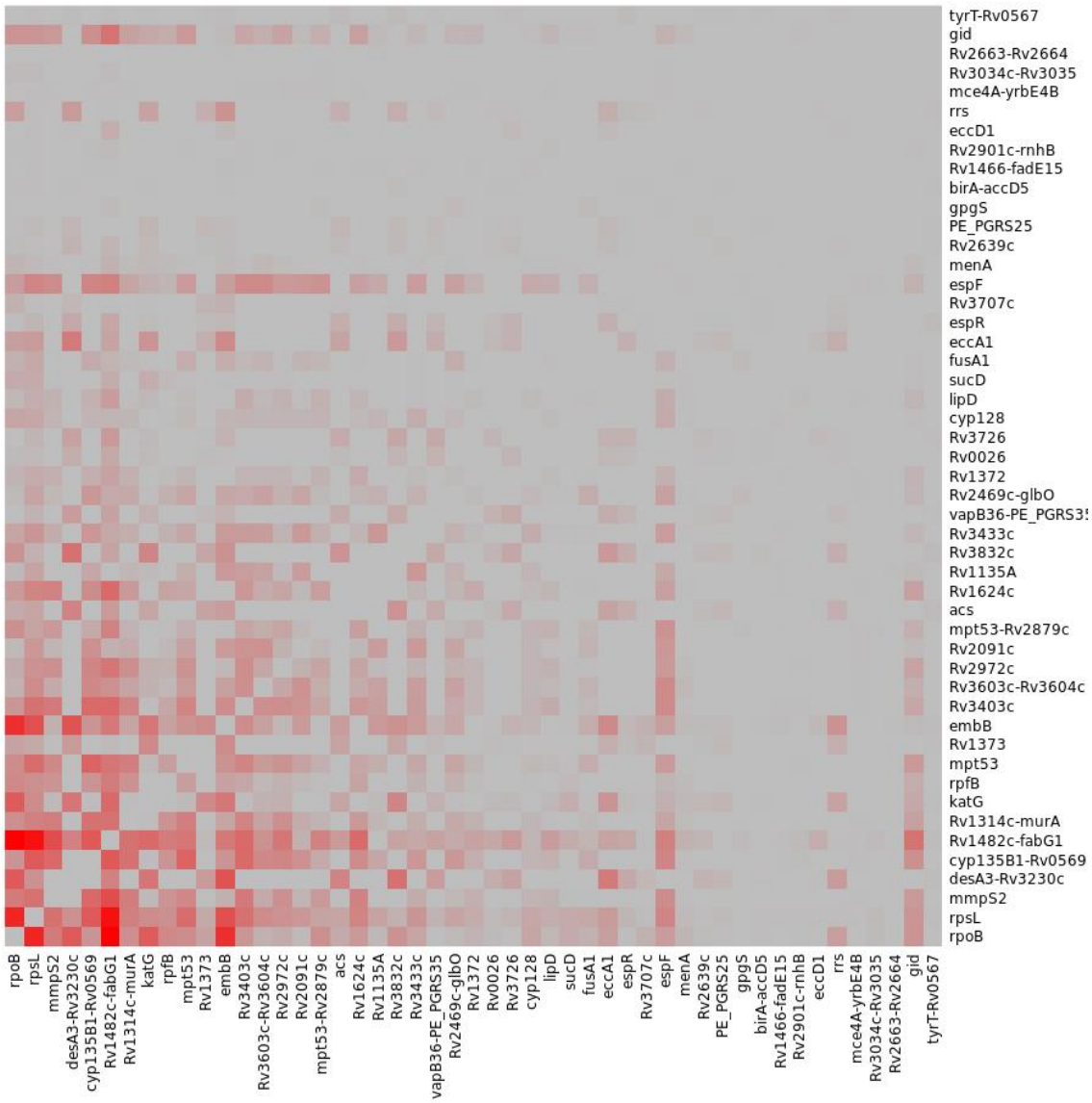
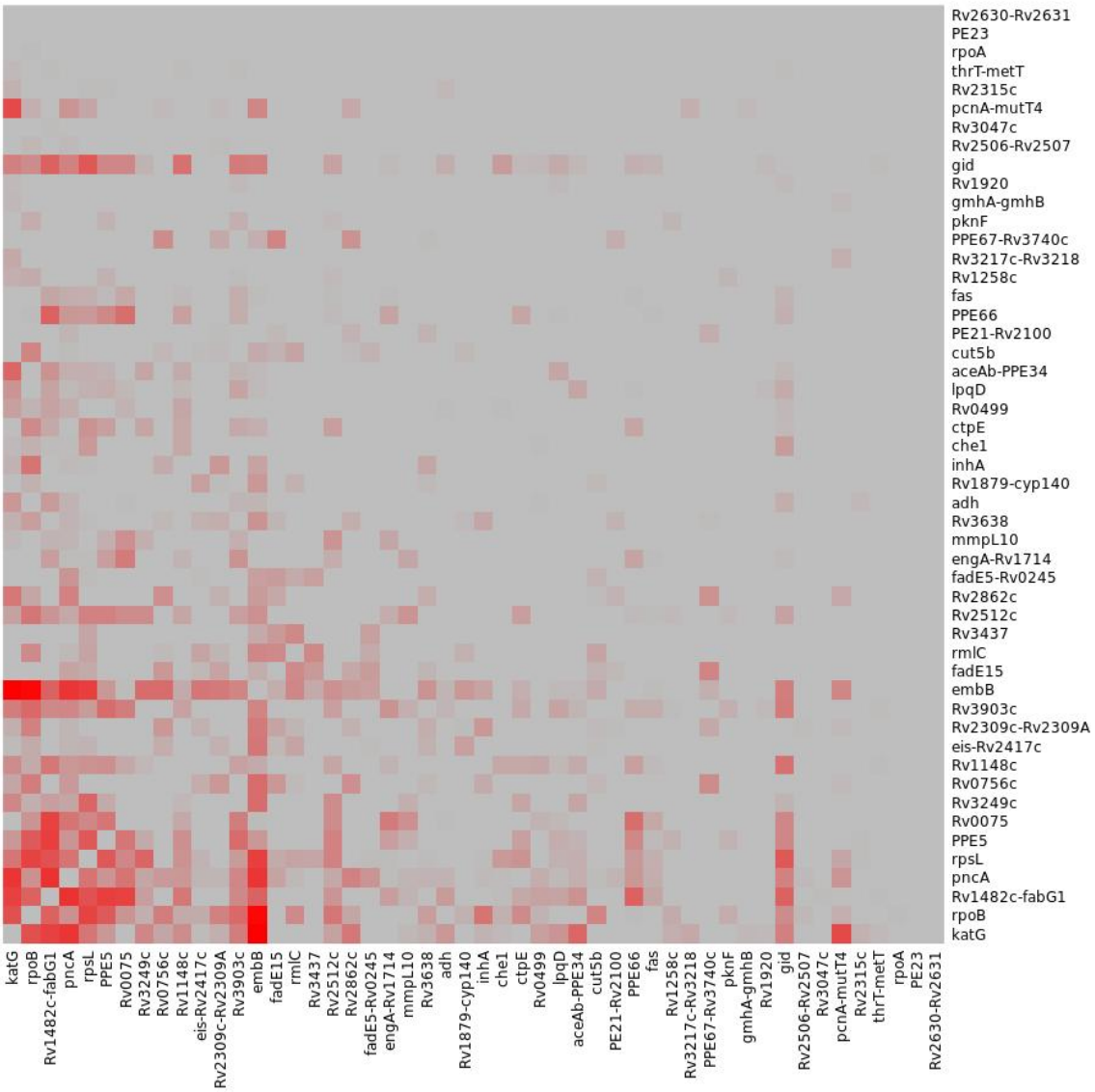


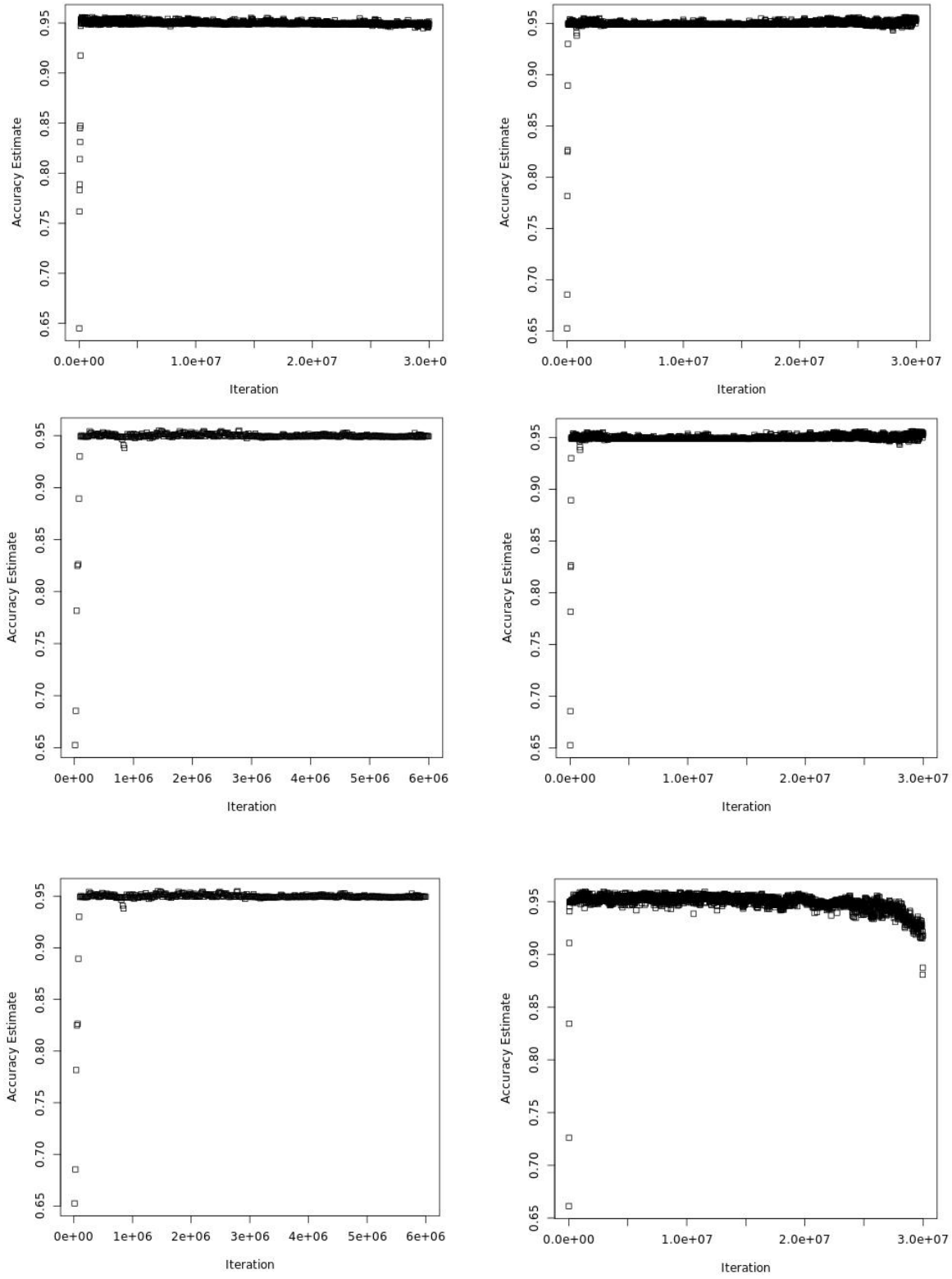
Figure 4

Co-occurrence heat map for MDR-TB: Heat map showing co-occurring pairs of loci found in rules predictive of MDR-TB.



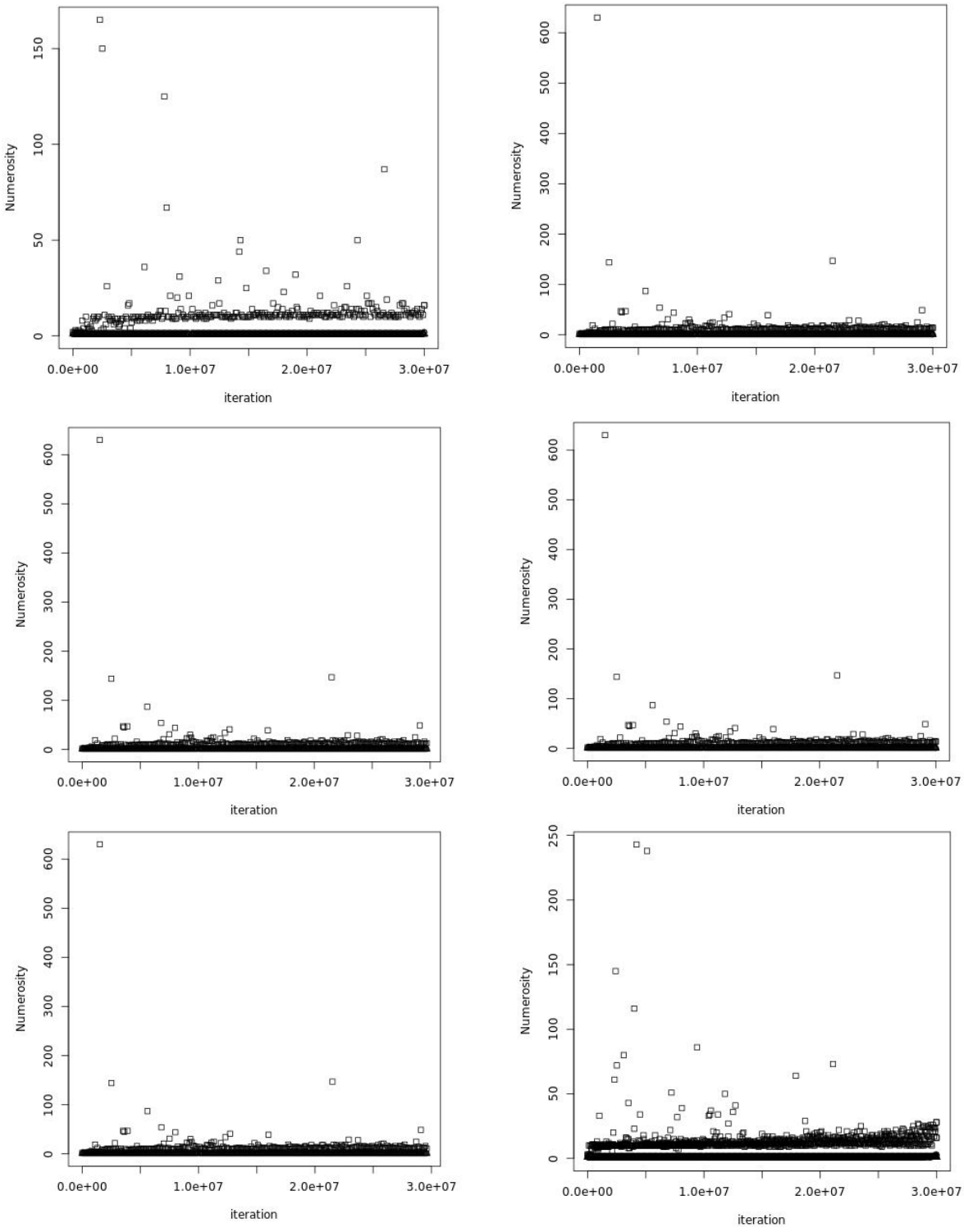
Supplementary Figure 1

Accuracy Plots for different Rule Population Sizes (A) N=1000, (B) N=2000, (C) N=3000, (D) N=4000, (E) N=5000, (F) N=2000 and sub_acc=0.95



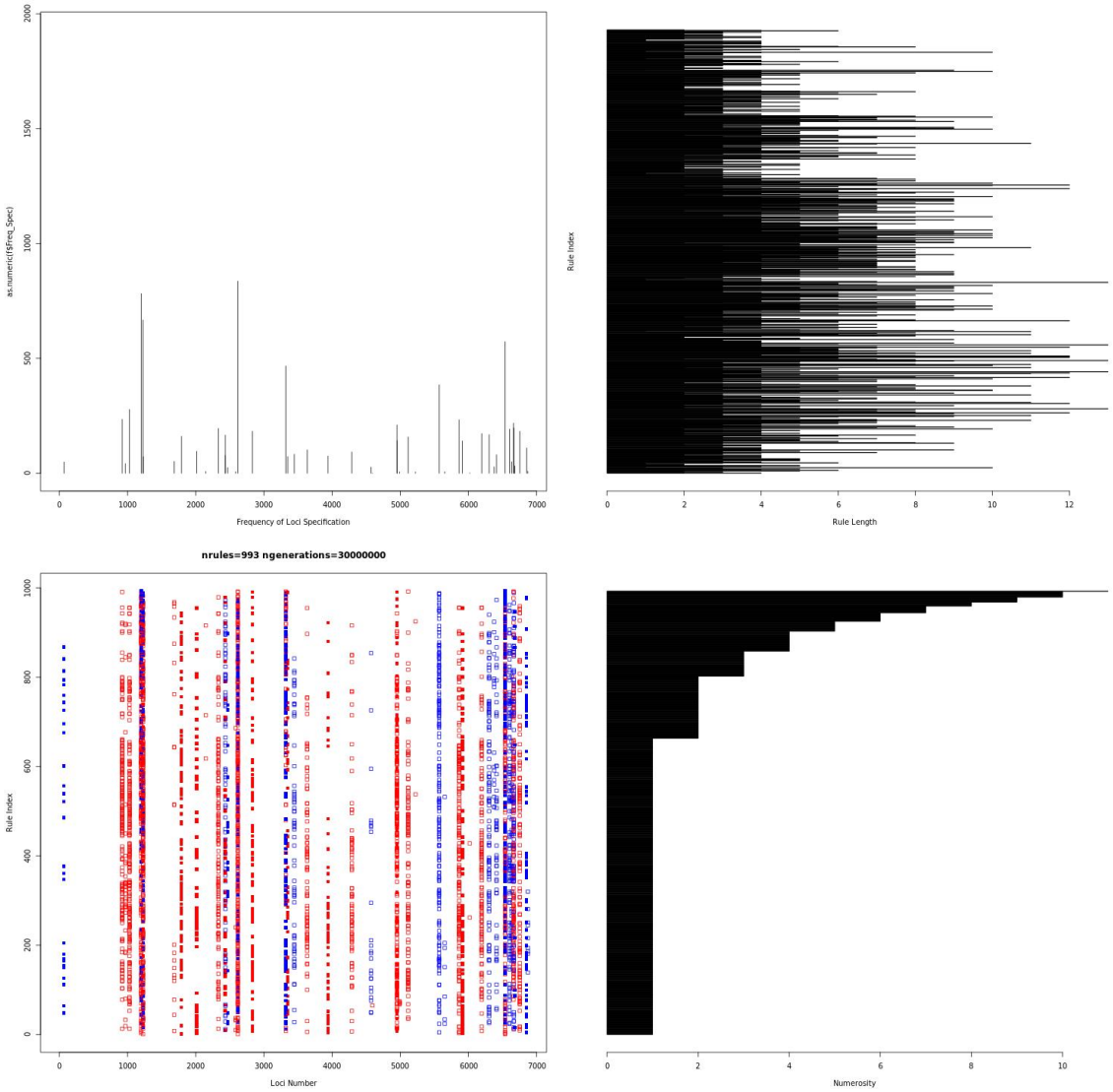
Supplementary Figure 2

Numerosity Plots for different Rule Population Sizes (A) N=1000, (B) N=2000, (C) N=3000, (D) N=4000, (E) N=5000, (F) N=2000 and sub_acc=0.95



Supplementary Figure 3

Rule plot for isoniazid: clockwise showing from top left; Frequency of locus against locus number; Rule index (the highest index is the most frequent rule) against rule length; Rule index (the highest index is the most frequent rule) against locus number- red indicates resistance, blue indicates susceptibility, hollow points represent reference, filled points represent non-reference; Rule index against frequency (also referred to as 'numerosity').



Supplementary Figure 4

Rule plot for MDR-TB: clockwise showing from top left; Frequency of locus against locus number; Rule index (the highest index is the most frequent rule) against rule length; Rule index (the highest index is the most frequent rule) against locus number- red indicates resistance, blue indicates susceptibility, hollow points represent reference, filled points represent non-reference; Rule index against frequency (also referred to as 'numerosity').

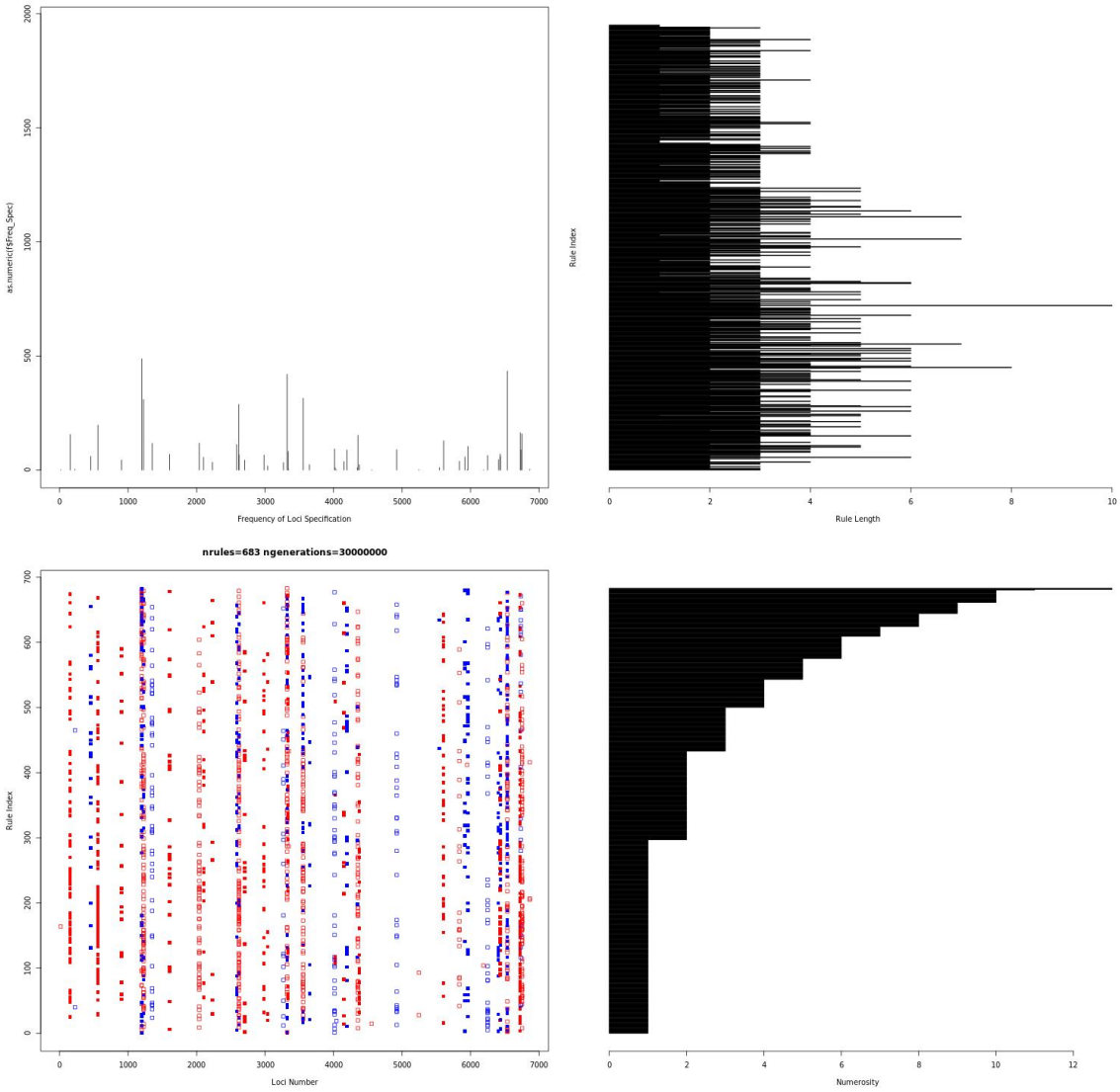


Table 1

Phenotype	Sensitivity	Specificity	Number in Test Set	Number of Resistant in Test Set	Rule Population Size	No. Iterations	Sub_acc	TBProfiler Benchmark for Phenotype-Sensitivity [14]	TBProfiler Benchmark for Phenotype-Specificity [14]
rifampicin	93.9%	94.9%	3259	876	1000	30000000	0.99	95.9%	98.2%
MDRvPAN	93.5%	97.6%	2864	742	2000	30000000	0.99	94.1%	98.3%
rifampicin	93.7%	95.8%	3259	876	2000	30000000	0.99	95.9%	98.2%
isoniazid	86.8%	94.2%	3242	1025	2000	30000000	0.99	93.7%	98.1%
rifampicin	93.6%	92.9%	3259	876	2000	30000000	0.95	95.9%	98.2%
rifampicin	93.4%	92.5%	3259	876	2000	30000000	0.90	95.9%	98.2%
rifampicin	93.7%	95.8%	3259	876	4000	30000000	0.99	95.9%	98.2%
rifampicin	91.4%	96.5%	3259	876	3000	30000000	0.99	95.9%	98.2%
rifampicin	93.7%	95.8%	3259	876	5000	30000000	0.99	95.9%	98.2%

Table 2

Frequency	Macro Frequency	Locus Name	Known	Min. Rule Length	Median Rule Length	Max Rule Length	Companions
850	253	rpoB	Y	1	3	5	*sigA;katG;Rv3818;Rv2395;pncA;embB;proC;Rv2230c;Rv1482c-fabG1;Rv3193c;purA;celA1;rpoC;gltB;Rv2576c-Rv2577;esxV;hns;Rv0338c;sdhA;rrs;pyrF;Rv1726;Rv1215c;Rv0140;rpsL;Rv2913c;Rv1129c-prpD;PE_PGRS30;uspC;Rv0303;rplS;Rv2424c;Rv1254;Rv3142c;rpe;leuU-parE2;pbpB;otsB2*
588	352	pncA	Y	1	5	14	*rpoB;Rv2395;accD5;Rv1482c-fabG1;purA;guaB2;embB;fadE10-Rv0874c;Rv1129c-prpD;Rv3818;celA1;katG;uspC;Rv2929-fadD26;rrs;Rv1254;Rv2576c-Rv2577;sigA;pbpB;Rv2913c;hns;pyrF;Rv3193c;rpoC;proC;otsB2;Rv1215c;Rv1726;Rv2230c;Rv3142c;gltB;PE_PGRS30;sdhA;Rv0140;esxV;Rv0338c;rpsL;Rv0690c;Rv1765A-Rv1766;rplS;Rv0303*
478	245	katG	Y	1	4	10	*Rv1482c-fabG1;Rv3142c;uspC;rpoB;Rv3818;Rv2395;sdhA;celA1;pncA;proC;PE_PGRS30;pbpB;embB;otsB2;Rv1215c;guaB2;rrs;Rv2230c;Rv0338c;esxV;gltB;purA;rpoC;pyrF;Rv1129c-prpD;rpsL;Rv1726;accD5;hns;sigA;Rv2913c;Rv0140;inhA-hemZ;Rv3193c;rplS;Rv0303;rpe;tgs1-Rv3131*
461	283	Rv1482c-fabG1	Y	1	4	14	*katG;Rv3142c;uspC;guaB2;Rv2913c;sdhA;rpoB;Rv2230c;accD5;purA;pncA;Rv2929-fadD26;rrs;Rv1254;Rv2395;Rv2576c-*

							Rv2577;sigA;pbpB;hns;proC;Rv3193c;Rv1215c;pyrF;otsB2;Rv3818;Rv1726;PE_PGRS30;embB;esxV;Rv0338c;celA1;gltB;lipL;inhA-hemZ;Rv1129c-prpD;rpoC;Rv1765A-Rv1766;rplS;Rv0303
454	233	embB	Y	1	4	13	*rpoB;proC;guaB2;sigA;Rv3193c;Rv2395;celA1;pncA;fadE10-Rv0874c;Rv1129c-prpD;Rv3818;pbpB;katG;rpoC;gltB;pyrF;Rv2230c;rpsL;rrs;Rv0338c;esxV;uspC;sdhA;PE_PGRS30;Rv2913c;Rv1482c-fabG1;otsB2;Rv0140;accD5;purA;Rv1726;Rv2576c-Rv2577;Rv2929-fadD26;Rv3142c;hns;inhA-hemZ;rplS;rpe;Rv1254*
314	191	rrs	Y	1	5	14	*celA1;Rv1482c-fabG1;pncA;uspC;Rv2929-fadD26;Rv1254;Rv2395;Rv2576c-Rv2577;sigA;purA;pbpB;Rv2913c;hns;Rv3193c;accD5;pyrF;Rv2230c;guaB2;Rv1215c;katG;otsB2;proC;rpoB;gltB;Rv3142c;PE_PGRS30;Rv1726;rpoC;embB;Rv3818;Rv1129c-prpD;esxV;Rv0338c;rpsL;lipL;Rv0140;sdhA;Rv0303;rplS*
305	164	Rv2395	N	1	5	14	*rpoB;pncA;accD5;katG;uspC;sdhA;embB;Rv3193c;Rv1482c-fabG1;Rv2929-fadD26;rrs;Rv1254;Rv2576c-Rv2577;sigA;purA;pbpB;Rv2913c;hns;Rv1215c;Rv1726;guaB2;Rv3142c;PE_PGRS30;Rv0338c;proC;esxV;otsB2;Rv0690c;Rv0303*
253	119	Rv2230c	N	2	4	8	*rpoB;Rv1482c-fabG1;gltB;celA1;rrs;pyrF;rpoC;Rv1129c-prpD;katG;Rv3818;embB;pncA;Rv0140;rpsL;rplS;tgs1-Rv3131*
233	166	pbpB	N	1	6	14	*Rv1482c-fabG1;pncA;uspC;Rv2929-fadD26;rrs;Rv1254;Rv2395;Rv2576c-Rv2577;sigA;purA;Rv2913c;hns;Rv3193c;embB;katG;acc*

							D5;Rv0338c;esxV;guaB2;proC;PE_PGRS30;Rv1726;Rv3142c;Rv1215c;lipL;otsB2;sdhA;Rv0303;rpoB
225	167	uspC	N	2	6	14	*katG;Rv1482c-fabG1;Rv3142c;Rv2395;sdhA;pncA;Rv2929-fadD26;rrs;Rv1254;Rv2576c-Rv2577;sigA;purA;pbpB;Rv2913c;hns;proC;accD5;otsB2;Rv1726;Rv0338c;esxV;embB;guaB2;Rv1215c;Rv3193c;PE_PGRS30;rpoB;Rv1765A-Rv1766;rpe*
209	98	celA1	N	2	4	8	*rrs;embB;rpoB;rpoC;gltB;Rv2230c;katG;pncA;Rv1129c-prpD;pyrF;Rv3818;rpsL;Rv1482c-fabG1;Rv0140;rplS;tgs1-Rv3131*
196	148	guaB2	N	1	6	14	*Rv1482c-fabG1;Rv2913c;sdhA;embB;accD5;purA;pncA;rrs;Rv1215c;Rv2576c-Rv2577;katG;otsB2;proC;Rv0338c;esxV;pbpB;Rv3142c;PE_PGRS30;Rv1726;Rv3193c;hns;uspC;Rv2395;sigA;Rv1254;Rv2929-fadD26;Rv1765A-Rv1766*
160	91	PE_PGRS30	N	1	4	14	*sdhA;katG;purA;Rv1726;pbpB;Rv3142c;accD5;guaB2;rrs;Rv1482c-fabG1;Rv3193c;hns;Rv1215c;embB;pncA;Rv2913c;otsB2;esxV;Rv1254;uspC;rpoB;Rv2395;sigA;Rv0690c;Rv2929-fadD26;proC;Rv2576c-Rv2577*
159	96	rpoC	Y	2	4	8	*rpoB;celA1;embB;pncA;Rv2230c;Rv1129c-prpD;Rv3818;pyrF;rrs;gltB;rpsL;katG;Rv1482c-fabG1;rplS;Rv0140*
152	81	esxV	N	1	4	9	*rpoB;purA;proC;Rv0338c;pbpB;guaB2;katG;embB;uspC;PE_PGRS30;sdhA;Rv1482c-fabG1;otsB2;pncA;rrs;Rv2395;Rv2913c;Rv1726;Rv3193c;*

							Rv1215c;Rv3142c;Rv1765A-Rv1766;sigA;Rv1254;Rv0303
149	116	Rv1215c	N	1	6	14	*Rv1482c-fabG1;Rv3193c;hns;purA;Rv2395;guaB2;Rv2576c-Rv2577;rrs;katG;otsB2;proC;pncA;Rv3142c;pbpB;PE_PGRS30;Rv1726;rpoB;uspC;sigA;accD5;lipL;sdhA;esxV;Rv2913c;Rv2929-fadD26;Rv0303*
144	95	sdhA	N	1	4	13	*Rv1482c-fabG1;guaB2;Rv2913c;katG;uspC;Rv2395;PE_PGRS30;proC;otsB2;Rv2576c-Rv2577;rpoB;embB;pncA;esxV;Rv0338c;Rv1726;pbpB;Rv2929-fadD26;Rv3193c;hns;Rv1215c;sigA;Rv3142c;purA;rrs;Rv1254;rpe*
140	114	Rv3142c	N	1	7	14	*katG;Rv1482c-fabG1;uspC;purA;Rv1215c;guaB2;rrs;PE_PGRS30;Rv1726;pbpB;accD5;pncA;hns;Rv2395;Rv2576c-Rv2577;sigA;Rv3193c;Rv1254;lipL;embB;Rv2929-fadD26;Rv0338c;proC;sdhA;otsB2;esxV;Rv2913c;rpoB*
139	88	accD5	N	1	5	13	*pncA;Rv2395;Rv1482c-fabG1;purA;guaB2;uspC;pbpB;rrs;PE_PGRS30;Rv1726;Rv3142c;Rv0338c;katG;embB;hns;Rv1215c;proC;Rv2576c-Rv2577;sigA;Rv2929-fadD26;Rv2913c;rpe;Rv3193c*
132	67	rpsL	Y	1	4	7	*rpoC;celA1;embB;Rv3818;rpoB;katG;rrs;gltB;pncA;Rv1129c-prpD;Rv2230c;rplS;pyrF;Rv0140*
115	69	Rv3193c	N	2	4	13	*rpoB;sigA;embB;Rv2395;purA;rrs;Rv1482c-fabG1;hns;pbpB;pncA;Rv2929-fadD26;Rv1215c;PE_PGRS30;guaB2;Rv1726;uspC;Rv2576c-*

							Rv2577;Rv3142c;Rv2913c;otsB2;sdhA;Rv0690c;Rv1254;esxV;katG;proC;rpe;accD5
114	67	Rv2913c	N	1	5	13	*Rv1482c-fabG1;guaB2;sdhA;pncA;uspC;Rv2929-fadD26;rrs;Rv1254;Rv2395;Rv2576c-Rv2577;sigA;purA;pbpB;hns;embB;PE_PGRS30;rpoB;proC;otsB2;Rv0338c;esxV;Rv3193c;Rv1726;katG;Rv3142c;Rv1215c;accD5*
106	70	Rv1726	N	2	7	14	*PE_PGRS30;purA;pbpB;Rv3142c;accD5;guaB2;rrs;Rv1482c-fabG1;uspC;Rv2395;pncA;hns;rpoB;Rv1215c;katG;Rv2576c-Rv2577;sigA;Rv3193c;Rv2913c;proC;otsB2;sdhA;embB;Rv2929-fadD26;esxV*
102	77	proC	N	1	4	13	*embB;rpoB;katG;Rv1482c-fabG1;uspC;pncA;otsB2;sdhA;Rv2576c-Rv2577;Rv0338c;esxV;Rv1215c;guaB2;rrs;pbpB;Rv2395;Rv2913c;Rv1726;accD5;hns;purA;sigA;Rv2929-fadD26;Rv3142c;Rv1254;PE_PGRS30;Rv3193c;rpe*
102	60	Rv3818	N	2	4	8	*rpoB;katG;embB;pncA;fadE10-Rv0874c;Rv1129c-prpD;gltB;Rv1482c-fabG1;pyrF;rpoC;Rv2230c;celA1;rpsL;rrs;Rv0140*
97	84	hns	N	2	8	14	*Rv1482c-fabG1;pncA;uspC;Rv2929-fadD26;rrs;Rv1254;Rv2395;Rv2576c-Rv2577;sigA;purA;pbpB;Rv2913c;Rv3193c;Rv1215c;rpoB;Rv1726;PE_PGRS30;guaB2;Rv3142c;accD5;katG;embB;otsB2;sdhA;Rv0338c;proC;rpe*
95	67	sigA	N	2	7	14	*rpoB;embB;Rv3193c;purA;Rv1482c-fabG1;pncA;uspC;Rv2929-*

							fadD26;rrs;Rv1254;Rv2395;Rv2576c-Rv2577;pbpB;Rv2913c;hns;Rv1726;guaB2;Rv1215c;Rv3142c;Rv0338c;accD5;proC;PE_PGRS30;katG;sdhA;otsB2;Rv0303;esxV;Rv1765A-Rv1766
76	57	pyrF	N	2	4	8	*pncA;rrs;Rv1482c-fabG1;Rv2230c;Rv3818;rpoB;celA1;rpoC;embB;gltB;katG;Rv1129c-prpD;rpsL;rplS;tgs1-Rv3131;Rv0140*
75	51	gltB	N	2	4	7	*rpoB;celA1;Rv2230c;embB;Rv3818;rrs;rpoC;katG;pncA;pyrF;Rv1129c-prpD;Rv1482c-fabG1;rpsL;rplS;Rv0140*
66	49	Rv1129c-prpD	N	2	5	8	*embB;pncA;fadE10-Rv0874c;Rv3818;Rv2230c;rpoC;celA1;rrs;katG;rpoB;gltB;pyrF;rpsL;Rv0140;Rv1482c-fabG1;rplS*
61	30	Rv1254	N	1	4	13	*Rv1482c-fabG1;pncA;uspC;Rv2929-fadD26;rrs;Rv2395;Rv2576c-Rv2577;sigA;purA;pbpB;Rv2913c;hns;guaB2;PE_PGRS30;Rv3142c;Rv3193c;proC;esxV;Rv1765A-Rv1766;sdhA;rpoB;embB*
46	32	otsB2	N	1	4.5	8	*proC;uspC;sdhA;Rv2576c-Rv2577;katG;Rv1482c-fabG1;pncA;Rv1215c;guaB2;rrs;PE_PGRS30;embB;esxV;Rv2913c;Rv0338c;Rv2395;Rv1726;pbpB;Rv2929-fadD26;Rv3193c;hns;sigA;Rv3142c;inhA-hemZ;rpoB*
45	36	Rv0338c	N	2	6	13	*rpoB;proC;esxV;pbpB;guaB2;katG;embB;uspC;pncA;rrs;Rv1482c-fabG1;accD5;Rv2395;otsB2;sdhA;Rv2913c;hns;purA;Rv2576c-Rv2577;sigA;Rv2929-fadD26;Rv3142c*
42	37	purA	N	2	8	14	*rpoB;sigA;Rv3193c;Rv1482c-fabG1;accD5;pncA;guaB2;uspC;Rv2929-fadD26;rrs;Rv1254;Rv2395;Rv2576c-*

							Rv2577;pbpB;Rv2913c;hns;esxV;Rv1215c;Rv3142c;PE_PGRS30;Rv1726;katG;embB;Rv0338c;proC;sdhA
32	20	rplS	N	2	4	7	*embB;Rv1482c-fabG1;Rv2230c;celA1;rpsL;gltB;rpoC;rpoB;katG;rrs;pncA;pyrF;Rv0140;Rv1129c-prpD*
25	23	Rv2929-fadD26	N	4	11	13	*Rv1482c-fabG1;pncA;uspC;rrs;Rv1254;Rv2395;Rv2576c-Rv2577;sigA;purA;pbpB;Rv2913c;hns;Rv3193c;embB;Rv1726;Rv3142c;otsB2;sdhA;Rv0338c;accD5;proC;guaB2;Rv1215c;PE_PGRS30*
21	21	Rv2576c-Rv2577	N	2	12	13	*rpoB;Rv1482c-fabG1;pncA;uspC;Rv2929-fadD26;rrs;Rv1254;Rv2395;sigA;purA;pbpB;Rv2913c;hns;proC;otsB2;sdhA;guaB2;Rv1215c;Rv1726;Rv3142c;Rv3193c;embB;Rv0338c;accD5;PE_PGRS30*
17	11	Rv0303	N	2	3	6	*rpoB;sigA;Rv2395;katG;Rv1482c-fabG1;esxV;rrs;Rv1215c;pncA;pbpB*
16	15	Rv0140	N	2	6	8	*rpoB;pncA;embB;Rv2230c;rrs;Rv1129c-prpD;katG;celA1;Rv3818;gltB;rpoC;pyrF;rpsL;rplS*
8	5	rpe	N	2	3	5	*katG;uspC;accD5;hns;proC;Rv3193c;sdhA;embB;rpoB*
2	2	Rv1765A-Rv1766	N	7	7	7	*uspC;esxV;Rv1482c-fabG1;pncA;sigA;guaB2;Rv1254*
2	1	leuU-parE2	N	2	2	2	*rpoB*
1	1	Rv0690c	N	5	5	5	*Rv2395;PE_PGRS30;pncA;Rv3193c*
1	1	fadE10-	N	5	5	5	*embB;pncA;Rv1129c-prpD;Rv3818*

		Rv0874c					
1	1	inhA-hemZ	N	5	5	5	*katG;otsB2;Rv1482c-fabG1;embB*
1	1	lipL	N	6	6	6	*Rv1482c-fabG1;Rv1215c;pbpB;Rv3142c;rrs*
1	1	Rv2424c	N	2	2	2	*rpoB*
1	1	tgs1-Rv3131	N	5	5	5	*katG;celA1;Rv2230c;pyrF*

For ‘Known’, ‘Y’= Yes/Known, ‘N’=No/Unknown

Supplementary Table 1

Project	Rifampi-cin tested total	Rifampi-cin resistant %	Isoniazid tested total	Isoniaizid resistant %	MDR	Pan-suscept-ible
cryptic_nejm_2018	7964	32.3%	7842	35.7%	2344	4946
PRJEB10385	610	70.0%	611	78.7%	295	98
PRJEB11653	125	44.0%	125	42.4%	35	14
PRJEB14199	123	95.1%	123	74.0%	14	0
PRJEB15857	38	39.5%	38	42.1%	15	18
PRJEB2138	34	64.7%	34	70.6%	13	8
PRJEB2221	355	2.0%	355	5.1%	6	331
PRJEB2358	321	0.6%	321	8.7%	2	289
PRJEB2424	45	88.9%	45	93.3%	40	3
PRJEB2777	93	0.0%	93	0.0%	0	93
PRJEB2794	1258	0.6%	1257	6.5%	7	1175
PRJEB5162	185	1.6%	190	5.8%	2	174
PRJEB6276	3	0.0%	3	0.0%	0	3
PRJEB6945	46	0.0%	46	0.0%	0	46
PRJEB7056	1087	4.8%	1086	13.3%	41	873
PRJEB7281	95	45.3%	95	56.8%	41	38
PRJEB7669	228	100.0%	231	100.0%	217	0
PRJEB7727	28	25.0%	28	32.1%	5	12
PRJEB9680	1019	24.3%	1019	28.0%	246	700
PRJNA183624	329	66.0%	329	69.9%	138	83
PRJNA187550	157	72.6%	157	72.6%	91	43
PRJNA200335	124	78.2%	125	78.4%	43	23
PRJNA235852	208	9.6%	208	21.6%	20	155
PRJNA282721	1778	6.5%	1807	16.5%	87	1452
PRJNA376471	13	38.5%	13	38.5%	5	8
PRJNA49659	30	0.0%	30	0.0%	0	30

Supplementary Table 2

Phenotype	Min.	Median	Max.
rifampicin	1	4	14
MDRvPAN	1	3	10
isoniaizid	1	4	13

Chapter 5:

Genome-wide Learning Classifier System applied to Extensively Drug Resistant *Mycobacterium tuberculosis* discovers novel resistance mechanisms

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1500052	**Title**	Ms
First Name(s)	Yaa Emily Adoma		
Surname/Family Name	Oppong		
Thesis Title	Investigating the genomic basis of antimicrobial resistance in *Mycobacterium tuberculosis* (*Mtb*) using genome-wide methodologies		
Primary Supervisor	Martin Hibberd		

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

<u>SECTION B – Paper already published</u>

Where was the work published?	
When was the work published?	
If the work was published prior to registration for your research degree, give a brief rationale for	

its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	

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

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	BMC Genomics
Please list the paper's authors in the intended authorship order:	Yaa E A Oppong, Jody E Phelan, Martin L Hibberd, Taane G Clark
Stage of publication	Not yet submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	**YO conceived the project and performed bioinformatic and statistical analyses under the supervision of TC and MH, and wrote the first draft of the manuscript. JPh generated the sequencing dataset.**

SECTION E

Student Signature	
Date	

Supervisor Signature	
Date	

Application of a genome-wide learning classifier system to identify mutations involved in extensively drug resistant *Mycobacterium tuberculosis* extensively drug Resistance

Yaa E A Oppong[1, §], Jody E Phelan[1], Martin L Hibberd[1,*], Taane G Clark[1,2,*]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

[2] Faculty of Epidemiology and Population Health, LSHTM, London, United Kingdom

* Joint authors

[§] Corresponding author

Yaa E A Oppong

Department of Infection Biology

Faculty of Infectious and Tropical Diseases

London School of Hygiene and Tropical Medicine, London, United Kingdom

yaa.oppong@lshtm.ac.uk

ABSTRACT

Background

Tuberculosis, caused by *Mycobacterium tuberculosis* bacteria, is a major global public health issue, with drug resistance making disease control more difficult. Of major concern is the evolution of extensively drug resistant *M. tuberculosis* (XDR-TB), which is multidrug-resistant tuberculosis to isoniazid and rifampicin plus resistance to fluoroquinolones and injectable second-line drugs. The analysis of *M. tuberculosis* (XDR-TB) whole genome sequencing (WGS) and drug susceptibility test data can improve the understanding of the complex combination of *M. tuberculosis* genetic mutations involved in XDR-TB, and thereby have important positive implications for surveillance, diagnosis and treatment.

Results

A machine learning classifier system was applied on a locus-based resolution to a global dataset (n=13,270; 308 XDR-TB), consisting of two *M. tuberculosis* lineages (2 and 4). The analyses identified known resistance loci (9 for lineage 2, 13 for lineage 4), as well as potential novel ones (107 for lineage 2, 116 for lineage 4). The constructed models included loci to adjust for the confounding effect of *M. tuberculosis* strain-types, as well as avoided overfitting. They predicted XDR-TB with sensitivity of 93.9% and specificity of 98.6% for lineage 2 and sensitivity of 100% and specificity of 99.5% for lineage 4, which is similar to other recent machine learning applications.

Conclusions

The learning classifier system approach is an effective predictor of XDR *Mtb*, requiring no prior knowledge of *M. tuberculosis* genomics and could be used to inform outbreak control and drug resistance, through the prediction of phenotype and discovery of biological mechanisms.

INTRODUCTION

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*Mtb*), is a leading global cause of mortality and morbidity, with the evolution of drug resistance threatening disease control. Extensively drug-resistant *Mtb* (XDR-TB) is defined as multidrug resistant (MDR-TB, resistance to two first-line treatments, isoniazid and rifampicin) plus additional resistance to fluoroquinolones and one second-line injectable. XDR-TB has been observed in at least 123 countries, with around 8.5% of MDR TB cases being XDR in 2017 [1]. The transmission of XDR-TB has been observed [2] and the complexity of underlying genomics of XDR-TB strains has been described [3] (Oppong et al., submitted), with suggestions of epistatic interactions contributing to fitness in resistant strains [4, 5] and differences in genomics between *Mtb* lineages [6].

The increased availability of whole genome sequencing (WGS) data for clinical *Mtb* isolates, raises the possibility of gaining a deeper understanding of XDR-TB outbreak dynamics to inform outbreak control strategies, diagnosis and treatment [7]. Current analytical approaches include the application of phylogenetic based characterisation of transmission clusters, and the use of GWAS and convergent evolution methods to identify mutations associated with drug resistance and transmissibility [8–10]. Such methods have detected individual known mutations in the context of MDR-TB, but there is a need to identify novel mechanisms to explain the XDR-TB phenotype, including any epistasis at play. The application of machine learning approaches has the potential to uncover new resistance loci [11–19]. Learning classifier systems (LCS) iteratively search for combinations of loci in populations of models using mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection. These approaches have the potential to explore large datasets and epistatic effects in relation to complex “resistance”

traits (Oppong et al, in prep). Here we apply an LCS approach to predict *Mtb* XDR-TB in a global dataset (n=13,270; 308 XDR-TB), with paired WGS and drug susceptibility testing data.

METHODS

Data and Processing

A global dataset of WGS data for 18,255 *Mtb* isolates from lineages 1 (12.1%), 2 (25.3%), 3 (15.2%) and 4 (47.3%) was collated, alongside phenotypic drug susceptibility testing data across 16 anti-TB drugs (Oppong et al., under review, 2019). XDR-TB was assigned as combined resistance to isoniazid and rifampicin plus resistance to a fluoroquinolone (ciprofloxacin, levofloxacin, moxifloxacin) and to a second line injectable (amikacin, kanamycin, capreomycin). Pan-susceptibility was assigned as susceptibility to rifampicin and isoniazid alongside no other known resistance (see **Supplementary Table 1, Oppong et al., submitted**) (n=7,063, 53.2%). Only analyses where the number of XDR-TB isolates in the test set was >10 were included in downstream analyses, leaving lineages 2 (n=4,642) and 4 (n=8,628). Overall the number of XDR-TB cases was 308 (lineage 2, n=155, 3.34%; lineage 4, n=153, 1.77%).

Sequence data was aligned to the H37rv reference and variants were called, resulting in a total of 613,821 SNPs. A binary 'locus-type' was assigned by collapsing SNP data by its gene or intergenic region. The loci containing one or more non-synonymous SNPs were coded as '1' and any loci that did not contain one or more non-synonymous SNPs were coded as '0', producing a dataset of 6,195 variant loci for lineage 2 and 6,677 for lineage 4.

Learning classifier systems (LCS)

The ExSTraCS Learning Classifier System [20] was applied to each lineage (2 and 4) to construct populations of models that predict XDR-TB (vs. pan-susceptible). All parameters were default except the rule population size (n=2,000), which was recommended for complex and noisy problems [21]. For each analysis, the data was split into a training (80%) and test (20%) set (see **Figure 1**). The frequencies of entire rules, co-occurring pairs and individual loci were quantified in the final rule populations. Frequency is defined as the total number across the total rule population. Macro frequency is defined as the total number across unique rules. The number of loci in each rule (rule length) was also enumerated. Phenotypes, testing and training sets and misclassified isolates were plotted on lineage specific maximum likelihood phylogenetic trees, created using ExaML [22], as described in (Oppong et al., submitted). The functionality of loci identified in the LCS analysis were taken from STRING [23] and Mycobrowser [24].

RESULTS

The lineage 2 and lineage 4 analyses identified nine loci in common as predictive of XDR-TB: *embB* (ethambutol), *ethA* (ethionamide), *lppC*, *pncA* (pyrazinamide), *rpoB* (rifampicin), *rrs* (streptomycin), *Rv0458*, *Rv1482c-fabG1* (isoniazid) and *Rv2434c.* Of these *lppC*, *rv0458* and *Rv2434c* have not previously been implicated in drug resistance (see **Table 2 and Table 3**). *lppC* is a putative lipoprotein*, Rv0458* is a putative aldehyde dehydrogenase and *Rv2434c* is a putative conserved transmembrane protein [23]. Assuming the drug susceptibility test result as the gold standard, the LCS achieved a predictive performance of 93.9% sensitivity and 98.6% specificity for lineage 2 and 100.0% sensitivity and 99.5% specificity for lineage 4 (see **Table 1**). Performing the final models on the test set resulted in a low misclassification error in both lineages (6/310 lineage 2, 6/1163 lineage 4) (see **Table 1, Supplementary Figure 1**).

Lineage 2 analysis

Nine known resistance involved loci were identified by lineage 2 analysis: *Rv1482c-fabG1*, *rpoB*, *ethA*, *rrs*, *embB*, *pncA*, *rpoC*, *rpoA* and *alr* (see **Table 2**). 107 novel loci identified by lineage 2 analysis (see **Table 2**), where these included markers to stratify by different strain-types. The most frequent loci across the population of 2,000 models were: *Rv0575c* (frequency 186), *Rv1823* (f126), *hycE-Rv0088* (121)*, narJ* (110), *Rv3115-moeB2* (110) and *ponA1* (105). The most frequently co-occurring pair of loci was *Rv0575C* and *narJ* (frequency 45) (see **Supplementary Table 2**). Whilst, the most frequently occurring rules were non-reference *Rv1922* (frequency 11)*, fadE1 (11),* and *Rv2897c* (11) in prediction of pan-susceptibility (**Supplementary Table 3).** Based on an analysis of gene function, a number of novel loci are candidates for directly conferring resistance, including *Rv1877* which shows similarity to drug efflux proteins [23] (see **Table 2**).

Lineage 4 analysis

Thirteen known resistance involved loci were identified by lineage 4 analysis: *rpoB (*rifampicin*)*, *embB* (ethambutol), *katG* (isoniazid), *ethA-ethR* (ethionamide), *rrs* (streptomycin), *pncA* (pyrazinamide), *Rv1482c-fabG1* (isoniazid), *drrA*, *gyrB* (fluorquinolones), *rpsL* (streptomycin*)*, *embC-embA* (ethambutol), *inhA* (isoniazid) and *ethA* (ethionamide) (see **Table 3**). 116 novel loci were identified by lineage 4 analysis (see **Table 3**), where the most frequent across the 2,000 models were *fadD30* (226), *Rv2059* (166), *Rv0158* (125), *vapB34* (125) and *glnA3* (124) (see **Table 3**). *Rv0158* and *glnA3* was the most frequently occurring pair of loci (frequency 43) (see **Supplementary Table 4**). The most frequent rule in the rule population was non-reference fadE15-PE_PGRS29 in prediction of pan-susceptibility (12) (see **Supplementary Table 5**). A number of novel loci identified were potentially involved in efflux such as (i) *Rv2025c,* where *Rv2024c-Rv2025c* is a probable

cation efflux system; (ii) *mctB* involved in efflux of copper, and (iii) *Rv0194* of *Rv0193c-Rv0194*, which is a multidrug efflux protein [23] (see **Table 3**).

DISCUSSION

We have applied LCS to a global dataset of XDR and pan-susceptible *Mtb* isolates from lineages 2 and 4, achieving high accuracy of prediction and identifying known loci and novel candidates for involvement in these phenotypes.

The high accuracy of XDR-TB prediction, as observed here for lineage 4, is especially interesting considering the model presented is based solely on loci, rather than single nucleotide polymorphisms or indels. It would appear even lower resolution locus-level data can provide enough information to predict XDR-TB phenotype, when multiple loci are being considered in tandem. The idea that many loci are of importance in predicting the XDR-TB phenotype is supported by the lower and more evenly distributed rule frequencies that make up the rule populations described here, compared to previous analyses predicting single resistance to rifampicin (Oppong et al., in prep). It follows that phenotype prediction through locus-level, as opposed to mutation-level, LCS may work better when the phenotype in question is highly complex, involving multiple loci, rather than a simple Mendelian trait.

One factor contributing towards the higher prediction accuracy observed for lineage 4 compared to lineage 2 could be the increased diversity in lineage 4 (Oppong et al., submitted). Perhaps the lineage 2 analyses could benefit from the inclusion of SNP and indel data or from greater numbers to capture more low frequency variants involved in the phenotype. Indeed, as it becomes more computationally feasible, applying the methodology presented here to mutation-level data could enable real-time monitoring of

clinical *Mtb* isolates to predict XDR-TB outbreaks before they occur, allowing more targeted resource allocation and thus more effective intervention to prevent further transmission. Furthermore, the high accuracy XDR-TB prediction, as achieved here, could be useful in diagnosis of clinical *Mtb* isolates in order to inform treatment regimen.

There is potential that the LCS is picking up on variation unrelated to fitness of XDR *Mtb* and rather identifying markers related to population structure. Thus, it might be interesting to further test the LCS models developed here on new data. However, the use of only non-synonymous mutations at locus level means any loci identified may well have functional importance and at the very least be descriptive of clinically relevant XDR strains. Moreover, it is interesting to note that both true and false predictions appear to be evenly distributed throughout the phylogenetic tree for each lineage.

A number of understudied loci were identified, some in prediction of XDR-TB and some in prediction of pan-susceptibility, and thus warrant further biological investigation. These include potential mechanisms of drug resistance such as probable drug efflux pumps.
The high number of novel candidate loci identified here is in keeping with the nature of this method. Unlike other machine learning methods previously applied to detect epistasis in drug resistant *Mtb* [11], LCS is able to identify associated loci across the entire genome, without the need for prior knowledge.

CONCLUSIONS

LCS is a potentially useful methodology for predicting the complex XDR-TB phenotype. It requires no prior knowledge, but seems to demonstrate high predictive accuracy, account for population structure, and identify plausible and biologically interesting loci as candidates for further study. Large-scale use of this approach could be useful in the real-

time monitoring of *Mtb* genomics in outbreak settings, informing more targeted public health intervention, especially with the extended application of LCS to individual mutation level data. Furthermore, its high predictive accuracy could be useful in diagnosing XDR-TB in order to inform treatment.

ACKNOWLEDGEMENTS

YO and JPh were funded by a BBSRC PhD studentships. TGC received funding from the MRC UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC UK (BB/R013063/1).

REFERENCES

1. World Health Organisation. Global Tuberculosis Report 2018. Geneva; 2018.

2. Dheda K, Esmail A, Mcnerney R, Theron G, Streicher EM, van Helden P, et al. Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. Lancet Respir Med. 2017;2600:1–13. doi:10.1016/S2213-2600(16)30433-7.

3. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018;50:307–16. doi:10.1038/s41588-017-0029-0.

4. Meftahi N, Namouchi A, Mhenni B, Brandis G, Hughes D, Mardassi H. Evidence for the critical role of a secondary site rpoB mutation in the compensatory evolution and successful transmission of an MDR tuberculosis outbreak strain. J Antimicrob Chemother. 2016;71:324–32.

5. Merker M, Barbier M, Cox H, Rasigade J-P, Feuerriegel S, Kohl TA, et al. Compensatory

evolution drives multidrug-resistant tuberculosis in Central Asia. Elife. 2018;7:1–31. doi:10.7554/eLife.38200.

6. Oppong YEA, Phelan J, Perdigão J, Machado D, Miranda A, Viveiros M, et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics. 2019;:1–15. doi:10.1186/s12864-019-5615-3.

7. Dheda K, Gumbo T, Maartens G, Dooley KE, Murray M, Furin J, et al. The Lancet Respiratory Medicine Commission: 2019 update: epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant and incurable tuberculosis. Lancet Respir Med. 2019;7:820–6. doi:10.1016/S2213-2600(19)30263-2.

8. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet. 2013;45:1183–9. doi:10.1038/ng.2747.

9. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018;50.

10. Hicks ND, Yang J, Zhang X, Zhao B, Grad YH, Liu L, et al. Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. Nat Microbiol. 2018;3:1032–42. doi:10.1038/s41564-018-0218-3.

11. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, et al. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nat Commun. 2018;9:4306. doi:10.1038/s41467-018-06634-y.

12. Chowdhury AS, Khaledian E, Broschat SL, Science C, States U. Capreomycin Resistance Prediction in Two Species of Mycobacterium Using a Stacked Ensemble Method. 2019;0059.

13. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Sequence analysis

Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. 2018;34 December 2017:1666–71.

14. Chen ML, Doddi A, Royer J, Freschi L, Ezewudo M, Kohane IS, et al. Deep Learning Predicts Tuberculosis Drug Resistance Status from Whole-Genome Sequencing Data. 2018.

15. Huang H, Ding N, Yang T, Li C, Jia X, Wang G, et al. Cross-sectional Whole-genome Sequencing and Epidemiological Study of Multidrug-resistant Mycobacterium tuberculosis in China. Clin Infect Dis. 2019;69:405–13.

16. Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics. 2018; November:1–7. doi:10.1093/bioinformatics/bty949.

17. Yang Y, Walker TM, Iqbal Z, Walker AS, Daniel J, Peto TEA, et al. DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis. 2018;:1–8.

18. Sergeev RS, Kavaliou I, Sataneuski U, Gabrielian A, Rosenthal A, Tartakovsky M, et al. Genome-wide Analysis of MDR and XDR Tuberculosis from Belarus: Machine-learning Approach. IEEE/ACM Trans Comput Biol Bioinforma. 2019;16:1–1. doi:10.1109/TCBB.2017.2720669.

19. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial Resistance Prediction in PATRIC and RAST. Sci Rep. 2016;6 May:1–12. doi:10.1038/srep27930.

20. Urbanowicz RJ, Moore JH. ExSTraCS 2.0: description and evaluation of a scalable learning classifier system. Evol Intell. 2015;8:89–116. doi:10.1007/s12065-015-0128-8.

21. Urbanowicz RJ, Moore JH. ExSTraCS User's Guide Version 2.0.2 Beta. 2014;:0–43.

22. Kozlov AM, Aberer AJ, Stamatakis A. ExaML version 3: A tool for phylogenomic analyses on supercomputers. Bioinformatics. 2015;31:2577–9.

23. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic

Acids Res. 2015;43:D447–52. doi:10.1093/nar/gku1003.

24. Kapopoulou A, Lew JM, Cole ST. The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis. 2011;91:8–13. doi:10.1016/j.tube.2010.09.006.

25. http://tbdr.lshtm.ac.uk/download.html

FIGURES AND TABLES

Figure 1- Rule plots lineage 2 and lineage 4 XDR

Figure 2- Co-occurrence heatmaps for lineage 2 and lineage 4

Supplementary Figure 1- Phylogenetic trees for lineage 2 and lineage 4

Table 1- Prediction Accuracies

Table 2- Locus table lineage 2 XDR

Table 3- Locus table lineage 4 XDR

Supplementary Table 1- Phenotype Frequencies by lineage

Supplementary Table 2 (see **[25]**)- Co-occurrence table lineage 2 XDR

Supplementary Table 3 (see **[25]**)- Rule table lineage 2 XDR

Supplementary Table 4 (see **[25]**)- Co-occurrence table lin4 XDR

Supplementary Table 5 (see **[25]**)- Rule table lineage 4 XDR

Figure 1

Rule plots lin2 and 4 XDR:(A) lineage 2 and (B) lineage 4. clockwise showing from top left; Frequency of locus against locus number; Rule index (the highest index is the most frequent rule) against rule length; Rule index (the highest index is the most frequent rule) against locus number- red indicates resistance, blue indicates susceptibility, hollow points represent reference, filled points represent non-reference; Rule index against frequency (also referred to as 'numerosity').

A-lineage 2

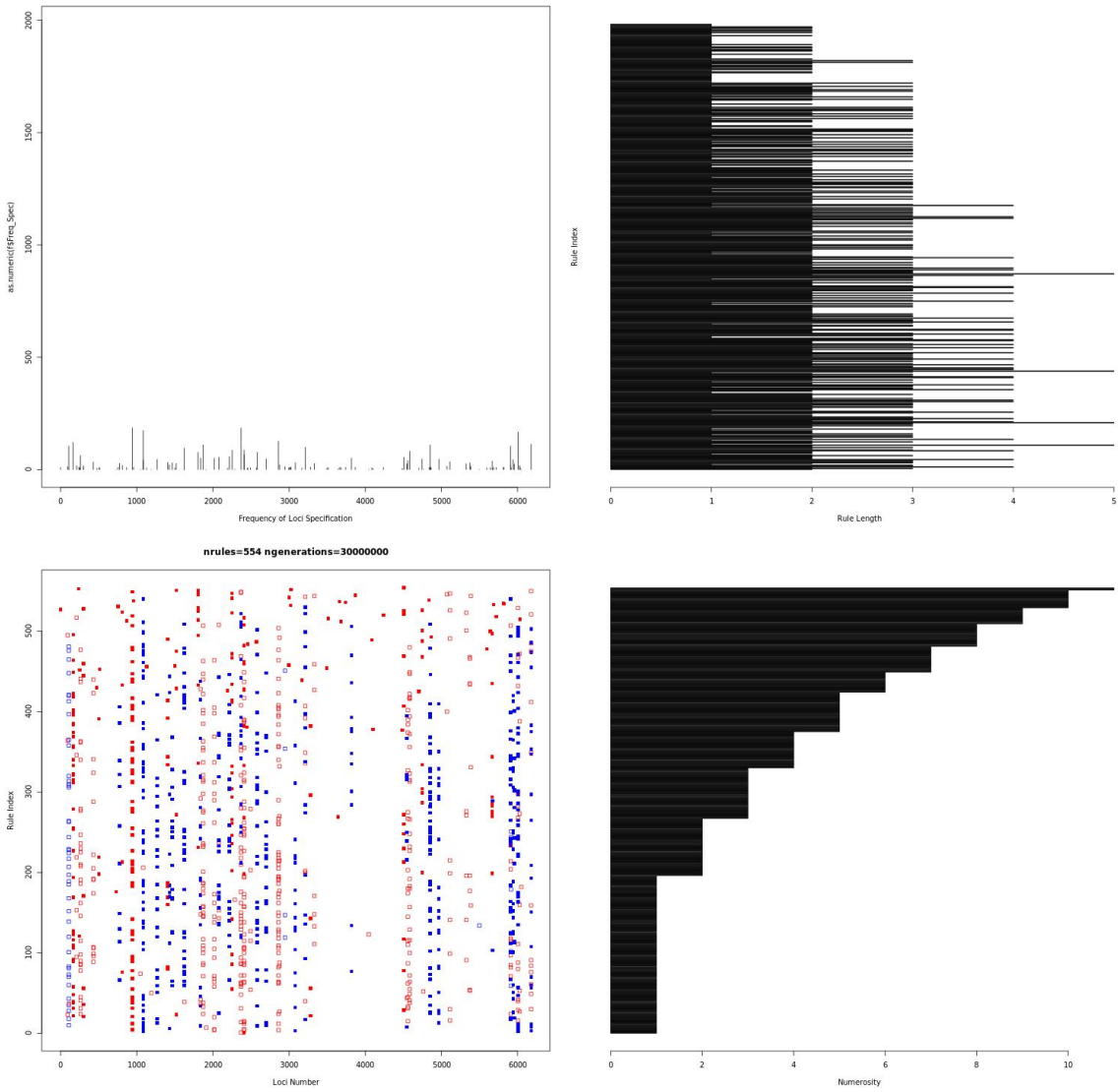


Figure 1 (cont.)

B- lineage 4

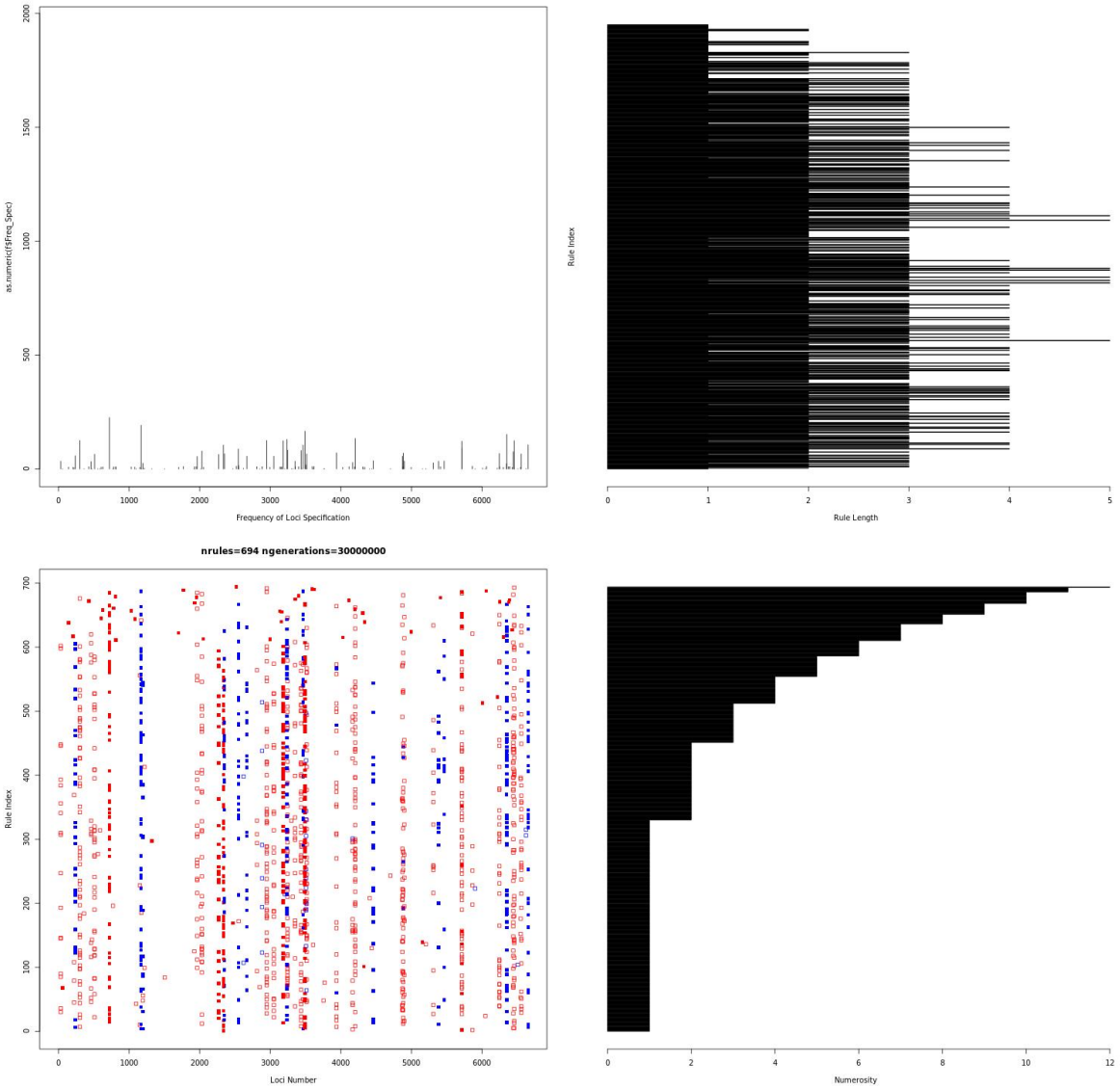


Figure 2

Heat maps showing co-occurring loci found in rules predictive of XDR for (A) lineage 2 and (B) lineage 4 XDR.

A-lineage 2

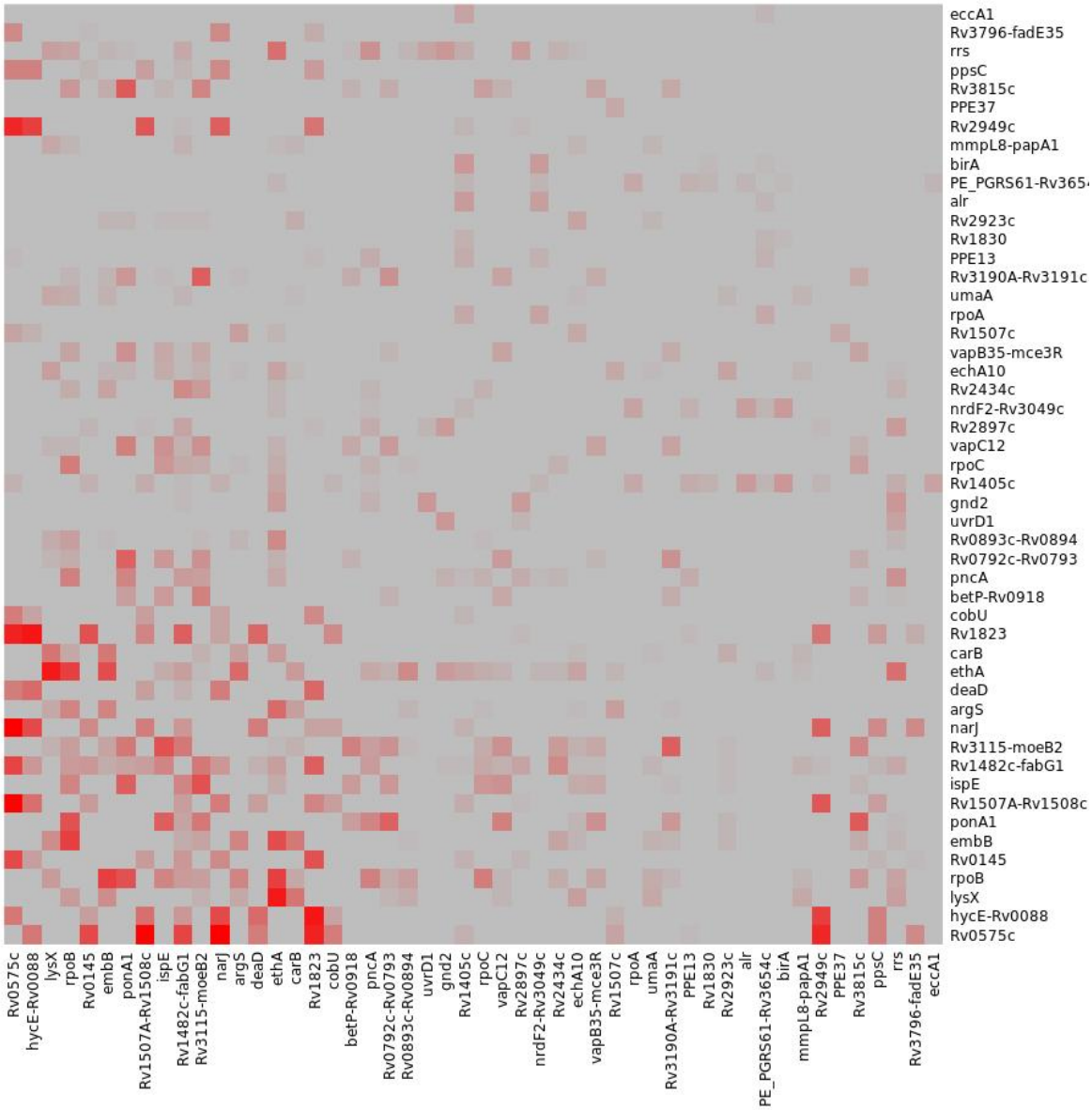
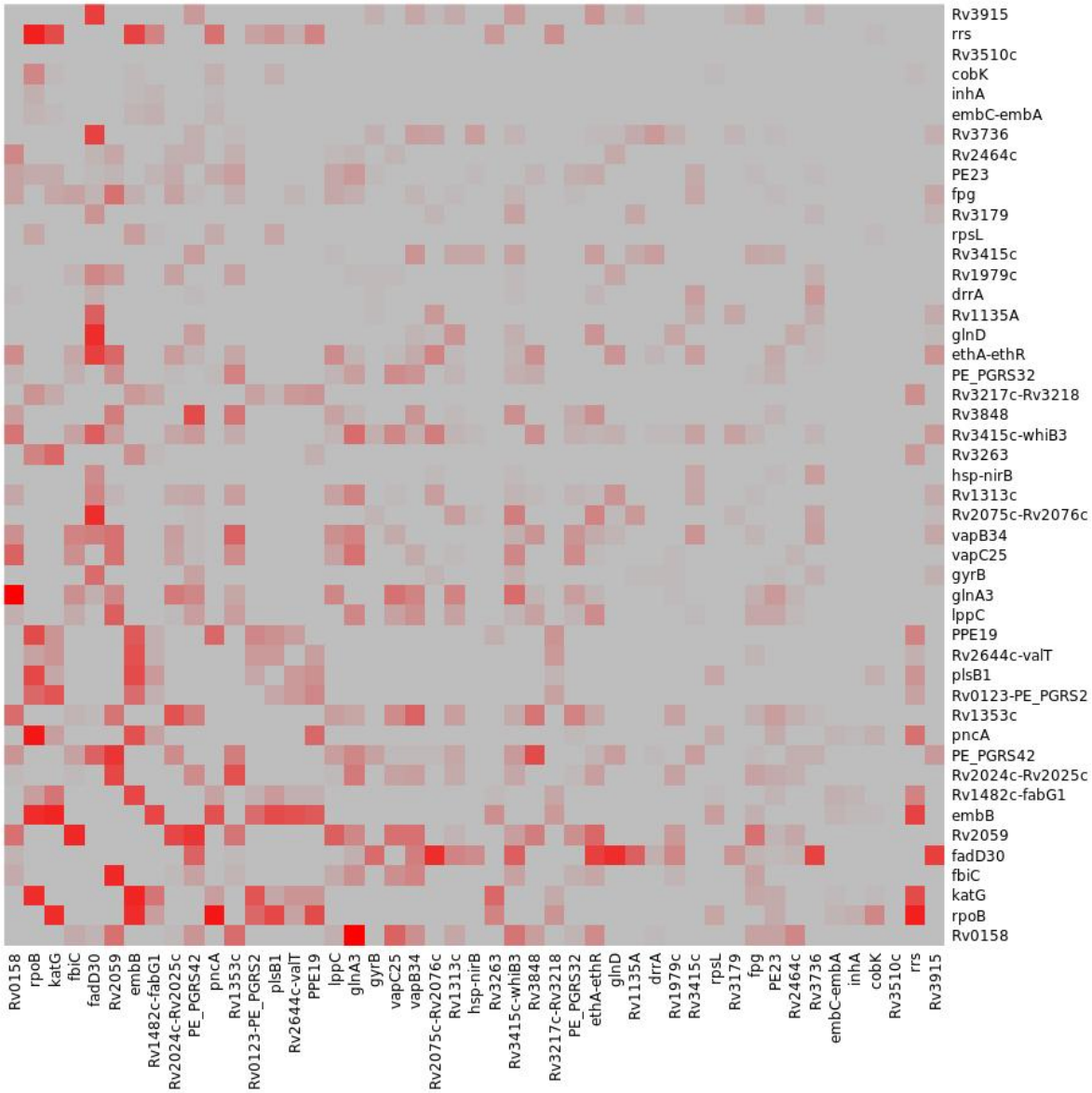


Figure 2 (cont.)

B-lineage 4



Supplementary Figure 1

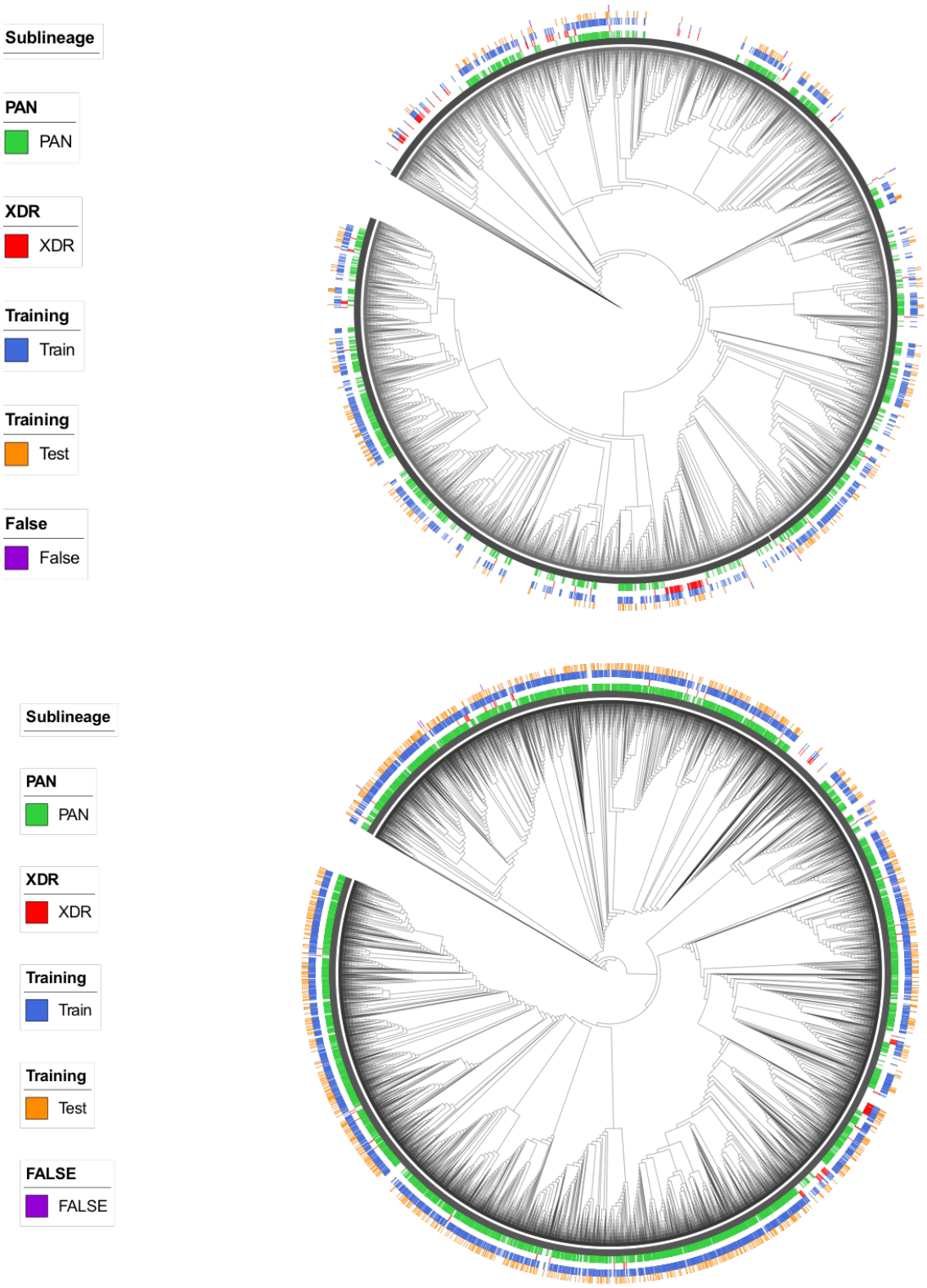


Table 1

Phenotype	Sensitivity	Specificity	Number in Test Set	Number of Resistant in Test Set	Lineage
XDRvPAN	100.0%	99.5%	1163	31	4
XDRvPAN	93.9%	98.6%	310	33	2

Table 2

Locus table lin2 XDR

Frequency	**Macro Frequency**	**Locus Name**	**Known**	**Min. Rule Length**	**Median Rule Length**	**Max. Rule Length**	**Companions**	**Function**
186	60	Rv0575c	N	1	3	5	*deaD;Rv1823;Rv1507A-Rv1508c;mce1D;ppsC;hycE-Rv0088;Rv0145;Rv1482c-fabG1;narJ;Rv2949c;Rv1405c;Rv2897c;cobU;gmhB;Rv3796-fadE35;dhaA;Rv1566c-Rv1567c;PPE13;Rv1507c*	Possible oxidoreductase, similar to many diverse oxidoreductases and monooxygenases

186	55	Rv1482c-fabG1	Y	1	2	5	*Rv1507c;Rv1507A-Rv1508c;ppsC;hycE-Rv0088;Rv0145;Rv0575c;deaD;rrs;Rv1823;Rv2949c;Rv1405c;narJ;Rv1877;ispE;lysX;carB;PPE37;rskA;vapB35-mce3R;gnd2;Rv2923c;echA10;cyp141-Rv3122;umaA;mmpL8-papA1;Rv3115-moeB2;vapC12;embB;Rv2434c;rpoC;ponA1;Rv2897c;ethA;rpoB;pncA*	Uncharacterized protein; Catalyzes the NADPH-dependent reduction of beta-ketoacyl-ACP substrates to beta-hydroxyacyl-ACP products, the first reductive step in the elongation cycle of fatty acid biosynthesis

174	47	rpoB	Y	1	3	4	*rpoC;ethA;Rv0893c-Rv0894;ponA1;vapC12;Rv3190A-Rv3191c;Rv0792c-Rv0793;Rv3115-moeB2;Rv3815c;vapB35-mce3R;embB;lysX;argS;ispE;carB;mmpL8-papA1;Rv1877;Rv2434c;umaA;rrs;pncA;Rv1482c-fabG1*	DNA-directed RNA polymerase subunit beta- DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates
167	51	ethA	Y	1	2	4	*rpoC;rpoB;Rv0893c-Rv0894;echA10;lysX;rrs;birA;Rv2897c;pncA;PPE13;Rv1830;PPE37;argS;carB;embB;Rv079*	FAD-containing monooxygenase-monooxygenase able to convert a wide range of ketones to the corresponding esters or lactones via a

							2c-Rv0793;vapC12;Rv3115-moeB2;mmpL8-papA1;Rv1507c;PE_PGRS61-Rv3654c;nrdF2-Rv3049c;Rv2434c;ispE;Rv1405c;Rv1482c-fabG1;gnd2	Baeyer-Villiger oxidation reaction. C
126	51	Rv1823	N	1	3	5	*deaD;Rv0575c;Rv1507A-Rv1508c;Rv3796-fadE35;narJ;Rv0145;ppsC;Rv1482c-fabG1;Rv2949c;PPE13;cobU;hycE-Rv0088;dhaA;Rv1566c-Rv1567c;Rv2897c*	Conserved protein

121	40	hycE-Rv0088	N	1	3	5	*pepE;ppsC;Rv0145;Rv0575c;Rv1482c-fabG1;narJ;Rv2949c;pks16-pth;deaD;cobU;Rv1507A-Rv1508c;Rv1823;Rv3796-fadE35;dhaA;Rv1566c-Rv1567c;PPE13;Rv1507c*	Possible formate hydrogenase HycE; Possible polyketide cyclase/dehydrase. Belongs to the SRPBCC ligand-binding domain superfamily. Predicted to be an outer membrane protein.
113	33	rrs	Y	1	2	4	*vapB35-mce3R;Rv2923c;ethA;rpoC;birA;Rv2949c;echA10;ispE;ponA1;Rv3815c;Rv1482c-fabG1;senX3-regX3;Rv1830;alr;carB;embB;*	Ribosomal RNA 16S

							Rv3466;betP-Rv0918;Rv0893c-Rv0894;lysX;pncA;Rv2434c;Rv1405c;rpoB;uvrD1;Rv2897c;gnd2	
110	42	narJ	N	1	3	4	*Rv3796-fadE35;Rv0145;Rv1823;ppsC;Rv0575c;hycE-Rv0088;Rv2949c;deaD;Rv1507A-Rv1508c;Rv1482c-fabG1;Rv2897c;cobU;PPE13;Rv1405c*	Probable respiratory nitrate reductase delta chain

110	44	Rv3115-moeB2	N	2	3	5	*Rv3190A-Rv3191c;argS;Rv3815c;ponA1;rpoB;echA10;embB;ispE;accD4;lysX;carB;vapB35-mce3R;Rv0792c-Rv0793;rpoC;Rv0893c-Rv0894;ethA;Rv2923c;betP-Rv0918;pncA;vapC12;Rv2434c;Rv1482c-fabG1*	Probable IS1081 transposase; Probable moeB2,molybdopterin cofactor biosynthesis protein
105	32	embB	Y	1	3	4	*pncA;mmpL8-papA1;Rv3115-moeB2;Rv3815c;ponA1;Rv0893c-Rv0894;rpoB;rrs;Rv0792c-*	Probable arabinosyltransferase B; Arabinosyl transferase responsible for the polymerization of arabinose into the arabinan of arabinogalactan

							Rv0793;echA10;argS;lysX;carB;ethA;rpoC;Rv2923c;Rv3190A-Rv3191c;umaA;Rv1482c-fabG1;Rv1877;Rv2434c	
105	41	ponA1	N	2	3	5	*vapC12;rpoB;Rv0792c-Rv0793;Rv3190A-Rv3191c;Rv3115-moeB2;Rv3815c;ispE;rrs;embB;lysX;betP-Rv0918;rpoC;pncA;vapB35-mce3R;accD4;Rv2923c;echA10;Rv1482c-fabG1*	Penicillin-binding protein 1A- cell wall formation. Synthesis of cross-linked peptidoglycan from the lipid intermediates. The enzyme has a penicillin-insensitive transglycosylase N-terminal domain (formation of linear glycan strands) and a penicillin-sensitive transpeptidase C-terminal domain (cross-linking of the peptide

								subunits) (By similarity). Has little peptidoglycan hydrolytic activity, however it inhibits the synergistic peptidoglycan hydrolysis of RipA plus RpfB
100	23	pncA	Y	1	2	3	*embB;ethA;ispE;argS;Rv1877;ponA1;nrdF2-Rv3049c;Rv1405c;Rv3115-moeB2;Rv2434c;gnd2;rpoC;rrs;PPE13;Rv2897c;rpoB;Rv1482c-fabG1*	Nicotinamidase/pyrazinamidase- catalyzes the deamidation of nicotinamide (NAM) into nicotinate. Likely functions in the cyclical salvage pathway for production of NAD from nicotinamide
96	34	ispE	N	1	3	5	*vapB35-mce3R;Rv0792c-Rv0793;betP-*	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase- catalyzes the

							Rv0918;accD4;rrs;ponA1;Rv3815c;vapC12;Rv3115-moeB2;Rv2434c;Rv3190A-Rv3191c;pncA;Rv1877;Rv1482c-fabG1;Rv2923c;rpoB;rpoC;ethA;echA10	phosphorylation of the position 2 hydroxy group of 4-diphosphocytidyl-2C-methyl-D-erythritol. Belongs to the GHMP kinase family.
87	42	Rv1507A-Rv1508c	N	1	3	5	*deaD;Rv1823;Rv0575c;Rv1482c-fabG1;ppsC;Rv2949c;Rv3796-fadE35;narJ;Rv0145;cobU;hycE-Rv0088;Rv1566c-Rv1567c;dhaA;PPE13;Rv1405*	Uncharacterized protein; Probable membrane protein

							c;Rv2897c	
86	21	Rv1405c	N	1	2	3	*Rv0575c;Rv2949c;Rv0145;Rv1482c-fabG1;pncA;narJ;Rv1507A-Rv1508c;PE_PGRS61-Rv3654c;nrdF2-Rv3049c;cobU;Rv1830;rrs;PPE13;rpoA;ethA;eccA1;alr;birA*	Uncharacterized protein
82	30	Rv2949c	N	1	3	4	*Rv0575c;hycE-Rv0088;narJ;Rv0145;rrs;Rv1405c;deaD;Rv1823;Rv1507A-Rv1508c;Rv1482c-*	Chorismate pyruvate-lyase; Removes the pyruvyl group from chorismate to provide 4- hydroxybenzoate (4HB). Involved in the synthesis of

							fabG1;Rv3796-fadE35;ppsC;PPE13;cobU;Rv2897c	glycosylated p-hydroxybenzoic acid methyl esters (p-HBADs) and phenolic glycolipids (PGL) that play important roles in the pathogenesis of mycobacterial infections
77	10	gnd2	N	1	2	2	*Rv1482c-fabG1;pncA;senX3-regX3;Rv0042c;Rv2897c;ethA;alkB-Rv3253c;rrs;uvrD1*	Probable gnd2,6-phosphogluconate dehydrogenase, decarboxylating
77	29	lysX	N	2	3	4	*echA10;ethA;carB;umaA;Rv0792c-Rv0793;vapC12;mmpL8-papA1;Rv1482c-fabG1;argS;rpoB;ponA1;Rv31*	Lysylphosphatidylglycerol biosynthesis bifunctional protein- catalyzes the production of L-lysyl-tRNA(Lys)transfer and the transfer of a lysyl group from L-lysyl-tRNA(Lys) to membrane-bound

							15-moeB2;embB;rpoC;Rv3656c;rrs;Rv0893c-Rv0894	phosphatidylglycerol (PG), which produces lysylphosphatidylglycerol (LPG), one of the components of the bacterial membrane with a positive net charge. LPG synthesis contributes to the resistance to cationic antimicrobial peptides (CAMPs)
66	12	Rv1507c	N	1	2	2	*Rv1482c-fabG1;Rv0302;ethA;hycE-Rv0088;pepE;PPE37;echA10;mce1D;Rv0575c;argS;gmhB*	Conserved protein
63	28	Rv0145	N	1	3	5	*Rv3796-fadE35;narJ;Rv1823;ppsC;hyc*	Putative S-adenosyl-L-methionine-dependent methyltransferase

							E-Rv0088;Rv0575c;Rv1482c-fabG1;Rv2949c;Rv2897c;Rv1507A-Rv1508c;Rv1405c;Rv1566c-Rv1567c;deaD;cobU	
58	22	carB	N	1	3	4	*lysX;umaA;rrs;Rv1482c-fabG1;Rv3115-moeB2;embB;ethA;rpoB;argS;Rv2923c;echA10;mmpL8-papA1;cyp141-Rv3122*	Carbamoyl-phosphate synthase large chain
55	23	argS	N	1	3	4	*Rv3190A-Rv3191c;Rv3115-moeB2;echA10;ethA;pncA;lysX;rpoB;pepE;gmhB;embB;bet*	Arginine--tRNA ligase

							P-Rv0918;Rv0893c-Rv0894;rpoC;carB;Rv1507c	
55	13	Rv2897c	N	1	2	3	*ethA;Rv0575c;Rv0145;narJ;deaD;senX3-regX3;Rv1507A-Rv1508c;Rv2949c;Rv1823;uvrD1;pncA;Rv1482c-fabG1;rrs;gnd2*	Conserved hypothetical protein, possibly Mg-chelatase
51	23	echA10	N	1	2	4	*ethA;lysX;mce1D;Rv3115-moeB2;rrs;argS;PPE37;embB;rpoC;carB;Rv1482c-fabG1;umaA;mmpL8-papA1;Rv2923c;ponA1;ispE;Rv1507c*	Probable enoyl-CoA hydratase

51	12	Rv2434c	N	1	2	3	*ispE;fadE28;pncA;ethA;rpoC;Rv3115-moeB2;rrs;rpoB;Rv1482c-fabG1;embB*	Probable conserved transmembrane protein
50	23	deaD	N	1	3	4	*Rv1823;Rv0575c;Rv1507A-Rv1508c;hycE-Rv0088;Rv1482c-fabG1;Rv2949c;narJ;Rv3796-fadE35;Rv2897c;ppsC;Rv0145*	ATP-dependent RNA helicase- DEAD-box RNA helicase involved in various cellular processes at low temperature, including ribosome biogenesis, mRNA degradation and translation initiation
47	9	nrdF2-Rv3049c	N	1	2	2	*pncA;PE_PGRS61-Rv3654c;ethA;Rv1405c;PPE13;rpoA;alr;birA*	Ribonucleoside-diphosphate reductase subunit beta- provides the precursors necessary for DNA synthesis. Catalyzes the biosynthesis

								of deoxyribonucleotides from the corresponding ribonucleotides. Two genes for this protein are present in M.tuberculosis- this is the active form; Probable monooxygenase
47	23	vapC12	N	1	3	4	*ponA1;rpoB;Rv3190A-Rv3191c;Rv0792c-Rv0793;betP-Rv0918;ispE;lysX;Rv3815c;vapB35-mce3R;ethA;Rv3115-moeB2;Rv1482c-fabG1*	Ribonuclease- toxic component of a type II toxin-antitoxin (TA) system. An RNase
46	23	Rv3190A-	N	1	3	4	*vapC12;rpoB;Rv0792c-Rv0793;Rv3115-*	Conserved protein; Probable transposase

		Rv3191c					*moeB2;argS;ponA1;Rv3815c;ispE;vapB35-mce3R;betP-Rv0918;accD4;embB*	
45	24	Rv0792c-Rv0793	N	1	3	5	*vapC12;Rv3190A-Rv3191c;rpoB;ponA1;Rv3815c;vapB35-mce3R;ispE;betP-Rv0918;accD4;lysX;embB;Rv3115-moeB2;ethA*	Probable transcriptional regulatory protein (Probably GntR-family); Putative monooygenase that might be involved in antibiotic biosynthesis, or may act as reactive oxygen species scavenger that could help in evading host defenses
44	22	Rv3815c	N	2	3	4	*Rv3115-moeB2;ponA1;rpoB;Rv0792c-Rv0793;Rv3190A-*	Possible acyltransferase

							Rv3191c;ispE;rrs;embB;betP-Rv0918;vapC12;vapB35-mce3R;rpoC	
42	17	rpoC	Y	2	3	4	*rpoB;ethA;rrs;lysX;argS;Rv0893c-Rv0894;embB;Rv3815c;ponA1;Rv3115-moeB2;echA10;ispE;Rv2434c;pncA;Rv1482c-fabG1*	DNA-directed RNA polymerase subunit beta'- DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates
40	19	ppsC	N	1	3	5	*Rv1507A-Rv1508c;hycE-Rv0088;Rv0145;Rv0575c;Rv1482c-fabG1;narJ;Rv1823;Rv1566c-*	Phthiocerol synthesis polyketide synthase type I PpsC- involved in the elongation of either C22-24 fatty acids by the addition of malonyl-CoA and

							Rv1567c;deaD;cobU;Rv2949c	methylmalonyl-CoA extender units to yield phthiocerol derivatives
38	10	PE_PGRS61-Rv3654c	N	1	2	2	*birA;Rv1830;eccA1;alr;ethA;nrdF2-Rv3049c;Rv1405c;PPE13;rpoA*	PE-PGRS family protein- mediates Ca(2+)-dependent up-regulation of the anti- inflammatory cytokine IL-10; Apoptosis inhibitor- effector protein that participates in the suppression of macrophage apoptosis by blocking the extrinsic pathway
35	9	birA	N	1	2	2	*ethA;rrs;cobU;alr;PE_PGRS61-Rv3654c;Rv1830;nrdF2-Rv3049c;Rv1405c*	Possible bifunctional protein: biotin operon repressor and biotin--[acetyl-CoA-carboxylase] synthetase
33	15	cobU	N	1	3	4	*hycE-*	Bifunctional cobinamide

							Rv0088;Rv0575c;birA;Rv1507A-Rv1508c;Rv1823;narJ;ppsC;Rv0145;Rv2949c;Rv1405c	kinase/cobinamide phosphate guanylyltransferase
32	11	PPE13	N	1	2	4	*ethA;Rv1823;Rv3796-fadE35;eccA1;hycE-Rv0088;Rv1507A-Rv1508c;Rv0575c;Rv2949c;narJ;nrdF2-Rv3049c;PE_PGRS61-Rv3654c;pncA;Rv1405c*	Uncharacterized PPE family protein
31	18	vapB35-	N	1	3	5	*rrs;ispE;Rv0792c-Rv0793;betP-*	Antitoxin component of a type II toxin-antitoxin (TA) system; Transcriptional

		mce3R					*Rv0918;accD4;rpoB;Rv3190A-Rv3191c;vapC12;Rv3815c;Rv3115-moeB2;Rv1482c-fabG1;ponA1*	repressor- represses the transcription of mce3 operon and downregulates its own expression, but does not affect the transcription of mce1, mce2 and mce4 operons
30	16	betP-Rv0918	N	1	3	5	*vapB35-mce3R;ispE;Rv0792c-Rv0793;accD4;vapC12;Rv3190A-Rv3191c;Rv3815c;argS;ponA1;rrs;Rv3115-moeB2*	Uncharacterized transporter; Conserved protein
30	8	rpoA	Y	1	2	2	*eccA1;Rv1566c-Rv1567c;Rv1830;alr;Rv1405c;PE_PGRS61-Rv3654c;nrdF2-*	DNA-directed RNA polymerase subunit alpha- DNA-dependent RNA polymerase catalyzes the transcription

							Rv3049c	of DNA into RNA using the four ribonucleoside triphosphates as substrates
28	11	umaA	N	1	2	3	*lysX;carB;mmpL8-papA1;echA10;Rv1482c-fabG1;Rv2923c;embB;rpoB*	S-adenosylmethionine-dependent methyltransferase- methyltransferase that modifies short-chain fatty acids. In vitro, catalyzes the transfer of the methyl group from S- adenosyl-L-methionine (SAM) to the double bond of phospholipid- linked oleic acid to produce tuberculostearic acid (10-methylstearic-acid or TSA)
27	7	PPE37	N	1	2	2	*ethA;Rv1482c-*	Uncharacterized PPE family protein

							fabG1;echA10;mce1D;Rv1507c;Rv3466	
27	5	uvrD1	N	1	2	2	*Rv0042c;Rv2897c;rrs;gnd2*	ATP-dependent DNA helicase- DNA-dependent ATPase, acting on dsDNA with a 3'-ssDNA tail, unwinding with 3'-to 5'-polarity. A minimal tail of 18 nt is required for activity. Also highly efficient on nicked DNA. Involved in the post-incision events of nucleotide excision repair, as well as in nitrosative and oxidative stress response and possibly in persistence in the host. Inhibits RecA-mediated

								DNA strand exchange
26	7	alr	Y	1	2	2	*rrs;birA;rpoA;PE_PGRS61-Rv3654c;nrdF2-Rv3049c;Rv1405c*	Alanine racemase; Catalyzes the interconversion of L-alanine and D-alanine. D-alanine plays a key role in peptidoglycan cross- linking
26	10	Rv2923c	N	1	2	3	*rrs;ispE;carB;Rv3115-moeB2;Rv1482c-fabG1;embB;ponA1;echA10;umaA*	Conserved protein

25	10	mmpL8-papA1	N	1	2	3	*embB;lysX;umaA;ethA;echA10;rpoB;carB;Rv1482c-fabG1*	Sulfolipid-1 exporter- required for the biosynthesis and the transport across the inner membrane of sulfolipid-1 (SL-1), which is a major cell wall lipid of pathogenic mycobacteria; SL659 acyltransferase
21	9	Rv0893c-Rv0894	N	2	3	4	*rpoB;ethA;embB;argS;rpoC;Rv3115-moeB2;rrs;lysX*	Putative S-adenosyl-L-methionine-dependent methyltransferase; Uncharacterized protein
21	8	Rv1830	N	1	2	2	*eccA1;rrs;ethA;rpoA;birA;PE_PGRS61-Rv3654c;Rv1405c*	Uncharacterized HTH-type transcriptional regulator
21	11	Rv379	N	1	3	4	*narJ;Rv0145;Rv1823;deaD;PP*	Conserved protein; Probable acyl-CoA

		6-fadE35					*E13;Rv1507A-Rv1508c;Rv0575c;hycE-Rv0088;Rv2949c*	dehydrogenase
18	6	eccA1	N	1	2	2	*Rv1830;rpoA;PPE13;PE_PGRS61-Rv3654c;Rv1405c*	ESX-1 secretion system protein- EccA1 exhibits ATPase activity and may provide energy for the export of ESX-1 substrates
18	4	senX3-regX3	N	1	2	2	*rrs;Rv2897c;gnd2*	Sensor-like histidine kinase- probably forms part of a two-component regulatory system senX3/regX3. Phosphorylates regX3 (Probable); Sensory transduction protein regX3- probably forms part of a two-

								component regulatory system regX3/senX3
17	5	gmhB	N	1	2	2	*Rv0575c;argS;Rv0302;Rv1507c*	D-glycero-alpha-D-manno-heptose-1,7-bisphosphate 7-phosphatase-converts the D-glycero-alpha-D-manno-heptose 1,7- bisphosphate intermediate into D-glycero-alpha-D-manno-heptose 1- phosphate by removing the phosphate group at the C-7 position
17	6	mce1D	N	1	2	2	*Rv0575c;echA10;pepE;PPE37;Rv1507c*	Mce-family protein
15	2	alkB-	N	1	1.5	2	*gnd2*	Probable transmembrane alkane-1-

		Rv3253c						monooxygenase; Possible cationic amino acid transporter, integral membrane protein
15	3	Rv3466	N	1	2	2	*rrs;PPE37*	Uncharacterized protein
13	3	Rv0042c	N	1	2	2	*uvrD1;gnd2*	Possible transcriptional regulatory protein, MarR-famil
13	4	Rv1877	N	2	2.5	3	*ispE;Rv1482c-fabG1;pncA;rpoB;embB*	Uncharacterized MFS-type transporter, similar to many antibiotic and drug efflux proteins
11	1	fadE1	N	1	1	1	*NA*	Probable acyl-CoA dehydrogenase
11	5	pepE	N	1	2	2	*hycE-Rv0088;mce1D;argS;Rv1507c*	Probable dipeptidase

11	1	Rv1922	N	1	1	1	*NA*	Probable conserved lipoprotein, possibly peptidase similar to many peptidases
10	1	cysE	N	1	1	1	*NA*	Serine acetyltransferase- catalyzes the acetylation of serine by acetyl-CoA to produce O-acetylserine (OAS)
10	1	dppB	N	1	1	1	*NA*	Probable dipeptide-transport integral membrane protein ABC-transporter (see citation below), similar to many peptide permeases
10	1	lppC	N	1	1	1	*NA*	Putative lipoprotein LppC- probably involved in bacterial recognition and uptake by its host

10	1	lppF	N	1	1	1	*NA*	Probable conserved lipoprotein
10	1	mbtI	N	1	1	1	*NA*	Salicylate synthase- involved in the incorporation of salicylate into the virulence-conferring salicylate-based siderophore mycobactin
10	1	moaA 1	N	1	1	1	*NA*	GTP 3',8-cyclase 1- catalyzes the cyclization of GTP to (8S)-3',8-cyclo-7,8- dihydroguanosine 5'-triphosphate
10	1	rpiB	N	1	1	1	*NA*	Ribose-5-phosphate isomerase B- catalyzes the interconversion of ribulose-5-P and ribose-5-P
10	1	Rv045 8	N	1	1	1	*NA*	Probable aldehyde dehydrogenase

10	1	Rv0584	N	1	1	1	*NA*	Uncharacterized glycosidase, possible conserved exported protein
10	1	Rv3740c	N	1	1	1	*NA*	Putative diacyglycerol O-acyltransferase
10	1	Rv3742c-ctpJ	N	1	1	1	*NA*	Possible oxidoreductase, probably combines with product of downstream ORF MTV025.090c to form a functional monooxygenase; Probable cation-transporting P-type ATPase
9	1	aceE	N	1	1	1	*NA*	Pyruvate dehydrogenase E1 component
9	1	AS172	N	1	1	1	*NA*	Putative small regulatory RNA

		6						
9	2	cyp141-Rv3122	N	2	2	2	*Rv1482c-fabG1;carB*	Putative cytochrome P450 141; Hypothetical unknown protein
9	1	dgt-Rv2345	N	1	1	1	*NA*	Deoxyguanosinetriphosphate triphosphohydrolase-like protein; Probable dgt, deoxyguanosine triphosphate triphosphohydrolase,
9	1	eccCb1-PE35	N	1	1	1	*NA*	ESX-1 secretion system protein - EccCb1 may link the cytosolic components of the system with the membrane components; PE family

								immunomodulator- plays a major role in RD1-associated pathogenesis, and may contribute to the establishment and maintenance of M.tuberculosis infection. Together with PPE68, stimulates the secretion of IL-10 and MCP-1 from human macrophages, via the interaction with human Toll-like receptor 2 (TLR2)
9	1	mce1B	N	1	1	1	*NA*	Mce-family protein
9	3	Rv0302	N	1	2	2	*Rv1507c;gmhB*	Probable transcription regulatory protein, TetR family
9	1	Rv052	N	1	1	1	*NA*	Probable conserved transmembrane

		8						protein
9	1	Rv2717c	N	1	1	1	*NA*	UPF0678 fatty acid-binding protein-like protein, may play a role in the intracellular transport of hydrophobic ligands
9	1	Rv3686c	N	1	1	1	*NA*	Uncharacterized protein
8	6	accD4	N	3	4	5	*vapB35-mce3R;ispE;Rv0792c-Rv0793;betP-Rv0918;Rv3115-moeB2;Rv3190A-Rv3191c;ponA1*	Probable accD4,propionyl-CoA carboxylase beta chain 4
8	1	lprI	N	1	1	1	*NA*	Lipoprotein- strongly binds and inhibits lysozyme, may help bacteria

								survive in lysozyme-producing host cells
8	1	moaR1-PPE49	N	1	1	1	*NA*	Transcriptional regulatory protein-acts as a positive transcriptional regulator of the molybdopterin biosynthesis moa1 locus, promoting the expression of the moaA1B1C1D1 genes. Binds directly to the moaA1 promoter; Uncharacterized PPE family protein
8	5	Rv1566c-Rv156	N	2	3	4	*rpoA;Rv0145;Rv1507A-Rv1508c;ppsC;Rv1823;Rv0575c;hycE-Rv0088*	Possible Inv protein; Probable membrane protein

		7c						
8	1	Rv1632c-uvrB	N	1	1	1	*NA*	Uncharacterized protein; UvrABC system protein B- the UvrABC repair system catalyzes the recognition and processing of DNA lesions
8	1	Rv3639c	N	1	1	1	*NA*	Uncharacterized protein
8	1	snoP	N	1	1	1	*NA*	Pyridoxal 5'-phosphate synthase subunit- catalyzes the hydrolysis of glutamine to glutamate and ammonia as part of the biosynthesis of pyridoxal 5'-phosphate. The resulting ammonia molecule is channeled to the active

								site of PdxS
7	1	ftsH	N	1	1	1	*NA*	ATP-dependent zinc metalloprotease FtsH; Acts as a processive, ATP-dependent zinc metallopeptidase for both cytoplasmic and membrane proteins. Plays a role in the quality control of integral membrane proteins
7	1	katG-furA	N	1	1	1	*NA*	Catalase-peroxidase; Bifunctional enzyme with both catalase and broad-spectrum peroxidase activity, oxidizing various electron donors including NADP(H). Protects M.tuberculosis

								against toxic reactive oxygen species (ROS) including hydrogen peroxide as well as organic peroxides and thus contributes to its survival within host macrophages by countering the phagocyte oxidative burst; Transcriptional regulator- represses transcription of the catalase-peroxidase gene katG and its own transcription by binding to the promoter region in a redox-dependent manner
7	1	lldD1-	N	1	1	1	*NA*	Putative mycofactocin system

		Rv0695						heme/flavin oxidoreductase MftD; Putative mycofactocin system creatinine amidohydrolase family protein
7	1	Rv0140	N	1	1	1	*NA*	Conserved protein
7	1	Rv0308	N	1	1	1	*NA*	Probable conserved integral membrane protein, with C-terminus highly similar to C-terminus of other integral membrane proteins or phosphatases
7	1	Rv0939	N	1	1	1	*NA*	Possible bifunctional enzyme, including 2-hydroxyhepta-2,4-diene-

								1,7-dioate isomerase activity, and cyclase/dehydrase activity
7	1	Rv2230c	N	1	1	1	*NA*	GTP cyclohydrolase 1 type 2 homolog;
6	1	Rv0281	N	1	1	1	*NA*	Putative S-adenosyl-L-methionine-dependent methyltransferase
6	1	Rv1366	N	1	1	1	*NA*	Uncharacterized protein Rv1366; Rv1366, (MTCY02B10.30), len: 273 aa. Hypothetical unknown protein
6	1	Rv2013	N	1	1	1	*NA*	Transposase
6	1	Rv3023c	N	1	1	1	*NA*	Probable IS1081 transposase

5	1	frr-pyrH	N	1	1	1	*NA*	Ribosome-recycling factor- responsible for the release of ribosomes from messenger RNA at the termination of protein biosynthesis. May increase the efficiency of translation by recycling ribosomes from one round of translation to another; Uridylate kinase- catalyzes the reversible phosphorylation of UMP to UDP
5	1	PE_PGRS45-Rv261	N	1	1	1	*NA*	PE-PGRS family protein PE_PGRS45; Conserved protein

		6						
5	1	Rv1534-Rv1535	N	1	1	1	*NA*	Probable transcriptional regulator; Uncharacterized protein
4	2	Rv3656c	N	1	1.5	2	*lysX*	Uncharacterized protein
3	1	Rv2324-Rv2325c	N	1	1	1	*NA*	Probable transcriptional regulatory protein; Uncharacterized protein
1	1	cyp138	N	1	1	1	*NA*	Putative cytochrome P450 138

1	1	dhaA	N	5	5	5	*Rv1507A-Rv1508c;Rv0575c;hycE-Rv0088;Rv1823*	Haloalkane dehalogenase 3- catalyzes hydrolytic cleavage of carbon-halogen bonds in halogenated aliphatic compounds
1	1	espB	N	1	1	1	*NA*	ESX-1 secretion-associated protein- required for host-cell death and may support an EsxA- independent virulence function. Secreted processed form of EspB binds to phosphatidic acid and phosphatidylserine. Inhibits IFN-gamma-induced autophagy in murine macrophages

1	1	fadE28	N	2	2	2	*Rv2434c*	Acyl-CoA dehydrogenase- involved in the third cycle of side chain dehydrogenation in the beta-oxidation of cholesterol catabolism. May play an important role for the initial macrophage invasion, possibly in response to the acidification of phagosome. It contributes partly to the virulence by increasing the efficiency of beta-oxidation.
1	1	mmaA 2-mmaA	N	1	1	1	*NA*	Cyclopropane mycolic acid synthase. Cyclopropanated mycolic acids are key factors participating in cell envelope

		1						permeability, host immunomodulation and persistence; Mycolic acid methyltransferase
1	1	PE16	N	1	1	1	*NA*	Member of the Mycobacterium tuberculosis PE family of proteins
1	1	pks16-pth	N	2	2	2	*hycE-Rv0088*	Putative ligase; Peptidyl-tRNA hydrolase- the natural substrate for this enzyme may be peptidyl- tRNAs which drop off the ribosome during protein synthesis; Belongs to the PTH family
1	1	rskA	N	2	2	2	*Rv1482c-fabG1*	An anti-sigma factor for extracytoplasmic function (ECF) sigma

								factor SigK. ECF sigma factors are held in an inactive form by an anti-sigma factor until released by regulated intramembrane proteolysis (RIP)
1	1	Rv0052	N	1	1	1	*NA*	Conserved protein
1	1	Rv0740-Rv0741	N	1	1	1	*NA*	Uncharacterized protein; Probable transposase
1	1	Rv1188	N	1	1	1	*NA*	Probable proline dehydrogenase
1	1	Rv305	N	1	1	1	*NA*	Probable oxidoreductase, probably

		7c						short-chain alcohol dehydrogenase/reductase

Table 3

Locus table for lineage 4

Freq-uency	Macro Frequen-cy	Locus Name	Known	Min. Rule Length	Median Rule Length	Max. Rule Length	Companions	Function
226	67	fadD30	N	1	3	5	fpg;Rv1313c;Rv3415c-whiB3;Rv3179;glnD;gyrB;PE23;cmr-Rv1676;lppC;Rv2075c-Rv2076c;vapB34;Rv3915;Rv3415c;ethA-ethR;Rv0158;glnA3;fbiC;Rv2024c-Rv2025c;drrA;Rv1353c;vapC25;Rv2059;PE_PGRS42;Rv3736;Rv3848;Rv11	Long-chain-fatty-acid--AMP ligase- catalyzes the activation of long-chain fatty acids as acyl-adenylates

							35A;hsp-nirB;Rv1979c;Rv2464c;smc;PE_PGRS30	
192	63	rpoB	Y	1	3	4	pncA;rpsL;Rv3217c-Rv3218;rrs;katG;embB;Rv2644c-valT;eccE2;plsB1;PPE19;Rv0123-PE_PGRS2;cobK;Rv3263;mctB;PPE61-PPE62;mcr7;embC-embA;inhA;PE23;Rv1482c-fabG1	DNA-directed RNA polymerase subunit beta-DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates
166	77	Rv2059	N	1	3	5	Rv3848;ethA-ethR;Rv1353c;fpg;fbiC;Rv0158;PE_PGRS32;vapC25;Rv2024c-Rv2025c;Rv1313c;Rv3915;Rv3415c;v	Uncharacterized protein

							apB34;Rv1979c;Rv3415c-whiB3;PE23;Rv2464c;PE_PGRS42;glnA3;lppC;Rv1135A;fadD30	
152	60	embB	Y	2	3	5	Rv1482c-fabG1;Rv2644c-valT;rpsL;Rv3263;rpoB;plsB1;rrs;Rv2074;katG;PPE19;Rv0123-PE_PGRS2;pncA;cobK;Rv3217c-Rv3218;mctB;mcr7;inhA;fpg;embC-embA	Probable arabinosyltransferase B
134	62	PE_PGRS42	N	1	3	5	Rv2024c-Rv2025c;Rv1353c;gyrB;Rv3736;drrA;glnD;PE_PGRS30;Rv0158;lppC;Rv3415c-whiB3;fbiC;fpg;Rv1313c;Rv2075c-	PE-PGRS family protein

							Rv2076c;glnA3;Rv0258c;vapC25;Rv2464c;Rv3848;Rv2059;vapB34;Rv3915;Rv3510c;Rv1135A;Rv0308-Rv0309;fadD30;ethA-ethR;Rv3910;Rv3415c	
130	45	katG	Y	2	3	5	Rv0123-PE_PGRS2;PPE19;Rv3217c-Rv3218;rpoB;rpsL;embB;eccE2;cobK;Rv2644c-valT;fpg;Rv1889c-Rv1890c;Rv1482c-fabG1;Rv2464c;plsB1;mcr7;rrs;embC-embA;Rv3263;PE23;pncA	Catalase-peroxidase- protects M.tuberculosis against toxic reactive oxygen species (ROS) including hydrogen peroxide as well as organic peroxides and thus contributes to its survival within host macrophages by countering the phagocyte

								oxidative burst
125	58	Rv0158	N	1	3	5	PE23;Rv2059;fpg;fbiC;Rv2464c;glnA3;Rv3910;vapC25;Rv1313c;Rv1353c;Rv3736;Rv2024c-Rv2025c;lppC;vapB34;Rv3848;PE_PGRS42;fadD30;ethA-ethR;Rv2075c-Rv2076c;Rv3415c-whiB3;PE_PGRS32;drrA	Probable transcriptional regulatory protein (Possibly TetR-family)
125	62	vapB34	N	1	3	5	Rv3848;Rv3415c-whiB3;PE23;Rv2024c-Rv2025c;PE_PGRS32;pncA;lppC;drrA;Rv1353c;panB-	Putative antitoxin VapB34-antitoxin component of a possible type II toxin- antitoxin (TA) system

							Rv2226;vapC25;Rv2059;fadD30;gyrB;Rv0158;gca-gmhA;fbiC;glnD;ethA-ethR;Rv3915;PE_PGRS42;glnA3;Rv1135A;Rv3179;fpg;Rv2028c;Rv3510c;Rv3736;Rv3415c	
124	46	ethA-ethR	Y	1	3	4	Rv3848;Rv2059;glnA3;drrA;Rv3915;glnD;PE_PGRS32;vapC25;lppC;Rv3736;fadD30;Rv0158;hsp-nirB;vapB34;Rv2075c-Rv2076c;Rv3415c-whiB3;Rv3510c;Rv1313c;fbiC;PE_PGRS42;Rv1353c;Rv2024c-Rv2025c;PE23;Rv3415c	FAD-containing monooxygenase; HTH-type transcriptional regulator involved in the repression of the monooxygenase EthA which is responsible of the formation of the active metabolite of ethionamide (ETH)

124	58	glnA3	N	1	3	5	ethA-ethR;Rv3910;Rv0158;vapC25;Rv1313c;Rv3415c-whiB3;Rv1353c;PE_PGRS32;fadD30;fbiC;PE_PGRS42;fpg;Rv3736;Rv2075c-Rv2076c;Rv2059;vapB34;Rv2024c-Rv2025c;Rv3848;lppC;Rv2464c;Rv1979c;PE23	Probable glutamine synthetase class I
122	64	Rv3415c-whiB3	N	1	3	5	fadD30;Rv3179;glnD;vapB34;Rv1979c;Rv3736;Rv3915;hsp-nirB;cmr-Rv1676;fpg;lppC;PE_PGRS42;fbiC;glnA3;Rv1353c;PE_PGRS32;PE23;Rv2059;vapC25;Rv2075c-Rv2076c;ethA-	Uncharacterized protein; Redox- and pH-responsive transcriptional regulator, leads to respiratory alterations and bioenergetic deficiencies that

							ethR;Rv3910;Rv2464c;smc;Rv1313c;Rv2024c-Rv2025c;Rv0158;Rv2028c;Rv1135A;Rv0326;gyrB;Rv3848;drrA;Rv3415c	negatively impact virulence
106	39	rrs	Y	2	3	4	Rv0123-PE_PGRS2;rpoB;Rv3217c-Rv3218;PE23;plsB1;embB;Rv2074;rpsL;mctB;pncA;embC-embA;cobK;PPE19;Rv2644c-valT;Rv1482c-fabG1;katG;Rv3263	Ribosomal RNA 16S
105	27	pncA	Y	1	3	5	rpoB;rpsL;PE_PGRS32;vapB34;rrs;embB;embC-embA;Rv3217c-Rv3218;Rv2644c-valT;mctB;cobK;inhA;Rv1482c-	Nicotinamidase/pyrazinamidase-catalyzes the deamidation of nicotinamide (NAM) into nicotinate

							fabG1;mmpL12-Rv1523;PPE19;PE23;katG	
105	59	Rv1353c	N	1	3	5	PE_PGRS42;Rv2024c-Rv2025c;Rv2059;PE23;Rv1313c;PE_PGRS32;lppC;Rv3848;vapB34;Rv0158;Rv3736;vapC25;glnA3;Rv3415c-whiB3;fbiC;fpg;Rv2464c;Rv1979c;fadD30;ethA-ethR	Probable transcriptional regulatory protein, belongs to the TetR/AcrR family of transcriptional regulators
91	19	Rv3415c	N	1	2	3	Rv3510c;Rv2059;fadD30;Rv3915;Rv3113-Rv3114;Rv2024c-Rv2025c;PE_PGRS30;lppC;Rv1135A;mkl-vapC6;PE23;hsp-nirB;fpg;Rv1313c;Rv3415c-	Uncharacterized protein

							whiB3;drrA;PE_PGRS42;ethA-ethR;vapB34	
88	28	Rv1482c-fabG1	Y	1	3	3	embB;Rv2644c-valT;Rv3263;Rv0123-PE_PGRS2;Rv3217c-Rv3218;katG;Rv2464c;rrs;PPE19;rpsL;pncA;mmpL12-Rv1523;plsB1;PE23;inhA;embC-embA;rpoB	Uncharacterized protein; 3-oxoacyl-[acyl-carrier-protein] reductase FabG1- catalyzes the NADPH-dependent reduction of beta- ketoacyl-ACP substrates to beta-hydroxyacyl-ACP products, the first reductive step in the elongation cycle of fatty acid biosynthesis
83	44	lppC	N	1	3	5	Rv1353c;vapB34;drrA;Rv3848;Rv1313c;fadD30;Rv2075c-	Putative lipoprotein, probably involved in bacterial recognition

							Rv2076c;PE_PGRS32;vapC25;ethA-ethR;Rv3415c-whiB3;fpg;Rv0158;PE_PGRS42;PE23;Rv0258c;Rv2024c-Rv2025c;fbiC;parE1;Rv2059;glnA3;Rv2464c;Rv3415c	and uptake by its host
81	45	Rv2024c-Rv2025c	N	2	3	5	PE_PGRS42;Rv1353c;PE23;Rv1313c;PE_PGRS32;vapB34;Rv2059;fpg;Rv0158;Rv3736;drrA;fbiC;lppC;vapC25;Rv2464c;Rv3415c;fadD30;glnA3;Rv3415c-whiB3;Rv1979c;gyrB;ethA-ethR	Uncharacterized protein; Probable cation efflux system protein
79	37	fbiC	N	1	3	5	Rv2059;fpg;Rv0158;Rv2024c-Rv2025c;gca-	FO synthase

							gmhA;vapB34;Rv3415c-whiB3;PE_PGRS42;vapC25;fadD30;glnA3;PE_PGRS32;Rv1353c;Rv3910;lppC;Rv2075c-Rv2076c;ethA-ethR;Rv0326;Rv1979c	
76	39	Rv3848	N	1	3	5	Rv2059;ethA-ethR;vapB34;Rv0021c-whiB5;Rv1353c;lppC;Rv0158;vapC25;Rv1979c;PE23;PE_PGRS32;PE_PGRS42;glnA3;fpg;Rv3915;fadD30;Rv3736;Rv3415c-whiB3;Rv3910;Rv1313c	Probable conserved transmembrane protein
71	31	PE23	N	1	2	4	Rv0158;Rv1353c;Rv2024c-Rv2025c;fadD30;gyrB;vapB34;rrs;gca-	Uncharacterized PE family protein

							gmhA;Rv2464c;PE_PGRS32;Rv3415c-whiB3;lppC;Rv2059;fpg;Rv3848;Rv3736;hsp-nirB;ethA-ethR;glnA3;Rv1482c-fabG1;Rv3415c;katG;rpoB;pncA	
70	39	fpg	N	1	3	4	Rv2059;fbiC;fadD30;Rv0158;glnD;Rv2464c;Rv3736;Rv2024c-Rv2025c;Rv3415c-whiB3;lppC;Rv1313c;PE_PGRS42;glnA3;Rv1353c;PE23;Rv3910;katG;Rv2644c-valT;gyrB;Rv3915;Rv3179;PE_PGRS32;Rv3848;vapB34;Rv1979c;embB;Rv	Formamidopyrimidine-DNA glycosylase 1, involved in base excision repair of DNA damaged by oxidation or by mutagenic agents

							3415c	
68	37	Rv3736	N	1	3	5	Rv1979c;Rv3415c-whiB3;Rv3915;glnD;hsp-nirB;fpg;Rv2464c;Rv0158;Rv1353c;drrA;Rv2024c-Rv2025c;PE_PGRS42;Rv3179;Rv1135A;Rv0696;ethA-ethR;Rv1313c;Rv2075c-Rv2076c;glnA3;gyrB;fadD30;Rv3848;PE23;vapB34	Transcriptional regulatory protein (Probably AraC/XylS-family)
67	29	PPE19	N	2	3	5	katG;Rv0123-PE_PGRS2;embB;Rv2644c-valT;rpoB;plsB1;rrs;Rv1482c-	Uncharacterized PPE family protein

							fabG1;Rv3217c-Rv3218;Rv3263;pncA	
66	32	Rv2075c-Rv2076c	N	1	3	4	hsp-nirB;Rv2219;Rv0681;lppC;fadD30;Rv3736;Rv1313c;PE_PGRS42;glnD;ethA-ethR;Rv3415c-whiB3;gyrB;Rv1135A;glnA3;fbiC;Rv3179;Rv0158;Rv3915;Rv0696	Uncharacterized protein; Uncharacterized protein
66	33	Rv3915	N	1	3	4	Rv1313c;ethA-ethR;Rv1979c;Rv3736;Rv3415c-whiB3;PE_PGRS32;Rv2059;fadD30;Rv3415c;vapB34;gyrB;fpg;PE_PGRS42;glnD;Rv3179;Rv1135A;Rv2075c-	N-acetylmuramoyl-L-alanine amidase CwlM- cell-wall hydrolase that hydrolyzes the amide bond between N-acetylmuramic acid and L-

							Rv2076c;Rv3848	alanine in cell-wall glycopeptides
65	36	vapC25	N	2	3	5	PE_PGRS32;Rv2059;glnA3;Rv0158;Rv1313c;Rv1353c;vapB34;lppC;ethA-ethR;Rv3848;fbiC;Rv3415c-whiB3;Rv2464c;Rv2024c-Rv2025c;PE_PGRS42;fadD30;Rv1979c;Rv0326	Ribonuclease VapC25- toxic component of a type II toxin-antitoxin (TA) system. An RNase (By similarity)
64	34	Rv1313c	N	1	3	4	fadD30;Rv3915;Rv1353c;PE_PGRS32;Rv2024c-Rv2025c;Rv2059;vapC25;glnA3;Rv0158;lppC;fpg;Rv3736;Rv2075c-Rv2076c;PE_PGRS42;glnD;drrA;ethA	Probable transposase for insertion sequence element IS1557

							-ethR;Rv3415c-whiB3;Rv1135A;Rv3848;Rv3415c	
58	29	Rv0123-PE_PGRS2	N	2	3	5	rrs;katG;PPE19;mctB;rpoB;embB;Rv1482c-fabG1;Rv3217c-Rv3218;Rv3263;cobK;Rv1889c-Rv1890c;plsB1;Rv2644c-valT;rpsL	Uncharacterized protein; Member of the Mycobacterium tuberculosis PE family, PGRS subfamily of gly-rich proteins
56	28	glnD	N	1	3	4	fpg;fadD30;Rv3415c-whiB3;Rv3736;Rv2464c;PE_PGRS32;Rv3179;hsp-nirB;ethA-ethR;PE_PGRS42;PE_PGRS30;Rv1313c;Rv2075c-Rv2076c;smc;vapB34;Rv1979c;Rv3915	Bifunctional uridylyltransferase/uridylyl-removing enzyme- modifies, by uridylylation and deuridylylation, the PII regulatory protein (GlnB), in response to the nitrogen status

								of the cell that GlnD senses through the glutamine level
56	33	PE_PGRS32	N	1	3	5	Rv2059;vapC25;Rv1353c;Rv1313c;Rv2024c-Rv2025c;Rv2464c;glnD;Rv3179;vapB34;pncA;Rv3915;lppC;ethA-ethR;PE23;glnA3;Rv3415c-whiB3;fbiC;Rv3848;fpg;Rv0158;Rv0326	PE-PGRS family protein
56	24	plsB1	N	2	3	5	embB;rrs;Rv2074;rpsL;rpoB;PPE19;cobK;Rv0123-PE_PGRS2;Rv2644c-valT;Rv3217c-Rv3218;katG;Rv1482c-fabG1	Putative acyltransferase

55	22	Rv1135A	N	1	2	3	Rv3736;Rv0696;Rv2075c-Rv2076c;Rv2464c;Rv3179;Rv2059;vapB34;Rv3915;PE_PGRS42;fadD30;Rv3415c-whiB3;PE_PGRS30;Rv1313c;gyrB;Rv3415c	Possible acetyl-CoA acetyltransferase (possible gene fragment)
36	25	Rv2644c-valT	N	2	3	5	Rv1482c-fabG1;embB;rpoB;PPE19;katG;fpg;Rv0123-PE_PGRS2;plsB1;cobK;Rv3217c-Rv3218;mctB;pncA;rrs	Uncharacterized protein; tRNA-Val, anticodon cac
35	16	drrA	Y	1	2	4	ethA-ethR;lppC;vapB34;Rv3736;Rv2024c-	Doxorubicin resistance ATP-binding protein- part of the ABC

							Rv2025c;PE_PGRS42;inhA-hemZ;Rv1313c;typA-lpqW;fadD30;Rv3179;Rv3415c-whiB3;Rv0158;gyrB;Rv3415c	transporter complex DrrABC involved in doxorubicin resistance.
35	13	Rv3263	N	2	3	4	embB;Rv1482c-fabG1;rpoB;Rv0123-PE_PGRS2;rrs;katG;PPE19	Probable DNA methylase
34	18	gyrB	Y	1	3	4	PE_PGRS42;fadD30;PE23;vapB34;smc;Rv3915;Rv2075c-Rv2076c;fpg;Rv1979c;Rv3736;hsp-nirB;Rv2024c-Rv2025c;Rv3415c-whiB3;Rv1135A;drrA	DNA gyrase subunit B
34	18	Rv1979c	N	2	3	4	Rv3736;Rv3415c-whiB3;Rv3915;Rv2059;Rv3848;glnD;	Uncharacterized transporter, probable amino-acid or

							Rv1353c;gyrB;Rv2024c-Rv2025c;vapC25;fpg;fadD30;glnA3;fbiC	metabolite transport protein
34	18	Rv3217c-Rv3218	N	2	3	5	rpoB;rrs;katG;Rv0123-PE_PGRS2;Rv1482c-fabG1;cobK;Rv2644c-valT;embB;mctB;pncA;plsB1;PPE19	Probable conserved integral membrane protein; Conserved protein
33	13	hsp-nirB	N	1	2	4	Rv2075c-Rv2076c;Rv3736;Rv3415c-whiB3;glnD;ethA-ethR;gyrB;PE23;fadD30;PE_PGRS30;Rv1945-lppG;Rv3415c	heat-stress-induced ribosome-binding protein A; Probable nitrite reductase [NAD(P)H] large subunit
29	19	Rv2464c	N	1	3	5	PE_PGRS32;glnD;Rv3179;Rv0158;fpg;Rv3736;gca-	Endonuclease 8 1- involved in base excision repair of DNA

							gmhA;PE23;vapC25;Rv2024c-Rv2025c;PE_PGRS42;Rv3415c-whiB3;Rv2059;Rv1353c;Rv1135A;glnA3;lppC;Rv1482c-fabG1;katG;fadD30	damaged by oxidation or by mutagenic agents
27	15	Rv3179	N	1	3	4	fadD30;Rv3415c-whiB3;Rv2464c;PE_PGRS32;glnD;Rv3736;Rv1135A;Rv2075c-Rv2076c;fpg;Rv3915;vapB34;drrA	Conserved protein
25	15	rpsL	Y	1	2	4	rpoB;pncA;embB;katG;plsB1;rrs;cobK;Rv0123-PE_PGRS2;Rv1482c-fabG1	30S ribosomal protein S12- with S4 and S5 plays an important role in translational accuracy
23	7	embC-embA	Y	1	2	3	pncA;rrs;katG;rpoB;embB;Rv1482c-fabG1	Probable arabinosyltransferase C; Probable

								arabinosyltransferase A
18	13	cobK	N	2	3	4	rpoB;katG;rpsL;plsB1;embB;Rv3217c -Rv3218;rrs;Rv0123-PE_PGRS2;Rv2644c-valT;mctB;pncA	Precorrin-6A reductase-catalyzes the reduction of the macrocycle of precorrin- 6X into precorrin-6Y
15	6	inhA	Y	1	2	2	mmpL12-Rv1523;pncA;embB;Rv1482c-fabG1;rpoB	Enoyl-ACP reductase of the type II fatty acid syntase (FAS-II) system, catalyzes the NADH-dependent reduction of the double bond of 2-trans-enoyl-[acyl-carrier protein], an essential step in the fatty acid elongation cycle of the FAS-II

								pathway
12	1	fadE15-PE_PGRS29	N	1	1	1	NA	Probable acyl-CoA dehydrogenase; PE-PGRS family protein
12	6	Rv3510c	N	1	2	2	Rv3415c;ethA-ethR;mkl-vapC6;PE_PGRS42;vapB34	Conserved protein
11	1	hisG	N	1	1	1	NA	ATP phosphoribosyltransferase
11	1	mfd	N	1	1	1	NA	Transcription-repair-coupling factor- couples transcription and DNA repair by recognizing RNA

								polymerase (RNAP) stalled at DNA lesions
11	1	Rv2136c	N	1	1	1	NA	Undecaprenyl-diphosphatase-catalyzes the dephosphorylation of undecaprenyl diphosphate (UPP). Confers resistance to bacitracin
11	1	Rv3632-Rv3633	N	1	1	1	NA	Possible conserved membrane protein; Conserved protein
10	1	echA2	N	1	1	1	NA	Enoyl-CoA hydratase EchA2
10	1	gabD1	N	1	1	1	NA	Succinate-semialdehyde

								dehydrogenase
10	6	mctB	N	2	3	5	Rv0123-PE_PGRS2;rrs;rpoB;cobK;Rv3217c-Rv3218;Rv2644c-valT;embB;pncA	Copper transporter- pore-forming protein, which is involved in efflux of copper across the outer membrane. Essential for copper resistance and maintenance of a low intracellular copper concentration
10	1	mpt64	N	1	1	1	NA	Immunogenic protein
10	6	PE_PGRS30	N	1	2	3	PE_PGRS42;glnD;Rv3415c;Rv1135A;hsp-nirB;Rv1945-lppG;fadD30	PE-PGRS family protein-mediates suppression of proinflammatory immune

								response in macrophages via modulation of host cytokine response
10	1	pirG	N	1	1	1	NA	Exported repetitive protein-surface-exposed protein required for multiplication and intracellular growth
10	1	PPE66	N	1	1	1	NA	Uncharacterized PPE family protein PPE66
10	1	ppk2	N	1	1	1	NA	Polyphosphate:GDP phosphotransferase- uses inorganic polyphosphate (polyP) as a donor to convert GDP to

								GTP and modulates nucleotide triphosphate synthesis catalyzed by the nucleoside diphosphate kinase (Ndk) in favor of GTP production over CTP or UTP
10	1	Rv1112	N	1	1	1	NA	Ribosome-binding ATPase, YchF-ATPase that binds to both the 70S ribosome and the 50S ribosomal subunit in a nucleotide-independent manner
10	1	Rv1128c-Rv112	N	1	1	1	NA	Uncharacterized protein; HTH-type transcriptional regulator PrpR- plays a key role in

		9c						regulating expression of enzymes involved in the catabolism of short chain fatty acids (SCFA)
10	1	Rv2005c	N	1	1	1	NA	Universal stress protein family protein, predicted possible vaccine candidate
10	1	Rv2434c	N	1	1	1	NA	Probable conserved transmembrane protein
10	1	Rv3818	N	1	1	1	NA	Putative Rieske 2Fe-2S iron-sulfur protein
9	1	plsC	N	1	1	1	NA	Possible transmembrane phospholipid biosynthesis

								bifunctional enzyme
9	1	Rv0349	N	1	1	1	NA	Uncharacterized protein
9	1	Rv0443	N	1	1	1	NA	Conserved protein
9	1	Rv0585c	N	1	1	1	NA	Probable conserved integral membrane protein
9	1	Rv1868	N	1	1	1	NA	Uncharacterized protein
9	1	Rv2554c	N	1	1	1	NA	Putative pre-16S rRNA nuclease-could be a nuclease involved in processing of the 5'-end of pre-16S rRNA

9	5	smc	N	1	2	3	gyrB;glnD;Rv3415c-whiB3;fadD30	Chromosome partition protein Smc- required for chromosome condensation and partitioning
9	1	ureG	N	1	1	1	NA	Urease accessory protein UreG- facilitates the functional incorporation of the urease nickel metallocenter
8	1	adhA	N	1	1	1	NA	Probable adhA,alcohol dehydrogenase
8	1	aspS	N	1	1	1	NA	Aspartyl-tRNA synthetase with relaxed tRNA specificity since it is able to aspartylate not only its cognate tRNA(Asp) but also

								tRNA(Asn)
8	1	PE6-Rv0336	N	1	1	1	NA	Member of the Mycobacterium tuberculosis PE family; Conserved 13E12 repeat family protein
8	1	vapC1-icd2	N	1	1	1	NA	Ribonuclease- toxic component of a type II toxin-antitoxin (TA) system; Isocitrate dehydrogenase
8	1	vapC29	N	1	1	1	NA	Ribonuclease VapC29- toxic component of a type II toxin-antitoxin (TA) system
7	1	fadD2	N	1	1	1	NA	Putative fatty-acid--CoA ligase

		1						
7	1	mbtB	N	1	1	1	NA	Phenyloxazoline synthase, involved in the initial steps of the mycobactin biosynthetic pathway
7	3	mkl-vapC6	N	1	2	2	Rv3510c;Rv3415c	Probable ribonucleotide transport ATP-binding protein; Toxic component of a type II type II toxin-antitoxin (TA) system, an RNase
7	1	mscL	N	1	1	1	NA	Large-conductance mechanosensitive channel-channel that opens in response

								to stretch forces in the membrane lipid bilayer
7	1	Rv0106	N	1	1	1	NA	Uncharacterized protein
7	1	Rv0458	N	1	1	1	NA	Probable aldehyde dehydrogenase
7	1	Rv1771	N	1	1	1	NA	L-gulono-1,4-lactone dehydrogenase
7	1	Rv2990c	N	1	1	1	NA	Uncharacterized protein
7	1	Rv3768-Rv376	N	1	1	1	NA	Uncharacterized protein; Uncharacterized protein

		9						
7	1	Rv3840	N	1	1	1	NA	Possible transcriptional regulator, highly similar in part to PSR proteins (penicillin binding protein repressors)
7	5	Rv3910	N	2	3	5	glnA3;Rv0158;fbiC;Rv3415c-whiB3;fpg;PE_PGRS42;Rv3848	Probable peptidoglycan biosynthesis protein MviN
5	4	gca-gmhA	N	1	2.5	3	Rv2464c;PE23;vapB34;fbiC	Possible GDP-mannose 4,6-dehydratase; Phosphoheptose isomerase
4	1	lysS	N	1	1	1	NA	Lysine--tRNA ligase 1
4	1	Rv3722c	N	1	1	1	NA	Conserved protein

3	2	mmpL 12-Rv152 3	N	2	2.5	3	inhA;Rv1482c-fabG1;pncA	Probable conserved transmembrane transport protein; Probable methyltransferase
3	2	Rv069 6	N	3	3	3	Rv1135A;Rv3736;Rv2075c-Rv2076c	Putative mycofactocin biosynthesis glycosyltransferase, MftF
2	2	mcr7	N	2	2.5	3	embB;rpoB;katG	Putative small regulatory RNA
2	2	Rv202 8c	N	1	2	3	vapB34;Rv3415c-whiB3	Universal stress protein family protein
1	1	cmr-Rv167 6	N	3	3	3	fadD30;Rv3415c-whiB3	HTH-type transcriptional regulator- positively regulates the expression of at least

								groEL2; Uncharacterized protein
1	1	eccE2	N	3	3	3	katG;rpoB	ESX-2 secretion system protein
1	1	ethA	Y	1	1	1	NA	FAD-containing monooxygenase
1	1	glnQ	N	1	1	1	NA	Probable glutamine-transport ATP-binding protein ABC transporter
1	1	inhA-hemZ	N	2	2	2	drrA	Enoyl-ACP reductase of the type II fatty acid syntase (FAS-II) system, catalyzes the NADH-dependent reduction of the double bond of 2-trans-enoyl-[acyl-carrier protein], an essential step in the fatty acid

								elongation cycle of the FAS-II pathway; Ferrochelatase-involved in the biosynthesis of heme
1	1	lipR	N	1	1	1	NA	Putative acetyl-hydrolase-required for maintaining the appropriate mycolic acid composition and permeability of the envelope on its exposure to acidic pH
1	1	lpqG-hpt	N	1	1	1	NA	Probable conserved lipoprotein; hypoxanthine-guanine phosphoribosyltransferase

1	1	lytB2-Rv1111c	N	1	1	1	NA	4-hydroxy-3-methylbut-2-enyl diphosphate reductase 2-catalyzes the conversion of 1-hydroxy-2-methyl-2-(E)- butenyl 4-diphosphate (HMBPP) into a mixture of isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP); Conserved hypothetical protein
1	1	panB-Rv2226	N	2	2	2	vapB34	3-methyl-2-oxobutanoate hydroxymethyltransferase-catalyzes the reversible reaction

								in which hydroxymethyl group from 5,10-methylenetetrahydrofolate is transferred onto alpha-ketoisovalerate to form ketopantoate; Conserved protein
1	1	parE1	N	2	2	2	lppC	Toxic component of a type II toxin-antitoxin (TA) system
1	1	phoR	N	1	1	1	NA	Possible two component system response phosphate sensor kinase membrane-associated
1	1	PPE61	N	2	2	2	rpoB	Uncharacterized PPE family

		- PPE62						protein; Member of the Mycobacterium tuberculosis PPE protein family
1	1	psd- moeA 2	N	1	1	1	NA	Phosphatidylserine decarboxylase proenzyme-catalyzes the formation of phosphatidylethanolamine (PtdEtn) from phosphatidylserine (PtdSer); Molybdopterin molybdenumtransferase 2-catalyzes the insertion of molybdate into adenylated

								molybdopterin with the concomitant release of AMP
1	1	rpoC-Rv0669c	N	1	1	1	NA	DNA-directed RNA polymerase subunit beta- DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates; Neutral ceramidase- catalyzes the cleavage of the N-acyl linkage of the ceramides (Cers) to yield sphingosine (Sph) and

								free fatty acid
1	1	Rv0021c-whiB5	N	2	2	2	Rv3848	Uncharacterized protein; Transcriptional regulator- a transcription factor that is probably redox- responsive, probably plays a role in immunomodulation and reactivation after chronic infection. Its induction results in transcription of a number of

								genes including sigM, and the genes for 2 type VII secretion systems ESX-2 and ESX-4
1	1	Rv0193c-Rv0194	N	1	1	1	NA	Uncharacterized protein; Multidrug efflux ATP-binding/permease protein-overexpression in M. smegmatis increases resistance to erythromycin, ampicillin, novobiocin and vancomycin. It also reduces accumulation of ethidium bromide in the cell.

1	1	Rv0258c	N	3	3	3	lppC;PE_PGRS42	Uncharacterized protein
1	1	Rv0308-Rv0309	N	2	2	2	PE_PGRS42	Probable conserved integral membrane protein
1	1	Rv0326	N	5	5	5	PE_PGRS32;vapC25;fbiC;Rv3415c-whiB3	Uncharacterized protein
1	1	Rv0628c	N	1	1	1	NA	Uncharacterized protein
1	1	Rv0681	N	3	3	3	Rv2075c-Rv2076c;Rv2219	Probable transcriptional regulatory protein (Possibly TetR-family)

1	1	Rv0870c-cspB	N	1	1	1	NA	Possible conserved integral membrane protein; Probable cold shock-like protein B
1	1	Rv1435c-gap	N	1	1	1	NA	Probable conserved Pro-, Gly-, Val-rich secreted protein; Glyceraldehyde-3-phosphate dehydrogenase- catalyzes the oxidative phosphorylation of glyceraldehyde 3-phosphate (G3P) to 1,3-bisphosphoglycerate (BPG) using the cofactor NAD
1	1	Rv188	N	3	3	3	Rv0123-PE_PGRS2;katG	Uncharacterized protein;

		9c-Rv1890c						Uncharacterized protein
1	1	Rv1945-lppG	N	3	3	3	hsp-nirB;PE_PGRS30	Uncharacterized protein; Possible conserved lipoprotein
1	1	Rv2074	N	4	4	4	plsB1;embB;rrs	F420H(2)-dependent biliverdin reductase- catalyzes the F420H(2)-dependent reduction of biliverdin-IXalpha at C10 position, leading to bilirubin-IXalpha, a potent antioxidant. As biliverdin-IXalpha is produced in

								high amounts in macrophages infected with M.tuberculosis, its reduction by Rv2074 may play a role in protecting mycobacteria against oxidative stress, aiding the persistence of M.tuberculosis infection
1	1	Rv2129c	N	1	1	1	NA	Probable oxidoreductase
1	1	Rv2219	N	3	3	3	Rv2075c-Rv2076c;Rv0681	Uncharacterized protein
1	1	Rv2616	N	1	1	1	NA	Conserved protein

1	1	Rv2633c	N	1	1	1	NA	Uncharacterized protein
1	1	Rv2809	N	1	1	1	NA	Uncharacterized protein
1	1	Rv3113-Rv3114	N	2	2	2	Rv3415c	Possible phosphatase; Conserved hypothetical protein
1	1	typA-lpqW	N	2	2	2	drrA	GTP-binding translation elongation factor; Probable monoacyl phosphatidylinositol tetramannoside-binding protein

Supplementary Table 1

Phenotype Frequencies by lineage

Lineage	Drug-resistant	MDR	Pan-Susceptible	XDR
2	469	1546	1399	155
4	934	1601	5664	153

Chapter 6:
Discussion and Conclusions

Discussion

The work in this thesis has utilised three distinct analysis technologies to explore the relationship between the *Mtb* genome and the development of drug resistance. Due to the clonal nature and distinct lineages of *Mtb*, I hypothesised that the specific genomic background would influence the specific mutations that develop in response to therapy; and that as a result, lineage specific GWAS, or LCS techniques would be more revealing of the genetic causes of resistance (and spread) than traditional approaches. Thus, my work provides; methodological insights regarding the GWAS, phyC and LCS methodologies, biological insights gained through their application, and implications for surveillance, diagnosis and treatment of tuberculosis and beyond.

Methodological Insights

Differences in the loci identified between each methodology demonstrates the complementarity between GWAS, phyC and LCS; there is utility in each approach (see **Chapters 2-5**). PhyC relies on independent evolution events and, as expected, appears more suited to highly diverse *Mtb* genomic datasets, considering all lineages together (see **Chapter 2**). On the other hand, GWAS is better able to identify lineage-specific associations with drug resistance (see **Chapters 2** and **3**). In such cases, there may be a higher relative importance of transmission than multiple independent evolution events. This might be expected when variants are highly transmissible; hence, GWAS may be a powerful tool in identifying variants

associated with highly transmissible resistant strains, as further indicated in **Chapter 3**. It is feasible that high transmissibility could coevolve with drug resistance. This is because more transmissible strains would likely experience more drug therapy and thus greater selection for resistance (see **Chapter 3**).

It is less clear how LCS deals with structure in *Mtb* populations, as LCS takes an evolutionary approach to machine learning and thus a less structured approach to lineage diversity that may not be enhanced by controlling for lineage. Despite this, the importance of lineage background is revealed by its identification of lineage specific loci (see **Chapters 4** and **5**).

While both phyC and GWAS approaches identify single variants or loci associated with a phenotype, LCS can identify patterns of one or more loci predictive of phenotype. This means LCS has the potential to discover epistatic relationships where GWAS and phyC do not (see **Chapters 4** and **5**).

Here GWAS and phyC have been applied to both loci and genomic variants (SNPs and short indels) (see **Chapters 2** and **3**); conversely, due to computational intensity, LCS was applied only to locus-based genomic variants. The ability of LCS to make predictions from such lower resolution locus-based data may be due to the power increase resulting from simultaneous consideration of multiple loci (see **Chapters 4 and 5**). It is important to consider whether use of a binary locus-type, reference or non-reference, could introduce biases into LCS. For example, loci with a greater length may be generally more likely to have a non-synonymous mutation, and some loci may be able to withstand more non-synonymous variation with little effect on protein function. However, as computing power increases and there is

further algorithmic development, it would be interesting to apply LCS to SNP-level data.

More generally, it is important to note that all three methodologies currently rely on mapping to a reference and thus introduce biases. This means that structural variation in the *Mtb* genome may be overlooked.

All three methodologies could lead to predictions of resistance phenotype from *Mtb* genome. For GWAS and phyC this may be through variant discovery to inform a database of predictive variants within a predictor tool [1–8], whilst for LCS prediction is an inherent part of the algorithm and can be performed without the requirement of running the full learning algorithm [9].

Biological Insights

The success of lineage-specific GWAS highlights potential differences in the genomics of drug resistance between lineages (see **Chapters 2** and **3**). Whilst some variants were identified by lineage-combined approaches, others were identified in only one lineage (see **Chapters 2** and **3**). Differences in evolutionary trajectory between lineages are compatible with recent evolution, since lineage divergence, as would be expected to result from selection pressures arising due to the introduction of antibiotics. Lineages 1 and 3 have the smallest sample sizes in the data analysed here, new larger datasets for these lineages could provide further insight here.

Chapter 3 highlights the complexity of XDR *Mtb*, revealing a number of novel variants associated with the XDR phenotype. The application of LCS to XDR *Mtb* in **Chapter 5** further delves into this complexity. Perhaps the ability of LCS to achieve

high accuracy of XDR prediction is due to its ability to deal with such complexity (see **Chapter 5**). Future work is required to determine the importance of selection compared to genetic drift in linking the loci identified to resistance phenotypes; nonetheless identifying such variants present in clinical XDR TB populations could inform public health response strategies (see **Chapter 3**).

Loci identified in **Chapter 2-5** warrant further investigation to unearth more detail on how they might functionally relate to resistance and transmission or fitness in general, especially those that lack knowledge at present. This demonstrates the utility of the GWAS, phyC and LCS methodologies in creating a more targeted approach to hypothesis generation for wet lab studies.

Implications for Surveillance, Treatment and Diagnosis

The GWAS, phyC and LCS as presented here (see **Chapters 2-5**) could have important implications for surveillance; as these approaches could be deployed to understand new clonal spreads of TB disease. As new whole genome sequence data is generated, it can be analysed to see if there are new genetic factors that are driving these outbreaks. This could be automated and facilitated through online *Mtb* monitoring systems that would enable precision public health, facilitating better allocation of resources and design of research strategies.

In addition, the findings from the work presented can be used to improve the current online tools (such as TB-profiler) to give more accurate drug resistance predictions.

However, it is important to consider feasibility of such proposals in light of growing *Mtb* genomic datasets and changing phenotypic methods [10]. For example, phyC

becomes increasingly difficult as datasets grow, due to limitations on phylogenetic tree construction. This indicates a need for new methods for tree construction or sampling of isolates from which to build trees. Similarly, as dataset size increases LCS may take longer to run.

Furthermore, there are implications for diagnosis and treatment. Accurate prediction of resistance status from whole genome sequence can allow better selection of therapy regimens for individual patients without the need to wait for drugs susceptibility testing which can be lengthy. Moreover, an improved understanding of the biological mechanisms of resistance present in global *Mtb* populations could inform drug design.

Future Avenues of Work

There are a number of future avenues that could follow from this work. For example, the application of methodologies used here to new *Mtb* datasets; for example, using new genome sequencing methodologies such as long read methods [11], using datasets exploring within patient *Mtb* genomic diversity and exploring new phenotypes such as latent TB or further exploring drug resistance through MIC values.

Conclusions

GWAS, phyC and LCS are complementary methodologies that can each provide insight into the genomic basis of drug resistance in *Mtb* and the prediction of drug resistance from the *Mtb* genome. This work proposes a number of loci as candidates for further study, that may be involved in drug resistance or in particular facilitate spread of drug resistant strains. This has important implications for surveillance, diagnosis and treatment of tuberculosis.

References

1. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. N Engl J Med. 2018;379:1403–15. doi:10.1056/NEJMoa1800474.

2. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med. 2019;11:41. doi:10.1186/s13073-019-0650-x.

3. Mahé P, El Azami M, Barlas P, Tournoud M. A large scale evaluation of TBProfiler and Mykrobe for antibiotic resistance prediction in *Mycobacterium tuberculosis* . PeerJ. 2019;7:e6857.

4. van Beek J, Haanperä M, Smit PW, Mentula S, Soini H. Evaluation of whole genome sequencing and software tools for drug susceptibility testing of *Mycobacterium tuberculosis*. Clin Microbiol Infect. 2019;25:82–6.

5. Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, et al. TGS-TB: Total genotyping solution for *Mycobacterium tuberculosis* using short-read whole-genome sequencing. PLoS One. 2015;10:1–12.

6. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun. 2015;6.

7. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: Targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics. 2014;15:1–12.

8. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al.

PhyResSE: A web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. J Clin Microbiol. 2015;53:1908–14.

9. Urbanowicz RJ, Moore JH. ExSTraCS User's Guide Version 2.0.2 Beta. 2014;:0–43.

10. Xie YL, Chakravorty S, Armstrong DT, Hall SL, Via LE, Song T, et al. Evaluation of a rapid molecular drug-susceptibility test for tuberculosis. N Engl J Med. 2017;377:1043–54.

11. Phelan J, de Sessions PF, Tientcheu L, Perdigao J, Machado D, Hasan R, et al. Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. Sci Rep. 2018;8:160.