Mycobacterium tuberculosis genome mutations and fitness cost: molecular and epidemiological modelling of functional implications

MALANCHA KARMAKAR

(ORCID ID: 0000-0003-0303-2895)

Doctor of Philosophy

January 2021

Faculty of Medicine, Dentistry and Health Sciences

Department of Microbiology and Immunology

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

ABSTRACT

Identification of *Mycobacterium tuberculosis* (Mtb) increasingly involves characterising large sections of genetic material, such as through whole genome sequencing. While some mutations identified through these techniques are well-characterised and strongly associated with anti-tuberculous drug resistance, such molecular methods frequently identify mutations with unknown significance or limited understanding of associated functional biological pathways. In this PhD, I have developed computational protein structural tools and mathematical models of TB transmission, that use genomic data to understand the impact of genomic changes and predict the consequences with regards to transmissibility and drug susceptibility of Mtb.

Drug resistant mutations often carry both a selective advantage and a fitness cost, which can be reflected by the changes in protein structure and function. I developed a pipeline that captured the molecular consequences of coding mutations on protein stability, dynamics and interactions. Using my pipeline to evaluate the mechanistic consequences of mutations, I applied it to the real-time genomic analysis of a Victorian tuberculosis patient. The analysis led to identification of a novel resistant strain and altered patient treatment – the first reported use of structural information to guide clinical resistance detection. The information was then used to inform a compartmental epidemiological model of Mtb transmission in order to understand the rise of drug resistance in two high TB-incidence setting. Using a adaptive metropolis algorithm, I estimated drug resistance amplification proportions for two first-line anti-tuberculosis drugs, and explored how structural changes may alter the fitness landscape and transmission dynamics.

The work highlighted the power of combining genomic, epidemiological and structural information in the fight against tuberculosis, and presents examples of application across the spectrum from laboratory, clinical and programmatic contexts. This work has further laid the foundation to rapidly apply and translate this approach to other infectious and non-infectious diseases.

DECLARATION

This is to certify that:

1. the thesis comprises only my original work towards the Doctor of Philosophy, except where indicated in the Preface;

2. due acknowledgement has been made in the text to all other materials used;

3. the thesis is fewer than 100,000 words in length, exclusive of tables, figures, references and appendices; and

4. the thesis complies with the stipulations set out for the degree of Doctor of Philosophy by the University of Melbourne

Signed

Malancha Karmakar

January 2021

PREFACE

I would like to acknowledge the following people for their contributions towards the thesis:

Chapter 3.1 consists of the article "Structure guided prediction of Pyrazinamide resistance mutations in pncA" published in Scientific Reports in February 2020. This work conceived by David Ascher. I worked on the study design, data curation, formal analysis, investigation, methodology and validation and wrote the original draft of the manuscript. Along with Carlos Rodrigues, I developed the computational analysis tool. Kristy Horan and Justin Denholm contributed to data collection and analysis respectively. All authors contributed in manuscript revisions.

Chapter 3.2 consists of the article "Analysis of a novel pncA mutation for susceptibility to Pyrazinamide therapy" published in American Journal of Respiratory Medicine and Critical Care in April 2018. For this research paper, I was involved in the study design, execution, data analysis, and writing of all versions of the manuscript along with Justin Denholm and David Ascher. Maria Globan, Janet Fyfe, Timothy Stinear, Paul Johnson and Natasha Holmes were involved in clinical and laboratory aspects of investigation. All authors contributed to article revisions.

Chapter 4 consists of the article "Empirical ways to identify novel Bedaquiline resistance mutations in AtpE" published in PlosOne in May 2019. David Ascher and Justin Denholm conceived the project. I worked on the data curation, performed the analysis and validation and wrote the original draft of the manuscript. Carlos Rodrigues helped in developing the webserver. Kathryn Holt and Sarah Dunstan provided resources for validating the tool. All authors contributed in article revisions.

Chapter 5 consists of the article "Hyper transmission of Beijing lineage *Mycobacterium tuberculosis*: Systematic review and Meta-analysis" published in Journal of Infection in October 2019. This review was initially conceptualised after a thorough discussion with Justin Denholm. I refined and executed the search strategy independently, screened titles in duplicate, independently screened full texts, extracted the

data, wrote the first draft of the manuscript, performed the meta-analysis and revised subsequent drafts. James Trauer helped with data interpretation and analysis. Justin Denholm, David Ascher and James Trauer helped in revising the manuscript.

Chapter 6 consists of the article "Estimating tuberculosis drug resistance amplification rates in high-burden settings" which has been submitted to International Journal of Epidemiology on May 2021. The work was initially conceptualized by Romain Ragonnet and Justin Denholm. With Romain's help I developed the epidemiological model to capture drug resistance amplification dynamics. I calibrated and optimized the model and generated the figures required for the manuscript. James Trauer helped with additional feedback on the methodology to make the model realistic. I wrote the first draft of the manuscript and James Trauer, Justin Denholm, Romain Ragonnet and David Ascher helped in revising the manuscript.

ACKNOWLEDGEMENTS

I would like to thank The University of Melbourne for supporting this PhD with the Melbourne Research Scholarship and my doctoral supervisors, Associate Professor Justin Denholm and Associate Professor David Ascher, for their constant guidance and support throughout my PhD.

Justin has been a great source of encouragement and a persistent advocate of my work. He has always given me a great deal of freedom to determine the course of this PhD and his expertise on tuberculosis has been invaluable. David introduced structural bioinformatics to me and helped me navigate it smoothly. He has always pushed me to do a little better each day and helped me to become a more confident researcher.

Several individuals have provided assistance and support without which this work would not have been possible. I would especially like to acknowledge Carlos Rodrigues for helping me develop the webservers and improving my coding skills, Michael Silk for helping me learn R and make beautiful figures for the manuscript, Romain Ragonnet for his immense patience while teaching me the concepts of mathematical modelling and James Trauer for being a great critic which has always improved my work.

I am thankful to Stephanie Portelli for being an amazing friend and supporting me emotionally throughout my PhD and Katie Dale for inspiring me to do better evidence-based research.

Finally, I would like to thank my family and my extended family here in Melbourne who have always believed in me and my dreams. My parents, Kanta and Dibakar and my sister Bratati for their continuous moral support. My husband Arindam who has endured me and helped me stay sane and calm. My friends and colleagues – Rishabh, Pavneet, Daniella, Christina, Uzma, Joel, Sharjeel, Geet, Bhavna, Bhavya, Yoo Chan, Binh, Noa, Vittoria, Marialena, Alex, Bruna, Moshe, Raghad and Elston for making my PhD journey stress free.

PUBLICATIONS IN PEER REVIEWED JOURNALS

1. **Karmakar, M.**, Global, M., Fyfe, J.A.M., Stinear, T.P., Johnson, P.D.R., Holmes, N.E., Denholm, J.T., Ascher, D.B. (2018). Analysis of a novel pncA mutation for susceptibility to Pyrazinamide therapy. *American Journal of Respiratory and Critical Care Medicine*, 198(4):541-544.

2. Shaweno, D., **Karmakar, M.**, Alene, K.A., Ragonnet, R., Clements, A., Trauer, J., Denholm, J.T., McBryde, E., (2018). Methods used in the spatial analysis of tuberculosis epidemiology: a systematic review. *BMC Medicine* **16,** 193

3. **Karmakar, M.**, Rodrigues C.H.M., Holt K.E., Dunstan S.J., Denholm J., Ascher D.B. (2019). Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLoS ONE* 14(5): e0217169

4. **Karmakar, M.**, Ascher, D.B., Trauer, M.J., Denholm, J.T. (2019). Hyper transmission of Beijing lineage *Mycobacterium tuberculosis*: Systematic review and Meta-analysis. *Journal of Infection* Volume 79, Issue 6, Pages 572-581

5. Pires, D.E.V., Rodrigues, C.H., Albanaz, A.T.S., **Karmakar, M.**, Myung, Y., Xavier, J.S., Michanetzi, E., Portelli, S., Ascher, D.B. (2019). Exploring Protein Supersecondary Structure Through Changes in Protein Folding, Stability, and Flexibility. *Methods Mol Biol Springer Netherlands*; 1958:173-185

6. **Karmakar, M.**, Rodrigues, C.H.M., Horan, K., Denholm, J.T., Ascher, D.B. (2020). Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Scientific Reports* **10,** 1875.

7. Xavier, J., Nguyen, T.B., **Karmakar, M.**, Portelli, S., Rezende, P., Velloso, J., Ascher, D.B., Pires, D.E.V. (2021). ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Research*, gkaa925

8. Pires, D.E.V., Portelli, S., Rezende, P.M., Veloso, W.N.P., Xavier, J.S., Velloso, J.P.L., **Karmakar, M.**, Myung, Y., Rodrigues, C.H.M., Silk, M., Ascher, D.B. (2020). A comprehensive computational platform to guide drug development using graph-based signature methods. *Methods in Molecular Biology*; 2112:91-106.

9. Airey, E., Portelli, S., Xavier, J.S., Myung, Y., Silk, M., **Karmakar, M.**, Velloso, J.P.L., Rodrigues, C.H.M., Parate, H.H., Al-Jarf, R., Barr, L., Geraldo, J.A., Rezende, P.M., Pires, D.E.V., Ascher, D.B. (2021). Identifying genotype-phenotype correlations via integrative mutation analysis. *Methods Mol Biol Springer Netherlands*; 2190:1-32.

10. Dale, K.D., **Karmakar, M.,** Snow, K.J., Menzies, D., Trauer J.M., Denholm, J.T. (2021). Quantifying the rates of late reactivation tuberculosis: A systematic review. *Lancet Infectious Diseases*, S1473-3099(20)30728-3

ARTICLES SUBMITTED FOR PUBLICATION

1. **Karmakar M.**, Ragonnet R., Ascher, D.B., Trauer J.M., Denholm, J.T. (2021). (In Press). Modelling tuberculosis drug resistance amplification rates in high-burden settings. *International Journal of Epidemiology*

2. **Karmakar M.***, Cicaloni V.*, Rodrigues, C.H.M., Santucci A., Spiga O. and Ascher D.B. (2021). HGDiscovery: an online tool providing functional and phenotypic information on novel variants of homogentisate 1,2- dioxygenase. *Briefings in Bioinformatics* (*joint first author)

3. Cicaloni V.*, **Karmakar M.***, Frusciante L., Pettini F., Trezza A., Orlandini M., Galvagni F., Nardi F., Mongiat M., Ascher D.B., Santucci A. and Spiga O. (2021). (In Press). Bioinformatics approaches to predict mutation effects in the binding site of the proangiogenic molecule CD93. *Nature Computational Biology* (*joint first author)

CONTENTS

Abstract **2**

Declaration **3**

Preface **4**

Acknowledgements **6**

Publications in Peer Reviewed Journals **7**

Articles Submitted For Publication **9**

List of Figures **12**

Chapter 1: Literature Review **13**

1.1 Introduction 13

1.2 Drug Resistance: An increasing global public health problem 14

1.3 Experimental methods to quantify drug resistance 17

1.4 Understanding drug resistance mutations using protein structures 18

1.5 Epidemiological modeling and fitness cost 25

1.5.1 Population-based studies 26

1.5.2 Secondary attack rates 32

1.5.3 Strain-specific differences in fitness and transmissibility 32

1.5.4 Mathematical models of Mtb transmission 35

Chapter 2: Methodology **39**

2.1 Structure based Analysis 41

2.2 Sequence-based Analysis 44

2.3 Homology modeling 46

2.2 Molecular Docking 49

2.4 Novel methodological pipeline to build the empirical classifier 50

2.5 Mathematical Modeling 58

Chapter 3.1: Structure guided prediction of pyrazinamide resistance in tuberculosis **63**

Chapter 3.2: Analysis of a novel PncA mutation for susceptibility to pyrazinamide therapy **84**

Chapter 4: Empirical ways to identify novel Bedaquiline resistance mutations **90**

Chapter 5: Hyper transmission of Beijing lineage *Mycobacterium tuberculosis*: Systematic review and Meta-analysis **115**

Chapter 6: Estimating the risk of tuberculosis drug resistance amplification in high-burden settings **129**

Chapter 7: Conclusion **152**

References **157**

Appendix 1: Methods used in the spatial analysis of tuberculosis epidemiology: a systematic review **167**

Appendix 2: Exploring Protein Supersecondary Structure Through Changes in Protein Folding, Stability, and Flexibility **186**

Appendix 3: ThermoMutDB: a thermodynamic database for missense mutations **200**

Appendix 4: A comprehensive computational platform to guide drug development using graph-based signature methods **206**

Appendix 5: Identifying genotype-phenotype correlations via integrative mutation analysis **223**

Appendix 6: Quantifying the rates of late reactivation tuberculosis: A systematic review **256**

Appendix 7: HGDISCOVERY: an online tool providing functional and phenotypic information on novel variants of homogentisate 1,2- dioxigenase **310**

Appendix 8: Bioinformatic approaches to predict mutation effects in the binding site of the proangiogenic molecule CD93 ..... **326**

LIST OF FIGURES

Figure 1: Methodological pipeline to develop the empirical classifier.. 21

Figure 2: Protein Structure of PncA (pyrazinamidase). 23

Figure 3: The crystal structure of Bedaquiline bound to AtpE.. 25

Figure 4: Molecular Genotyping Methods.. 32

Figure 5: Mycobacterium tuberculosis lineage classification. 34

Figure 6: Thermodynamic cycle of protein folding.. 41

Figure 7: Example of leave-one-out cross-validation where the number of instances is 5. 55

Figure 8: Confusion Matrix.. 56

Figure 9: A prototypical TB transmission model. 61

CHAPTER 1: LITERATURE REVIEW

1.1 Introduction

Tuberculosis (TB) is an ancient deadly airborne disease caused by organisms of the "*Mycobacterium tuberculosis* complex", which includes *M. bovis*, *M. africanum*, and most commonly *M. tuberculosis*. *Mycobacterium tuberculosis* (Mtb) is an obligate human pathogen, primarily infects the lungs, but can also cause disease in almost any tissue of the body. Infection with Mtb can progress from containment in the host, in which the bacteria are isolated within granulomas (latent TB infection), to an active and potentially contagious state, in which the patient may show symptoms that can include cough, fever, night sweats and weight loss [1].

Mtb has been an irrepressible pathogen since its discovery by Robert Koch in 1882 [2]. The global statistics for TB are substantial, with over 10 million new cases and 1.4 million deaths in 2019 [3]. Even though TB is a serious life-threatening illness, it is curable as long as it is diagnosed early and effective chemotherapy applied, since one of the greatest risks of mortality in TB is delayed treatment. Principle objective to chemotherapy in TB patients is the eradication of the whole bacillary load [4]. Treatment involves usage of anti-tuberculous (anti-TB) drugs for a prolonged period to avoid bacterial resistance and persistence. The anti-TB drugs available to treat TB [5] are isoniazid (INH) [6], rifampicin (RIF) [7], ethambutol (EMB) [8], pyrazinamide (PZA), fluoroquinolones (Levofloxacin, Moxifloxacin, Gatifloxacin) [9, 10], streptomycin [11], amikacin [12], kanamycin [12, 13], capreomycin [14], bedaquiline (BDQ) [15], delamanid [16], pretomanid [17] and linezolid [18].

Effective TB therapy should ideally include early bactericidal action against rapidly growing organisms and subsequent sterilization of dormant populations of bacilli. The current therapy for drug-sensitive tuberculosis recommended by (WHO), is a combination of four first-line drugs, viz., rifampin (RIF), isoniazid (INH), pyrazinamide (PZA), and ethambutol (EMB) [19]. The first-line drugs, INH, RIF and

EMB, help in wiping out the actively metabolizing bacilli, while the non-replicating persisters bacilli are taken care by PZA [20], a unique drug which acts at an acidic pH. Second-line drugs are reserved to strengthen treatment when resistance arises in the first-line drugs. Second-line drugs include levofloxacin, moxifloxacin, bedaquiline, delamanid, linezolid along with pretomanid which was recommended in 2019 for the treatment of DR-TB [21].

According to World Health Organization (WHO) definition, the five main categories of drug resistance in TB are - mono-resistance TB, poly-resistance TB, rifampicin resistance TB (RR-TB), multi-drug resistant TB (MDR-TB) and extremely drug resistance (XDR-TB). To elaborate further, resistance to one first-line anti-TB drug only is referred to as mono-resistance TB, whereas, resistance to more than one first-line anti-TB drug, other than both INH and RIF is referred to as poly-resistance TB. MDR-TB is TB that is resistant to RIF and INH, the two most powerful anti-TB drugs. RR-TB refers to resistance to RIF detected using phenotypic or genotypic methods, with or without resistance to other anti-TB drugs. It includes any resistance to RIF, in the form of mono-resistance, poly-resistance, MDR or XDR. XDR strains of Mtb are MDR strains with additional resistance to fluoroquinolones and to at least one of the three injectable second-line tuberculosis drugs --- amikacin, capreomycin or kanamycin [22]. In 2019, there were an estimated 465,000 new cases of MDR-TB/RR-TB [3]. As mentioned above, both MDR-TB and RR-TB require treatment with a second-line drug regimen. MDR-TB is associated with lengthy, expensive and toxic therapy and high rates of mortality [3]. In 2019, 12,350 cases of XDR-TB were reported globally. Till now, 123 countries have reported at least one case of XDR-TB. On average, an estimated 6.2% of people diagnosed with MDR-TB have XDR-TB [22]. Drug-resistant TB (DR-TB) threatens global TB care and prevention, hence remains a major public health concern in many countries.

1.2 Drug Resistance: An increasing global public health problem

The first vaccine for TB, Bacille Calmette Guerin (BCG) was introduced in 1921 [23] and the discovery of the “magic bullet” (antibiotics) in the 1940’s revolutionized the treatment of infectious diseases

including TB. The last 70 years has seen the emergence of resistant strains to almost every anti-TB drug which was introduced for mainstream use to treat the disease [24]. Evidence shows that drug resistance in TB emanated at the same time when the first anti-TB drugs were introduced [4].

Drug resistance is a derivate of bacterial evolution which can occur either via the modification of vertically inherited genes (duplication of genes or neo-functionalization) or acquisition of new genes (transformation, transduction and conjugation). The process of acquiring new genes is also referred to as horizontal gene transfer and enables the micro-organism to exploit new conditions such as a pathogenic lifestyle [25, 26]. Mtb is an exception with no evidence for horizontal gene transfer, as this pathogen is devoid of plasmids, as well as transfer of genomic DNA [27]. In Mtb, genetically encoded drug resistance arises exclusively through spontaneous *de novo* chromosomal mutations [28], which comprises of single nucleotide polymorphisms (SNP's) or nucleotide insertions and deletions (indels) [4].

Mtb has a clonal mode of reproduction and a very low mutation rate, which make it an *a priori* unlikely resistance threat [29]. Despite this, it has been a challenge to understand the dynamics and survival of this pathogen and there have been more questions than answers with the emergence of MDR and XDR strains in recent years. Comparative genomic analysis [30] reveals that high-level of drug resistance in TB is most exclusively through chromosomal mutations in genes required for antibiotic action i.e. the drug binding site (target protein) or the enzyme required for pro-drug activation. Additional intrinsic mechanisms that contribute to drug resistance in mycobacteria include the production of drug-modifying and drug-inactivating enzymes [31], low cell wall permeability [32], and efflux-related mechanisms [33]. This suggests that drug resistance in Mtb may be more complex and drug resistance can be attributed to spontaneous mutations in drug targets genes and /or upregulation of efflux pumps.

For most bacterial species, resistance-conferring mutation often confer a biological cost that presents a selective growth disadvantage relative to the growth capability of drug susceptible isogenic strains in the absence of the drug [34]. Anti-TB drugs target essential genes of *Mycobacterium* which are functionally

and physiologically important for the growth of the organism and therefore imposes a strong selective pressure associated with antibiotic resistance and is referred to as "fitness cost" [34]. There are multiple parameters that are involved with fitness and influence the short-term competitiveness of specific mutants/lineages and, in turn, long term evolution within different host and host population. "Success" of a strain/lineage is considered as the longevity of the pathogen within a specific environment or host population. Therefore, for an obligate pathogen the success of a specific lineage is defined as the ability to establish an infection, to replicate and persist within a host, and capacity to transmit [35]. Furthermore, compensatory mutations can reduce the initial fitness defects caused by a specific drug resistance-conferring mutation [36]. Gagnuex *et al*. measured the growth rates of RIF resistant Mtb mutants relative to drug-susceptible parental strains and showed that competitive fitness was dependent on both the nature of the mutation and the strain genotype. Therefore, presence or absence of compensatory mutations is directly correlated to the strain fitness [37].

Although, the number of cases of DR-TB is relatively small compared to drug susceptible TB, drug resistant TB poses a greater threat and a disproportionate burden on the public health systems. The emergence is significantly higher in endemic regions because of chaotic treatment models. Physicians in these settings are often forced to choose among the available anti-tubercular agents depending on the patient's disease and financial status, the cost of drugs, and the tolerability profile [38]. Since most of these alternatives have poor tolerability and are moderately effective at best, the treatment outcomes are hardly encouraging. Moreover, test to determine the effectivity of the drug are too expensive and time-consuming. All these factors are responsible for the rapid spread of drug resistance. Therefore, we need multiple interventions to control the global spread of drug resistance, one of which will plausibly include individualized therapy based on rapid comprehensive drug susceptibility testing (DST) [39].

1.3 Experimental methods to quantify drug resistance

DST for Mtb is usually determined by either observing growth or metabolic inhibition in a medium containing anti-TB drugs or it could be detected at the molecular level by looking into the mutations in the genes responsible for drug action. From a technical standpoint DST involves a) macroscopic observation of Mtb growth in a drug-free or/and drug- containing media b) lysing with a mycobacteriophage c) measuring metabolic activity or generation of products and d) using molecular techniques to detect genetic mutation [40].

Culture-based phenotypic DSTs are currently the gold standards for determining drug resistance in TB. Traditionally, DST relies on a single critical concentration which is used to differentiate between a susceptible and resistant Mtb isolate and is specific for each anti-TB drug and test method. Laboratory testing to determine the susceptibility profile of Mtb serves three main purposes: 1) they help determine the chemotherapeutic regimen to be given to the patient; 2) helps to confirm emergence of drug resistance when a patient fails treatment or fails to show satisfactory recovery; 3) useful to conduct surveys to study emergence of drug resistance [41]. In clinical practice, confirmation of DR-TB is primarily by phenotypic drug-susceptibility testing on slowly-growing Mtb cultures. It's a time-consuming process, and the delay results in improper treatment leading to higher mortality and transmission rates of drug-resistant strains [42].

Current molecular genetic based tests, such as the Gene® Xpert MTB/RIF [43] and GenoType® MTBDRplus [44], have accelerated the clinical detection of known mutations causing RIF and/or INH resistance. But these genotypic susceptibility testing techniques for Mtb can only elucidate resistance profiles based on known mutations [45]. Therefore, using high-throughput sequencing to diagnose patients and identify drug resistance mutations is gathering more interest in recent times as it is fast, accurate, sensitive and economic. This helps with correct treatment strategies for patients and even with public health policy guidance by following the spread of resistance [46]. Direct sequencing involves drug-

resistant loci amplification and sequencing to detect mutations from smear-positive samples [47]. The entire process, from extracting DNA from the sputum samples to reporting of results, can be accomplished in 3 days. Comparing this to standard culture-based DST testing which takes around one to three months. Therefore, this method has many advantages, especially in determining novel mutations.

The rapid developments in high-throughput sequencing have created vast opportunities to understand the link between our genomes and phenotypes especially with the dramatic drops in the cost of these screening processes. This helped in the expansion of multiple avenues such as targeted therapies, personalized medicines and public health policies. To fully exploit the potential of these recent developments and to bridge the gap between the genotype and the corresponding phenotype, we need further understanding of the molecular consequences of novel mutation [46] and how do they impact strain fitness.

1.4 Understanding drug resistance mutations using protein structures

Although several new diagnostics or diagnosis methods have been introduced by WHO since 2007, there is still a need for simpler, rapid and readily applicable tools. The advent of high-throughput techniques like whole genome sequencing and saturation mutagenesis comes as a relief as it provides wealth of information related to phenotype and mutations, but the susceptibility associated with the novel mutations is generally unknown and therefore, cannot guide clinical therapeutic decisions. While experimental and clinical knowledge on new mutations and the cost they exert on the reproductive fitness of the organism will always provide the gold standard for predicting and identifying drug resistance; robust, accurate and scalable computational structure-based approaches can be used to complement this limited available information by providing the power to look at novel mutations [46]. Mtb is an ideal pathogen to apply structural and sequence-based mutational analysis approaches as it has a clonal population structure. It has been recently seen experimental approaches like saturation mutagenesis being integrated to a

biological assay output tool to predict high-resolution, functional dissection of mutations [48]. Although complete understanding of the functional consequence of these mutation is still a challenge because the effects are multifactorial and complex [49].

Significant progress has been made in the past few years with respect to innovative tools to understand and quantify the various ways in which a mutation or a series of mutation would give rise to a phenotype and though we can relate few mutations identified through these techniques are strongly associated with anti-TB drug resistance, frequently mutations with unknown significance or limited understanding of associated functional biological pathways are identified. To overcome this challenge, we need an effective computational approach where protein structural information can be used to decipher the complex genomic background patterns to shed light on the molecular mechanism of resistance and emergence of a phenotype. We have seen initial efforts in building predictors and databases for certain proteins and diseases, but they can be too cumbersome to be used by a geneticist to complement experimental evidence.

To address the current issues, I developed a novel methodological pipeline (Figure 1) to build a robust and user-friendly empirical classifier that can be used to determine novel drug resistance in TB. It is a computational tool which could be used to rapidly translate sequencing data into clinical application. The three main steps involved in developing the classifier are:

1) **Data set collection and curation** - collection and curation of experimental and clinical data on mutational effects linked to phenotype in comprehensive databases. This information forms the evidence set necessary for the proposal of novel computational methods as well as the improvement of current approaches.

2) **Feature generation** – multiple *in-silico* approaches based on evolutionary and physiochemical evidence have been used to build tools to predict the effect of mutation on protein stability and function. These methods include sequence and structure-based tools to understand the effect of amino acid

substitution in a specific protein, which is usually the drug target. The different sequence-based approaches which are well established amongst others are SIFT [50], PROVEAN [51] and SNAP2 [52]. These sequence-based tools can be applied to naturally occurring non-synonymous (nsSNP) polymorphisms and laboratory-induced missense mutations. Pioneering structure-based approaches, SDM [53], uses environment-specific substitution tables of protein families to derive a statistical potential energy function. mCSM-Stability [54], which uses graph-based signatures to represent the three-dimensional environment of the wild-type residue can be used to quantitatively predict the changes upon mutation in the Gibbs free energy ($\Delta\Delta G$) of folding. A newer method to predict changes in protein stability is DUET [55], which takes advantage of the relative strengths of the two different approaches (SDM and mCSM-Stability) mentioned above. Other methods based on graph signatures that consider properties to understand the effects of mutations on the recognition of binding partners, including proteins, nucleic acid and other ligands are: protein-protein (mCSM-PPI) [54], protein-nucleic acid (mCSM-NA) affinities [54], and protein-ligand affinity (mCSM-Lig) [56]. ENCoM [57] is a tool which helps predict the effect of mutations on the thermos-stability and dynamics as well as to generate geometrically realistic conformational ensembles. DynaMut, integrates graph-based signatures along with normal mode dynamics to generate a consensus prediction of the impact of a mutation on protein stability and flexibility [58]. Thus, the information generated from these features helps us understand the underlying molecular consequences of drug resistant mutations. The scores generated from these methods help in distinguishing between susceptible and resistant TB mutations.

3) **Supervised machine learning and webserver development** – In the final step the best performing features (features which can differentiate susceptible from resistant mutations) are chosen and the data set is trained, tested and validated using a supervised machine learning algorithm to accurately predict the susceptibility profile of the variant. The predictor is deployed as a user-friendly freely available webserver referred to as SUSPECT (Structural Susceptibility Prediction).

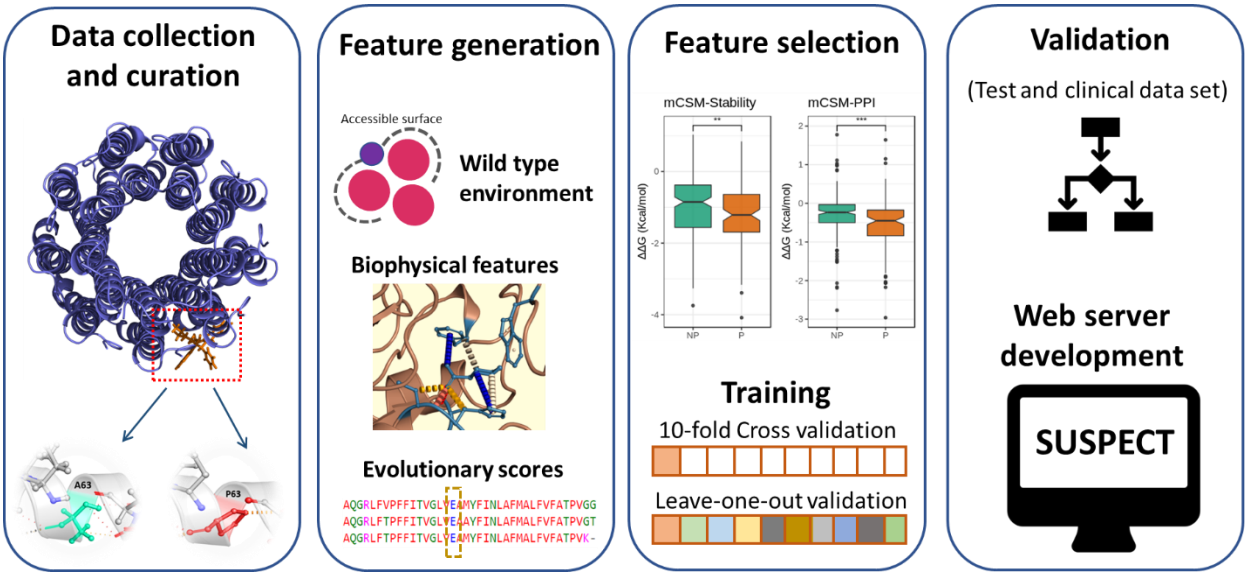


Figure 1: Methodological pipeline to develop the empirical classifier. The first step is data collection from different TB databases and experimental evidence available from high-precision laboratory studies. The mutations are mapped on a high-resolution protein structure which is generally available from protein data bank or can be homology modelled. Curated mutations are mapped on to the protein structure to observe the spread of resistance and susceptible mutations. Second step involves generating the score for the various in-silico tools. These features help us understand the functional and molecular consequences of the resistant mutations. The third step is to evaluate these features and identify underlying patterns which can distinguish resistant from a susceptible variant using supervised machine learning algorithm. The algorithm is tested and validated using clinical datasets to assess its robustness. Finally, it is deployed as a user friendly freely available webserver called SUSPECT (Structural Susceptibility Prediction).

I used the above novel pipeline to understand and develop the predictive tool for two anti-TB drugs -

1. Pyrazinamide: PZA is an important first-line sterilizing drug [59] for TB treatment as it can kill dormant Mtb bacilli at an acidic pH and shorten treatment duration for patients diagnosed with drug-susceptible TB (DS-TB), MDR-TB and XDR-TB and reduces TB relapse rates [60, 61]. PZA usage is reported to have higher success rates in treating MDR-TB patients [62]. PZA is probably the only drug

which could be part of new regimens for shortening treatment courses for all forms of TB. Being such a crucial first-line drug, culture-based methods to perform PZA susceptibility testing is difficult and produces unreliable results. It requires an acidic pH to inhibit bacterial growth and a larger inoculum volume which interferes with PZA activity [63]. The current method recommended by the WHO is the automated Bactec MGIT 960 liquid culture system (Sparks, MD) for phenotypic-PZA susceptibility testing. This method needs a proper laboratory set up, which is difficult in high TB burden countries, and produces a high rate of false-positive resistance results.

PZA is a pro-drug which is converted into its active form pyrazinoic acid with the help of the enzyme pyrazinamidase (PncA). Resistance in PZA is mostly associated with mutations in PncA, which lead to a reduction or loss of PncA's activity. However, several other mechanisms of actions has been reported [64-67], among which many resistance mutations mapped to the panD gene [68, 69]. PanD, part of the pantothenate biosynthetic pathway, is an aspartate decarboxylase responsible for the formation of β-alanine from L-aspartate, which essential for vitamin B5 and coenzyme A biosynthesis in Mtb [70].

Mapping clinical resistance mutations on to the structure of PncA revealed the mutations were spread throughout the protein structure (Figure 2). This explains the high rate of false positive resistance associated with the drug susceptibility testing (DST) for PZA.

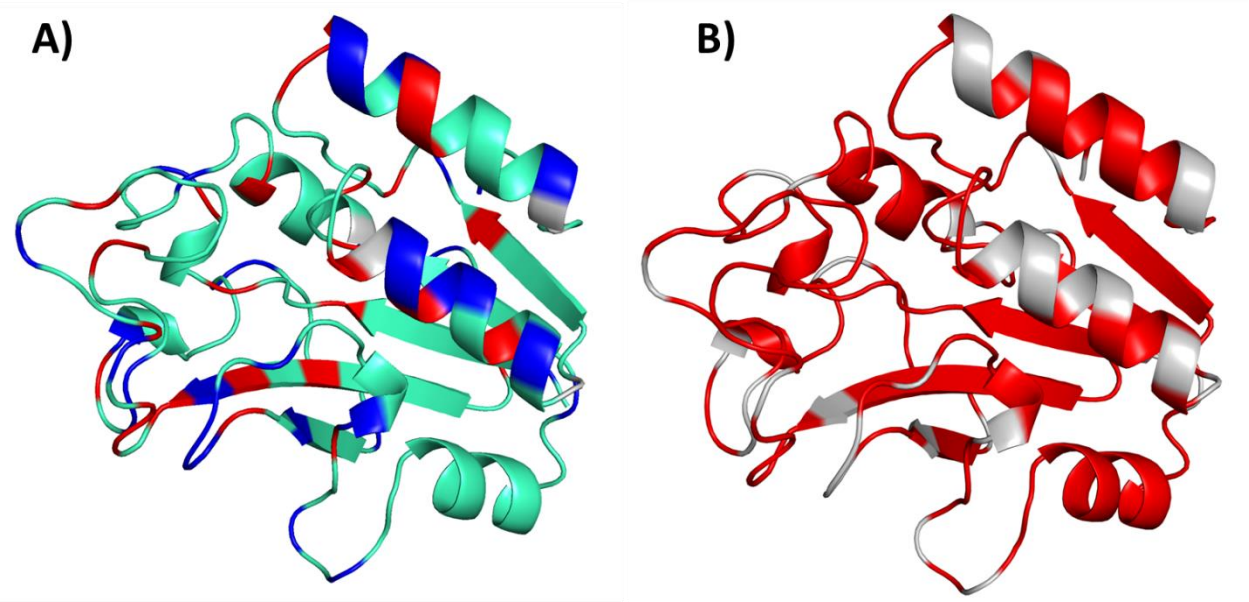


Figure 2: Protein Structure of PncA (pyrazinamidase). A) After data curation, the resistant (red) and susceptible (blue) variants are mapped on the protein structure. Cyan represents the amino acid positions which were reported to have both resistant and susceptible variation. B) Clinical resistant (red) mutations were mapped on to the protein structure. This highlights the complexity associated with determining resistance and the higher rates of false resistance detection.

To overcome issue of unreliable DST for PZA, structural information was used to guide the genetic detection of resistance. The supervised machine learning model was implemented as a web-server SUSPECT-PZA [71] (http://biosig.unimelb.edu.au/suspect_pza/), which would enable the rapid structural evaluation of the functional and phenotypic consequences of any *pncA* nsSNP mutation to support informed clinical decisions. The pipeline was further used to evaluate the mechanistic consequences of a frameshift mutations of a Victorian tuberculosis patient in real-time. The analysis led to identification of a novel resistant strain, and altered patient treatment – which was the first reported use of structural information to guide clinical resistance detection [72]. This work has been published as two papers and forms Chapter 3 of my thesis.

Bedaquiline: BDQ is a diarylquinoline with a new mechanism of action. It binds to the c-subunit (AtpE) of ATP-synthase, an essential enzyme involved with the energy production in Mtb and inhibits its activity [73]. ATP hydrolysis by ATP synthase with truncated α-subunit is inhibited by BDQ in a concentration dependent manner [74]. Micromolar concentration of BDQ is required for bactericidal activity, although nanomolar concentration inhibits mycobacterial growth [75, 76]. It was observed, at these concentrations, BDQ appears to dissipate the proton motive force, causing proton leak by disrupting the interface between α- and c-subunits [74, 76].

Due to its high selectivity towards mycobacterial ATP synthase, it less likely to produce target-based toxicity compared to homologous eukaryotic enzyme (Selectivity Index >20 000) [77]. BDQ has activity against actively replicating and dormant bacilli, hence, its usage in the recent years has expanded considerably especially for MDR-TB where it has shown higher cure rates [78]. However, clinical failure was observed [79] which rings the bell for a better understanding of how resistant variants arise to aid in the early detection of resistance [80].

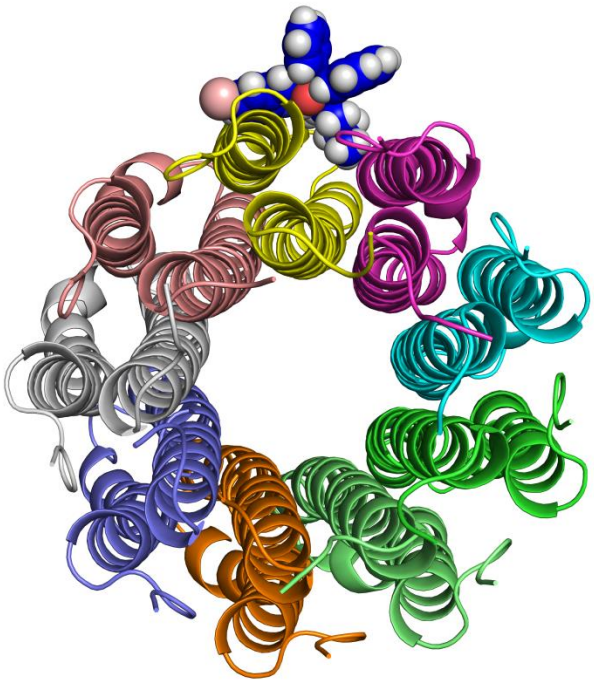


Figure 3: The crystal structure of Bedaquiline bound to AtpE. The c-subunit of ATP synthase (AtpE) assembles as a homo-nonamer and is a transmembrane protein. BDQ binds to the cleft between two adjacent monomers and interferes with the proton –binding residue Glu 61.

Being a new drug, the WHO strongly urges the development of an accurate and reproducible DST for BDQ. In the absence of specific DST, WHO recommends BDQ resistance should be monitored through MIC assessment with resistance development evaluated in patients with treatment failure or relapse [81]. Characterizing resistance mutations early would assist TB patient management and avoid treating individuals with ineffective toxic regimens [82, 83]. With few known resistance variants being identified [84], rapid genotypic prediction of BDQ resistance is limited.

Structural information was considered to support rapid identification of potential BDQ resistance mutations which could help guide clinical inference on genomic variants. Using the above novel methodology, in which comprehensive combination of structure and sequence-based tools were used to train a supervised machine learning algorithm to predict novel drug resistance mutations in BDQ. The model is deployed as free available user-friendly webserver - SUSPECT-BDQ (http://biosig.unimelb.edu.au/suspect_bdq/) [80]. This work has been published and forms chapter 4 of my thesis.

1.5 Epidemiological modeling and fitness cost

Epidemiology is the science of public health. The study relates to distribution and determinants of events or disease in a population with an overall aim to control public health problems. Epidemiological studies ranges from cluster examination at the individual level to building mathematical models to simulate disease dynamics at the population level [85]. With respect to the thesis, the molecular data generated from structural studies can be further deployed to study lineage specific transmission in TB using clustering data and understand drug amplification using mathematical models of tuberculosis transmission.

The ongoing debate on the extent to which MDR-TB is a global pandemic or a local problem which can be managed by the proper implementation of currently recommended strategies, centers on relative "fitness" of drug-resistant strains [86]. There are various approaches to estimate the fitness of TB strains. The concept of "fitness" is derived from the disciplines of ecology and evolutionary biology and implies the existence of heritable variation among individual members of a species [87]. For infectious pathogens, fitness is a composite measure of an organism's ability to survive, reproduce, and be transmitted. It indicates growth characteristics of an individual within the host, ability to withstand environmental stress within and between-host, and ability to disseminate and establish itself in a new host. Few of these traits can be quantified in the laboratory setting, which includes growth rate measurement, adaptations to withstand certain challenges and infectivity in animal models, but their empiric success in the real world is not very well translated. Epidemiological fitness could be an alternative approach to assess "fitness" of an organism. Epidemic potential is calculated by looking into the average number of secondary cases or secondary infections caused by a specific genotype after being introduced to a completely susceptible population. The information on estimates can be obtained from clustering studies, model-based studies and traditional epidemiological investigation studies. As these epidemic estimates are based on human population, they serve to be more reliable and realistic compared to microbial behavior from laboratory experiments [88].

1.5.1 Population-based studies

The underlying principle for epidemiological studies dealing with relative fitness is to compare the basic reproductive number (R_0) of resistant and sensitive organisms and thereby establish whether a person who harbored a resistant strain would cause the same number of secondary cases as a person with a sensitive strain. Thus, R_0 is a hypothetical construct representing the cases caused by a single infectious host in an entirely susceptible population. An alternative strategy to measure fitness / transmission involves looking for number of people who were either infected or developed a disease with sensitive and resistant strains; an approach where the frequency and size of "clusters" are compared. A cluster is defined as a group of

cases in a community whose isolates share similar or identical DNA fingerprints and are therefore presumably "epidemiologically" related—i.e., a cluster includes members of a transmission chain or network [88]. Molecular epidemiology is a field of study which helps in understanding Mtb transmission and outbreaks using clustering investigation. Polymorphisms present in the mycobacterial genome are exploited and are used as genetic markers. DNA typing methods utilizes these genetic markers to differentiate between strains and obtain evolutionary relationships. Commonly used molecular genotyping methods (Figure 4) are:

IS*6110* – RFLP Analysis: The mycobacterial genome contains a large amount of repetitive DNA elements, which vary in location, length and structure [89]. The two main groups are tandem repeats, which are, short monomeric sequences (up to 100 bps) organized as head-to-tail arrays, and interspersed repeats, which are, scattered as individual copies throughout the entire genome. The interspersed repeats could be mobile genetic elements and referred as insertion sequence (IS). Thierry et al. [90] in the early 1990's investigated the IS*6110* which is the best known IS. IS*6110* is 1355 bp long and belongs to the IS3 family with a unique 28 bp terminal inverted repeats. Two overlapping reading frames, orfA and orfB, encoding a transposase is located between the repeats. The transposase enzyme is responsible for the transposition of the insertion sequence [91]. The copy number of IS*6110* varies in between 0 to 25 and is dependent on the frequency of the transposition. IS*6110* can integrate anywhere in the chromosome, but the coding regions of the DNA have a higher frequency for transposition and are referred to as hot-spots [92]. Thus, difference in copy number and locations within the genome responsible for higher polymorphisms of IS*6110*, makes it a suitable candidate as a molecular marker for genotyping of mycobacterial strains [93].

The main advantages of IS*6110* are its high discriminatory power and reproducibility. Although IS*6110* is highly unstable, its transposition events are very rare [94], making it a reliable method to discriminate epidemiologically related from non-related strains. With the usage of the method from 1990's it has been standardized over time and fingerprint generated at different experimental labs can be compared and

catalogued [95]. A major limitation of the method is its applicability in strains with low copy number for IS*6110*. Another technical limitation of the tool includes need for 2-3 μg of high-quality DNA sample, which requires prior culture of the bacterial isolates. The method needs skilled personal and sophisticated computer soft wares. Despite these limitations, IS*6110* has been thoroughly used for genotyping mycobacterial strains.

Spoligotyping: Spacer oligonucleotide typing is a polymerase chain reaction (PCR) based approach to differentiate mycobacterial strains. It is based on the polymorphism of the direct repeat (DR) locus which belongs to clustered regularly interspersed short palindromic repeats (CRISPRs) family of repetitive DNA [96]. DR regions are composed of multiple direct variant repeat sequences, each comprising a series of well-conserved 36 bp DRs interspersed with unique, non-repetitive spacer sequences of 34–41 bp [97]. 43 types of spacer are revealed that separate DRs in a specific locus of the *Mtb* genome, of which 37 are typical to Mtb (spacers 1–19, 22–32, and 37–43) and the rest (spacers 20-21 and 33–36) is used to analyse *M. bovis* strains [98]. In practice, the DR locus is first amplified using PCR and then hybridized to a membrane with 43 covalently bound synthetic oligonucleotides representing the polymorphic spacers. The hybridization signals are detected by chemiluminescence and depending on the number of spacers that are missing from the complete 43-spacer set, individual strains are differentiated [99]. This presence or absence of spacer is read in a binary format and can be easily interpreted, computerized, and compared between different laboratories [100]. SpolDB4, an international spoligotyping database, released in 2006, describes 1,939 STs (shared types, i.e., spoligotype patterns shared by two or more isolates) and 3,370 orphan types (i.e., spoligotype patterns reported for only single isolates) from a total of 39,295 M. tuberculosis complex isolates, from 122 countries [101]. Recently, SITVIT, a publicly available multi-marker database was published. It consists of 7105 spoligotype patterns (corresponding to 58,180 clinical isolates) - grouped into 2740 shared types containing 53,816 clinical isolates and 4364 orphan patterns [102].

Spoligotyping is an accurate, cost-effective, simple, reproducible and high-throughput method where results are obtained in 2 days. Because it targets only a single genetic locus, covering less than 0.1% of the M. tuberculosis complex genome, this method has limited discriminatory powers. Spoligotyping is used to discriminate strains with low IS*6110* copy number.

MIRU-VNTR: stands for Mycobacterial interspersed repetitive units - variable number tandem repeat. Mtb was among the first bacterial species where tandem repeats resembling mini-satellites of eukaryotic genome was found. VNTR can be used a genetic marker which provides data in a format based structure on the number of repetitive polymorphic regions in the mini and micro- satellite regions [103]. The first described VNTRs were exact tandem repeat and major polymorphic tandem repeat [104]. Supply *et. al.* described a new VNTR element called MIRU which was as 46– 101 bp tandem repeats scattered at 41 loci throughout the mycobacterial genome [105]. Of these 41 MIRU loci, 12 are identified as hypervariable repetitive units. MIRU-VNTR analysis involves PCR amplification of certain MIRU loci followed by determination of amplicon size by gel electrophoresis. Alternatively, multiplex PCRs are run on an automated fluorescence-based sequencer. As the size of the repeat units is already known, calculated sizes reflect the number of MIRU-VNTR copies amplified. The result is a multi-digit numerical code also referred to as MIRU-VNTR code. The advantage of being of the result being in a digital format is that it can be shared across labs around the world and a global database has been established, (http://www.miru-vntrplus.org/), which can be used for for large-scale epidemiological and population genetic studies [106, 107].

MIRU-VNTR is a simple and efficient method whose discriminatory power is dependent on the number of loci being assessed. When 12 loci are being used for evaluation, the discriminatory power is higher for strains with a low copy number of IS*6110*; the power reduces with high copy number of IS*6110* [100]. 12 MIRU-VNTR cannot be used as a sole typing method for large population-based studies, as it may overestimate the number of true epidemiological links [108]. MIRU is generally combined with spoligotyping, when more than 12 loci are used, and the discriminatory power is equivalent to IS6110.

Current recommendation for molecular phylogenetic epidemiologic studies are 15 and 24 MIRU loci [109].

24 MIRU-VNTR, considered the gold standard of genotyping, is specifically recommended for typing Beijing strains [110]. Though this is further challenged by Allix-Béguec *et al* and he showed an additional seven hypervariable MIRU-VNTR loci are required to produce a higher resolution and lower clustering rate for the Beijing strain [111]. Overall, for large epidemiologic investigation, genotyping using mini-satellites is a fast, reliable and highly discriminatory approach [112].

Whole Genome Sequencing (WGS): WGS and next-generation sequencing (NGS) are the two new methods in molecular epidemiology, where the entire genome is sequenced, to enable identification of mycobacterial lineages and facilitate phylogenetic and evolutionary traits studies. WGS is a relatively fast and affordable method where the DNA is sheared into several sizes and sub-cloned into plasmids. Sub-clones are then oversampled to generate sequencing reads which provide the necessary information for performing whole genome assembly algorithms [113]. The classical genotyping techniques provide information on part of the transmission chain, whereas WGS helps in determining the chain of transmission events [114]. Frederick Sanger and his co-workers developed the Sanger sequencing method which is the most commonly used method for sequencing [115].

Upon molecular genotyping, clinical isolates of Beijing family had identical IS6110 RFLP pattern, spoligotyping pattern and MIRU-VNTR profile. Niemann *et al* used WGS to discriminate between the two isolates and found differences in 130 SNPs and a large deletion, suggesting epidemiological link between the isolates may have been remote [116]. This proves WGS will be the gold standard for typing strains in the near future with higher power of resolution.

The method does have certain limitations like the need of specialized software to analyse sequences and incomplete understanding associated with the various polymorphisms. Moreover, an important requirement are culture-positive samples to have sufficient material to perform WGS [117]. An

alternative is NGS which generates millions of short read of the entire genome to identify SNPs, perform comparative genomics and explore various aspects about transmission dynamics [118]. NGS does not require cloning of the DNA template into the bacterial vector and are optimally suited for re-sequencing. To name a few platforms available to perform NGS are: *Roche/454 FLX pyrosequencer* [118], *Illumina/Solexa Genome Analyzer [113], Pacific Biosciences Single Molecule Real Time (SMRT)*[118].

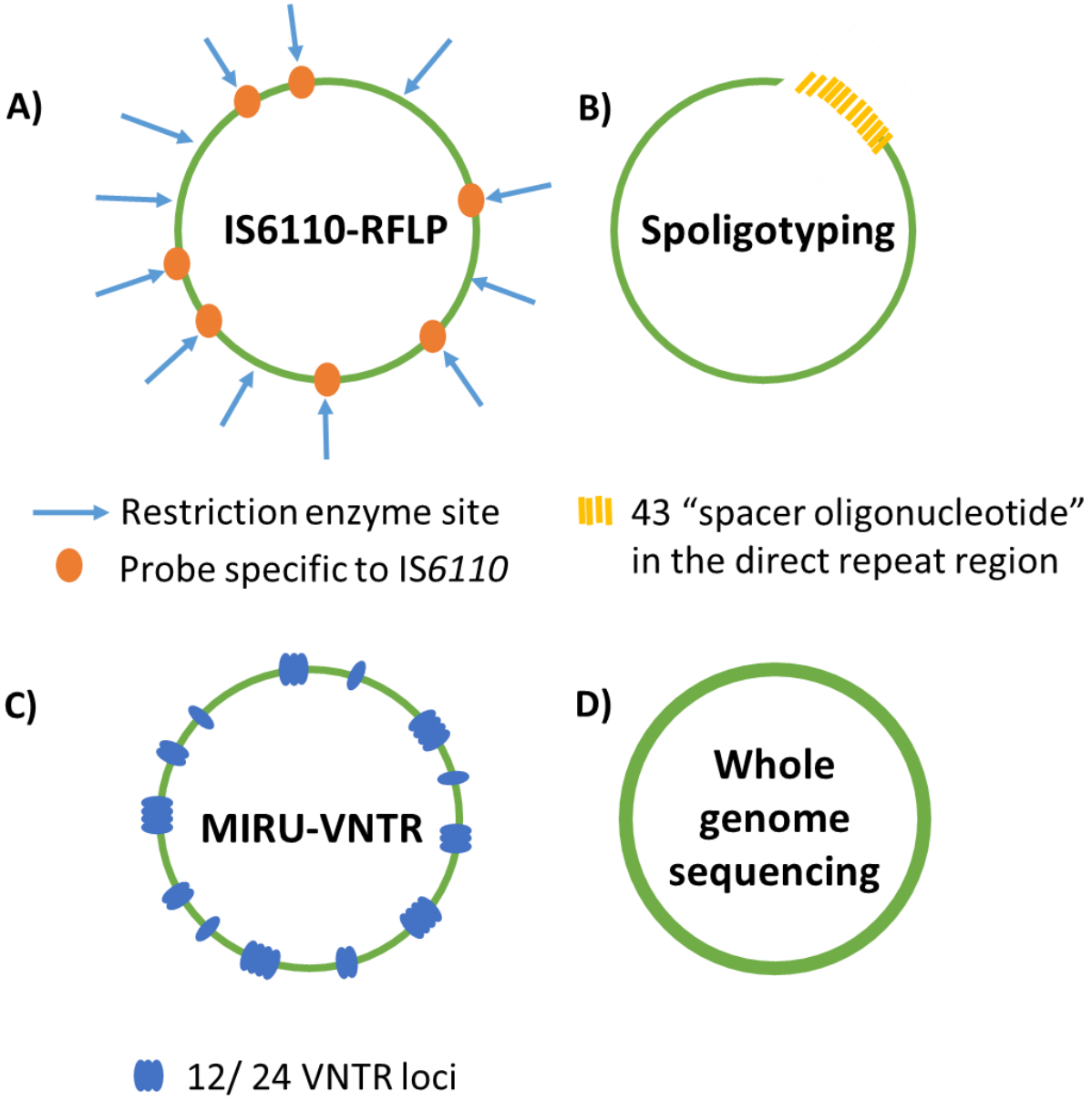


Figure 4: Molecular Genotyping Methods. The green circular band represents the MtB genome. A) IS6110-RFLP – the restriction enzyme cuts the genome (at places shown by the blue arrow) and the resulting fragments are visualized and separated using gel electrophoresis. The orange dots represent the probe specific to IS*6110* insertion element, and vary by position and count between isolates, resulting in distinct banding patterns. B) Spoligotyping – hybridization assay used to detect the presence or absence of the 43-spacer oligonucleotide present in the direct repeat region (hashed lines). The pattern is converted into a binary followed by an octal code. C) MIRU-VNTR – the 12 and 24 loci (blue ovals) are amplified using PCR and the product separated using gel electrophoresis. Repeats are calculated and converted to digital code to facilitate comparison against database. D) Whole genome sequencing (WGS) – The whole genome is analysed; SNPs helps in understanding relationship between isolates.

(The above figure is adapted from Guthrie JL, Gardy JL. A brief primer on genomic epidemiology: lessons learned from Mycobacterium tuberculosis. Ann N Y Acad Sci. 2017 Jan;1388(1):59-77. doi: 10.1111/nyas.13273. Epub 2016 Dec 23. PMID: 28009051 [119])

1.5.2 Secondary attack rates

A second way to estimate relative transmissibility of drug-resistant and drug-sensitive strains is to compare the secondary attack - rates of resistant and sensitive strains. This method requires the researcher to compare the number of secondary infections resulting from a single drug-resistant case with those resulting from a single drug-sensitive case [88] i.e., the number of people with a positive TST (presumably infected with tuberculosis) and/or cases of clinical tuberculosis among the household contacts of source cases.

1.5.3 Strain-specific differences in fitness and transmissibility

Prevalence of drug-resistant tuberculosis is dependent on the rate of acquisition of resistance-conferring mutations (acquired resistance) and the rate of transmission of drug-resistant strains (primary resistance).

Lower growth rate and transmissibility is observed initially in strains of mycobacteria that have acquired mutations conferring antibiotic resistance compared to their susceptible counterparts [120]. However, secondary site mutations help in ameliorated fitness costs of resistance mutations. These mutations are referred to as "compensatory mutations" and help in restoring fitness of the organism in the presence and/or absence of anti-TB drugs.

In addition to these direct effects like drug resistance-conferring mutation, strain's genetic background can significantly influence the fitness effects and transmissibility [121]. A specific mutation may hamper the relative fitness in one strain but when transferred to another strain background would be involved in increased fitness. Genomic analyses of strain collections from global sources have revealed that *M. tuberculosis* has a phylogeographic population structure (Figure 5), in which different strain lineages are associated with specific geographic regions [122]. Genotyping methods along with WGS analyses helped in reconstructing the evolutionary pathway of Mtb from a pool of recombinogenic *Mycobacterium canettii*-like strains [123] towards the clonal *M. tuberculosis* complex (MTBC) [124]. Seven main lineages were identified which causes TB in humans in different parts of the world – lineage 1 (Indo Oceanic), lineage 2 (East Asian), lineage 3 (East African Indian), lineage 4 (Euro American), lineage 5 (West African 1), lineage 6 (West African 2) and lineage 7 (Ethiopian). Lineage 1, lineage 2, lineage 3, lineage 4 and lineage 7 belong to Mtb and lineage 5 and lineage 6 belong to *M. africanum*. Additionally, animal adapted strains were identified which causes infection in different mammalian species and shares a common ancestor with *M. africanum* [125]. TbD1, defined as the Mtb specific deletion 1 region, represents the loss of 2153 bp genomic segment. It is seen that lineage 2, lineage 3 and lineage 4 diverged after a shared evolutionary bottleneck and have the TbD1 deleted [123, 126]. These lineages are referred to as the "modern" Mtb lineages and are widely spread. The lineages with an intact TbD1, also known as "ancient / ancestral" lineages are more often endemic and restricted to a given geographical region [127]. Modern Mtb sub-lineages include Beijing (lineage 2), CAS/Delhi (lineage 3) and the LAM and Haarlem (lineage 4) strains and are associated with global TB epidemics [128, 129].

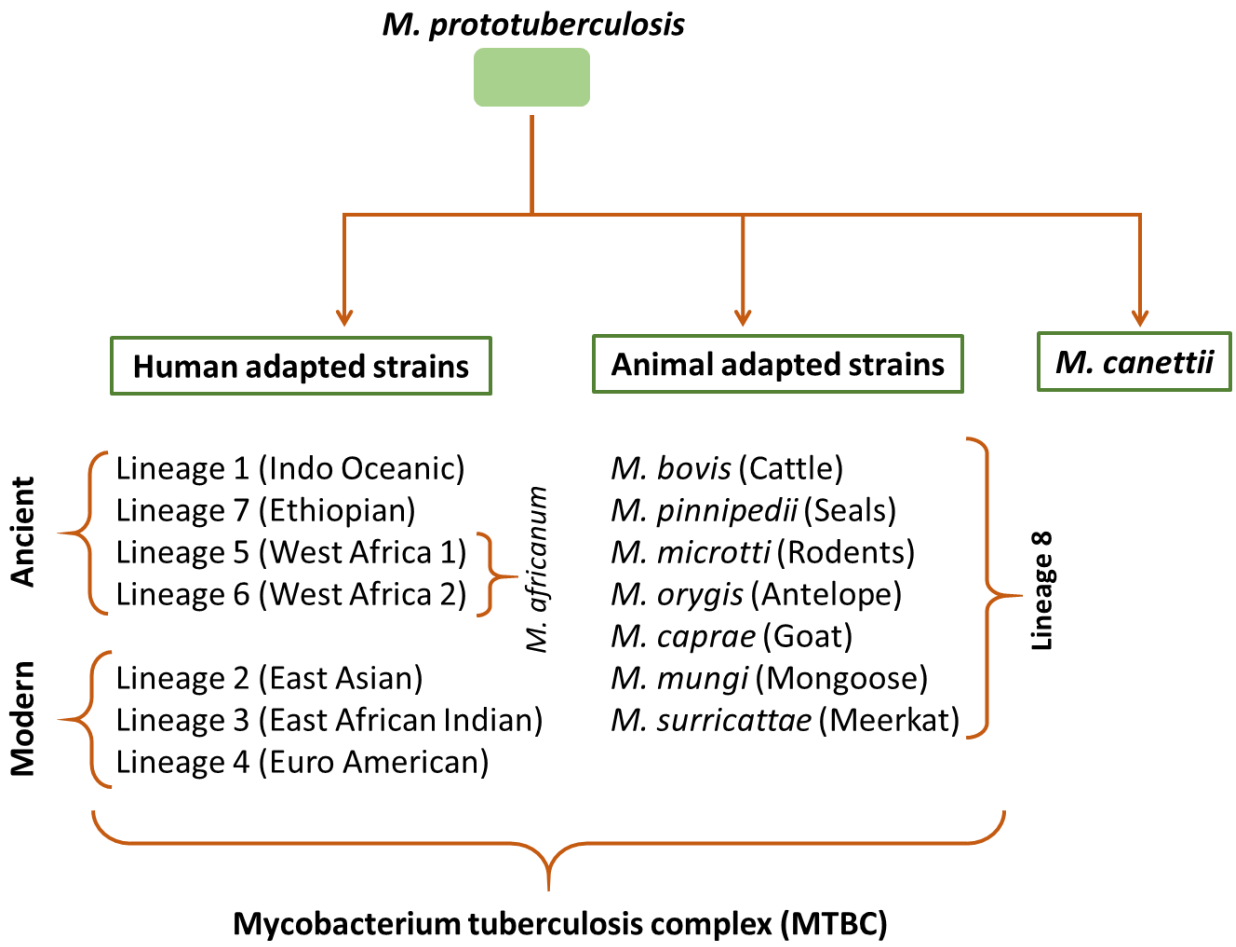


Figure 5: Mycobacterium tuberculosis lineage classification. An evolutionary scheme of MTBC, depicts adaptation of different human and animal species.

In the past few years, the Beijing lineage, a sub-lineage of lineage 2 has received much attention due to increased association with antibiotic resistance, hyper-virulence and fast progression to disease [130]. The precipitous rise of Beijing lineage in certain populations around the world has led researchers to speculate that Beijing may be more transmissible than other less widely distributed Mtb lineages. Using clustering information from genotyping studies, I explored the transmission dynamics of Beijing lineage compared to the non-Beijing lineages. A thorough systematic review along with meta-analysis was conducted to study the hyper-transmissibility of Beijing lineage. The work has been published and is included as Chapter 5 in the thesis.

1.5.4 Mathematical models of Mtb transmission

Mathematical models have used the fitness cost of resistance as the primary parameter that determines both the frequency of resistance at any given level of antibiotic use and the rate at which that frequency will change with alterations in antibiotic use patterns [131]. These models have been developed to predict tuberculosis dynamics and to examine how key parameters may affect the success or failure of current policy decisions. Acquired drug resistance patterns can be used to infer selection processes during treatment and mathematical models can help in generating information on the relative impacts of treatment parameters involved in the evolution of resistance which could lead to improved treatment protocols [132].

Ideally, robust models of Mtb transmission dynamics should be able to accurately predict future trends in drug resistance across a variety of scenarios. However, current modelling capacity is limited by several factors, including ongoing controversy regarding the extent to which transmission risk varies between drug susceptible and drug resistant strains of TB [133]. Partial explanations for this uncertainty are that Mtb is more genetically diverse than is often appreciated and drug-resistant strains can exhibit heterogeneous fitness compared to drug-susceptible strains [121]. Incorporation of comprehensive strain data, including information on specific drug resistance mutations and genetic background of the strain into epidemiological studies of transmission of DR-TB is urgently required, as the transmission success of certain drug resistance mutations are dependent on the interactions of these variables. Recent studies have also suggested that many MDR TB cases result from patient-to-patient transmission rather than from the de novo acquisition of resistance during treatment [134].

Molecular epidemiological data that allow classification of isolates into genotypic classes (clusters) has been used to measure relative fitness of resistant strains. The relative fitness of resistant strains compared with that of sensitive strains has thereby been quantified from a comparison of their genetic clustering [135]. But these are indirect methods as they do not consider the dynamics of tuberculosis transmission, evolution of resistance, and mutation of molecular markers. By mathematically modeling these stochastic

processes simultaneously and applying modern computational Bayesian methods of inference [136], estimates of the relative fitness can be improved. Additionally, the cost to transmission incurred by resistance, the rate of acquisition of drug resistance due to treatment failure can also be successfully estimated. Even the relative contributions of resistance evolution (acquired) versus transmission of resistant strains (primary) can be quantified [137]. Important epidemiological parameters such as detection rates and treatment success rates have been identified using these mathematical models [132, 138, 139]. Thus, accurate estimates of underlying parameters are of critical importance to predict the spread of drug resistance [140].

MDR-TB as explained above is defined as resistance towards rifampicin and isoniazid, two important first-line drugs involved in treating TB. Isoniazid is a prodrug, which enters the Mtb cytoplasm via passive diffusion and kills actively replicating bacteria [141]. INH requires cellular activation by the enzyme katG, a catalase and peroxidase, which produces the radical form of INH [142, 143]. This entity reacts with NAD^+ to yield an INH-NAD adduct, which binds to the active site of the NADH-dependent enoyl-ACP reductase InhA. InhA is part of the mycobacterial fatty acid elongation system, fatty acid synthase type II (FASII) [144] and is involved in the reduction of monounsaturated acyl-ACP to acyl-ACP [145]. The INH-NAD adduct binds to and inhibits InhA [146] leading to disruption of mycolic acid biosynthesis and cell death [147, 148]. KatG mutations are the major cause of INH resistance in clinical isolates [149]. The other genes responsible for resistance to INH are inhA, kasA, ndh and oxyR – ahpC [150, 151].

RIF inhibits bacterial RNA polymerase, an enzyme involved in DNA transcription. This enzyme is responsible for ribonucleoside triphosphates polymerization on a DNA template and aids in catalyzing the transcription of DNA to RNA [152]. RNA polymerase consists of five subunits $\alpha_2\beta\beta'\omega$ and RIF binds to the β subunit (rpoB gene) [153]. Mutations in RIF are exclusively observed in the β subunit. 95% of the RIF resistance is located within an 81-bp region (located between codons 507 - 533) of the rpoB gene, referred to as the rifampicin resistance determining region (RRDR) [154]. GeneXpert MTB/RIF, a rapid

molecular biology technique has been recommended by WHO, to be included in national programs. This diagnostic tool which is a cartridge-based rapid automated can determine RIF resistance in the RRDR region in clinical samples in less than 2 hours [155].

Jenkins et al [156, 157] in 2011 showed that the number of new TB cases with INH resistance is increasing in several disparate geographical settings. For example, the data was consistent in low burden settings like British Columbia and Canada [158], parts of western Europe like France [159], and United States of America [160]. Similar observation was made in high-burden TB settings like Tanzania [161], India [162], Georgia [163], Viet Nam [164, 165] and former Soviet Union [156]. Moreover, only a tiny proportion of TB patients in the world have access to INH drug susceptibility testing [166]. Therefore, RIF's resistance is used as a surrogate marker for MDR-TB, as more than 90% of RIF-resistant isolates are also resistant to INH [167]. As the process to test susceptibility for INH typically relies on culture-based methods which may not be routinely performed in many global settings, patients with INH mono-resistance not identified at baseline are put on a standard regimen which results in effective rifampicin monotherapy for the latter four months of the six month treatment course. This exposure to a single drug to the remaining MTB strains increases its risk of development of multi drug-resistant TB [166].

To predict the future behavior of DR-TB within a community, it is important to construct mathematical models which can distinguish the relative contribution of primary resistance (transmission) versus secondary resistance (amplification) to the occurrence of new cases of resistance [168]. Currently, majority of mathematical models constructed provide information on MDR transmission [169], few on MDR amplification [170] rates. One study highlights the emergence of MDR-TB is likely due to transmission rather than acquisition of these strains [134]. Knowledge on the transmission and amplification rates for RIF and INH mono-resistance is still rare. This intrigued me to construct a compartmental epidemiological model which would help estimate mono-drug resistant amplification rates for INH and RIF. It is important to understand whether the rates of transmission and amplification are same for both the first-line drugs. This would provide us with information regarding gaps in the current

diagnostic assays and whether we need different methods and approaches to control the MDR-TB epidemic. This work has been written up as Chapter 6 in the thesis.

CHAPTER 2: METHODOLOGY

In this thesis I have used two major approaches, structural bioinformatics and mathematical modeling to understand and develop novel methods to tackle the rise of DR-TB. In this section I will be elaborating on both the methodological approaches.

Structural Bio-informatics

The study of functional consequences associated with mutations can be broadly classified into those that exploit the extensive structural information which are currently available for many proteins, and those that seek to understand the effects of mutations from the amino acid sequence of a protein alone [54]. The amino acid sequence, coded by three consecutive bases, forms the primary structure of the protein. The primary structure first folds into the secondary structures namely alpha-sheets and beta-helix, which is further folded into its tertiary structure for it to be functional. Tertiary structure is an overall three-dimensional shape created by interactions between polar, nonpolar, acidic, and basic R group within the polypeptide chain. Sometimes, these tertiary structures form the subunit and need to come together to form the quaternary structure. Hence, "protein folding" is crucial for the optimal functionality of the protein and changes in the amino acid level could lead to disruption.

Figure 6 represents a thermodynamic cycle of protein folding of a wild-type and mutant protein. The nascent wild-type mRNA is translated into an amino acid sequence and folds into a functional protein. This can be thermodynamically shown as:

$$\Delta G = -\mathrm{RT}\ \mathrm{lnK} \qquad \text{---------- (Equation 1)}$$

$$\Delta G = -\mathrm{RT} \ln \frac{Folded\ Wildtype}{Unfolded\ Wildtype} \qquad \text{---------- (Equation 2)}$$

Where, R = 1.985× 10−3 kcal $K^{-1}mol^{-1}$, is the ideal gas constant, T is the temperature (in Kelvin) and K is the equilibrium dissociation constant of the reaction. ΔG is the Gibbs free energy of the protein folding reaction and is the sum of the entropy (ΔS) and enthalpy (ΔH).

$$\Delta G = \Delta H - T\Delta S \qquad \text{---------- (Equation 3)}$$

Every protein molecule represents a highly ordered macroscopic structure and this unique native conformation is disrupted when variations appear in the amino acid sequence. The thermodynamic difference ($\Delta\Delta$G) can be calculated using the following equation:

$$\Delta\Delta G_{folding} = \Delta G_{folded\ wild-type} - \Delta G_{folded\ mutant} \qquad \text{---------- (Equation 4)}$$

This equation can be used to deduce information on protein-protein interaction changes, ligand binding affinities and stability upon mutation.

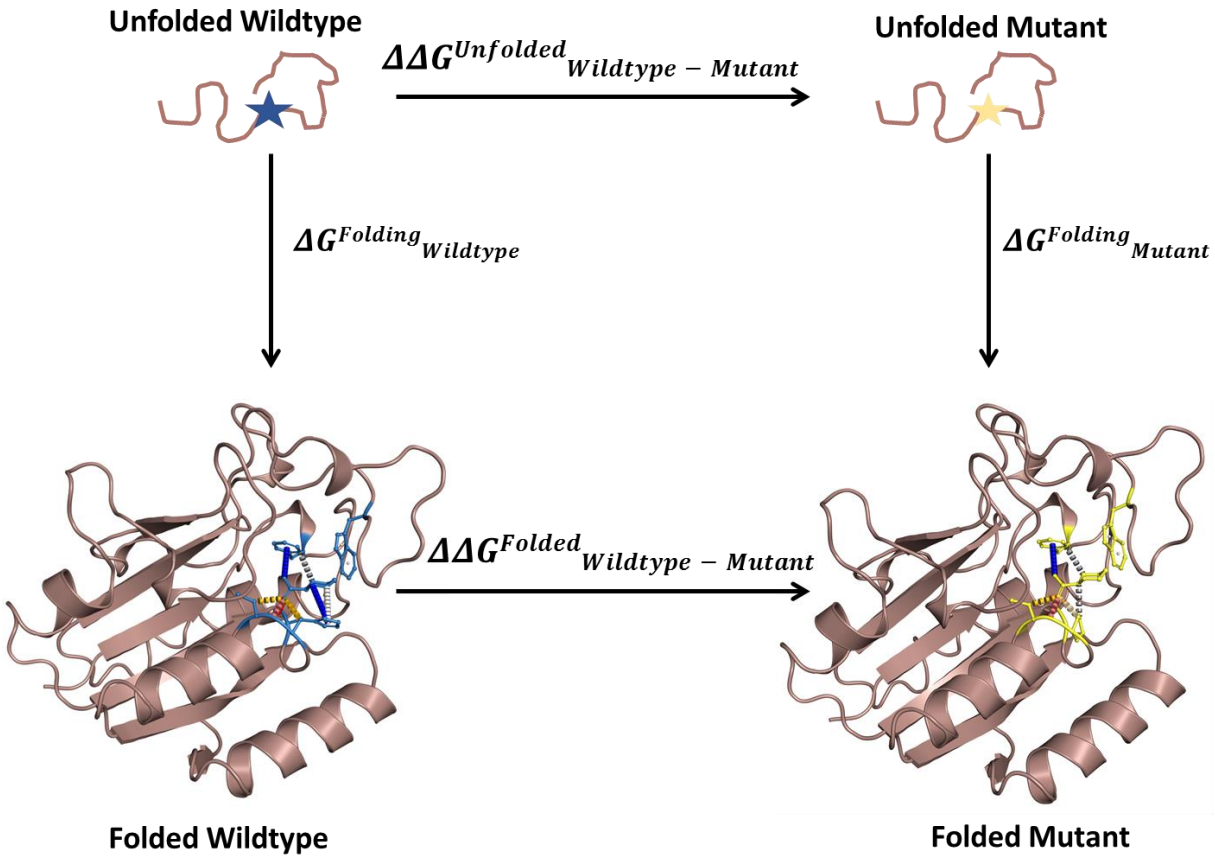


Figure 6: Thermodynamic cycle of protein folding. A schematic representation of the different states of protein (unfolded to folded) and Gibbs free equations to calculate structural variations upon mutation.

2.1 Structure based Analysis

Structure-based approaches, which may be categorized as machine learning methods and potential energy functions, typically attempt to predict either the direction of change in protein stability on mutation (as a classification task) or the actual free energy value ($\Delta\Delta G$) (as a regression task). Support vector machines have been extensively used to predict protein stability changes from both protein sequences and structures [171] [172] and more recently to predict disease-related mutations [173]. There have also been recent attempts to predict the stability changes on multisite mutations [174]. Machine learning methods have proven to be powerful predictive tools, even when data on which to train the methods have not been extensively available.

Environment-specific substitution tables, which describe the propensities of residues to mutate in a certain protein-structural environment during evolutionary time, have been used to derive a statistical potential energy function used by the method SDM [53, 175]. Empirical energy functions have also been used in a method that performed Monte Carlo optimization [176], which has also been used to study the role of conformational sampling as a way to assess the impact of single point mutations in protein structures [177]. The method Bongo [178] attempts to predict structural effects of nsSNPs by evaluating graph theoretic metrics and identifying key residues using a vertex cover algorithm. Therefore, an alternative approach to study mutations is to represent residue environments as graphs where nodes are the atoms and the edges are the physicochemical interactions established among them. From these graphs, distance patterns can also be extracted and summarized in a structural signature, which may then be used as evidence to train predictive models. The Cutoff Scanning Matrix (CSM) [179] is a protein structural signature successfully used in large-scale protein function prediction and structural classification tasks. The concept of graph-based structural signatures has been used by our group to study and predict the

impact of single-point mutations on protein stability, protein–protein interaction and protein–nucleic acid affinity. The approach, called mutation Cutoff Scanning Matrix (henceforth called mCSM), encodes distance patterns between atoms to represent protein residue environments. The suites of tools used to study nsSNPs are:

- **mCSM-Stability**: the concept of graph-based signatures is used to predict the effects of a mutation on the overall protein stability. The tool is available at http://biosig.unimelb.edu.au/mcsm/stability. To calculate the mCSM signatures for a given mutation, the wild-type environment by the atoms within a distance "r" from its geometric center is defined first. An atom distance matrix is generated by calculating the pairwise distance between the atoms of the environment and accounts for both short to long ranges of distance. From this matrix, distance patterns are then extracted and summarized as a "feature vector". This can be modelled as a contact graph, where the atoms are the nodes and the edges are defined by a cutoff distance. To consider the changes in the atom due to the mutation, "pharmacophore count vector" is introduced. The eight possible categories for the atom are - positive, negative, hydrophobic, hydrogen acceptor, aromatic, hydrogen donor, sulphur and neutral. Each one of the 20 amino acid residues are represented by a different vector, where each position denotes the frequency of a certain pharmacophore in that residue. The difference vector between the wild-type and mutant pharmacophore vectors is then appended to the signature. The ProTherm [180] data set was used assess the applicability of mCSM signatures in predicting the impact of mutations in protein stability.

$$\Delta\Delta G_{Stability} = \Delta\Delta G_{Wild-type} - \Delta\Delta G_{Mutant} \quad \text{---------- (Equation 5)}$$

Mutation was considered highly destabilizing if the ΔΔG value was ≥ 2 kcal/mol, destabilizing if ΔΔG value was between -2 kcal/mol and 0 kcal/mol, stabilizing if ΔΔG value was between 0 kcal/mol and +2 Kcal/mol and highly stabilizing if ΔΔG is ≥ +2 kcal/mol).

- **SDM**: a computational tool to predict changes in protein stability due to a single mutation using conformationally constrained environment-dependent amino acid substitution tables, available at http://marid.bioc.cam.ac.uk/sdm2. It analyses the variation in the amino acid replacements occurring at specific structural environment which are tolerated within the protein and converts them into substitution probability tables which are further used as quantitative measures for predicting the protein stability upon mutation.

- **DUET:** Uses two complementary approaches mCSM-Stability and SDM in order to create a consensus prediction to calculate the effects of a mutation on protein stability, http://biosig.unimelb.edu.au/duet/. The results of both the methods are consolidated using Support Vector Machines (SVMs) trained with Sequential Minimal Optimization.

- **mCSM-PPI**: Predicts the effects of a mutation within a specified protein on its impact with overall protein–protein interactions. The webserver is available at http://biosig.unimelb.edu.au/mcsm/protein_protein. mCSM-PPI2, creates a similar prediction to PPI but incorporates the effects of mutations on inter-residue noncovalent interaction network using graph kernels, evolutionary information, complex network metrics, and energetic terms., available at http://biosig.unimelb.edu.au/mcsm_ppi2/. PPI and PPI2 use graph-based structural signatures to represent the environment of the wild-type residue. This approach models both the geometry and physicochemical properties of the interactions and architecture of wild-type structure. The change in binding affinity upon mutation can be written as:

$$\Delta\Delta G_{PPI} = \Delta\Delta G_{Wild-type} - \Delta\Delta G_{Mutant} \quad \text{---------- (Equation 6)}$$

- **mCSM-Ligand**: a structure guided computational approach to predict the effects of single-point mutations on the stability of a protein–ligand complex, available at http://biosig.unimelb.edu.au/mcsm_lig/. Wild type environment and small molecule chemical

features are represented using graph-based signatures and changes in protein stability as evidence to train an algorithm using representative set of protein-ligand complexes from the Platinum database [56].

- **DynaMut**: is a novel method that considers molecular motions and combines graph-based signatures with coarse-grained normal mode analysis, to generates a consensus prediction of effects of mutations on the protein conformational repertoire. It is available at http://biosig.unimelb.edu.au/dynamut/. Normal Mode Analysis (NMA) is implemented in DynaMut using two different approaches Bio3D [181] and ENCoM, providing rapid and simplified access to powerful and insightful analysis of protein motions.

- **ENCoM**: is a coarse-grained normal mode analysis method to predict the effect of single point mutations on protein dynamics and thermostability resulting from vibrational entropy changes. The webserver is available at: http://bcb.med.usherbrooke.ca/encom.

- **Arpeggio**: this is webserver for calculating interactions within and between DNA, proteins and protein or small-molecule ligands. The 13 different types of interaction calculated between atoms include hydrogen bonds, carbonyl, specific atom–aromatic ring (cation–π, donor–π, halogen–π, and carbon–π), aromatic ring–aromatic ring (π–π), ionic, hydrophobic, van der Waal, halogen bonds, and metal. It is accessible at: http://biosig.unimelb.edu.au/arpeggioweb/. The server accepts user-submitted structures in addition to PDB accession codes and therefore can be used to calculate interactions for non-PDB structures such as homology models or docking poses.

2.2 Sequence-based Analysis

nsSNPs are the major contributor for development of drug resistance in TB. The sequence-based tools available to determine the how an amino acid substitution would alter protein functionality are:

- **SIFT**: Sorting Intolerant From Tolerant, is a sequence-based tool and uses sequence homology to predict whether an amino acid substitution will affect protein function and hence, potentially alter phenotype [50, 182]. The main methodology for SIFT is that it presumes important amino acids will be conserved in the protein and therefore any changes in the well-conserved regions can be predicted to be deleterious. For example, if a position in an alignment of a protein family only contains the amino acid isoleucine, it is presumed that substitution to any other amino acid is selected against and that isoleucine is necessary for protein function. Therefore, a change to any other amino acid will be predicted to be deleterious to protein function. If a position in an alignment contains the hydrophobic amino acids isoleucine, valine and leucine, then SIFT assumes, in effect, that this position can only contain amino acids with hydrophobic character. At this position, changes to other hydrophobic amino acids are usually predicted to be tolerated but changes to other residues (such as charged or polar) will be predicted to affect protein function. To predict whether an amino acid substitution in a protein will affect protein function, SIFT considers the position at which the change occurred and the type of amino acid change. Given a protein sequence, SIFT chooses related proteins and obtains an alignment of these proteins with the query. Based on the amino acids appearing at each position in the alignment, SIFT calculates the probability that an amino acid at a position is tolerated conditional on the most frequent amino acid being tolerated. If this normalized value is less than a cutoff, the substitution is predicted to be deleterious [182]. The webserver is available at: https://sift.bii.a-star.edu.sg/.

- **SNAP2**: Screening for Non-Acceptable Polymorphisms, is a method that combines many sequence analysis tools in a battery of neural networks [52]. SNAP could potentially classify all nsSNPs in all proteins into non-neutral (effect on function) and neutral (no effect) using sequence-based computationally acquired information alone. For each instance SNAP provides a reliability index, i.e. a well-calibrated measure reflecting the level of confidence of a specific

prediction. Information needed as input was obtained from protein sequence. The webserver is available at: https://rostlab.org/services/snap/.

- **PROVEAN:** Protein Variation Effect Analyzer, is a software tool which predicts whether an amino acid substitution or indel has an impact on the biological function of a protein [51]. It uses an alignment-based score approach. It is useful tool which can help filter sequence variants that are functionally important. It can predict for single nucleotide substitutions, multiple amino acid substitutions, insertions, and deletions using the same underlying scoring scheme. The webserver is available at: http://provean.jcvi.org/index.php.

- **ConSurf:** is a tool which analyses evolutionary patterns of amino acids of the protein in interest to reveal regions which are important for structure and function. The input is either a query sequence or structure; it automatically collects homologues, performs a multiple sequence alignment and constructs a phylogenetic tree which represents the evolutionary relationship. Next a probabilistic framework is used to determine evolutionary rates of each sequence position. The webserver is available at: http://consurf.tau.ac.il.

2.3 Homology modeling

The main goal of protein modeling is to predict a structure from its sequence with the accuracy that is comparable to the best results achieved experimentally. This allows users to safely use quickly generated *in silico* models in all the contexts like structure-based drug design, analysis of protein function, interactions, antigenic behavior, and rational design of proteins with increased stability or novel functions. One of the most widely used three-dimensional (3D) structure prediction approaches is homology modeling. It builds an atomic model based on experimentally determined known structures that have sequence homology of more than 40%. It is also known as **comparative modeling**.

The principle behind it is that if two proteins share a high sequence similarity, they are likely to have very similar three-dimensional structures. If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence.

Homology Modeling is moderately accurate for the positions of alpha carbons and inaccurate for side chain positions and loops. The other approaches are **threading** for <40% similarity and ***ab initio*** prediction for no homolog [183]. Homology modeling is a multi-step process that can be summarized into seven main steps.

1. **Template recognition:** The template selection involves searching the Protein Data Bank (PDB) for homologous proteins with determined structures. The search can be performed using a heuristic pairwise alignment search program like BLAST or FASTA. As a rule of thumb, a database protein should have at least 40% sequence identity, highest resolution and the most appropriate cofactors for it to be a template sequence. The protein sequence for whose 3D structure is to be predicted is the "target sequence".

2. **Sequence Alignment:** Once the template is identified, the full-length sequences of the template and target proteins need to be realigned using refined alignment algorithms to obtain optimal alignment. CLUSTALW is a very powerful multiple sequence alignment tool.

3. **Backbone Generation:** Once optimal alignment is achieved, the corresponding coordinates residues of the template proteins selected can be simply copied onto the target protein. If the two aligned residues are identical, coordinates of the side chain atoms are copied along with the main chain atoms. If multiple templates selected, then average coordinate values of the templates are used.

4. **Loop Modeling:** After the sequence alignment, there are often regions caused by insertions and deletions leads to gaps in sequence alignment. The gaps are modeled by loop modeling, which is a very problem and is also a major source of error. Currently, there are two main techniques used to approach the problem:

- The database searching method - this involves finding loops from known protein structures and aligning onto the two stem regions (main chains mostly) of the target protein. Some specialized programs like FREAD and CODA can be used.
- The ab initio method - this generates many random loops and searches for the one that has reasonably low energy and φ and ψ angles in the allowable regions in the Ramachandran plot.

5. **Side Chain Modeling:** After the main chain atoms are built, the positions of side chains should be determined. This is important in evaluating protein–ligand interactions at active sites and protein–protein interactions at the contact interface. A side chain can be built by searching every possible conformation by every torsion angle of the side chain to select the one that has the lowest interaction energy with neighboring atoms. A Rotamer library can also be used, which all the favorable side chain torsion angles extracted from has known protein crystal structures.

6. **Model Optimization:** This step includes the energy minimization procedure on the entire model, which moves the atoms in such a way that the overall conformation has the lowest energy potential. The goal of energy minimization is to relieve steric collisions without altering the overall structure. In these loop modeling and side chain modeling steps, potential energy calculations are applied to improve the model. Model refinement can also be done by Molecular Dynamic simulation which moves atoms toward a global minimum by applying various stimulation conditions (heating, cooling, considering water molecules) and has a better chance at finding the true structure.

7. **Model Validation:** There are two principally different ways to estimate errors in a structure:
    - Calculating the model's energy based on a force field
    - Determination of normality indices that describe how well a given characteristic of the model resembles the same characteristic in real structures.

2.2 Molecular Docking

Molecular docking is the process that involves placing molecules in appropriate configurations to interact with a receptor. Molecular docking is a natural process which occurs within seconds in a cell. In molecular modeling the term "molecular docking" refers to the study of how two or more molecular structures fit together. The information obtained from the docking technique can be used to suggest the binding energy, free energy and stability of complexes.

The main objective of molecular docking is to attain ligand-receptor complex with optimized conformation and with the intention of possessing less binding free energy. The net predicted binding free energy (ΔG_{bind}) is revealed in terms of various parameters, hydrogen bond (ΔG_{hbond}), electrostatic (ΔG_{elec}), torsional free energy (ΔG_{tor}), dispersion and repulsion (ΔG_{vdw}), desolvation (ΔG_{desolv}), total internal energy (ΔG_{total}) and unbound system's energy (ΔG_{unb}). Therefore, good understanding of the general ethics that govern predicted binding free energy (ΔG_{bind}) provides additional clues about the nature of various kinds of interactions leading to the molecular docking [184].

Approaches for Molecular Docking

For performing molecular docking, primarily two types of approaches are used.

1. **Simulation approach**

    Here the ligand and target are being separated by physical distance and then ligand is allowed to bind into groove of target after "definite times of moves" in its conformational space. The moves involve variations to the structure of ligand either internally (torsional angle rotations) or

externally (rotations and translations). The ligand in every move in the conformational limit releases energy, as "Total Energy". This approach is more advantageous in the sense that it is more compatible to accept ligand flexibility. Additionally, it is more real to assess the molecular recognition between ligand and target. However, this approach takes longer duration to estimate optimal docked conformer due to the large energy dissipating for each conformation. Recently, fast optimization method and grid-based tools have dominantly revolutionized this drawback to make simulation approach more user-friendly [185].

2. **Shape complementarity approach**

   This approach employs ligand and target as surface structural feature that provides their molecular interaction. Here, the surface of target is shown with respect to its solvent-accessible surface area and ligand's molecular surface is showed in terms of matching surface illustration. The complementarity between two surfaces based on shape matching illustration helps in searching the complementary groove for ligand on target surface. For example, in protein target molecules, hydrophobicity is estimated by employing number of turns in the main-chain atoms. This approach is rather quick and involves the rapid scanning of numerous thousands of ligands in a few seconds to find out the possible binding properties of ligand on target molecular surface [185, 186].

2.4 Novel methodological pipeline to build the empirical classifier

A crucial step towards establishing a genotype-phenotype correlation is the initial understanding of the molecular and functional mechanisms of the drug resistance mutations obtained from the scores of the biophysical and evolutionary tools. But looking into individual results manually can often miss underlying relationships among different mutational measurements, which can help relate them to the phenotype [187]. Supervised machine learning helps address this issue by providing a rigorous set of

algorithms, used to analyse the labelled train data set to obtain a binary classifier which is further applied to make predictions for unseen data. The identification of patterns and associations within the data helps the predictive model establish a distinction between mutations within the same gene leading to different phenotypes, and hence the development of an effective predictive tool that can be used to interpret novel clinical variants. My aim was to build a binary machine learning classifier to distinguish between susceptible and resistance mutations for pncA, a gene responsible for activation of the drug Pyrzinamide and atpE, the drug binding site for Bedaquiline. The steps involved in building a non-biased accurate classifier are:

1. Data Preparation: The first step is collecting good quality experimental data. The quality of a classifier is a direct reflection of the quality of the data used to build it. Therefore, accurate clinical sources are required to label mutations as susceptible or resistance. The clinical databases available for Mtb are:

- **GMTV**: Genome-wide Mycobacterium tuberculosis Variation database [188], a database which list genomic variations for Russian isolates. The database is available at (http://mtb.dobzhanskycenter.org).
- **TBDReaMDB**: Tuberculosis Drug Resistance Mutation Database [189], comprehensive and interactive database cataloguing mutations associated with TB drug resistance and the frequency of the most common mutations associated with resistance to specific drugs. The database is available at (http://www.tbdreamdb.com/).
- **tbvar**: database annotating potential functional and drug resistance variants of Mtb. Using a systematic computational pipeline they have created a comprehensive variome map of Mtb comprising >29,000 single nucleotide variations [190]. The database is available at (http://genome.igib.res.in/tbvar).
- **MUBII-TB-DB**: a highly structure text-based database of resistance mutations for first-line and second-line antibiotics [191]. It can detect mutations in seven

genes: *rpo*B, *pnc*A, *kat*G, *gyr*A, *gyr*B, *mab*A (*fab*G1)-*inh*A and *rrs*. The database is available at (http://umr5558-bibiserv.univ-lyon1.fr/mubii/mubii-select.cgi).

Not all mutations for all drugs used to treat TB are listed in these databases. Especially for newer drugs like Bedaquiline and Delamanid, one needs to manually look for information on resistance and susceptible mutations from the literature. Sometimes biologically relevant information like lineage or fitness cost is available for few mutations. But they need to be available for all the variants listed in the dataset to train and test the algorithms because supervised algorithms do not handle missing data labels. While curating and building the dataset we should opt for equal representation of all class labels whenever feasible.

2. Protein structure and homology modeling: Uniprot (https://www.uniprot.org/) [192] was used to obtain the sequence and functional information for the protein of interest (PncA and AtpE). The information on sequence can also be obtained from NCBI Gene (https://www.ncbi.nlm.nih.gov/gene/). To run the biophysical tools such as mCSM, a high resolution protein crystallographic structure is required which can be downloaded from Protein Data Bank (PDB: http://www.rcsb.org/) [193]. When a protein structure is not available, it is generated via homology modeling that builds an atomic model based on experimentally determined known structures that have sequence homology of more than 40%. The principle behind it is that if two proteins share a high sequence similarity, they are likely to have very similar three-dimensional structures. If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence. For tools like mCSM-PPI, mCSM-PPI2, mCSM-Lig and mCSM-NA, we might further require molecular docking. Molecular docking is the process that involves placing molecules in appropriate configurations to interact with a receptor. Molecular docking is a natural process which occurs within seconds in a cell. To mimic the process in-silico we need time and computational power. The main objective of molecular docking is to attain ligand-receptor complex with optimized conformation and with the intention of possessing less binding free energy.

Once we have both the curated dataset and the protein structure, we map the variants on to the protein structure to identify potential hotspots and can be done by using the visualizing software such as PyMol and Chimera.

3. Feature Generation: The structural and sequence-based features encompass a diverse range of mutational information as they provide descriptive knowledge for each mutation. The information can be categorized into:

- **Stability features**: Protein stability and conformational dynamics and flexibility (mCSM-Stability, DUET, SDM, DynaMut)
- **Functional features**: Changes in protein – protein interaction (mCSM-PPI, mCSM-PPI2), changes in affinity towards ligand binding (mCSM-Lig) and changes in nucleic acid association (mCSM-NA)
- **Wild-type residue environment**: Structural information of the wild-type residue, including relative solvent accessibility (RSA), residue depth, secondary structure and dihedral angles of the protein side chain φ (phi) and ψ (psi). Inter-residue contacts on wild-type and mutant structures were calculated using Arpeggio to model the effects of mutations on intra-molecular interactions.
- **Evolutionary features**: Sequence-level predictors (SIFT, Polyphen, SNAP2) and evolutionary-based predictors (ConSurf).
- **Distance features**: distance of the mutation from the drug / ligand binding site and distance of the mutation from the adjoining protein/ monomer interface.

4. Machine Learning: Once the data is generated using the different features, it is time for the data set to be trained to build the empirical classifier using supervised machine learning. Supervised machine learning is the search for algorithms using information from the supplied instances to produce a general hypothesis which is then used to make prediction for future instances [194]. The data is divided into non-redundant train and test datasets with respect to amino acid position. This is important to avoid over-

biasing the model as certain features used to build the classifier have similar values. Machine learning can be carried out using Weka [195] (https://www.cs.waikato.ac.nz/ml/weka/) or the python package Scikit-Learn [196] which are inclusive of the classification algorithms: Decision Tree, Nearest Neighbors (KNN), Support Vector Machines (SVM) and Ensemble Classifiers (Random Forest, Extra Trees, AdaBoost and GradientBoosting), Linear Classifiers (Gaussian, Multinomial and Complement Naïve Bayes, Stochastic Gradient Descent). Preliminary testing is performed on all algorithms with the train and test dataset to assess the generalization power of each classifier, that is, its ability to correctly predict on new data, and to ensure that it has not been over or under-trained [187]. Depending on the size of the dataset, it is split into 70:30, 80:20 or 90:10, train to test proportions. Cross-validation is used to determine the classifier's accuracy. It is a technique where the training set is divided into mutually exclusive and equal-sized subsets and for each subset the classifier is trained on the union of all the other subsets. The average of the error rate of each subset is a representation of an estimate of the error rate of the classifier [194]. In case of smaller dataset, a large proportion of the data is retained to provide sufficient data to accurately measure performance of the trained model. Leave-one-out validation, a special case of cross validation is when the test subsets consist of a single instance. This type of validation is, of course, more expensive computationally, but useful when the most accurate estimate of a classifier's error rate is required [194] (Figure 7).

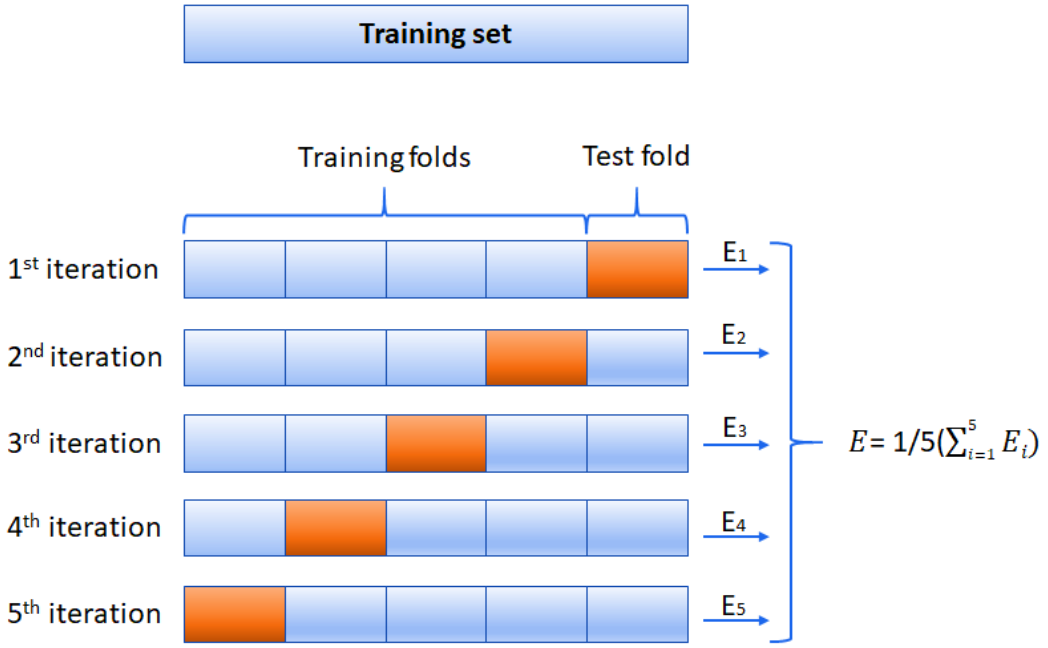


Figure 7: Example of leave-one-out cross-validation where the number of instances is 5. The number of instances equals the number of folds/ iterations. The blue boxes represent instances for train data set and the red box represents the instance for the test set.

Evaluation Metrics: For model evaluation, four metrics are used as each metric presents its own limitations and a broader analysis of all of them together is better suited for evaluating the models described in this work. The metrics are precision, recall, f-measure (also known as f-score), area under the ROC curve (AUC) and Matthew's Correlation Coefficient (MCC). These are well established and broadly used metrics for assessing the results of binary classification algorithms. Such measurements are expressed based on the values of a binary contingency table, also known as confusion matrix (Figure 8), where the classes are represented by convention with + (positive) and - (negative) signs. This 2x2 matrix (actual versus predicted class) uses the raw counts of the number of times each predicted label is associated with each real class.

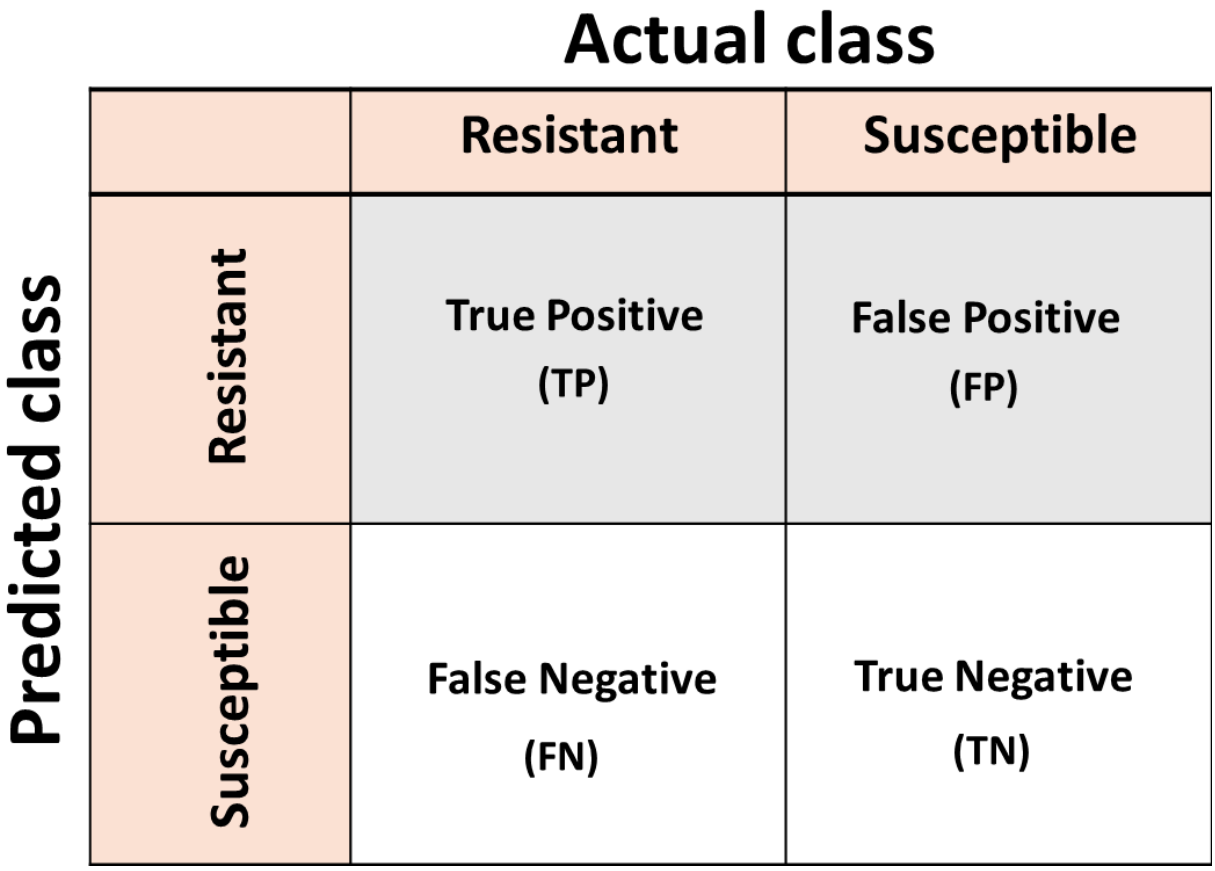


Figure 8: Confusion Matrix. True and False Positives (TP and FP) indicate the number of predicted positives that were correct and incorrect, respectively. Similarly, True and False Negatives (TN and FN) refer to correct and wrong predictions for negative class. The sum TP+FP+TN+FN is equal to the total amount number of instances in the data set being used.

Precision denotes the proportion of Predicted Positive cases that are Actual Positives. It is defined by TP/(TP+FP). On the other hand, Recall is defined as the proportion of Predicted Positives cases that are Actual Positives over all Predicted Positives. Using the convention described in Figure 3, it is defined as TP/(TP+FN). F_1-score is a combination of Precision and Recall in a harmonic mean between them. This measure is defined by the square of the geometric mean divided by the arithmetic mean. All these metrics present biases towards the predictions of positive class and ignore the performance in correctly predicting

the negative class. This is particularly true for data with classes that are not balanced, such as the ones presented in this work [197]. For a different perspective of analysis, given the bias problem with precision, recall and f-measure, use of the measure of Area under the ROC Curve (AUC or AUROC) is preferred. AUC considers the True Positive Rate (TPR), also known as sensitivity, which corresponds to the proportion of positive data points that are correctly considered as positive; and, the False Positive Rate (FPR) that corresponds to the proportion of negative data that are wrongly considered as positive, regarding all negative data points. A Receiver Operating Curve (ROC) is then plotted using TPR versus FPR and the AUC is the area under such curve [198]. Like precision, recall and f-measure, AUC has its best result is 1 and the worse is 0. A random binary classifier would generate an AUC of 0.5. An additional metric for rescue when we have classes of different sizes is MCC. It is regarded as a balanced measure as it takes into account true and false positives and negatives. The MCC is a correlation coefficient between the observed and predicted binary classifications and it returns a value between −1 and +1. A value of +1 represents a perfect correlation; value of 0 no better than random prediction and a value of −1 indicates total disagreement between prediction and observation [199].

$$Recall = \frac{TP}{TP+FN} \quad \text{---------- (Equation 7)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{---------- (Equation 8)}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{---------- (Equation 9)}$$

$$F1 - score = \frac{TP}{TP+1/2(FP+FN)} \quad \text{---------- (Equation 10)}$$

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{---------- (Equation 11)}$$

The metrics described above should be used in a combination to compare model performance for train dataset and even during cross-validation for better optimization. When the data set is imbalanced, MCC should be prioritized as other measures could possibly bias for an over-trained model.

5. Feature engineering: This is a crucial step in developing a good performing classifier as features selected for training strongly influences classifier accuracy. Therefore, it is important to choose informative features and discard non-informative / discriminative and irrelevant ones as they could only increase noise in the system. Choosing the right features prevents a model to be over-fit and computationally economic, as it aims to generate simpler, more concise models [187]. Greedy feature selection can be used where features are included iteratively, one at a time, based on their individual performance. The features can be ranked according to their performance for a chosen metrics. MCC and accuracy are the two main choices of metrics when we run the greedy algorithm. Feature selection methods provided by Scikit-Learn include feature importance, univariate selection, correlation matrix, and recursive feature elimination or addition.

6. Webserver development: Materialize CSS framework version 1.0.0 has been used to build the server front-end for SUSPECT-PZA AND SUSPECT-BDQ. The back end was built in Python via the Flask framework (version 0.12.2). Both the webservers is hosted on a Linux server running Apache.

2.5 Mathematical Modeling

Mathematical models have become important tools in analyzing the spread and control of infectious diseases. Modeling Mtb dynamics have improved our understanding of the natural history of TB infection and transmission, helped in projecting future disease burden and therefore revise policies [138, 200]. Models which study the evolution of infectious disease over time are therefore often described as dynamic epidemiological models.

2.5.1 Epidemiological elements of a mathematical model

TB progresses within the body of a susceptible individual (with no history of previous TB infection) starts with infection with Mtb. Individuals with latent TB can remain asymptomatic or can progress to active

disease by either exogenous reinfection or endogenous reactivation. Risk of progression to active disease can be affected by a range of factors, including age, duration since infection and co-morbid conditions such as HIV [201] or diabetes [202, 203]. TB is a treatable and curable disease and non-compliance to treatment leads to DR-TB. The different components TB epidemiology are:

- **Active and latent TB infection:** Once a susceptible individual is infected with Mtb, symptoms might not be immediately observed. The individual is said to have latent TB, but it is not infectious. People with latent TB infection do not show clinical symptoms, microbiological evidence or radiological abnormality [204]. The risk of progression to active TB is highest in the first few years after exposure to the pathogen, but it can persist lifelong. Sometimes, the bacteria remain inactive for a lifetime without causing any disease. Infectious period starts when latent TB progresses to active TB, and the individual can spread the disease.
- **Endogenous reactivation and exogenous reinfection:** Susceptible individuals often can mount an effective immune response upon infection which limits the spread of Mtb and helps in producing a long-term partial immunity. Reactivation of the latent TB bacilli which was acquired more than five years ago is referred to as endogenous reactivation. A secondary external infection which makes the individual infectious thereby causing active disease if referred to as exogenous re-infection [205, 206]. Infection with different strains of Mtb is referred to as mixed infection.
- **Treatment:** Treatment of active TB is called "chemotherapeutics" and treatment of latent TB is called "chemoprophylaxis". TB therapy typically consists of 3-4 medications for a period of 6 months, which may be longer depending on site of disease and drug resistance. In some contexts, treatment may be supervised ("Directly Observed Therapy") to enhance treatment adherence and monitoring. National programs may combine and deploy different control strategies to treat all types of TB infection.
- **Drug Resistance:** Due to ineffective treatment or intermittent compliance of the patient to the treatment, susceptible strains of Mtb may have a higher chance to gain resistance to

chemotherapeutics. Resistance can either be acquired during treatment or transmitted from individual with DR-TB strains. Acquired resistance initiates an epidemic of DR-TB whereas if the strain is more transmissible, risk of primary drug resistance increases over time.

2.5.2 Epidemiological models

The various types of mathematical models used to study transmission dynamics of TB are: compartmental models [207], agent-based models [208] and network models [209]. A compartmental model describes a population divided into mutually exclusive health states (compartments) and uses differential equations to represent the mechanisms of transition between these health states. Waaler et al [210], Ferebee et al [211] and ReVelle et al [212] were the first to develop mathematical models of tuberculosis transmission. Agent-based modelling and network models are less frequently used to study Mtb transmission, as we intend to model airborne transmission for a chronic infectious disease compared to other infectious diseases [213].

A prototypical "SIR" model divides the population into three compartments, each representing a mutually exclusive health states: susceptible (S), infected (I), and recovered (R). Transmission dynamics are described by defining rates of flow between compartments. TB pathology is complex given its potentially long latency period and hence, "SEIR" models are usually chosen to represent dynamics of TB transmission, where, S is susceptible, E is exposed, I is infected and R is recovered. The threshold for many epidemiology models is the basic reproduction number R_0, which is defined as the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible [214]. For many deterministic epidemiology models, an infection can get started in a fully susceptible population if and only if $R0 > 1$. Thus, the basic reproduction number R_0 is often considered as the threshold quantity that determines when an infection can invade and persist in a new host population.

Ragonnet et al [215] identified six different models structures and only those incorporating two latency compartments were capable to reproduce empirically observed dynamics of TB activation. The two compartments representing early and late latency could be either placed in series or in parallel. To develop my own mathematical model, I used a compartmental model (SEIR model), which includes five compartments with respect to the different states of an individual's disease status Figure 9.

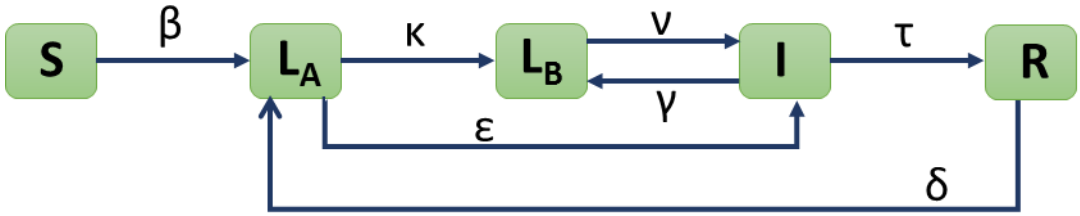


Figure 9: A prototypical TB transmission model. The different compartments represent different health states: Susceptible (S), Early latent (L_A), Late latent (L_B), Infection (I) and Recovered (R). By incorporating two latent compartments, this structure replicates the dynamics of TB latency accurately. The parameters described in the SEIR model are: β = transmission rate ($year^{-1}$), κ = transition to early latent compartment after being exposed to Mtb ($year^{-1}$), ε = progression rate ($year^{-1}$), ν = reactivation rate ($year^{-1}$), γ = self-recovery ($year^{-1}$), τ = treatment success rate, δ = risk of reinfection once recovered.

Compartmental models may be articulated either deterministically using systems of ordinary differential equations or stochastically via continuous-time Markov chains and stochastic differential equations [216]. Epidemic processes are stochastic in nature especially at an individual level and therefore stochastic models help in understanding these fluctuations involved in spread of the disease; though the consequence might not be due to difference in virulence or infectiousness [217]. Nevertheless, small populations are better suited for application of stochastic approach and for elucidating infection dynamics at initial stages [218]. Deterministic models on the other hand are suitable for an infinite population limit [219]. In a deterministic model, a heterogeneous population is subdivided into finite homogenous subpopulation representing the different disease states. Ordinary differential equations are used to model the epidemic

dynamics, capturing movement between subpopulations. The output provided is generally theoretical such reproductive number as described above. These models are commonly used to design proper strategies to reduce the spread of DR-TB. Hence, for our model construction a deterministic compartmental model was used to estimate mono drug resistant amplification rates.

CHAPTER 3.1: STRUCTURE GUIDED PREDICTION OF PYRAZINAMIDE RESISTANCE IN TUBERCULOSIS

Summary

Background: Pyrazinamide, a first-line drug with remarkable sterilizing activity, plays an important role in the treatment of tuberculosis, especially in multi-drug resistant strains. Pyrazinamide use, however, is complicated by its side-effects and challenges with reliable drug susceptibility testing. Resistance to pyrazinamide is largely driven by mutations in pyrazinamidase (pncA), responsible for drug activation, but large genetic diversity and heterogeneity has hindered the development of a comprehensive molecular diagnostic test.

Objective: Our objective was to use information from the proteins 3D structure to accurately identify resistance mutations in pncA.

Methods: To achieve this, we curated 610 pncA non-synonymous single nucleotide mutations with associated high confidence experimental and clinical information on pyrazinamide susceptibility. The molecular consequences of these mutations were assessed using the mCSM platform, which provided insights into changes in protein stability, conformation, and interactions for each mutation.

Findings: Using these structural and biophysical effects, we could correctly classify mutations as either susceptible or resistant with an accuracy of 78%. Our model was validated against a previously documented set of non-redundant clinically resistance mutations achieving 77% accuracy and 81% accuracy across all pncA missense mutations recently reported in the CRyPTIC dataset. Applying this structural analysis to a novel set of previously unreported Victorian clinical mutations with experimental drug susceptibility testing, our model showed clinical resistance in pyrazinamide could be predicted with 71% accuracy.

Web-server: We have made this model freely available through a user friendly web interface called SUSPECT-PZA, StrUctural Susceptibility PrEdiCTion for pyrazinamide, at: http://biosig.unimelb.edu.au/suspect_pza/. This will be a valuable resource to analyse any *pncA* missense mutation, providing structural insight to help guide patient treatment decisions and screening programs.

This chapter has been published in the *Scientific Reports* as a first author publication. "*Structure guided prediction of Pyrazinamide resistance mutations in pncA*", **Karmakar, M.**, Rodrigues, C.H.M., Horan, K., Denholm, J.T., Ascher, D.B. (2020). (doi: 10.1038/s41598-020-58635-x)

SCIENTIFIC REPORTS
nature research

OPEN

Structure guided prediction of Pyrazinamide resistance mutations in pncA

Malancha Karmakar[1,2,3], Carlos H. M. Rodrigues[1,2], Kristy Horan[4], Justin T. Denholm[3] & David B. Ascher[1,2,5*]

Pyrazinamide plays an important role in tuberculosis treatment; however, its use is complicated by side-effects and challenges with reliable drug susceptibility testing. Resistance to pyrazinamide is largely driven by mutations in pyrazinamidase (pncA), responsible for drug activation, but genetic heterogeneity has hindered development of a molecular diagnostic test. We proposed to use information on how variants were likely to affect the 3D structure of pncA to identify variants likely to lead to pyrazinamide resistance. We curated 610 pncA mutations with high confidence experimental and clinical information on pyrazinamide susceptibility. The molecular consequences of each mutation on protein stability, conformation, and interactions were computationally assessed using our comprehensive suite of graph-based signature methods, mCSM. The molecular consequences of the variants were used to train a classifier with an accuracy of 80%. Our model was tested against internationally curated clinical datasets, achieving up to 85% accuracy. Screening of 600 Victorian clinical isolates identified a set of previously unreported variants, which our model had a 71% agreement with drug susceptibility testing. Here, we have shown the 3D structure of pncA can be used to accurately identify pyrazinamide resistance mutations. SUSPECT-PZA is freely available at: http://biosig.unimelb.edu.au/suspect_pza/.

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, is the leading cause of infectious disease death worldwide. In 2017, 10 million people fell ill, and 1.6 million died, from tuberculosis[1]. While a range of antibiotics are available to treat TB, treatment is prolonged, and the increasing emergence of drug-resistant bacteria is a considerable threat to global health. In 2017 alone, an estimated 558,000 people developed multi-drug-resistant tuberculosis (MDR-TB), resistant to the two first-line drugs rifampicin and isoniazid[1].

Pyrazinamide (PZA) is a first-line drug that exhibits unique sterilizing activity towards both drug-susceptible and MDR-TB[2]. It is responsible for the killing of the persistent tubercle bacilli during the initial intensive phase of chemotherapy, allowing treatment to be shortened from 9 months to 6 months for drug susceptible cases[3]. PZA therapy has been linked to improved outcomes for both non-MDR and MDR-TB, and is being considered as part of the future regimens in combinations with bedaquiline, delamanid, PA-824 and moxifloxacin, which are currently in phase three trials[4,5].

Despite the highly important role of PZA in clinical outcomes, resistance has largely been underestimated, with up to 20% of non-MDR-TB patients PZA resistant[6]. Being a central drug in current and future regimens, it is important to be able to rapidly and accurately identify resistant isolates and track the emergence and spread of drug resistant strains. *In vitro* drug susceptibility testing (DST) is challenging, expensive and time-consuming as PZA is effective against *M. tuberculosis* only at acidic pH, leading to false resistance rates of up to 70%[7–13]. This has led to the WHO recommending the development of molecular genetics tests.

PZA is a structural analog of nicotinamide and is a pro-drug that needs to be converted into its active form, pyrazinoic acid (POA), by the non-essential enzyme pyrazinamidase, encoded by the *pncA* gene[14,15]. It has been

[1]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia. [2]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia. [3]Victorian Tuberculosis Program, Melbourne Health and Department of Microbiology and Immunology, University of Melbourne, Melbourne, Victoria, Australia. [4]Microbiological Diagnostic Unit Public Health Laboratory, University of Melbourne at The Peter Doherty Institute for Infection &Immunity, Melbourne, Victoria, Australia. [5]Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK. *email: david.ascher@unimelb.edu.au

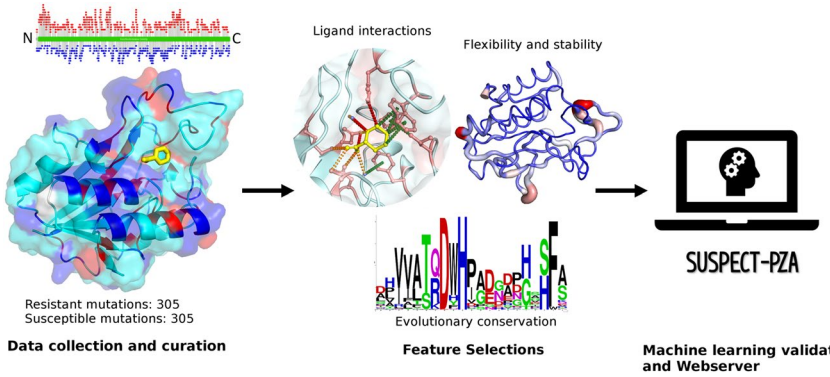


Figure 1. Methodology workflow. The methodology can be divided into three steps. In step 1, data is collected and curated from various tuberculosis databases and articles with experimental evidence like availability of DST results or high-precision laboratory screening study. The curated mutations are shown across both the protein sequence and 3D structure, respectively. The protein sequence and structure of PncA is colored by whether resistant (red) or susceptible (blue) mutations have been observed at that location. Highlighting the difficulty of genomic analysis of pncA, both resistant and susceptible mutations have been observed across many residue positions (cyan). In step 2, effects of mutations on protein stability, dynamics, complementary information regarding the environment characteristics of the wild-type residue (e.g. relative solvent accessibility, residue depth and secondary structure), PZA binding affinity are calculated using different *in-silico* tools. Step 3, all the features are used as evidence to train a supervised machine learning algorithm and after evaluating the performance of the predictive model, the consensus predictions are integrated into a server and can be used to guide clinical resistance detection.

postulated that the mechanism of action of PZA is through POA, which disrupts the bacterial membrane energetics and inhibits the membrane transport function which is necessary for the survival of the bacterium, at an acidic site of infection[16]. PZA resistance has been linked to mutations in a number of genes, including *pncA*, *rpsA*[17], *panD*[18], *clpC1*[19], and the putative efflux pumps *Rv0191*, *Rv3756c*, *Rv3008*, and *Rv1667c*[20], but mutations in *pncA* are the major mechanism for PZA resistance (70–97%)[21]. While sequencing the *pncA* gene can be a more reliable method to determine resistance than DST, which is prone to missing low-level pyrazinamide resistance caused by non-synonymous mutations in pncA[22], the development of a genetics based resistance screen is complicated as resistant and non-resistant mutations are found across the entire protein.

To solve the problem of a reliable DST for PZA, we previously showed that protein structural information can be used in a clinical setting to rapidly, accurately and pre-emptively predict drug resistant mutations in *pncA*[23]. This showed that mutations that affected protein folding, flexibility, stability and activity were strongly associated with resistance. Here we have used a comprehensive combination of structure and sequence-based features to develop a predictive tool to characterize novel PncA mutations, which we tested on novel mutations from the Victorian Tuberculosis Program, CRyPTIC[24] and Miotto *et al*. dataset[25]. This highlights the potential of using structural information to guide the genetic detection of resistance. We have implemented our model through the webserver SUSPECT-PZA (http://biosig.unimelb.edu.au/suspect_pza/), which will enable the rapid structural evaluation of the molecular and phenotypic consequences of any *pncA* nonsynonymous mutation to support informed clinical decisions.

Results

We used a structure-guided approach to understand the structural and functional consequences of variants in the drug target PncA, and machine learning to build an empirical tool that could identify likely resistant mutations. The workflow used to analyze the mutations and train a Random Forest algorithm is shown in Fig. 1 and it comprises three major steps: (1) data curation, which can be subdivided into mutational data set acquisition and protein structure curation; (2) feature analysis, which involves the generation and evaluation of features selected to develop the predictive model to determine novel drug resistance mutations in PncA; (3) machine learning and webserver development, which aims to train, test and validate a supervised machine learning algorithm to accurately predict the susceptibility of the variant followed by a database (SUSPECT-PZA) which has information for all possible variants of PncA.

Distribution of the mutations on the structure. We curated a dataset of 1322 nonsynonymous substitutions with high quality experimentally measured PZA susceptibility (71 susceptible mutations from GMTV[26], 12 resistant mutations from GMTV[26], 178 resistant mutations from TBdreamDB[27], Fig. 2A, 547 resistant and 514 susceptible mutations from experimental saturation mutagenesis[28]). After removal of duplicate mutations, we were left with a dataset of 610 mutations, which included 305 susceptible and 305 resistant mutations. Mapping the complete set of curated 610 nsSNVs (Fig. 1) and just the clinical variants only (Fig. 2B) onto the crystal structure of PncA revealed that variants were distributed throughout the entire protein structure, complicating resistance inference from sequence analysis. We also observed that the resistance mutations were not solely localized at the drug binding site but distributed throughout the protein (Fig. 2C).

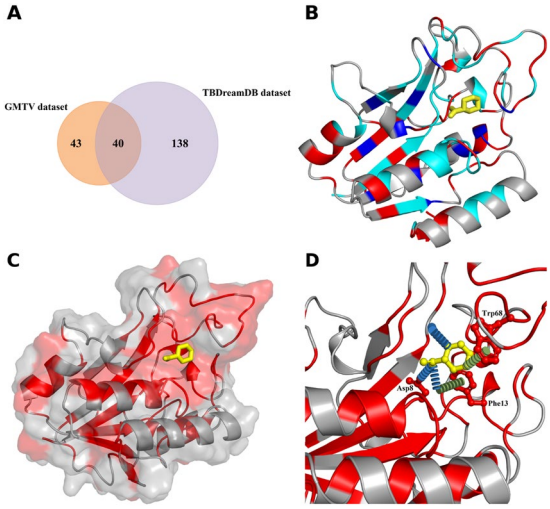


Figure 2. Distribution of clinical resistant and susceptible mutations in PncA. (**A**) Venn diagram representing the distribution of clinical mutations in the different datasets used to build the predictive model. (B) Clinical resistant and susceptible mutations mapped on the crystal structure. Amino acid positions where both susceptible and resistant mutations were seen are colored in cyan and emphasizes the need for a better and improved tool to classify them accurately. (**C**) Surface view of PncA with the docked PZA (yellow, ball and stick representation). Clinical resistant mutations, shown in red, are not just located at the PZA binding site, but are spread equally throughout the whole protein. (**D**) Molecular interactions between PZA (yellow sticks) and the surrounding amino acids which are part of the catalytic triad (Asp8) and substrate binding site (Trp68, Phe13). Hydrogen bonds are shown as blue dashes, and π-interactions as green dashes.

PncA is a small protein molecule which constitutes of 186 amino acids. The experimental crystal structure of the drug (PZA) bound to the enzyme (PncA) was unavailable. Therefore, PZA was *ab* initio docked into the experimental crystal structure of the holo-wild-type PncA protein (PDB ID: 3PL1[29]). The docked structure revealed that PZA formed key interactions within the proteins active site, which includes the catalytic triad (Asp8, Lys96, and Cys138), substrate-binding residues (Trp68 and Phe13), and the iron center (Asp49, His51, His57, and Fe 21). Analysis of the molecular interactions with Arpeggio[30] highlighted a strong network of polar and π- interactions between PZA and PncA (Fig. 2D).

Structural, biophysical and evolutionary consequences of PncA mutations. Looking at the SNAP2[31] and PROVEAN[32] scores, which consider evolutionary information to predict functionally important nonsynonymous mutations, we observed that resistant mutations were always associated with deleterious scores, while susceptible mutations were scored neutral (Table S1; Fig. 3). This suggest that although mutations were spread throughout the protein, mutations associated with resistance were having a stronger effect on the structure and function of the protein.

The wild-type environment also provided information to differentiate between resistant and susceptible mutations, which included relative solvent accessibility (RSA), residue depth and secondary structure of the wild-type residue (Table S1; Fig. 3). This showed that resistant mutations tended to be found at buried residues that were less solvent exposed (average RSA of 0.18 for resistant mutations compared to 0.39 for susceptible; average residue depth of 1.09 Å for resistant mutations compared to 0.75 Å for susceptible; Table S1). These values were consistent with susceptible mutations being in regions that have milder effects on protein stability and activity than the resistance mutations.

The impact of the resistant and susceptible mutations on protein folding, stability and conformation were assessed using biophysical tools which relies on graph-based signatures to calculate the change in Gibb's free energy, like mCSM-Stability[33], DUET[34] and DynaMut[35]. The effect of the mutations on the binding affinity for PZA were assessed using mCSM-Lig[36]. We observed that resistant mutations led to large decreases in PncA stability and conformational flexibility, while susceptible mutations were associated with milder changes (Table S1; Fig. 3). This is consistent with what we have observed previously for non-essential and drug activating proteins[37]. While resistant mutations, however, tended to be located closer to the PZA binding site (average < 10 Å from the PZA; Fig. 3), we did not see a significant difference in the distribution of the effects of resistant and susceptible mutations on PZA binding affinity (Table S1, Fig. S2), likely due to the importance of other molecular effects leading to resistance.

Machine learning to predict PZA resistance. Building on this structural and sequence-based analysis, we tested whether the information generated from these features could be used to train a supervised machine learning algorithm capable of accurately predicting resistant mutations in PncA. We grouped our features into five distinct categories: stability, dynamics, evolutionary conservation, ligand interactions and backbone geometry (structural environment). The performance of predictive models trained on each class of feature was evaluated

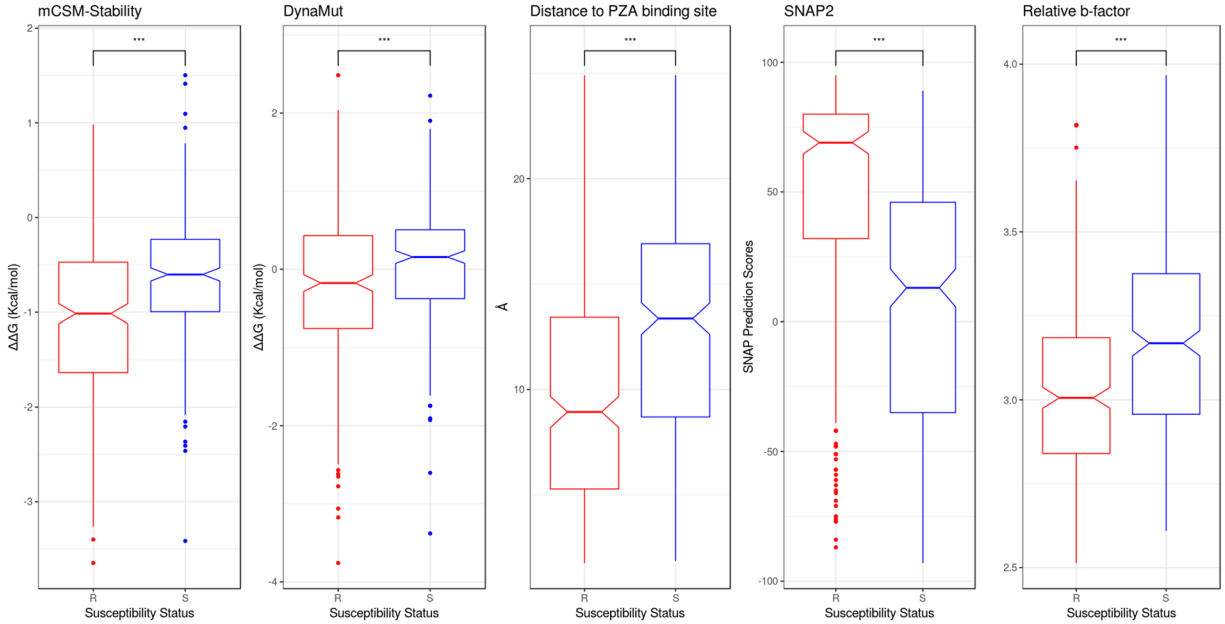


Figure 3. PCA analysis of key molecular features distinguishing resistant and susceptible mutations. Features used for model building are represented as boxplots for explanatory data analysis. The resistant associated mutations (R) are represented as red and the susceptible mutations (S) as blue. (***$p < 0.0001$, Welch two sample t-test).

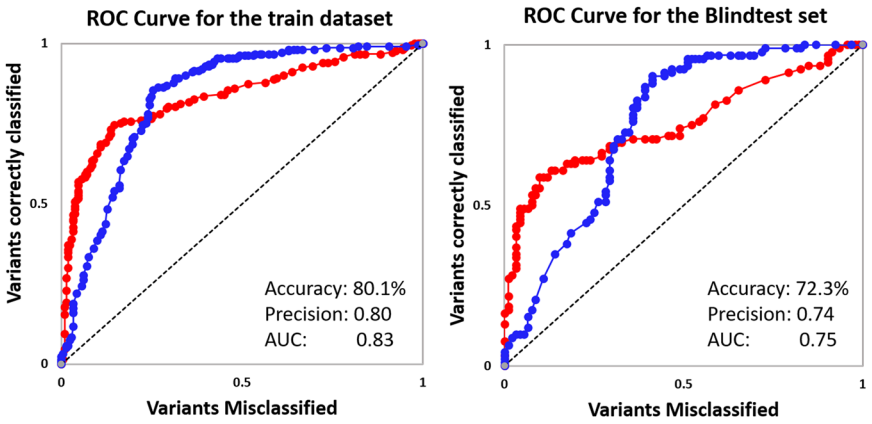


Figure 4. Evaluation Metric for machine learning. Receiver Operating Characteristic (ROC) curves of PZA classifier obtained using the structural and functional consequences of the mutations to accurately identify resistant (red) and susceptible (blue) mutations. (AUC = area under the curve).

separately to explore the contribution of each class to the predictive model (Table S2; Fig. S2). We were able to confirm that the individual categories of features did not yield a good metric for a reliable predictive model, but in combination using 10-fold cross-validation, models trained using Random Forest algorithm yielded a more balanced and accurate performance, highlighting the synergistic effect of these features. The final model correctly classified 80.1% and 72.3% of mutations in the training and blind datasets, respectively (Fig. 4; Table 1). The comparative performance across iterative non-redundant blind datasets suggested that the model was not overfitted.

Analysis of our model revealed that PncA-resistant mutations were associated with large changes in protein folding and stability (mCSM-Stability scores < -0.9 Kcal/mol; $p < 0.0001$, Welch Two Sample t-test) and conformational flexibility (DynaMut score < 0.78 Kcal/mol; $p < 0.0001$, Welch Two Sample t-test) or located in close proximity to the catalytic triad and substrate-binding site (< 10.8 Å; $p < 0.0001$, Welch Two Sample t-test). Alternatively, susceptible mutations had a relative b-factor value of ≥ 3.19 ($p < 0.0001$, Welch Two Sample t-test), residue depth of ≥ 0.9 ($p < 0.0001$, Welch Two Sample t-test), distance from PZA greater than 11.9 Å and mild effects on protein stability (SDM scores ≥ 2.68 Kcal/mol; $p < 0.0001$, Welch Two Sample t-test).

Validation using Clinical Datasets. We next validated our model using variants reported in the recently published CRyPTIC dataset[24]. 355 *pncA* nsSNVs associated with PZA resistance were reported, of which 75 were not present in our training dataset. Our model correctly classified 79.2% of the mutations across the whole dataset

	Total nsSNVs	Resistant nsSNVs	correctly classified variants SUSPECT-PZA (%)	Susceptible nsSNVs	correctly classified variants SUSPECT-PZA (%)	PPV (%) (95% CI)	Accuracy (%)
Training dataset (70%)	426	213	159 (74.5)	213	182 (85.5)	83.7 (78.6–87.8)	80.1
Blind test dataset (30%)	184	92	56 (60.8)	92	77 (83.7)	78.9 (69.5–85.9)	72.3
CRyPTIC dataset[24]	355	325	266 (81.8)	30	15 (50.0)	94.7 (92.5–96.2)	79.2
CRyPTIC novel nsSNVs	75	67	67 (74.6)	8	4 (50.0)	92.6 (86.0–96.2)	72.0
Miotto *et al.* dataset[25]	98	92	82 (89.1)	6	2 (33.3)	95.4 (92.1–97.3)	84.8
Miotto novel nsSNVs	44	43	35 (81.4)	1	0	97.2 (96.8–97.6)	79.5
Stellenbosch University and CDC, USA nsSNVs[38]	8	5	3 (60.0)	3	3 (100)	100	75.0
Victorian TB novel nsSNVs	7	4	4 (100)	3	1 (33.3)	66. 7 (47.3–81.7)	71.4

Table 1. Evaluation metrics across the train and blind test datasets. Accuracy = (TP + TN)/(TP + TN + FP + FN); TP: True positives, TN: True Negatives, FP: False Positives, FN: False Negatives PPV: Positive predictive value, predicting PZA resistance (nsSNVs - non-synonymous single nucleotide variant).

(355 mutations), and 72.0% of those non-redundant in amino acid position with the training data (75 mutations). The positive predictive value was 94.7% (95% CI [92.5% to 96.2%]).

We also validated our empirical classifier using the dataset reported by Miotto *et al.*[25], which contained 98 nsSNVs graded by the confidence of their association with phenotypic drug resistance. 44 out of the 98 nsSNVs reported in the paper were not present in our training dataset. We accurately predicted the drug susceptibility of 84.8% of the polymorphism across the whole dataset (98 mutations), with an accuracy of 79.5% for those mutations not included in the training data (44 mutations). The positive predictive value was 95.4% (95% CI [92.1% to 97.3%]). We observed mutations such as Q10P (21 cases reported), W68G (16 cases reported) and I133T (17 cases reported) with 0.98 probability associated with resistant phenotype[22] and categorized as high confidence for association with resistance, moderate confidence for association with resistance and minimal confidence for association with resistance respectively[25] were all classified as resistant by our predictive model, highlighting the sensitivity of the prediction.

Mutations reported by Miotto *et al.*[25] under the "no association with resistance" category, including I31T, L35R and T47A were predicted as resistant, and I6L as susceptible. This is consistent with the available experimental data[24,28], highlighting the advantage, accuracy and versatility of our approach. A closer look into the different biophysical scores for the resistant associated mutations revealed that they had large predicted destabilizing values for protein conformational flexibility (I31T, −2.49 Kcal/mol) and stability (I31T, −3.46 Kcal/mol) and one was located very close to the catalytic triad (T47A, <6 Å).

Our predictive model was further validated on PZA DST screening at 100 μg/ml of clinical isolates from culture collections at Stellenbosch University, South Africa (865 isolates) and the Centers for Disease Control and Prevention (CDC), Atlanta, USA (185 isolates)[38]. They identified 49 isolates with a susceptible phenotype containing 8 nsSNVs. All nsSNVs with an MIC < 50 μg/ml were correctly classified by our model as susceptible (E37V, D110G, T114M). Whitfield and colleagues suggest that those isolates with an MIC > 50 μg/ml should be considered clinically resistant, of which our model classified three as resistant (A170V, V130A and L35R) and two as susceptible (V163A and V180I). Overall, our model had a 75% agreement with the DST results and a positive predictive value of 100%

Application within a Clinical Setting. In a prospective genomic sequencing and DST analysis of over 600 Victorian clinical TB isolates, 7 *pncA* variants were detected in 11 variants phenotypically resistant to PZA, none of which were present in our training dataset. Our model correctly classified five out of seven variants as resistant (71.4% accuracy). The remaining two mutations, G108V and Q10H, which were susceptible according to the DST results were predicted to confer resistance and consistent with other experimental findings[24,25,28]. Both variants, had a SNV frequency of <0.5, which is known to impact upon the reliability of the DST results. This highlights the potential clinical power of our model.

Expanding our analysis, four additional *pncA* mutations (S104R, V128G, Y95R and E15A) were identified in Victorian clinical TB isolates lacking DST results. Both S104R and V128G were predicted as resistant by our model, consistent with previously reported DST results[24–28]. The remaining two mutations, Y95R and E15A, have not been reported previously. Our model suggests both mutations to confer susceptibility to PZA.

SUSPECT-PZA webserver. We have developed a user-friendly, freely available web server SUSPECT-PZA (StrUctural Susceptibility PrEdiCTion on PZA), http://biosig.unimelb.edu.au/suspect_pza/, which is a database for all possible variants of PncA. There are two different input options (Fig. S2): the first one is the "Single Mutation" option which allows the users to input one mutation for analysis. The basic format required by the server for this input option is that the mutation must be specified as a text string containing the wild-type residue one-letter amino acid code, its corresponding position on the structure and the mutant one-letter amino acid code. The second option is the "Mutation List", which allows the user to upload a list of mutations, in the same specified format as above but in a file for batch processing (Fig. S3). Sample submission entries are available to assist users to submit their mutations for analysis and an additional help page via the top navigation bar.

Figure 5 shows a snapshot of the output page for the "Single Mutation" option. The web server displays the prediction outcome (Resistant / Susceptible) along with details of the user input data, information on the wildtype residue environment and features used for prediction. In addition, there is an interactive 3D viewer, built using NGL[39], which allows analysis of non-covalent inter-residue interactions for the position specified in the input calculated using Arpeggio[30] for both wild-type and mutant structures. The results for the "Mutation List" option is summarized in a downloadable table. The users can access details of individual mutation as shown in Fig. S4. There is a 3D viewer at the bottom of the page in which the residues in the input list is colored according to the predicted effect (Fig. S5).

Discussion

PZA was discovered in 1948 in an *in vivo* screen of nicotinamide derivatives in a structure-activity relationship study[40] and used as anti-tuberculosis drug in 1952 for the first time. Till the 1970's PZA was used as a second-line drug to treat TB, until they discovered the sterilizing activity and reduction in treatment duration in combination with isoniazid and rifampicin. There has been a lot of studies conducted since then and with the continued usage of the drug to treat TB, there has been an increased incidence of resistance associated with it. Being an important first-line drug, accurate and rapid evaluation of PZA susceptibility is crucial for successful management of patients with either susceptible or drug-resistant TB. The existing molecular phenotypic tests are considered poorly reliable, expensive, and has a long turnaround time. To account for this situation there is an urgent requirement to develop a rapid, reliable and affordable molecular PZA DST. As resistance mutations are spread all over the length of the PncA protein, it is quite challenging to develop a new method. In this study, we establish a novel computational methodology to better understand the structural and functional consequences of drug resistance mutations by exploiting the protein's 3D structure. Using supervised machine learning algorithm, we developed an empirical tool to determine novel drug resistance in PncA followed by a database which has information on all possible variants of PncA.

The primary focus of our work is on missense non-synonymous mutations as these typically have more subtle molecular effects that can be harder to predict, than in-frame and frameshift indel mutations that have a much larger deleterious effect on PncA structure and function and are all classed as high-confidence resistant mutations. The structure-based tools implement the concept of graph-based signatures to predict the effect on single point mutations for protein stability. To assess changes in conformational flexibility, graph-based signatures were integrated with normal mode analysis to predict the impact on the protein structure. Scores for these features which were calculated as change in Gibb's free energy ($\Delta\Delta G$) provided important molecular information on resistant mutations, signifying larger effects on protein folding and dynamics and minimal effect on PZA binding affinity. Interpreting the results, we observed, resistance mutations were seen to affect protein activity and function through destabilization of the protein structure and conformation. It even helped in correlating earlier findings where resistant isolates were not associated with a loss of bacterial fitness[41] due to the fact that PncA was involved in nicotinamide recycling pathway rather than in its synthesis. These structural insights have been used to guide clinical decisions for novel PZA mutations[23].

Phenotypic DST which is the current "gold standard", which encompasses methods like Wayne and Bactec MGIT 960, suffers from poor reproducibility. Discrepancies among the results lead to considerable doubt over the clinical significance of the method. Next-generation sequencing based diagnostics can be an alternative for innovative tools to reduce false detection of PZA resistance cases and fast and accurate detection of drug resistance by molecular DST[42]. In the past couple of years researchers have used different techniques to come up with a better and consistent methodology to detect and determine resistance in PZA. Stoffels *et al.*[41] conducted an elaborate study on 14-year complete capture of clinical isolates, where he found frequency of spontaneous acquired resistance to be 10^{-5} bacilli *in vitro*. Miotto *et al.* 2014 work generated the minimum dataset of mutations that should be included in any molecular test for PZA, paving the way for predicting PZA resistance using new genome-based technologies[22]. This was followed by Farhat *et al.* 2016 comprehensive web-based dataset[43]. Though all these approaches were a step up from the existing phenotypic DST, they do not provide information on novel variants. The advantage with our database is it provides information on all possible variants for PncA. This data provides a basis for use as part of any molecular DST, needed for the valid interpretation of data generated by massive sequencing approaches.

Interestingly, comparing performance of SUSPECT-PZA across datasets used to train earlier methods, we observed that the weakest performance was across variants classified as susceptible. However, many of these mutations have been observed in clinically resistant isolates. Our biophysical analysis and SUSPECT-PZA predictions would be consistent with these mutations potentially being misclassified previously.

We also compared our empirical models output to the "revised DST" of Miotto *et al.*[22], where they accounted for enzymatic activity and structural analysis to adjust for possible errors in phenotypic DST. There were 178 missense mutations listed, of which 162 were labelled resistant (R) and 17 were labelled susceptible (S). Our model predicted 88.9% (144/162) of the resistant mutations and 58.8% (10/17) of the susceptible mutations accurately. The positive predictive value was 95.4% (95% CI [92.1% to 97.3%]). The primary divergence from the Miotto classifications was in predicting susceptible mutations. This is likely due to discrepancies in phenotypic and molecular DST results from different laboratory setups[16]. For example, mutations reported as susceptible in the "revised DST" like L159V, F81S, A102V, T135S, T168I and A46V were unanimously reported as resistant in other studies[24,26–28]. Our predictive tool also predicts them to be resistant and hence, proves to be more reliable, reproducible, free to use and a fast alternative to the existing gold standard methods.

This study highlights the power of using computational prediction of the structural consequences of variants in PncA to identify likely pyrazinamide resistance mutations, a critically important first-line drug in the treatment of tuberculosis. This approach, however, is not limited to pncA and has been developed for application to other antimicrobial agents like bedaquiline[44], a last line resort to treat multi-drug and extremely drug resistant

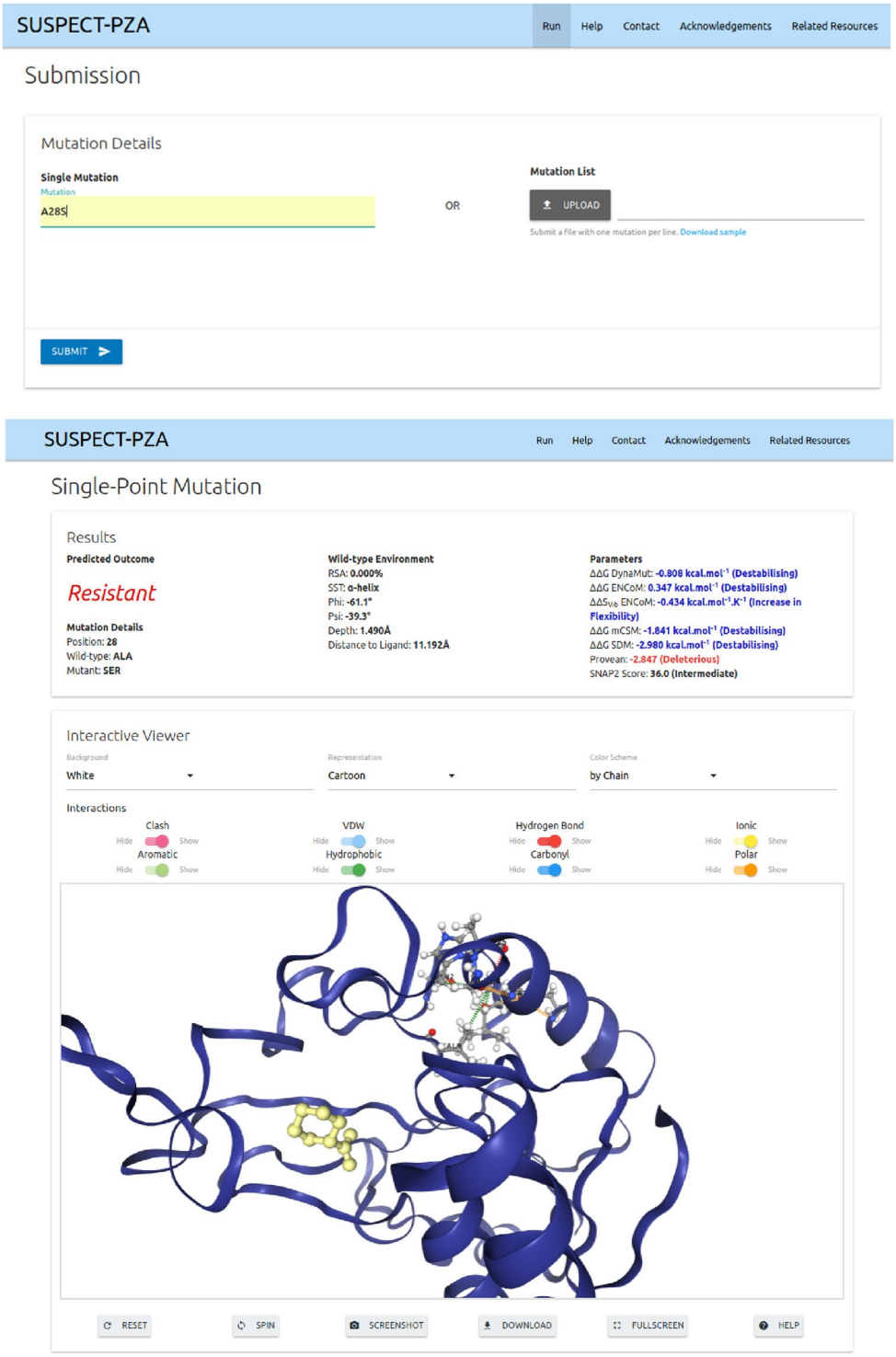


Figure 5. SUSPECT-PZA webserver Single point mutation prediction result page. The predicted outcome of the submitted mutation is displayed along with complimentary information on features used to aid in the development of the tool. The interactive 3D viewer allows user to further analyze non-covalent interactions for both wild type and mutant residues on the protein. A variety of controllers are provided to customize molecule representation.

TB. A major advantage of our tool is that it was built using a very well-balanced dataset. In case of mutations reported as both susceptible and resistant in the same or different datasets, we looked for frequency of occurrence and clinical information. We have extensively evaluated the method through both cross-validation and independent non-redundant blind tests, which provide a measure of a methods applicability and robustness. Across all test sets the method performed equally well, providing strong confidence in the approach. As with all machine

learning approaches, the availability of more phenotypic and clinical data will enable the development and validation of stronger approaches. This will be an iterative approach moving forward. The other aspect to improving our predictive model is through the inclusion of new features or parameters. We have shown previously that this approach can even capture strain dependent variations in resistant patterns[23]. While we did not have the data available to build into our current model, we next aim to integrate lineage specific information, which will enable more refined and personalized predictions. This comprehensive web server can be used in clinical settings as an improved diagnostic tool to help realize the power of whole genome sequencing diagnostic approaches.

Methods

Data set. A list of 610 nonsynonymous single-nucleotide mutations (nsSNVs) of pncA was obtained from the GMTV (Genome-wide Mycobacterium tuberculosis Variation) Database Project[26], Tuberculosis Drug Resistance Mutation Database[27], and saturation mutagenesis[28]. The clinical validation datasets used in the paper were from CRyPTIC[24] and Miotto *et al.*[25].

Modelling the biophysical consequences of missense mutations. We have developed a comprehensive *in silico* mutational analysis platform that uses graph-based signatures to represent the 3D structure of a protein and quantitatively predict the molecular consequences of point mutations on protein structure, function and interactions[30,33–36,45]. This has been used to characterize and preemptively identify likely resistance mutations in drug targets[23,37,46–54]. Using these tools, we assessed the molecular consequences of each mutation on the structure of PncA and drug activation.

The experimental crystal structure of holo-wild-type PncA (PDB ID: 3PL1)[29] was minimized in Prime, and PZA docked into the active site using Glide (Schrödinger Suite). The effects of mutations on PncA folding and stability were assessed using SDM[55], mCSM-Stability[33] and DUET[34], and their effects on protein flexibility and conformational was predicted using normal mode analysis by DynaMut[35]. The effect of the changes on the binding affinity of PZA towards PncA were predicted using mCSM-Lig[36,56]. These approaches are novel machine-learning algorithms. We also included structural information of the wild-type residue, including relative solvent accessibility, residue depth, secondary structure and dihedral angles of the PncA chain φ (phi) and ψ (psi). Additionally, SNAP2[31] and PROVEAN[32] were used to provide additional evolutionary information. Moreover, the scores calculated for the various structural and sequence-based features are independent of pH and temperature.

Machine learning. Here we used the Random Forest binary classifier using the Weka toolkit[57] to train our predictive models. Random Forest is an ensemble-learning robust classification algorithm, in which multiple decision trees are included over a random subset of features and decide the output via majority voting. The model was trained using 10-fold cross-validation and performance evaluated by area under the Receiver Operating Characteristic (AUROC) curve, precision and accuracy. Further validation of the models was performed using a blind-test set of 184 mutations, which were non-redundant at the position-level with mutations in the training set. Analysis of the final model revealed a set of structural features that distinguished between susceptible and resistant pncA point mutations.

Webserver development. The server front-end was built using materialize CSS framework version 1.0.0, while the backend was built in Python via the Flask framework (version 0.12.2). It is hosted on a Linux server running Apache.

Sequencing and DST of clinical isolates. Genomic DNA was extracted according to the mechanical cell disruption and ethanol precipitation method outlined in Votintseva 2015[58] with slight modifications. Briefly, no pre-treatment was used and approximately $3 \times 1\,\mu L$ loops of culture were dispersed in 700 µL TE buffer (Sigma Aldrich) as the starting material. The precipitated DNA pellet was only washed once and resuspended into 50 µL EB Buffer (Qiagen) at 55 °C for 10 minutes with regular vortexing. Finally, samples were centrifuged 3 min at 13,000 rpm and 45 µL of DNA extract was transferred into a clean tube for downstream processing. Each extract was interrogated for *Mycobacterium tuberculosis* viability by inoculating 15 µL of DNA extract into MGIT tube (Becton Dickinson, UK) and incubated in the Bactec MGIT 960 system (Becton Dickinson, UK). Unique dual indexed libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina). Libraries were sequenced on the Illumina NextSeq. 500 with 150-cycle paired end chemistry as described by the manufacturer's protocols.

Sequences were aligned to H37Rv (NC_0009623.3) and small nucleotide variations (SNV) mutations in *pncA* were identified using LoFreq (http://csb5.github.io/lofreq/). SNVs with a frequency > 0.6 were used to compare the genotype of isolates to the phenotype observed using standard laboratory methods for PZA susceptibility testing.

Received: 11 July 2019; Accepted: 28 November 2019;
Published online: 05 February 2020

References

1. WHO. Global Tuberculosis Report, Executive Summary, 2018. https://www.who.int/tb/publications/global_report/tb18_ExecSum_web_4Oct18.pdf?ua=1 (2018).
2. Heifets, L. & Lindholm-Levy, P. Pyrazinamide sterilizing activity *in vitro* against semidormant Mycobacterium tuberculosis bacterial populations. *The American review of respiratory disease* **145**, 1223–1225, https://doi.org/10.1164/ajrccm/145.5.1223 (1992).
3. Tarshis, M. S. & Weed, W. A. Jr. Lack of significant *in vitro* sensitivity of Mycobacterium tuberculosis to pyrazinamide on three different solid media. *American review of tuberculosis* **67**, 391–395 (1953).
4. Dawson, R. *et al.* Efficiency and safety of the combination of moxifloxacin, pretomanid (PA-824), and pyrazinamide during the first 8 weeks of antituberculosis treatment: a phase 2b, open-label, partly randomised trial in patients with drug-susceptible or drug-resistant pulmonary tuberculosis. *Lancet (London, England)* **385**, 1738–1747, https://doi.org/10.1016/s0140-6736(14)62002-x (2015).

5. Veziris, N. *et al*. A once-weekly R207910-containing regimen exceeds activity of the standard daily regimen in murine tuberculosis. *American journal of respiratory and critical care medicine* **179**, 75–79, https://doi.org/10.1164/rccm.200711-1736OC (2009).
6. Juma, S. P. *et al*. Underestimated pyrazinamide resistance may compromise outcomes of pyrazinamide containing regimens for treatment of drug susceptible and multi-drug-resistant tuberculosis in Tanzania. *BMC infectious diseases* **19**, 129, https://doi.org/10.1186/s12879-019-3757-1 (2019).
7. Chang, K. C., Yew, W. W. & Zhang, Y. Pyrazinamide susceptibility testing in Mycobacterium tuberculosis: a systematic review with meta-analyses. *Antimicrobial agents and chemotherapy* **55**, 4499–4505, https://doi.org/10.1128/aac.00630-11 (2011).
8. Chedore, P., Bertucci, L., Wolfe, J., Sharma, M. & Jamieson, F. Potential for erroneous results indicating resistance when using the Bactec MGIT 960 system for testing susceptibility of Mycobacterium tuberculosis to pyrazinamide. *Journal of clinical microbiology* **48**, 300–301, https://doi.org/10.1128/jcm.01775-09 (2010).
9. Hewlett, D. Jr., Horn, D. L. & Alfalla, C. Drug-resistant tuberculosis: inconsistent results of pyrazinamide susceptibility testing. *Jama* **273**, 916–917 (1995).
10. Hoffner, S. *et al*. Proficiency of drug susceptibility testing of Mycobacterium tuberculosis against pyrazinamide: the Swedish experience. *The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease* **17**, 1486–1490, https://doi.org/10.5588/ijtld.13.0195 (2013).
11. Miller, M. A., Thibert, L., Desjardins, F., Siddiqi, S. H. & Dascal, A. Testing of susceptibility of Mycobacterium tuberculosis to pyrazinamide: comparison of Bactec method with pyrazinamidase assay. *Journal of clinical microbiology* **33**, 2468–2470 (1995).
12. Pandey, S., Newton, S., Upton, A., Roberts, S. & Drinkovic, D. Characterisation of pncA mutations in clinical Mycobacterium tuberculosis isolates in New Zealand. *Pathology* **41**, 582–584 (2009).
13. Simons, S. O. *et al*. Validation of pncA gene sequencing in combination with the mycobacterial growth indicator tube method to test susceptibility of Mycobacterium tuberculosis to pyrazinamide. *Journal of clinical microbiology* **50**, 428–434, https://doi.org/10.1128/jcm.05435-11 (2012).
14. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nature medicine* **2**, 662–667 (1996).
15. Konno, K., Feldmann, F. M. & McDermott, W. Pyrazinamide susceptibility and amidase activity of tubercle bacilli. *The American review of respiratory disease* **95**, 461–469, https://doi.org/10.1164/arrd.1967.95.3.461 (1967).
16. Zhang, Y., Wade, M. M., Scorpio, A., Zhang, H. & Sun, Z. Mode of action of pyrazinamide: disruption of Mycobacterium tuberculosis membrane transport and energetics by pyrazinoic acid. *The Journal of antimicrobial chemotherapy* **52**, 790–795, https://doi.org/10.1093/jac/dkg446 (2003).
17. Shi, W. *et al*. Pyrazinamide inhibits trans-translation in Mycobacterium tuberculosis. *Science (New York, N.Y.)* **333**, 1630–1632, https://doi.org/10.1126/science.1208813 (2011).
18. Shi, W. *et al*. Aspartate decarboxylase (PanD) as a new target of pyrazinamide in Mycobacterium tuberculosis. *Emerging microbes & infections* **3**, e58, https://doi.org/10.1038/emi.2014.61 (2014).
19. Yee, M., Gopal, P. & Dick, T. Missense Mutations in the Unfoldase ClpC1 of the Caseinolytic Protease Complex Are Associated with Pyrazinamide Resistance in Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy* **61**, https://doi.org/10.1128/aac.02342-16 (2017).
20. Zhang, Y., Zhang, J., Cui, P., Zhang, Y. & Zhang, W. Identification of Novel Efflux Proteins Rv0191, Rv3756c, Rv3008, and Rv1667c Involved in Pyrazinamide Resistance in Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy*, **61**, https://doi.org/10.1128/aac.00940-17 (2017).
21. Hirano, K., Takahashi, M., Kazumi, Y., Fukasawa, Y. & Abe, C. Mutation in pncA is a major mechanism of pyrazinamide resistance in Mycobacterium tuberculosis. *Tubercle and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease* **78**, 117–122 (1997).
22. Miotto, P. *et al*. Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study. *mBio* **5**, e01819–01814, https://doi.org/10.1128/mBio.01819-14 (2014).
23. Karmakar, M. *et al*. Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy. *American journal of respiratory and critical care medicine* **198**, 541–544, https://doi.org/10.1164/rccm.201712-2572LE (2018).
24. Allix-Beguec, C. *et al*. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *The New England journal of medicine* **379**, 1403–1415, https://doi.org/10.1056/NEJMoa1800474 (2018).
25. Miotto, P. *et al*. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. *The European respiratory journal*, **50**, https://doi.org/10.1183/13993003.01354-2017 (2017).
26. Chernyaeva, E. N. *et al*. Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. *BMC genomics* **15**, 308, https://doi.org/10.1186/1471-2164-15-308 (2014).
27. Sandgren, A. *et al*. Tuberculosis drug resistance mutation database. *PLoS medicine* **6**, e2, https://doi.org/10.1371/journal.pmed.1000002 (2009).
28. Yadon, A. N. *et al*. A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. *Nature communications* **8**, 588, https://doi.org/10.1038/s41467-017-00721-2 (2017).
29. Petrella, S. *et al*. Crystal structure of the pyrazinamidase of Mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide. *PloS one* **6**, e15785, https://doi.org/10.1371/journal.pone.0015785 (2011).
30. Jubb, H. C. *et al*. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of molecular biology* **429**, 365–371, https://doi.org/10.1016/j.jmb.2016.12.004 (2017).
31. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC genomics* **16**(Suppl 8), S1, https://doi.org/10.1186/1471-2164-16-s8-s1 (2015).
32. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PloS one* **7**, e46688, https://doi.org/10.1371/journal.pone.0046688 (2012).
33. Pires, D. E., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics (Oxford, England)* **30**, 335–342, https://doi.org/10.1093/bioinformatics/btt691 (2014).
34. Pires, D. E., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research* **42**, W314–319, https://doi.org/10.1093/nar/gku411 (2014).
35. Rodrigues, C. H., Pires, D. E. & Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic acids research* **46**, W350–w355, https://doi.org/10.1093/nar/gky300 (2018).
36. Pires, D. E., Blundell, T. L. & Ascher, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Scientific reports* **6**, 29575, https://doi.org/10.1038/srep29575 (2016).
37. Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G. & Furnham, N. Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Scientific reports* **8**, 15356, https://doi.org/10.1038/s41598-018-33370-6 (2018).
38. Whitfield, M. G. *et al*. Mycobacterium tuberculosis pncA Polymorphisms That Do Not Confer Pyrazinamide Resistance at a Breakpoint Concentration of 100 Micrograms per Milliliter in MGIT. *Journal of clinical microbiology* **53**, 3633–3635, https://doi.org/10.1128/jcm.01001-15 (2015).
39. Rose, A. S. *et al*. NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics (Oxford, England)* **34**, 3755–3758, https://doi.org/10.1093/bioinformatics/bty419 (2018).
40. Kushner, S. *et al*. Experimental chemotherapy of tuberculosis; substituted nicotinamides. *The Journal of organic chemistry* **13**, 834–836, https://doi.org/10.1021/jo01164a008 (1948).

41. Stoffels, K., Mathys, V., Fauville-Dufaux, M., Wintjens, R. & Bifani, P. Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy* **56**, 5186–5193, https://doi.org/10.1128/aac.05385-11 (2012).
42. Koser, C. U. *et al.* Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS pathogens* **8**, e1002824, https://doi.org/10.1371/journal.ppat.1002824 (2012).
43. Farhat, M. R. *et al.* Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. *American journal of respiratory and critical care medicine* **194**, 621–630, https://doi.org/10.1164/rccm.201510-2091OC (2016).
44. Karmakar, M. *et al.* Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PloS one* **14**, e0217169, https://doi.org/10.1371/journal.pone.0217169 (2019).
45. Pires, D. E., Chen, J., Blundell, T. L. & Ascher, D. B. *In silico* functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Scientific reports* **6**, 19848, https://doi.org/10.1038/srep19848 (2016).
46. Ascher, D. B. *et al.* Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for. *RNA. Scientific reports* **4**, 4765, https://doi.org/10.1038/srep04765 (2014).
47. Kano, F. S. *et al.* The Presence, Persistence and Functional Properties of Plasmodium vivax Duffy Binding Protein II Antibodies Are Influenced by HLA Class II Allelic Variants. *PLoS Negl. Trop. Dis.* **10**, e0005177, https://doi.org/10.1371/journal.pntd.0005177 (2016).
48. Phelan, J. *et al.* Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **14**, 31, https://doi.org/10.1186/s12916-016-0575-9 (2016).
49. Silvino, A. C. *et al.* Variation in Human Cytochrome P-450 Drug-Metabolism Genes: A Gateway to the Understanding of Plasmodium vivax Relapses. *PloS one* **11**, e0160172, https://doi.org/10.1371/journal.pone.0160172 (2016).
50. Albanaz, A. T. S., Rodrigues, C. H. M., Pires, D. E. V. & Ascher, D. B. Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. *Expert Opin. Drug Discov* **12**, 553–563, https://doi.org/10.1080/17460441.2017.1322579 (2017).
51. Park, Y. *et al.* Essential but Not Vulnerable: Indazole Sulfonamides Targeting Inosine Monophosphate Dehydrogenase as Potential Leads against Mycobacterium tuberculosis. *ACS infectious diseases* **3**, 18–33, https://doi.org/10.1021/acsinfecdis.6b00103 (2017).
52. Singh, V. *et al.* The Inosine Monophosphate Dehydrogenase, GuaB2, Is a Vulnerable New Bactericidal Drug Target for Tuberculosis. *ACS infectious diseases* **3**, 5–17, https://doi.org/10.1021/acsinfecdis.6b00102 (2017).
53. Hawkey, J. *et al.* Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. *Microbial. Genomics* **4**, -, https://doi.org/10.1099/mgen.0.000165 (2018).
54. Holt, K. E. *et al.* Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.* **50**, 849–856, https://doi.org/10.1038/s41588-018-0117-9 (2018).
55. Worth, C. L., Preissner, R. & Blundell, T. L. SDM–a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research* **39**, W215–222, https://doi.org/10.1093/nar/gkr363 (2011).
56. Pires, D. E. & Ascher, D. B. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic acids research* **44**, W557–561, https://doi.org/10.1093/nar/gkw390 (2016).
57. Hall, M. *et al.* The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10–18, https://doi.org/10.1145/1656274.1656278 (2009).
58. Votintseva, A. A. *et al.* Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *Journal of clinical microbiology* **53**, 1137–1143, https://doi.org/10.1128/jcm.03073-14 (2015).

Acknowledgements

M.K. and C.M.H.R. were funded by the Melbourne Research Scholarship. Funding for genomic sequencing was provided by the Department of Health and Human Services, Victoria. D.B.A. was funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (MR/M026302/1), the Jack Brockhoff Foundation (JBF 4186, 2016), and an Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia (GNT1174405). This work was supported in part by the Victorian Government's OIS Program.

Author contributions

M.K. performed the analysis and along with C.H.M.R. developed the analysis tool. K.H. and J.D. contributed to data collected and analysis. D.B.A. conceived, designed and supervised the project. All authors contributed to manuscript writing and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58635-x.

Correspondence and requests for materials should be addressed to D.B.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Supplementary Materials

Structure guided prediction of Pyrazinamide resistance mutations in pncA

Malancha Karmakar[1,2,3], Carlos H. M. Rodrigues[2,3], Kristy Horan[4], Justin T. Denholm[1] , David B. Ascher[2,3,5]

[1]Victorian Tuberculosis Program, Melbourne Health and Department of Microbiology and Immunology, University of Melbourne, Melbourne, Victoria, Australia

[2]Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[3]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia

[4]Microbiological Diagnostic Unit Public Health Laboratory, University of Melbourne at The Peter Doherty Institute for Infection &Immunity, Melbourne, Victoria, Australia

[5]Department of Biochemistry, University of Cambridge, CB2 1GA, UK

Table S1: The list of different features used to analyze and build the empirical model for predicting novel resistance mutations in PZA.

Parameters	Effect measured	Technique	Mean of R mutations	Mean of S mutations	Mean	p-value*	95% CI
DUET (Kcal/mol)	Protein Stability	Graph-based signatures	-1.13	-0.57	-0.85	$5.28\ e^{-14}$	[-0.77 to -0.92]
Distance from Ligand (Å)	Distance of the mutation from the drug (PZA) binding site	Perl script (in-house)	9.48	13.03	11.25	$1.41\ e^{-15}$	[10.81 to 11.65]
DynaMut (Kcal/mol)	Conformational flexibility	Normal mode analysis	-0.24	0.06	-0.08	$6.55\ e^{-06}$	[-0.02 to -0.16]
mCSM-Stability (Kcal/mol)	Protein Stability	Graph-based signatures	-1.10	-0.64	-0.87	$3.83\ e^{-12}$	[-0.80 to -0.93]
RSA (Å)	Environmental characteristics	Python script (in-house)	0.18	0.39	0.28	$< 2.2\ e^{-16}$	[0.27 to 0.31]
SNAP	Functional effect of single nucleotide substitution	Neural Networks	49.83	4.81	27.73	$< 2.2\ e^{-16}$	[23.25 to 31.39]
Ligand binding affinity (mCSM-Lig)	Ligand binding affinity	Graph-based signatures	-0.98	-0.84	-0.90	0.14	[-0.82 to -0.99]
PROVEAN	Functional effect of single nucleotide substitution	alignment-based score approach	-5.42	-3.04	-4.23	$< 2.2\ e^{-16}$	[-4.01 to -4.45]
SDM (Kcal/mol)	Protein Stability	Graph-based signatures	-1.15	-0.30	-0.72	$1.14\ e^{-07}$	[-0.57 to -0.88]
Dihedral angle (Phi)	Environmental characteristics	Python script (in-house)	-73.18	-71.36	-71.71	0.58	[-67.30 to -77.22]
Dihedral angle (Psi)	Environmental characteristics	Python script (in-house)	50.65	36.98	44.35	0.11	[36.23 to 51.41]

Residue Depth	Environmental characteristics	Python script (in-house)	1.09	0.74	0.92	$< 2.2\ e^{-16}$	[0.89 to 0.95]
ENCoM	Conformational flexibility	Normal mode analysis	0.09	0.09	0.09	0.87	[0.06 to 0.13]
Relative b-factor	Environmental characteristics	Python script (in-house)	3.03	3.18	3.10	$2.57\ e^{-10}$	[3.09 to 3.13]

*p-value calculated using Welch two-sample t-test

Table S2: List of performances for predictive models trained on individual classes of attributes and all attributes combined using 10-fold cross validation.

Attributes	Class label	Accuracy	MCC	Precision	Recall	F-measure	ROC AUC
Stability	R	0.57	0.21	0.61	0.57	0.59	0.62
	S	0.64	0.21	0.59	0.64	0.62	0.62
	Weighted Avg.	0.61	0.21	0.60	0.60	0.60	0.62
Dynamics	R	0.55	0.14	0.57	0.55	0.56	0.62
	S	0.58	0.14	0.57	0.58	0.57	0.62
	Weighted Avg.	0.56	0.14	0.57	0.57	0.57	0.62
Evolutionary Conservation	R	0.66	0.32	0.66	0.66	0.66	0.70
	S	0.65	0.32	0.66	0.65	0.66	0.70
	Weighted Avg.	0.65	0.32	0.66	0.66	0.66	0.70
Ligand interactions	R	0.62	0.26	0.63	0.62	0.62	0.68
	S	0.64	0.26	0.62	0.64	0.63	0.68
	Weighted Avg.	0.63	0.26	0.63	0.63	0.63	0.68
Backbone geometry (Structural environment)	R	0.61	0.27	0.64	0.62	0.63	0.70
	S	0.64	0.27	0.63	0.64	0.64	0.70
	Weighted Avg.	0.63	0.27	0.63	0.63	0.63	0.70
Predictive Model	**R**	**0.75**	**0.60**	**0.84**	**0.75**	**0.79**	**0.83**
	S	**0.85**	**0.60**	**0.77**	**0.85**	**0.81**	**0.83**
	Weighted Avg.	**0.80**	**0.60**	**0.80**	**0.80**	**0.80**	**0.83**

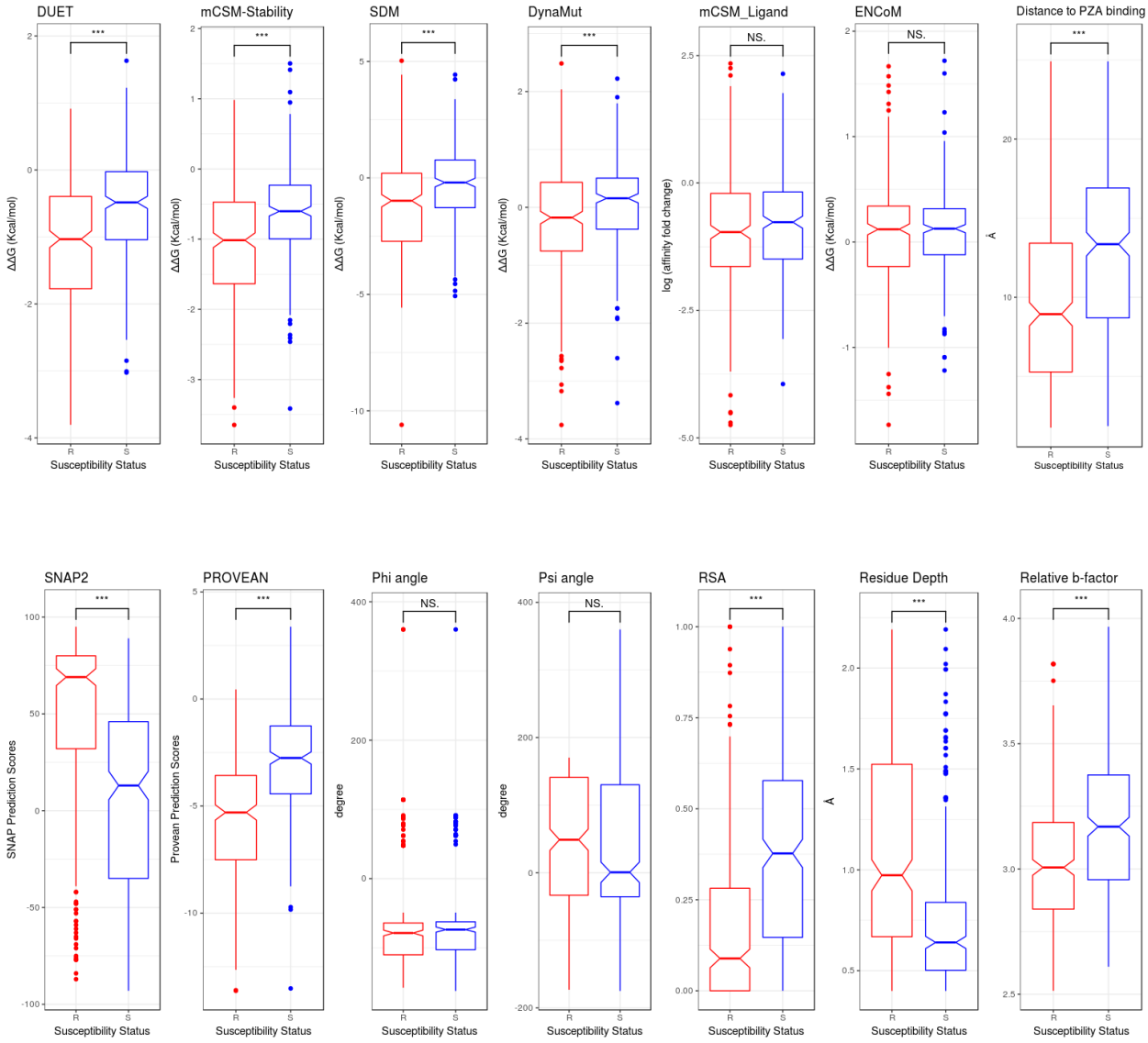


Figure S2: Boxplots comparing features calculated for resistant and susceptible variants. The resistant associated mutations (R) are represented as red and the susceptible mutations (S) as blue. NS- non-significant; *** $p < 0.0001$ by Welch two sample t-test.

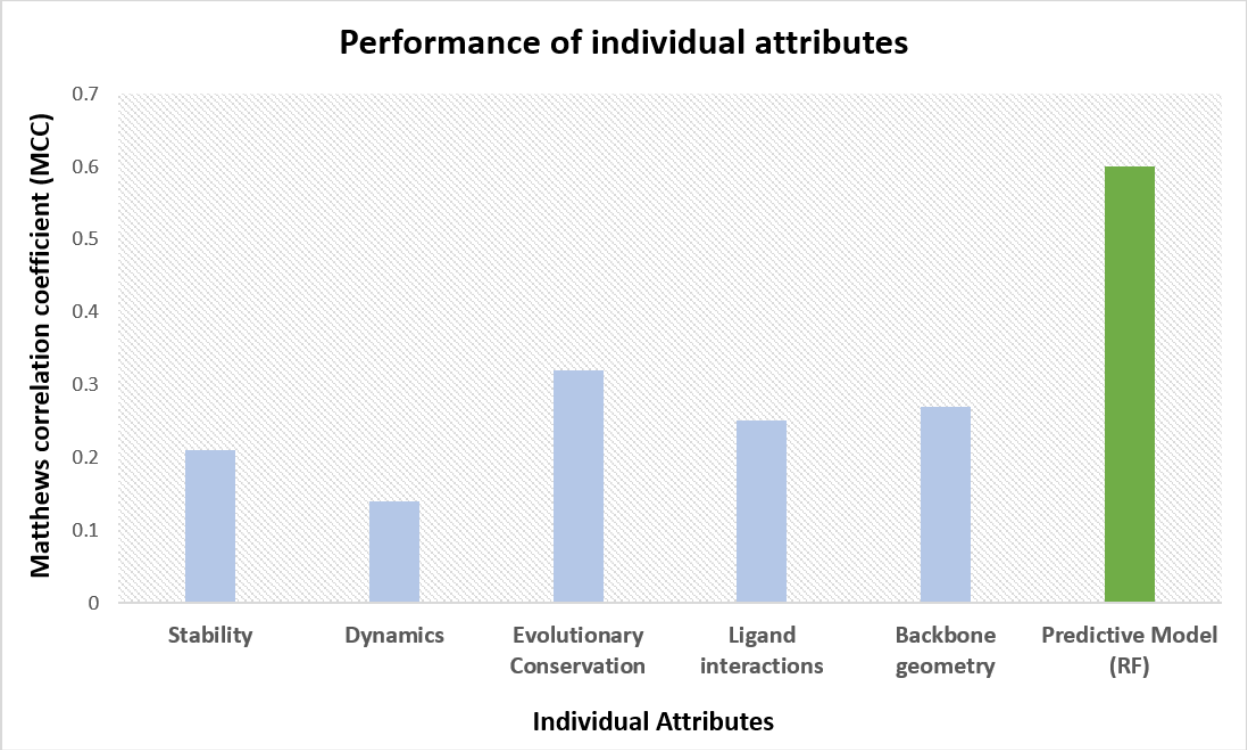


Figure S1: Performance of predictive model trained on single class of features. The Random Forest algorithm was trained using 10-fold cross validations using each single class of features (first five bars from left to right; blue bars) and with the combination of all features (green bar). We observe the predicted MCC score is low when we use only a single class of feature for training. However, a significant increase is observed when different features are combined for the predictive model.

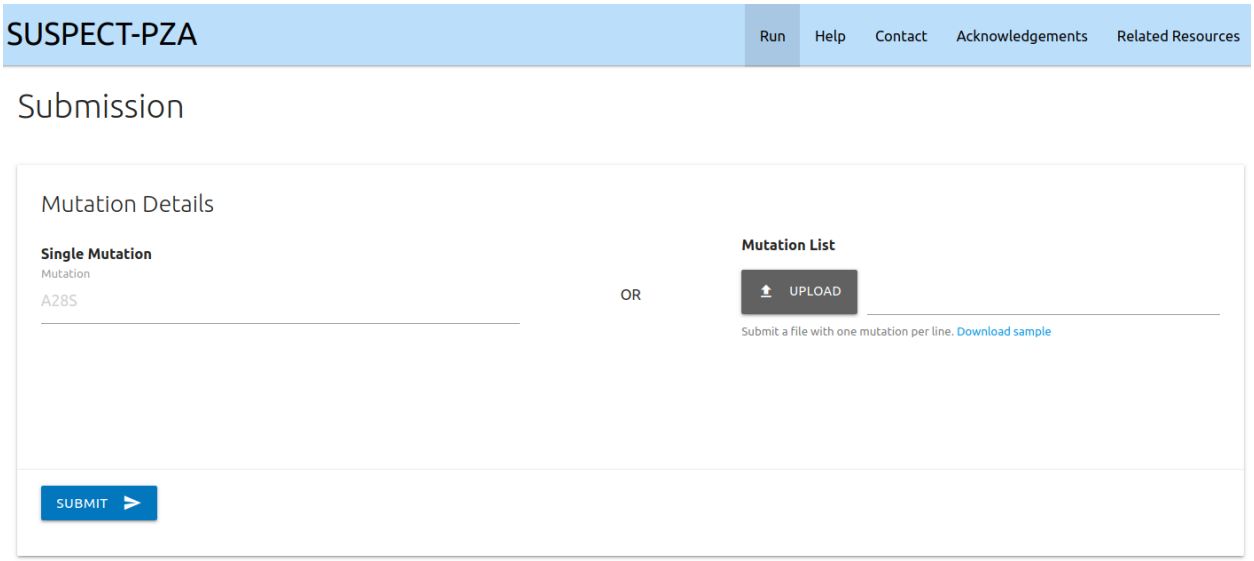


Figure S3: SUSPECT-PZA submission page. The submission page for single point mutation or to upload a list of single point mutations. This can be accessed via the menu item Run on the top bar.

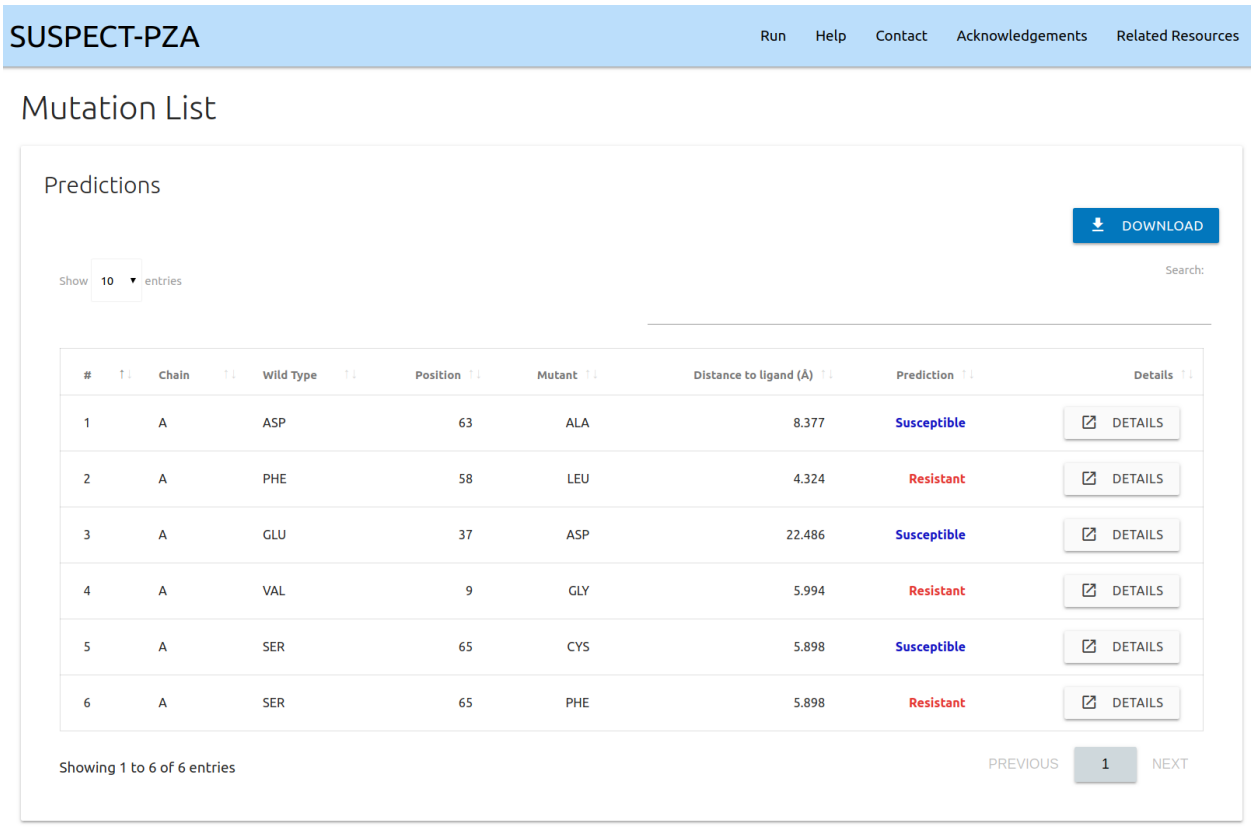


Figure S4: SUSPECT-PZA results page for a list of single point mutations. The predictions for every single point mutation will be displayed as a table in the order of input as in the mutation list. The results can be downloaded as a .csv file by clicking on the Download button on the top right corner. All the analysis discussed for the single mutation option can be analysed for each single mutation on the table through the Details button of each row.

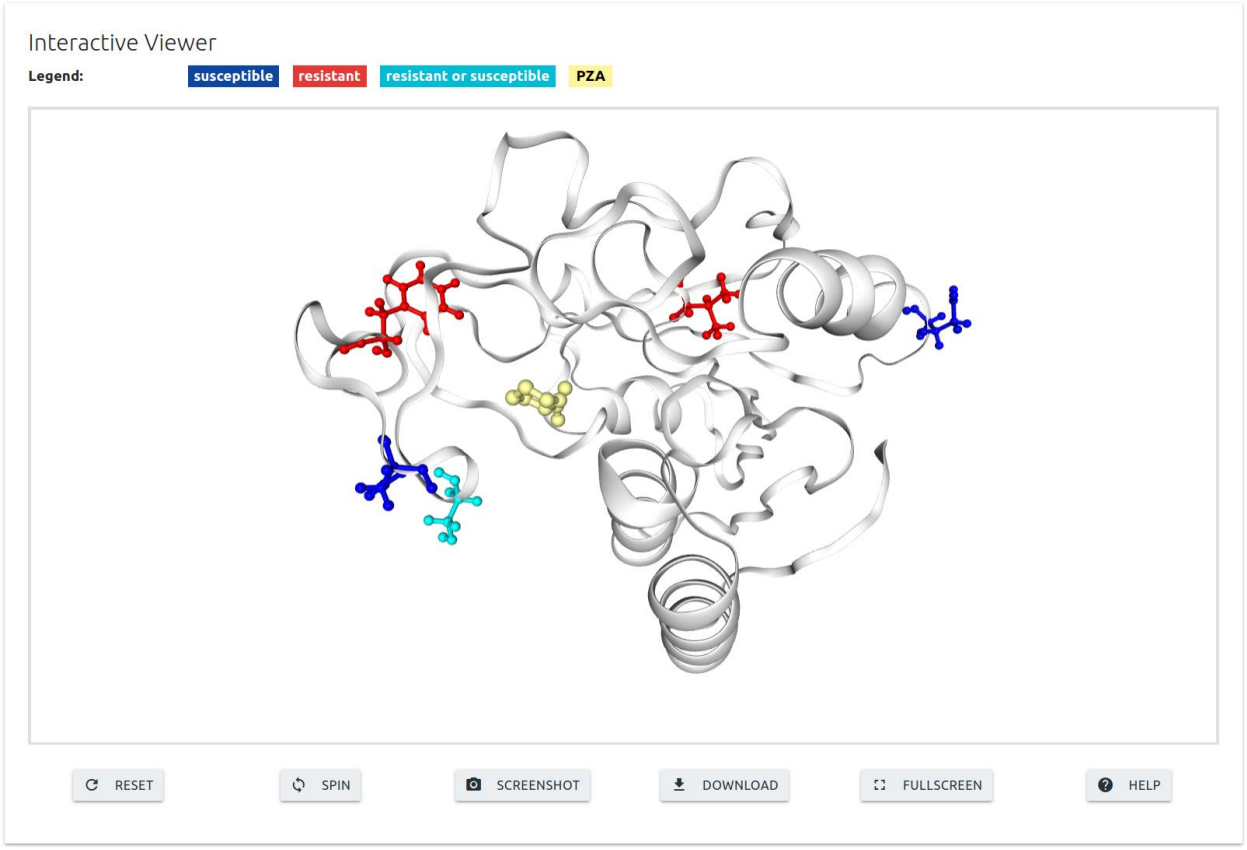


Figure S5: SUSPECT-PZA interactive viewer showing the mutations on the secondary structure. Result page displaying the location of the susceptible (blue, ball and stick representation) and resistant (red, ball and stick representation) mutations. For the amino acid position which harbors both susceptible and resistant mutation is shown in cyan (ball and stick representation).

CHAPTER 3.2: ANALYSIS OF A NOVEL PNCA MUTATION FOR SUSCEPTIBILITY TO PYRAZINAMIDE THERAPY

Summary: In early 2017, a 42-year-old woman, originally from Vietnam, presented with right upper lobe pneumonia; she was diagnosed with pulmonary tuberculosis. Phenotypic drug susceptibility testing identified resistance to isoniazid, rifampicin, pyrazinamide, and ethambutol. Although drug susceptibility testing suggested the patient was phenotypically resistant to PZA, consistent with World Health Organization recommendations, PZA treatment was continued as part of a multidrug-resistant tuberculosis regimen. Amplicon sequencing identified a novel frameshift mutation in the pncA gene of M. tuberculosis (c.85_86insG). Given the uncertain impact of this mutation, we went on to consider whether computational analysis of protein structure could provide insight into the potential efficacy of PZA. The structure of the mutant was generated using homology and *ab initio* modeling using the experimental crystal structure of the wild type. Structural insights revealed the frameshift mutation resulted in the generation of a truncated and incomplete protein that lacked the active site pocket, including most of the catalytic residues and iron coordination residues necessary for activity. This strongly suggests that the pncA c.85_86insG frameshift mutation would lead to a total loss of catalytic activity of the protein, and hence PZA treatment would be completely ineffective in this case, as the mutant PncA could not activate the prodrug. This is reflected in the structure of the mutant protein, which is incomplete and would lack any activity. This result was consistent with phenotypic testing, and accordingly, pyrazinamide treatment was ceased.

This work was published in the journal *American Journal of Respiratory and Critical Care Medicine* as a first author publication. "Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide

Therapy", **Karmakar, M.**, Globan, M., Fyfe, J.A.M., Stinear, T.P., Johnson, P.D.R., Holmes, N.E., Denholm, J.T., Ascher, D.B. (2018) (doi: 10.1164/rccm.201712-2572LE)

of general control nonderepressible 2 (GCN2) in pulmonary veno-occlusive disease. *J Heart Lung Transplant* 2018;37: 647-655.

Copyright © 2018 by the American Thoracic Society

Analysis of a Novel *pncA* Mutation for Susceptibility to Pyrazinamide Therapy

To the Editor:

Pyrazinamide (PZA), which is an analog of nicotinamide, is an important first-line drug used in the short-course treatment of tuberculosis. PZA is a prodrug devoid of significant antibacterial activity. It is metabolized into its active form, pyrazinoic acid, by the amidase activity of the *Mycobacterium tuberculosis* nicotinamidase/pyrazinamidase, encoded by the *pncA* gene. Mutations in *pncA* that prevent activation of the prodrug represent the major mechanism of PZA resistance in *M. tuberculosis* (1). This antibiotic plays a key role in shortening the duration of antituberculous treatment because of its activity against the persisting tubercle bacilli at acidic pH.

Current phenotypic testing for PZA drug susceptibility is problematic. Culture-based methods such as Wayne's method are used as screening assays with confirmation of resistant strains via the BD BACTEC MGIT 960 system (Becton Dickinson) (2). Results obtained from phenotypic laboratory testing have poor reproducibility. Sequencing of the *pncA* gene to determine the presence of mutations may be a more reliable method for confirmation of phenotypic PZA resistance (3). International recommendations suggest continued usage of PZA irrespective of susceptibility results, particularly in the treatment of multidrug-resistant disease (4). This is despite the adverse effects associated with PZA treatment.

Case Report

In early 2017, a 42-year-old woman, originally from Vietnam, presented with right upper lobe pneumonia; she was diagnosed with pulmonary tuberculosis. Phenotypic drug susceptibility testing identified resistance to isoniazid, rifampicin, pyrazinamide, and ethambutol. Although drug susceptibility testing suggested the patient was phenotypically resistant to PZA, consistent with World Health Organization recommendations, PZA treatment was continued as part of a multidrug-resistant tuberculosis regimen. Amplicon sequencing identified a novel frameshift mutation in the *pncA* gene of *M. tuberculosis* (c.85_86insG). Given the uncertain impact of this mutation, we went on to consider whether computational analysis of protein structure (5) could provide insight into the potential efficacy of PZA.

Methods

We have developed an *in silico* mutational analysis platform that is able to characterize the molecular consequences of mutations on protein structure and function (5). This has been used to preemptively identify likely resistance mutations in drug targets (6, 7). Using these tools, we assessed the biophysical changes on mutation on the structure of PncA and drug activation.

A list of 617 nonsynonymous single-nucleotide variants (nsSNVs) of *pncA* was obtained from the GMTV (Genome-wide *Mycobacterium tuberculosis* Variation) Database Project, Tuberculosis Drug Resistance Mutation Database, and saturation mutagenesis (8). Mapping nsSNVs associated with resistance onto the crystal structure of PncA revealed that they were distributed throughout the entire protein structure (Figure 1A), complicating resistance inference from sequence analysis. The structural and functional effects of these mutations were assessed using our graph-based signature pipeline (5). This provided insight into how the curated nsSNVs altered protein folding, stability, conformation, and PZA-binding affinity. This information was used to train a Random Forest (machine-learning algorithm) binary classifier, using the Weka toolkit. Random Forest is an ensemble-learning robust classification algorithm, in which multiple decision trees are included over a random subset of features and decide the output via majority voting. The model was trained by 10-fold cross-validation and performance evaluated by area under the receiver operating characteristic curve, precision, and accuracy. Further validation of the models was performed using two subsets of 93 mutations, which were nonredundant at the position-level mutations in the training set. Analysis of the final model revealed a set of structural features that distinguished between susceptible and resistant *pncA* point mutations.

Building on this structural analysis, the functional consequence of the novel clinical frameshift mutation was analyzed in the context of the protein structure. The experimental crystal structure of holo-wild-type PncA (PDB ID: 3PL1) (9) was minimized in Prime, and PZA docked into the active site using Glide, two exclusive packages of the comprehensive homology modeling software Schrödinger Suites. The docking revealed that PZA formed key interactions within the pocket, including with the catalytic triad (Asp8, Lys96, and Cys138), substrate-binding residues (Trp68 and Phe13), and the iron center (Asp49, His51, His57, and Fe^{2+}) (10). The wild-type and mutant protein sequences were manually aligned and displayed with ESPript 3.0 (Figure 2A), and the structure of the mutant (Figure 2C) was generated by homology and *ab initio* modeling, using the experimental structure of the wild type (Figure 2B) (Schrödinger Suites).

Results

Using the structural and biophysical effects of the mutations on the protein structure, we were able to classify mutations as either susceptible or resistant with an accuracy of 77% (Figure 1C). This approach performed equally well in the identification of either class, correctly classifying all mutations previously associated with conferring PZA resistance at high confidence, and mutations not involved in PZA resistance (100% accuracy) (11).

Supported by a Melbourne Research Scholarship from the University of Melbourne (M.K.). D.B.A was funded by a Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (MR/M026302/1), the Jack Brockhoff Foundation (JBF 4186, 2016), and a C. J. Martin Research Fellowship from the National Health and Medical Research Council of Australia (APP1072476).

Author Contributions: M.K., J.T.D., and D.B.A. were involved in the study design, execution, data analysis, and writing of all versions of this work. M.G., J.A.M.F., T.P.S., P.D.R.J., and N.E.H. were involved in clinical and laboratory aspects of investigation. All authors contributed to preparation of this manuscript and approve the final version.

Originally Published in Press as DOI: 10.1164/rccm.201712-2572LE on April 25, 2018

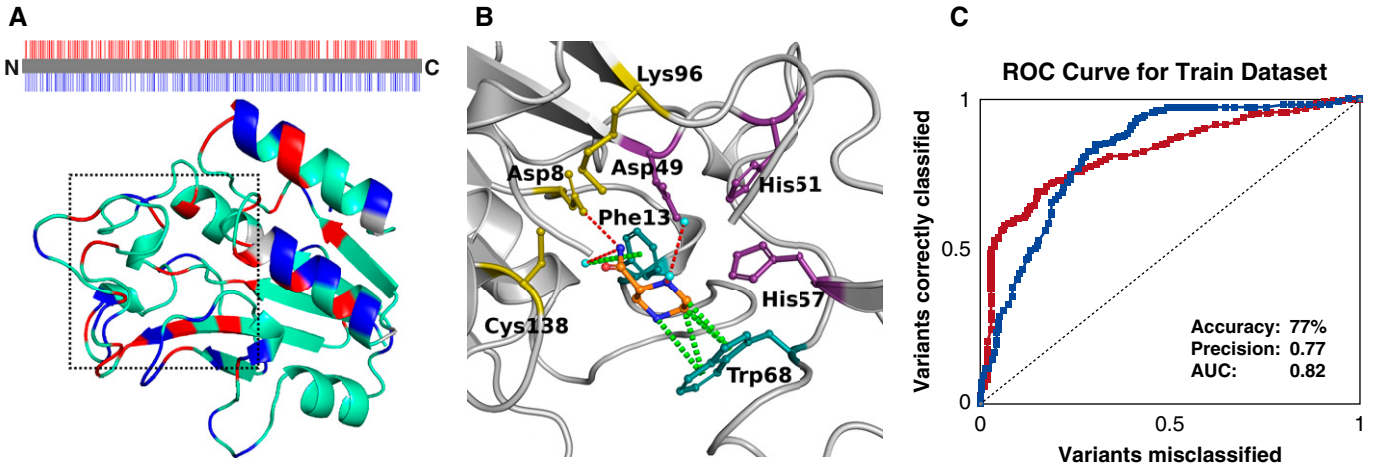


Figure 1. Identification of resistant and susceptible missense mutations in *pncA*. (*A*) The protein sequence and structure of PncA is colored by whether resistant (red) or susceptible (blue) variants have been observed at that location. Highlighting the difficulty of genomic analysis of *pncA*, both resistant and susceptible variants have been observed across many residue positions (cyan). The catalytic site in which pyrazinamide (PZA) was docked is located inside the dashed box. (*B*) The key molecular interactions between PZA (orange segments) and the catalytic triad (yellow), substrate-binding site (teal), and iron center (purple). Hydrogen bonds are shown as red dashes, and π interactions as green dashes. (*C*) The ROC curve shows that, using the structural and functional consequences of the variants, we were able to accurately identify resistant (red) and susceptible (blue) variants. AUC = area under the curve; ROC = receiver operating characteristic.

Strain-specific differences in variants with conflicting experimental data (12) could be detected by our tool, using homology models of the corresponding strain's PncA protein.

Analysis of our model revealed that PncA-resistant mutations were associated with large changes in protein folding and stability (mCSM-Stability scores ≥ -1.72 kcal/mol) ($P < 0.0001$) or located in close proximity to the catalytic triad and substrate-binding site (<8.54 Å) ($P < 0.0001$). Therefore, these freely available biophysical measurements could provide useful information to help guide genomic analysis of novel *pncA* variants.

We next considered the patient's frameshift mutation in light of these structural insights. As shown in Figure 2, the frameshift mutation resulted in the generation of a truncated and incomplete protein that lacked the active site pocket, including most of the catalytic residues and iron coordination residues necessary for activity. This strongly suggests that the *pncA* c.85_86insG frameshift mutation would lead to a total loss of catalytic activity of the protein, and hence PZA treatment would be completely ineffective in this case, as the mutant PncA could not activate the prodrug. This is reflected in the structure of the mutant protein, which is incomplete and would lack any activity (Figure 2). This result was consistent with phenotypic testing, and accordingly, pyrazinamide treatment was ceased.

Discussion

This case study demonstrates the power of using structural information to quantitatively evaluate novel variants in real time, providing invaluable insight to help guide therapy. Although existing recommendations may suggest continuing treatment of multidrug-resistant tuberculosis with pyrazinamide irrespective of phenotype testing, our approach suggests that using structural information to guide analysis of genomic sequencing may offer useful tools for clinicians to consider. These structural insights also assist in informing the mechanisms for drug activity and the development of resistance. Our approach is not limited only to analysis of variants in *pncA* but could be applied to any protein associated with resistance for infectious and noninfectious disease treatment. ■

Author disclosures are available with the text of this letter at www.atsjournals.org.

Malancha Karmakar, B.Sc., M.Sc.
University of Melbourne
Melbourne, Victoria, Australia

Maria Globan, B.Sc.
Janet A. M. Fyfe, B.Sc. (Hons.), Ph.D.
University of Melbourne
Melbourne, Victoria, Australia
and
Melbourne Health
Melbourne, Victoria, Australia

Timothy P. Stinear, B.Sc. (Hons.), Ph.D.
University of Melbourne
Melbourne, Victoria, Australia

Paul D. R. Johnson, M.B. B.S., Ph.D., F.R.A.C.P.
University of Melbourne
Melbourne, Victoria, Australia
and
World Health Organization Collaborating Centre for Mycobacterium ulcerans
Melbourne, Victoria, Australia

Natasha E. Holmes, M.B. B.S., Ph.D., F.R.A.C.P., Grad.Cert.Clin.Teach., Grad.Cert.Clin.Ed.
University of Melbourne
Melbourne, Victoria, Australia
and
Austin Health
Heidelberg, Victoria, Australia

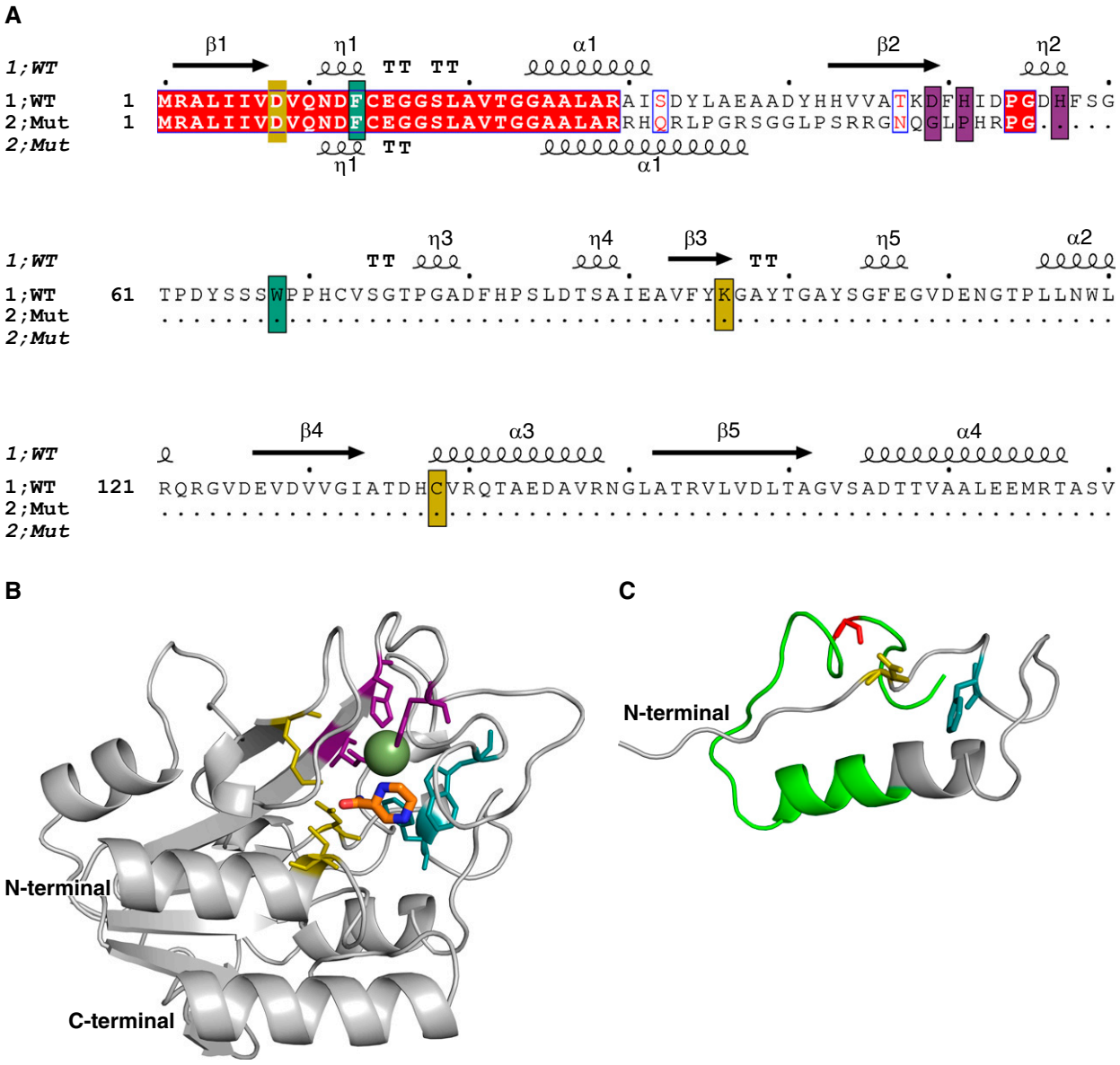


Figure 2. Structural analysis of a novel *pncA* frameshift mutation. (*A*) The sequence alignment between the wild-type and mutant protein sequences shows that only the first 29 residues are conserved (red), and that the frameshift leads to the introduction of a premature stop codon. The catalytic triad, substrate-binding site, and iron center are highlighted in yellow, teal, and purple, respectively. The secondary structure of the wild-type PncA protein is shown above the sequence (β = β sheet, α = α helix, and η = loop). (*B*) The structure of the wild-type PncA protein is represented as a ribbon (gray), bound to the drug pyrazinamide (in orange segments). The Fe^{2+} ion is shown as a green sphere. (*C*) The modeled structure of the mutant PncA protein highlights that most of the catalytic site and structure of the wild-type protein is absent in the mutant. The region not conserved with the wild-type sequence is shown in green. Both wild-type and mutant structures are shown from the same perspective. Mut = mutant; WT = wild type.

Justin T. Denholm, B.Med., M.Bioethics, M.P.H.+T.M., Ph.D., F.R.A.C.P.
University of Melbourne
Melbourne, Victoria, Australia

David B. Ascher, B.Biotech., B.Sc. (Hons.), L.L.B., Ph.D., M.R.A.C.I. C.Chem.
University of Melbourne
Melbourne, Victoria, Australia
and
University of Cambridge
Cambridge, United Kingdom

ORCID ID: 0000-0003-2948-2413 (D.B.A.).

References

1. Zhang Y, Shi W, Zhang W, Mitchison D. Mechanisms of pyrazinamide action and resistance. *Microbiol Spectr* 2013;2:1–12.
2. Cui Z, Wang J, Lu J, Huang X, Zheng R, Hu Z. Evaluation of methods for testing the susceptibility of clinical *Mycobacterium tuberculosis* isolates to pyrazinamide. *J Clin Microbiol* 2013;51:1374–1380.
3. Chang KC, Yew WW, Zhang Y. Pyrazinamide susceptibility testing in *Mycobacterium tuberculosis*: a systematic review with meta-analyses. *Antimicrob Agents Chemother* 2011;55:4499–4505.
4. World Health Organization. WHO treatment guidelines for drug-resistant tuberculosis: 2016 update. Geneva, Switzerland: World Health Organization; 2016.

5. Pires DE, Chen J, Blundell TL, Ascher DB. *In silico* functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci Rep* 2016;6:19848.
6. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, *et al*. Essential but not vulnerable: indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against *Mycobacterium tuberculosis*. *ACS Infect Dis* 2017;3:18–33.
7. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, *et al*. The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. *ACS Infect Dis* 2017;3:5–17.
8. Yadon AN, Maharaj K, Adamson JH, Lai YP, Sacchettini JC, Ioerger TR, *et al*. A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. *Nat Commun* 2017;8:588.
9. Petrella S, Gelus-Ziental N, Maudry A, Laurans C, Boudjelloul R, Sougakoff W. Crystal structure of the pyrazinamidase of *Mycobacterium tuberculosis*: insights into natural and acquired resistance to pyrazinamide. *PLoS One* 2011;6:e15785.
10. Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J Mol Biol* 2017;429: 365–371.
11. Miotto P, Cabibbe AM, Feuerriegel S, Casali N, Drobniewski F, Rodionova Y, *et al*. *Mycobacterium tuberculosis* pyrazinamide resistance determinants: a multicenter study. *MBio* 2014;5: e01819-14.
12. Baddam R, Kumar N, Wieler LH, Lankapalli AK, Ahmed N, Peacock SJ, *et al*. Analysis of mutations in *pncA* reveals non-overlapping patterns among various lineages of *Mycobacterium tuberculosis*. *Sci Rep* 2018;8:4628.

Copyright © 2018 by the American Thoracic Society

Overfitting and Use of Mismatched Cohorts in Deep Learning Models: Preventable Design Limitations

To the Editor:

We read with great interest the study by González and colleagues (1) in which they used deep learning models to learn from the computed tomography (CT) scans of 7,983 participants in the COPDGene (Genetic Epidemiology of COPD) study (2). Their objective was to learn from visual data present in these CT scans and subsequently study the model's ability to diagnose chronic obstructive pulmonary disease (COPD) and predict respiratory events and mortality in a validation cohort (1,000 COPDGene scans) and a test cohort (1,672 ECLIPSE [Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points] [3] scans). The validation and test cohorts differed significantly in terms of their COPD severity (lower FEV_1% predicted and higher Global Initiative for Chronic Obstructive Lung Disease [GOLD] stage in ECLIPSE [3]).

The model performed very well in terms of COPD detection as well as prediction of acute respiratory events in the validation cohort of 1,000 COPDGene participants (i.e., it correctly identified COPD in 773/1,000 scans, and there was a strong correlation between the actual FEV_1 and the predicted FEV_1 [1]). However, the model's performance in the test cohort was inferior in both detection of COPD and prediction of acute respiratory events (only 29% of individuals were correctly staged, and the model was unable to identify patients at higher risk of respiratory events [1, 4]).

We believe that there are two significant design limitations in the authors' approach toward execution of the deep learning process and selection of the cohorts.

A significant proportion of the CT scan data (7,983 out of a total of 8,983 COPDGene scans, 88.9%) were used for training the deep learning model (1). This leads to a potential overfitting of the learning model. Overfitting is the consequence of the model learning from a high volume of details that incorporate both noise and signal existing in the training datasets. This leads to a superior performance in the internal validation cohort and inferior performance in an external test dataset (5). In other words, such models do not explain test cohorts, but they explain the training data very well (5). This suspicion is supported by the study's superior results in the smaller internal validation cohort and inferior performance in the external test cohort. This is particularly relevant because the validation cohort (n = 1,000 scans, 10% of the COPDGene cohort) likely does not represent most of the variance existing in the COPDGene cohort (a cohort of smokers with and without COPD [2]).

The second limitation arises as an indirect consequence of inherent differences between the COPDGene and ECLIPSE cohorts. The authors do acknowledge in their discussion that there are significant differences between the validation and test cohorts (1). In this study, a predominantly GOLD stage 0–1 cohort (1, 2) served as the training set for a model that was tested in a GOLD stage $\geqslant$2 cohort (3). In our opinion, selecting COPDGene scans with established GOLD stages of $\geqslant$2 (representing 36% of the COPDGene cohort [1], n = 3,600 scans) for teaching and internal validation purposes would have improved the external performance. An ideal deep learning strategy would have allocated 50–70% (n = 1,800–2,500) of these scans to the learning model and the remaining 30–50% scans (n = 1,080–1,800) to the internal validation effort. This would have resulted in a true enumeration of the model's performance in the internal validation phase.

In conclusion, the findings could simply represent the performance of a potentially overfitted model (5) and likely do not reflect the suggested superior performance of the tool in the COPDGene validation dataset. The lack of use of appropriate cohorts for training and validation is another significant limitation and can explain the inferior performance in the test cohort (ECLIPSE). ■

Author disclosures are available with the text of this letter at www.atsjournals.org.

Srinivas R. Mummadi, M.D., M.B.I.
Metro Health-University of Michigan Health
Wyoming, Michigan

Akrum Al-Zubaidi, D.O.
National Jewish Health
Denver, Colorado

Peter Y. Hahn, M.D., M.B.A.
Metro Health-University of Michigan Health
Wyoming, Michigan

ORCID IDs: 0000-0002-8806-445X (S.R.M.); 0000-0001-6143-1144 (A.A.-Z.); 0000-0003-2410-6152 (P.Y.H.).

Originally Published in Press as DOI: 10.1164/rccm.201802-0350LE on April 11, 2018

CHAPTER 4: EMPIRICAL WAYS TO IDENTIFY NOVEL BEDAQUILINE RESISTANCE MUTATIONS

Summary

Background: Clinical resistance against Bedaquiline, the first new anti-tuberculosis compound with a novel mechanism of action in over 40 years, has already been detected in *Mycobacterium tuberculosis*. As a new drug, however, there is currently insufficient clinical data to facilitate reliable and timely identification of genomic determinants of resistance.

Objective: Here we investigate the structural basis for *M. tuberculosis* associated bedaquiline resistance in the drug target, AtpE.

Methods: Together with the 9 previously identified resistance-associated variants in AtpE, 54 non-resistance-associated mutations were identified through comparisons of bedaquiline susceptibility across 23 different mycobacterial species.

Results: Computational analysis of the structural and functional consequences of these variants revealed that resistance associated variants were mainly localized at the drug binding site, disrupting key interactions with bedaquiline leading to reduced binding affinity. This was used to train a supervised predictive algorithm, which accurately identified likely resistance mutations (98.7% accuracy).

Interpretation: Application of this model to circulating variants present in the Asia-Pacific region suggests that current circulating variants are likely to be susceptible to bedaquiline. This tool could be useful for the rapid characterization of novel clinical variants, to help guide the effective use of bedaquiline, and to minimize the spread of clinical resistance.

This chapter has been published in the *Plos one* as a first author publication. "*Empirical ways to identify novel Bedaquiline resistance mutations in AtpE*", **Karmakar, M.**, Rodrigues, C.H.M., Holt, K.E., Dunstan, S.J., Denholm, J.T., Ascher, D.B. (2019) (doi: 10.1371/journal.pone.0217169)

PLOS ONE

RESEARCH ARTICLE

Empirical ways to identify novel Bedaquiline resistance mutations in AtpE

Malancha Karmakar[1,2,3,4], Carlos H. M. Rodrigues[2,4], Kathryn E. Holt[2], Sarah J. Dunstan[5], Justin Denholm[1,3], David B. Ascher[2,4,6]*

1 Victorian Tuberculosis Program, Melbourne Health, Victoria, Australia, **2** Department of Biochemistry and Molecular Biology, University of Melbourne, Melbourne, Victoria, Australia, **3** Department of Microbiology and Immunology, University of Melbourne, Melbourne, Victoria, Australia, **4** Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia, **5** The Peter Doherty Institute for Infection and Immunity, University of Melbourne, Victoria, Australia, **6** Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

* david.ascher@unimelb.edu.au

Check for updates

OPEN ACCESS

Citation: Karmakar M, Rodrigues CHM, Holt KE, Dunstan SJ, Denholm J, Ascher DB (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS ONE 14(5): e0217169. https://doi.org/10.1371/journal.pone.0217169

Editor: Igor Mokrousov, St Petersburg Pasteur Institute, RUSSIAN FEDERATION

Received: January 31, 2019

Accepted: May 1, 2019

Published: May 29, 2019

Copyright: © 2019 Karmakar et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: M.K was funded by the Melbourne Research Scholarship. D.B.A was funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (MR/M026302/1), the Jack Brockhoff Foundation (JBF 4186, 2016), and a C. J. Martin Research Fellowship from the National

Abstract

Clinical resistance against Bedaquiline, the first new anti-tuberculosis compound with a novel mechanism of action in over 40 years, has already been detected in *Mycobacterium tuberculosis*. As a new drug, however, there is currently insufficient clinical data to facilitate reliable and timely identification of genomic determinants of resistance. Here we investigate the structural basis for *M. tuberculosis* associated bedaquiline resistance in the drug target, AtpE. Together with the 9 previously identified resistance-associated variants in AtpE, 54 non-resistance-associated mutations were identified through comparisons of bedaquiline susceptibility across 23 different mycobacterial species. Computational analysis of the structural and functional consequences of these variants revealed that resistance associated variants were mainly localized at the drug binding site, disrupting key interactions with bedaquiline leading to reduced binding affinity. This was used to train a supervised predictive algorithm, which accurately identified likely resistance mutations (93.3% accuracy). Application of this model to circulating variants present in the Asia-Pacific region suggests that current circulating variants are likely to be susceptible to bedaquiline. We have made this model freely available through a user-friendly web interface called SUSPECT-BDQ, StrUctural Susceptibility PrEdiCTion for bedaquiline (http://biosig.unimelb.edu.au/suspect_bdq/). This tool could be useful for the rapid characterization of novel clinical variants, to help guide the effective use of bedaquiline, and to minimize the spread of clinical resistance.

Introduction

Tuberculosis (TB) is the leading cause of infectious disease death worldwide, with over 10 million new cases and 1.6 million deaths in 2017 [1]. A disproportionate burden arises from the estimated 558,000 annual cases of rifampicin resistant TB (RR-TB) with 82% being multi-drug resistant (MDR), which is associated with lengthy, toxic therapy and high rates of mortality [1]. With limited therapeutic options available, especially for MDR-TB and extensively drug-

Health and Medical Research Council (NHMRC) of Australia (APP1072476). The Vietnam genomic dataset was funded by a NHMRC Australia grant (APP1056689) to SJD and KEH. This work was supported in part by the Victorian Government's OIS Program. No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

resistant (XDR) TB, the introduction of new treatment options is urgently required. Bedaquiline, a new anti-TB drug with a novel mechanism of action, targeting the c-ring of ATP synthase (AtpE) [2], was approved for treatment for MDR-TB in 2012 [3, 4]. This innovative drug is potent against both actively replicating and dormant bacilli and has been shown to increase culture conversion in patients with MDR-TB [5]. The use of bedaquiline has expanded considerably in recent years, and has been recommended for more routine use in MDR-TB regimens [6], however clinical failures have already been observed [7, 8]. This necessitates a better understanding of how variants result in resistance to aid in the early detection of resistance.

Phenotypic, and increasingly genotypic, drug susceptibility testing (DST) is recognized as essential for effective individualization of TB therapy. However, while progress has been made in strengthening laboratory diagnostics, the TB community is still struggling to build up laboratory networks with the needed capacity for routine culture and DST [7, 9]. The World Health Organization (WHO) has strongly urged the development of accurate and reproducible DST for bedaquiline and recommended that in the absence of specific DST, bedaquiline resistance should be monitored through MIC assessment [10] with resistance development evaluated in patients with treatment failure or relapse. Early characterization of drug resistance mutations would assist TB patient management and avoid treating individuals with ineffective toxic regimens [11, 12], but capacity for rapid genotypic prediction of bedaquiline resistance is limited by the identification of few known resistance associated variants [13].

In an era of rapidly expanding use of molecular technologies, including whole genome sequencing, tools for evaluating the impact of novel mutations are increasingly vital, particularly for drug resistance to novel and emerging medications such as bedaquiline. Though culture-based detection of resistance will remain the gold standard, *in silico* analyses can support informed decision-making. We have previously shown that the analysis of how variants can affect protein structure and function can be used to reliably characterize how variants lead to drug resistance [14–18]. Using this approach, we have shown that drug resistant mutations can be rapidly, accurately and pre-emptively predicted, guiding drug development [19–22] and clinical diagnosis [23].

In-vitro selection [24] and clinical studies [25] have shown that variants in the *atpE* gene can lead to bedaquiline resistance. To support rapid identification of potential bedaquiline resistance mutations, we considered whether structural information of the drug target could help guide clinical inference on genomic variants. Using a suite of well-established computational tools for characterizing the molecular consequences of mutations on protein structure and function, we have assessed the effects of mutations on the biophysical changes of AtpE folding, stability and on drug binding affinity. This was used to characterize how mutations in AtpE lead to resistance, and to train a predictive multilayer perceptron (feedforward artificial neural network) algorithm to characterize novel AtpE variants.

Methods

Data sets

Resistant variants from *in-vitro* selection studies were curated [13, 24, 26] along with a natural variant [4, 27] and used for model development. Susceptible variants were identified using a novel homology approach, where the genomes of all mycobacteria species sensitive to the drug [28] were aligned, therefore inferring that any present variants were likely to be susceptible. Clinically observed bedaquiline resistant *atpE* variants were curated from published reports [25]. The Vietnam dataset consists of whole genome sequences of 1635 *Mycobacterium tuberculosis* (*Mtb*) strains isolated from patients with pulmonary TB in Ho Chi Minh City, Vietnam.

The *Mtb* genome data is available in NCBI BioProject [ID: PRJNA355614; http://www.ncbi.nlm.nih.gov/bioproject/355614]. Details of the clinical study and the whole genome dataset are found in Thai et al [29] and Holt et al [15].

Homology modeling of AtpE

The structure of *Mtb* AtpE was modelled with MODELLER [30] using the experimental crystal structure of *Mycobacterium phlei* (*M. phlei*) AtpE (PDB ID: 4V1F). The model was then minimized in Prime and bedaquiline docked into the apo structure using Glide (Schrödinger Suite).

Modelling the biophysical consequences of missense variants

The structural consequences of the AtpE polymorphisms were assessed to account for all the potential effects of the mutations. The effects of mutations on protein folding and stability were assessed using SDM [31], mCSM-Stability [32] and DUET [33], and their effects on protein flexibility and conformation was predicted using normal mode analysis by DynaMut [34]. The effect of the difference on the protein-protein interactions between the protomers of AtpE were predicted using mCSM-PPI [32]. The effect of the changes on the binding affinity of bedaquiline towards AtpE were predicted using mCSM-Lig [35–37]. These approaches are novel machine-learning algorithms that use graph-based signatures to represent the structural and chemical environment of the wild-type 3D structure of a protein to quantitatively predict the effects of point mutations. Additionally, SNAP2 [38] was used to provide additional evolutionary based information.

Machine learning

To build the binary classifier, a multilayer perceptron neural network algorithm was trained, based on the implementation available through the Weka toolkit [39]. The resistant variants were up-sampled to create a more balanced model [40]. The training dataset constituted of 50 non-resistant associated variants and 5 resistant associated variants, while the blind test dataset constituted of 4 non-resistant associated variants and 4 resistant associated variants. To avoid over-biasing, the train and blind test dataset were non-redundant with respect to residue position. The model was trained and evaluated using jackknife [41] and leave-one-residue-position-out validation. The classification model was evaluated based on metrics, including the Area Under the ROC curve (AUC), precision and accuracy. Statistical analysis was performed using RStudio (version 3.1.1).

Webserver development

The server front-end was built using materialize CSS framework version 1.0.0, while the back-end was built in Python via the Flask framework (version 0.12.2). It is hosted on a Linux server running Apache.

Results

We used a structure-guided approach to understand the protein structure of the drug target AtpE and machine learning to build an empirical tool that could identify likely resistant mutations. The pipeline used to analyze the variants and train a multilayer perceptron neural network algorithm is shown in Fig 1.

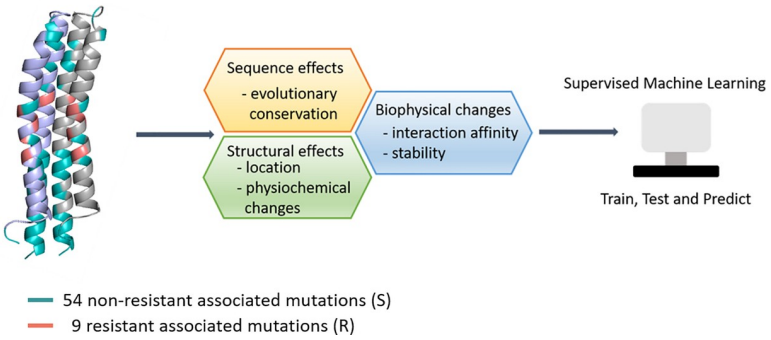


Fig 1. Methodology. This workflow highlights important steps in the methodology and how the main components of the algorithms are computed. In our analysis we used 54 non-resistant associated mutations and 9 resistant mutations for the biophysical analysis, followed by training and validation of our empirical model using a supervised machine learning algorithm.

https://doi.org/10.1371/journal.pone.0217169.g001

Structural information: The drug binding domain

A homology model of *Mtb* H37Rv AtpE was built using the existing experimental crystal structure of AtpE from *Mycobacterium phlei* (PDB ID: 4V1F) [42], which shares a high sequence identity with the *Mtb* protein (84.9%). The protomer model was an alpha helical hairpin structure comprising two membrane-spanning helices connected by a hydrophilic loop. The homo-oligomeric construct was built using the *M*. *phlei* structure as a guide, as the *Mtb* protein has been previously shown to assemble as a homo-nonamer [43] (Fig 2A and 2B). The cylindrical palisade model contained an internal hydrophobic cavity where phospholipid had been proposed to bind. The conserved proton binding residue (E61) was located sandwiched between adjacent protomers and equidistantly distributed along the center of the hydrophobic membrane bilayer.

The top docking poses of bedaquiline with the nonamer homology model identified a pose consistent with that observed in the *M*. *phlei* structure. The drug binding cleft was located at the interface of two protomers, with amino acid residues E61, A62, Y64, F65 from one protomer and I66 from the adjacent protomer defining the drug binding cleft. Analysis of the molecular interactions with Arpeggio [44] highlighted a strong network of polar interactions between the drug and AtpE (Fig 2C). Of particular interest, the diethylaminomethyl group of bedaquiline specifically interacted with the conserved proton binding residue E61, making

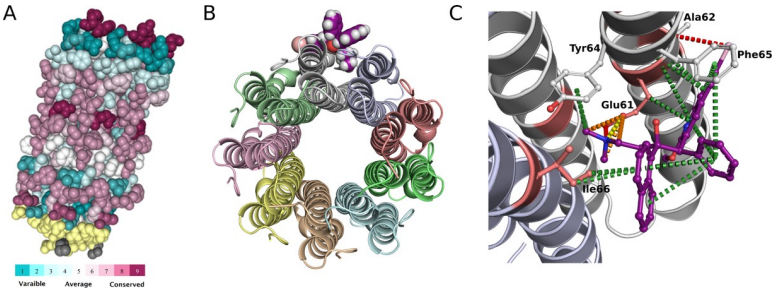


Fig 2. Structure and sequence information. (A) ConSurf analysis of AtpE (*M*. *tuberculosis*) where the evolutionary rates of conservation are color-coded on to the structure. (B) The experimental crystal structure of AtpE bound to Bedaquiline (purple). (C) The key molecular interaction between Bedaquiline (ball and stick representation; purple) and AtpE: ionic bond (yellow), π-interactions (green), proximal hydrogen bond (red) and weak polar van der Waal clashes (orange). The known resistance mutations are shown as salmon red (sticks) on the cartoon representation of the AtpE structure.

https://doi.org/10.1371/journal.pone.0217169.g002

tight ionic and hydrogen bonds with the carboxyl group of E61 (S1 Fig). In the docked model, bedaquiline also made strong π-interactions with residues Y64 and I66, and a hydrogen bond to A62.

Variant calling

We identified 9 previously published bedaquiline resistant non-synonymous single nucleotide variants (nsSNVs) from *in-vitro* selection experiments [4, 13, 24, 26]. To identify AtpE mutations not associated with drug resistance, we examined sequence variation amongst AtpE sequences from 23 mycobacterial species that have been shown to be phenotypically sensitive to the drug [27, 45–49] (Fig 3). Due to the high degree of sequence conservation across mycobacterial AtpE sequences (~ 66% sequence homology; Clustal Omega), variations between strains shown to be susceptible to bedaquiline were inferred to not be associated with drug resistance. Through comparison against the *Mtb* sequence (highlighted in yellow in Fig 3), 54 non-resistance-associated variants were identified (shown in teal in Fig 3).

Understanding the structural basis of resistance is important to facilitate the rapid identification of novel resistance variants, aiding efforts to minimize the rapid development of resistance [23]. The 54 non-resistance-associated variants ("S") and 9 resistant variants ("R") were mapped on the protein structure of AtpE (Fig 1). Most of the non-resistance-associated mutations were located on the N-terminal surface exposed inner loop of AtpE. Conserved regions (highlighted red in Fig 3) were evident, mainly on the C-terminal or the outer loop and embedded in the lipid bilayer of the membrane. All resistance-associated mutations were localized within 5 Å of the known drug binding site, which we refer to as the "resistance hotspot".

Structural and biophysical consequences of AtpE variants

The resistant associated variants were all predicted by SNAP2 [38] to be more functionally deleterious than the non-resistance associated variants, reflecting the resistant associated variants are in a more conserved region of the protein. In order to better understand the molecular consequences of the mutations on AtpE structure and function, the mutations were analyzed in the context of both the apo and complexed protomeric structures. The impact of resistant and non-resistant associated mutations on protein folding, stability and conformation were assessed using SDM [31], mCSM-Stability [32], DUET [33] and DynaMut [34]. The effect of the variants on the affinity of the protomers to form the cylindrical palisade homo-oligomer were examined using mCSM-PPI [32], and the effect of the variants on the binding affinity for bedaquiline were assessed using mCSM-Lig [37].

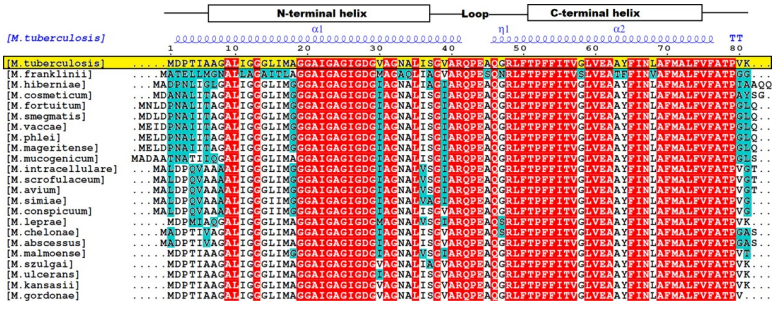


Fig 3. Non-resistant associated variant assignment. This image highlights the sequence alignment of 23 mycobacterial species sensitive to Bedaquiline. Residues that were different to the reference *M.tuberculosis* sequence (in yellow) are highlighted in teal, and were chosen as non-resistant associated variants for building the empirical model. The conserved residues are shown in red. The secondary structure of the AtpE protein is shown above the sequences in blue (α = alpha helix, η = loop). This image was created using ESPript 3 [56].

https://doi.org/10.1371/journal.pone.0217169.g003

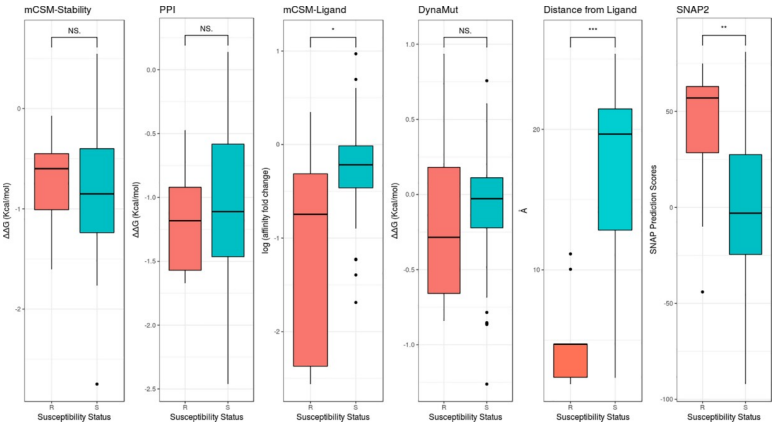


Fig 4. PCA analysis. Boxplot representation of all the features used to build the predictive model. The resistant associated mutations (R) are represented as red and the non-resistant associated mutations (S) as teal. (* $p<0.05$, ** $p<0.005$, *** $p<0.0001$, NS $p>0.5$ by Welch two sample t-test).

https://doi.org/10.1371/journal.pone.0217169.g004

Analysis of the variant effects on protomer stability and the formation of the cylindrical palisade did not reveal statistically significant differences between resistant and non-resistance-associated variants (Fig 4). This is consistent with recent work that showed in order to minimize fitness costs, resistant associated variants in drug targets tended to have mild effects on protein stability [50]. The largest destabilizing effect observed amongst the resistance-associated variants using mCSM-Stability and DUET was for the conservative mutation E61D (ΔΔG = -1.1 Kcal/mol), however normal mode analysis by DynaMut suggested that the E61D mutation would not destabilize the structure and was only associated with mild conformational changes (S1 Fig). Examination of residue conservation across 150 homologous sequences using ConSurf [51] showed the equivalent residue position in many species was an Asp, suggesting its introduction is unlikely to have a large structural or functional effect.

While all nine resistant variants were within 5 Å of the ligands, five in particular, A63M, A63P, E61D, L59V and I66M, were within 2.5 Å and making direct interactions with bedaquiline. Modelling of these mutations revealed that most of them would result in complete loss of these intermolecular interactions (S2 Fig). For example, E61 upon mutation to Asp would result in loss of these strong ionic and hydrogen bonds with bedaquiline. Interestingly, the mutation of I66 to Met and L59 to Val mutation revealed the formation of new interactions, although the overall binding affinity was predicted to be lower by CSM-lig. Most of the non-resistant associated variants were located distal to the bedaquiline binding site.

Analysis of predicted changes in bedaquiline binding affinity upon mutation using mCSM-Lig revealed a significant difference between variants associated with resistance or not associated with resistance (Fig 4). The non-resistance associated variants were associated with mild mCSM-Lig predicted changes in bedaquiline binding affinity (average of -0.25 log affinity fold change). This would be consistent with the mutations leading to minimal change in, or even increasing, drug binding affinity. The average predicted log fold change in binding affinity obtained for the 9 resistant mutations, by contrast, was -1.29 log affinity fold change, indicating that they would likely disrupt bedaquiline binding. Among them, all four D28 resistant variants were predicted to the largest destabilising effect on bedaquiline binding (-2.5 log affinity fold change on average). D28 is positioned on the inner helix of the protomer and is 4.7 Å from the drug binding site. When D28 was substituted with either Ala or Gly, a loss in inter-helical interactions and a gain in flexibility was observed, and when substituted to Pro and Val it led to a gain in intra-molecular interactions and rigidification of the AtpE structures (S2 Fig).

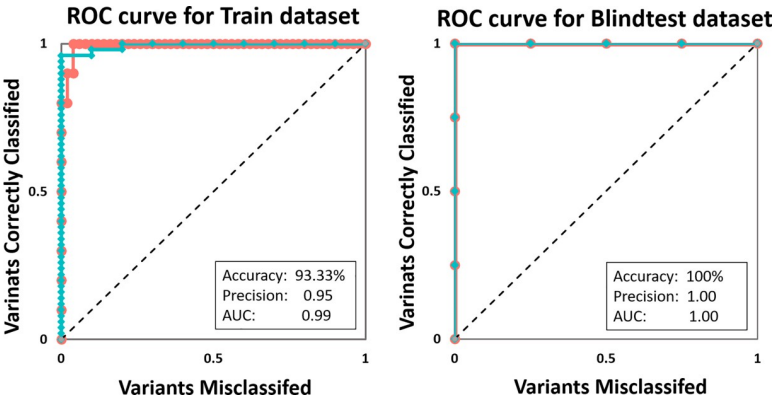


Fig 5. Evaluation metric. The ROC curve shows that using the structural and functional consequences of the variants, we were able to accurately identify resistant (red) and non-resistant associated (teal) variants.

https://doi.org/10.1371/journal.pone.0217169.g005

Machine learning algorithm: Multilayer perceptron network

Building on this structural analysis, we tested whether these structural features could be used to train a supervised machine learning algorithm capable of accurately predicting resistant associated variants. To avoid over-training, the 54 non-resistant and 9 resistant variants were split into a training and blind test dataset. Our training dataset constituted of 50 non-resistant associated variants and 5 resistant associated variants (A63V, A63P, I66M, L59V, E61D). Due to the small sample size, to balance the dataset, the resistant variants in the training dataset were oversampled (duplicated). The remaining 4 resistant (all D28 mutations) and 4 non-resistant associated (I11L, L15T, A34Q and A45S) variants in the blind test were positioned non-redundant with those in the training.

A list of features tested in method development is described in S1 Table. As discussed above, the features that best distinguished between the classes include distance from ligand binding site ("Distance from Ligand", $p < 0.0001$), mCSM-Lig ($p = 0.026$) and SNAP2 ($p < 0.0001$) (Fig 4). Using jackknife and leave-one-residue-position-out validation, models trained using multilayer perceptron neural networks yielded the strongest balanced performance. The final model correctly classified 93.33% and 100% of variants in the training and blind test datasets respectively (Fig 5, Table 1). The comparative performance across iterative non-redundant blind datasets suggested that the model was not over-fitted.

The classifier revealed that variants with mild effects on protein stability and conformation (DynaMut < 0.28 Kcal/mol and DUET < -1.65 Kcal/mol), located close to the docked bedaquiline (distance from ligand < 6.36 Å) were likely to be associated with resistance. A closer examination of the four incorrectly classified non-resistant associated variant in the train dataset revealed that three of them, G58S, A63T and L68V, were positioned very close to the bedaquiline binding site (< 2.5 Å) and N33A had a large predicted change in binding affinity (-1.4 log affinity fold change); indicating that these mutations might have direct consequences on bedaquiline binding.

Table 1. Evaluation metrics of the train and blind test dataset.

Multilayer Perceptron (MLP)	Precision score	Recall	F-measure	ROC area	PRC area
Train Dataset	0.952	0.933	0.938	0.970	0.967
Blind test Dataset	1.000	1.000	1.000	1.000	1.000

https://doi.org/10.1371/journal.pone.0217169.t001

Clinically identified resistance associated variants

Using a model trained without the D28 variants, we analyzed the recently reported clinical *atpE* bedaquiline resistant variants [25]. Both D28N and A63V were both predicted by the model to lead to bedaquiline resistance, consistent with the clinical data. Looking at these mutations within the structure, the mutation at D28 would disrupt interactions made by the wild-type residue to bedaquiline, consistent with the mCSM-Lig predictions that it would lead to a significant reduction in ligand binding affinity (S3 Fig; -1.87 log affinity fold change). Interestingly, while A63 did not make interactions directly with bedaquiline, the mutation to Val would lead to steric clashes with the bound ligand and prevent bedaquiline binding (S3 Fig).

Vietnam data analysis

We also used this approach to predict the sensitivity of two *atpE* nsSNVs, I16V and P52L, identified through whole genomic sequencing of *Mtb* strains isolated from 1635 TB patients in Vietnam [15]. The predictive tool classified the reported nsSNPs to be non-resistant associated variants. These variants were located approximately 10 Å away from the bedaquiline binding site, and mutations at these residues were not predicted to disrupt any interactions with bedaquiline (S4 Fig). As these samples had been collected from patients that had not been administered bedaquiline, it provided confidence that in our large analysis of patients in Vietnam there were no circulating strains likely to be resistant to bedaquiline.

SUSPECT-BDQ webserver

We have implemented SUSPECT-BDQ as a user-friendly, freely available web server http://biosig.unimelb.edu.au/suspect_bdq/. SUSPECT-BDQ provides two different input options. The "Single Mutation" option allows users to predict whether a mutation will be characterized as either Resistant or Susceptible. For this option, the server requires the point mutation to be specified as a text string containing the wild-type residue one-letter code, its corresponding position on the structure and the mutant one-letter code. The "Mutation List" option allows the user to upload a file with a list of mutations in a file for batch processing. In order to assist users to submit their mutations for analysis, sample submission entries are available for both input options and a help page is also available via the top navigation bar.

For the "Single Mutation" option, the web server displays the prediction outcome of SUSPECT-BDQ alongside with details of the user input data, information on the residue environment and parameters used on the prediction (S5 Fig). In addition, an interactive 3D viewer, built using NGL [52] allows for analysis of non-covalent inter-residue interactions for the position specified in the input calculated with Arpeggio [44] for wild-type and mutant structures. For the "Mutation List" option, the results are summarized in a downloadable table from which users can access details for each single mutation. A 3D viewer is also shown and each wild-type residue from the input list is colored according to the predicted effect.

Discussion

Early genomic detection of resistance is crucial for tailoring individual therapy and preventing the onward transmission of resistant infection. This is especially of importance to limit the spread of resistance to bedaquiline, one of the few treatment options for XDR-TB. While significant progress has been made in terms of innovative tools to understand and quantify the different range of effects in which a mutation or a set of mutations can give rise to a drug-resistant phenotype, a gap still exists when integrating these predictions and drawing conclusions

regarding causality and the strength of associations observed. This is compounded by the need for detailed information regarding the system/protein. The availability of scalable, effective computational methods to assess mutational effects creates new opportunities for developing integrated approaches and deciphering complex genomic background patterns, shedding light on their role in the emergence of a given phenotype and molecular mechanisms of action [19].

Here we have used a computational approach to better understand the molecular mechanism of drug resistance within the context of the protein's 3D structure. A machine learning algorithm was used to build a predictive tool which could pre-emptively determine novel bedaquiline resistant mutations within *atpE*. We began our investigation by studying the interaction dynamics between the c-ring of ATP synthase bound to bedaquiline. The correlations of conformational changes and Gibb's free energy provided novel molecular insights into how resistance variants affected bedaquiline binding but led to minimal disruption of protein folding and dynamics. Mapping of all the mutations on the crystal structure helped us identify the "mutational hotspot" for AtpE, which was in proximity to the drug binding site. We saw that resistance associated variants were more likely to be located within this resistance hotspot, and lead to a significant disruption in bedaquiline binding. Interestingly, the characterized resistant variants did not lead to large changes in protein folding, stability or oligomeric state, which would impose a larger fitness penalty [50].

This *in silico* biophysical information was used to build a predictive algorithm that accurately identified resistant mutations. We then prepared a comprehensive mutational dataset that contained the predictions of all possible mutations in AtpE, which we have made available through a web-based interface: SUSPECT_BDQ (http://biosig.unimelb.edu.au/suspect_bdq/). These analyses highlight the power of considering the structural environment of a mutation to understand the molecular and biological consequences [53]. As a relatively novel drug, there is still a paucity of reliable information regarding resistance mutations. While limited by the relatively small available datasets, repeated stratified non-redundant blind testing revealed the model was very robust. This associative approach thus helped us establish a set of guidelines which adds to the missing information in the database for new TB drugs like bedaquiline. It also provides a molecular understanding of how variants in AtpE affect ligand binding, leading to resistance, providing insight to guide development of second-generation inhibitors.

We intend further development of this tool through expanded genomic targets, and evaluation using additional clinical isolates. In particular we intend to extend SUSPECT_BDQ to include non-target based resistance to bedaquiline, which has been linked to mutations in *Rv0678* [54], a transcriptional repressor of the gene encoding the MmpS5-MmpL5 efflux pump, and *pepQ* (*Rv2535c*) [55], a putative Xaa-Pro aminopeptidase. Both are associated with low-level of resistance and therefore we did not include them in the study. However, low level resistance may have clinical significance in some settings, and future work will further evaluate other potentially important loci. Additionally, testing this tool on further clinical isolates will enhance the efficiency of the tool to predict the consequences of novel mutations.

Conclusion

This novel computational approach can enhance the impact of genome sequencing in identifying and characterizing variants more accurately and may therefore assist in guiding optimal usage of bedaquiline. The results obtained from our empirical tool is promising and should help facilitate routine genotypic drug susceptibility testing for bedaquiline and stimulate further research to help avoid the emergence of resistance to this new treatment through early detection.

Supporting information

S1 Table. The list of different features used to build the empirical model for predicting novel resistance associated mutations in bedaquiline.
(PDF)

S1 Fig. Detailed molecular interactions between the key proton binding residue E61, and upon its mutation to Asp, with bedaquiline. The wild-type residue is shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation). Hydrogen bonds are shown as orange dashes and ionic bond in yellow.
(TIF)

S2 Fig. Images of intermolecular interactions made by the wild-type residue (shown as cyan) and the mutant amino acid (shown as salmon red). Hydrogen bonds are shown in red, halogen bonds in blue, ionic bonds in yellow, hydrophobic bonds in green, π bonds in grey.
(TIF)

S3 Fig. Detailed molecular interactions between two clinically observed bedaquiline resistant variants, with the drug. The wild type residue is shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation). Halogen bonds are represented in blue dashes (amide-amide interaction) and π-bond as grey dashes.
(TIF)

S4 Fig. The localization of two circulating *atpE* variants relative to the bedaquiline binding pocket. The wild type residues are shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation).
(TIF)

S5 Fig. SUSPECT-BDQ webserver. Web-server results page for a single point mutation prediction. The predicted outcome is shown alongside with complementary information on the submitted mutation. An interactive 3D viewer allows for analysis of non-covalent interactions for both the wild type and mutant residue. In both cases controllers are provided in order to hide or show specific interactions and customize molecule representation.
(TIF)

Author Contributions

Conceptualization: Justin Denholm, David B. Ascher.

Data curation: Malancha Karmakar, Kathryn E. Holt, Sarah J. Dunstan, David B. Ascher.

Formal analysis: Malancha Karmakar, Kathryn E. Holt, Sarah J. Dunstan, Justin Denholm, David B. Ascher.

Funding acquisition: David B. Ascher.

Investigation: Malancha Karmakar, David B. Ascher.

Methodology: Malancha Karmakar, David B. Ascher.

Project administration: David B. Ascher.

Resources: Kathryn E. Holt.

Software: Carlos H. M. Rodrigues.

Supervision: David B. Ascher.

Validation: Malancha Karmakar, Kathryn E. Holt, Sarah J. Dunstan, Justin Denholm, David B. Ascher.

Writing – original draft: Malancha Karmakar, Carlos H. M. Rodrigues.

Writing – review & editing: Kathryn E. Holt, Sarah J. Dunstan, Justin Denholm, David B. Ascher.

References

1. WHO. Global Tuberculosis Report: Executive Summary. 2018; WHO/CDS/TB/2018.25.
2. Hards K, Robson JR, Berney M, Shaw L, Bald D, Koul A, et al. Bactericidal mode of action of bedaquiline. Journal of Antimicrobial Chemotherapy. 2015; 70(7):2028–37. https://doi.org/10.1093/jac/dkv054 PMID: 25754998
3. Koul A, Dendouga N, Vergauwen K, Molenberghs B, Vranckx L, Willebrords R, et al. Diarylquinolines target subunit c of mycobacterial ATP synthase. Nat Chem Biol. 2007; 3(6):323–4. Epub 2007/05/15. https://doi.org/10.1038/nchembio884 PMID: 17496888.
4. Petrella S, Cambau E, Chauffour A, Andries K, Jarlier V, Sougakoff W. Genetic basis for natural and acquired resistance to the diarylquinoline R207910 in mycobacteria. Antimicrob Agents Chemother. 2006; 50(8):2853–6. Epub 2006/07/28. https://doi.org/10.1128/AAC.00244-06 PMID: 16870785; PubMed Central PMCID: PMCPMC1538646.
5. Field SK. Bedaquiline for the treatment of multidrug-resistant tuberculosis: great promise or disappointment? Therapeutic Advances in Chronic Disease. 2015; 6(4):170–84. https://doi.org/10.1177/2040622315582325 PMC4480545. PMID: 26137207
6. WHO. Rapid Communication: Key changes to treatment of multidrug- and rifampicin-resistant tuberculosis (MDR/RR-TB). 2018. http://www.who.int/tb/publications/2018/rapid_communications_MDR/en/.
7. Salfinger M, Migliori GB. Bedaquiline: 10 years later, the drug susceptibility testing protocol is still pending. The European respiratory journal. 2015; 45(2):317–21. Epub 2015/02/06. https://doi.org/10.1183/09031936.00199814 PMID: 25653264.
8. Hoffmann H, Kohl TA, Hofmann-Thiel S, Merker M, Beckert P, Jaton K, et al. Delamanid and Bedaquiline Resistance in Mycobacterium tuberculosis Ancestral Beijing Genotype Causing Extensively Drug-Resistant Tuberculosis in a Tibetan Refugee. American journal of respiratory and critical care medicine. 2016; 193(3):337–40. Epub 2016/02/02. https://doi.org/10.1164/rccm.201502-0372LE PMID: 26829425.
9. Hoffmann H, Hofmann-Thiel S, Merker M, Kohl TA, Niemann S. Reply: Call for Regular Susceptibility Testing of Bedaquiline and Delamanid. American journal of respiratory and critical care medicine. 2016; 194(9):1171–2. Epub 2016/11/01. https://doi.org/10.1164/rccm.201605-1065LE PMID: 27797620.
10. WHO. The Use of Bedaquiline in the Treatment of Multidrug-Resistant Tuberculosis. 2013.
11. Coll F, McNerney R, Preston MD, Guerra-Assuncao JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome medicine. 2015; 7(1):51. Epub 2015/05/29. https://doi.org/10.1186/s13073-015-0164-0 PMID: 26019726; PubMed Central PMCID: PMCPMC4446134.
12. Nguyen TVA, Anthony RM, Banuls AL, Nguyen TVA, Vu DH, Alffenaar JC. Bedaquiline Resistance: Its Emergence, Mechanism, and Prevention. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. 2018; 66(10):1625–30. Epub 2017/11/11. https://doi.org/10.1093/cid/cix992 PMID: 29126225.
13. Segala E, Sougakoff W, Nevejans-Chauffour A, Jarlier V, Petrella S. New mutations in the mycobacterial ATP synthase: new insights into the binding of the diarylquinoline TMC207 to the ATP synthase C-ring structure. Antimicrob Agents Chemother. 2012; 56(5):2326–34. Epub 2012/02/23. https://doi.org/10.1128/AAC.06154-11 PMID: 22354303; PubMed Central PMCID: PMCPMC3346594.
14. Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, et al. Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom. 2018. Epub 2018/03/17. https://doi.org/10.1099/mgen.0.000165 PMID: 29547094; PubMed Central PMCID: PMCPMC5885017.
15. 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. Nature genetics. 2018; 50(6):849–56. Epub 2018/05/23. https://doi.org/10.1038/s41588-018-0117-9 PMID: 29785015.

16. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC medicine. 2016; 14:31. Epub 2016/03/24. https://doi.org/10.1186/s12916-016-0575-9 PMID: 27005572; PubMed Central PMCID: PMCPMC4804620.
17. Pires DE, Chen J, Blundell TL, Ascher DB. In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Scientific reports. 2016; 6:19848. Epub 2016/01/23. https://doi.org/10.1038/srep19848 PMID: 26797105; PubMed Central PMCID: PMCPMC4726175.
18. Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, et al. Structural Implications of Mutations Conferring Rifampin Resistance in Mycobacterium leprae. Scientific reports. 2018; 8(1):5016. Epub 2018/03/24. https://doi.org/10.1038/s41598-018-23423-1 PMID: 29567948; PubMed Central PMCID: PMCPMC5864748.
19. Albanaz ATS, Rodrigues CHM, Pires DEV, Ascher DB. Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert opinion on drug discovery. 2017; 12(6):553–63. Epub 2017/05/12. https://doi.org/10.1080/17460441.2017.1322579 PMID: 28490289.
20. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, et al. Essential but Not Vulnerable: Indazole Sulfonamides Targeting Inosine Monophosphate Dehydrogenase as Potential Leads against Mycobacterium tuberculosis. ACS infectious diseases. 2017; 3(1):18–33. Epub 2016/10/06. https://doi.org/10.1021/acsinfecdis.6b00103 PMID: 27704782; PubMed Central PMCID: PMCPMC5972394.
21. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, et al. The Inosine Monophosphate Dehydrogenase, GuaB2, Is a Vulnerable New Bactericidal Drug Target for Tuberculosis. ACS infectious diseases. 2017; 3(1):5–17. Epub 2016/10/12. https://doi.org/10.1021/acsinfecdis.6b00102 PMID: 27726334; PubMed Central PMCID: PMCPMC5241705.
22. Trapero A, Pacitto A, Singh V, Sabbah M, Coyne AG, Mizrahi V, et al. Fragment-Based Approach to Targeting Inosine-5'-monophosphate Dehydrogenase (IMPDH) from Mycobacterium tuberculosis. Journal of medicinal chemistry. 2018; 61(7):2806–22. Epub 2018/03/17. https://doi.org/10.1021/acs.jmedchem.7b01622 PMID: 29547284; PubMed Central PMCID: PMCPMC5900554.
23. Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, et al. Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy. American journal of respiratory and critical care medicine. 2018; 198(4):541–4. Epub 2018/04/26. https://doi.org/10.1164/rccm.201712-2572LE PMID: 29694240; PubMed Central PMCID: PMCPMC6118032.
24. Huitric E, Verhasselt P, Koul A, Andries K, Hoffner S, Andersson DI. Rates and mechanisms of resistance development in Mycobacterium tuberculosis to a novel diarylquinoline ATP synthase inhibitor. Antimicrob Agents Chemother. 2010; 54(3):1022–8. Epub 2009/12/30. https://doi.org/10.1128/AAC.01611-09 PMID: 20038615; PubMed Central PMCID: PMCPMC2825986.
25. Zimenkov DV, Nosova EY, Kulagina EV, Antonova OV, Arslanbaeva LR, Isakova AI, et al. Examination of bedaquiline- and linezolid-resistant Mycobacterium tuberculosis isolates from the Moscow region. The Journal of antimicrobial chemotherapy. 2017; 72(7):1901–6. Epub 2017/04/08. https://doi.org/10.1093/jac/dkx094 PMID: 28387862.
26. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, et al. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science (New York, NY). 2005; 307(5707):223–7. Epub 2004/12/14. https://doi.org/10.1126/science.1106753 PMID: 15591164.
27. Huitric E, Verhasselt P, Andries K, Hoffner SE. In vitro antimycobacterial spectrum of a diarylquinoline ATP synthase inhibitor. Antimicrob Agents Chemother. 2007; 51(11):4202–4. Epub 2007/08/22. https://doi.org/10.1128/AAC.00181-07 PMID: 17709466; PubMed Central PMCID: PMCPMC2151410.
28. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011; 7:539. Epub 2011/10/13. https://doi.org/10.1038/msb.2011.75 PMID: 21988835; PubMed Central PMCID: PMCPMC3261699.
29. Thai PVK, Ha DTM, Hanh NT, Day J, Dunstan S, Nhu NTQ, et al. Bacterial risk factors for treatment failure and relapse among patients with isoniazid resistant tuberculosis. BMC Infectious Diseases. 2018; 18(1):112. https://doi.org/10.1186/s12879-018-3033-9 PMID: 29510687
30. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology. 1993; 234(3):779–815. Epub 1993/12/05. https://doi.org/10.1006/jmbi.1993.1626 PMID: 8254673.

31. Worth CL, Preissner R, Blundell TL. SDM—a server for predicting effects of mutations on protein stability and malfunction. Nucleic acids research. 2011; 39(Web Server issue):W215–22. Epub 2011/05/20. https://doi.org/10.1093/nar/gkr363 PMID: 21593128; PubMed Central PMCID: PMCPMC3125769.

32. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics (Oxford, England). 2014; 30(3):335–42. Epub 2013/11/28. https://doi.org/10.1093/bioinformatics/btt691 PMID: 24281696; PubMed Central PMCID: PMCPMC3904523.

33. Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic acids research. 2014; 42(Web Server issue): W314–9. Epub 2014/05/16. https://doi.org/10.1093/nar/gku411 PMID: 24829462; PubMed Central PMCID: PMCPMC4086143.

34. Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic acids research. 2018; 46(W1):W350–w5. Epub 2018/05/03. https://doi.org/10.1093/nar/gky300 PMID: 29718330; PubMed Central PMCID: PMCPMC6031064.

35. Pires DE, Ascher DB. CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic acids research. 2016; 44(W1):W557–61. Epub 2016/05/07. https://doi.org/10.1093/nar/gkw390 PMID: 27151202; PubMed Central PMCID: PMCPMC4987933.

36. Pires DE, Blundell TL, Ascher DB. Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic acids research. 2015; 43(Database issue):D387–91. Epub 2014/10/18. https://doi.org/10.1093/nar/gku966 PMID: 25324307; PubMed Central PMCID: PMCPMC4384026.

37. Pires DE, Blundell TL, Ascher DB. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Scientific reports. 2016; 6:29575. Epub 2016/07/08. https://doi.org/10.1038/srep29575 PMID: 27384129; PubMed Central PMCID: PMCPMC4935856.

38. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC genomics. 2015; 16 Suppl 8:S1. Epub 2015/06/26. https://doi.org/10.1186/1471-2164-16-s8-s1 PMID: 26110438; PubMed Central PMCID: PMCPMC4480835.

39. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009; 11(1):10–8. https://doi.org/10.1145/1656274.1656278

40. Provost F. Machine learning from imbalanced data sets 101. Proceedings of the AAAI-2000 Workshop on Imbalanced Data Sets. 2000. citeulike-article-id:7616988.

41. Wager S, Hastie T, Efron B. Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. Journal of machine learning research: JMLR. 2014; 15(1):1625–51. Epub 2015/01/13. PMID: 25580094; PubMed Central PMCID: PMCPMC4286302.

42. Preiss L, Langer JD, Yildiz O, Eckhardt-Strelau L, Guillemont JE, Koul A, et al. Structure of the mycobacterial ATP synthase Fo rotor ring in complex with the anti-TB drug bedaquiline. Science advances. 2015; 1(4):e1500106. Epub 2015/11/26. https://doi.org/10.1126/sciadv.1500106 PMID: 26601184; PubMed Central PMCID: PMCPMC4640650.

43. Lu P, Lill H, Bald D. ATP synthase in mycobacteria: special features and implications for a function as drug target. Biochim Biophys Acta. 2014; 1837(7):1208–18. Epub 2014/02/12. https://doi.org/10.1016/j.bbabio.2014.01.022 PMID: 24513197.

44. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. Journal of molecular biology. 2017; 429(3):365–71. Epub 2016/12/15. https://doi.org/10.1016/j.jmb.2016.12.004 PMID: 27964945; PubMed Central PMCID: PMCPMC5282402.

45. Aguilar-Ayala DA, Cnockaert M, Andre E, Andries K, Gonzalez YMJA, Vandamme P, et al. In vitro activity of bedaquiline against rapidly growing nontuberculous mycobacteria. Journal of medical microbiology. 2017; 66(8):1140–3. Epub 2017/07/28. https://doi.org/10.1099/jmm.0.000537 PMID: 28749330; PubMed Central PMCID: PMCPMC5817190.

46. Chahine EB, Karaoui LR, Mansour H. Bedaquiline: a novel diarylquinoline for multidrug-resistant tuberculosis. The Annals of pharmacotherapy. 2014; 48(1):107–15. Epub 2013/11/22. https://doi.org/10.1177/1060028013504087 PMID: 24259600.

47. Ji B, Chauffour A, Andries K, Jarlier V. Bactericidal activities of R207910 and other newer antimicrobial agents against Mycobacterium leprae in mice. Antimicrob Agents Chemother. 2006; 50(4):1558–60. Epub 2006/03/30. https://doi.org/10.1128/AAC.50.4.1558-1560.2006 PMID: 16569884; PubMed Central PMCID: PMCPMC1426933.

48. Ji B, Lefrancois S, Robert J, Chauffour A, Truffot C, Jarlier V. In vitro and in vivo activities of rifampin, streptomycin, amikacin, moxifloxacin, R207910, linezolid, and PA-824 against Mycobacterium ulcerans. Antimicrob Agents Chemother. 2006; 50(6):1921–6. Epub 2006/05/26. https://doi.org/10.1128/AAC.00052-06 PMID: 16723546; PubMed Central PMCID: PMCPMC1479135.

49. Pang Y, Zheng H, Tan Y, Song Y, Zhao Y. In Vitro Activity of Bedaquiline against Nontuberculous Mycobacteria in China. Antimicrob Agents Chemother. 2017; 61(5). Epub 2017/03/01. https://doi.org/10.1128/aac.02627-16 PMID: 28242674; PubMed Central PMCID: PMCPMC5404590.

50. Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N. Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Scientific reports. 2018; 8(1):15356–. https://doi.org/10.1038/s41598-018-33370-6 PMID: 30337649.

51. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic acids research. 2016; 44(W1):W344–50. Epub 2016/05/12. https://doi.org/10.1093/nar/gkw408 PMID: 27166375; PubMed Central PMCID: PMCPMC4987940.

52. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlic A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics (Oxford, England). 2018; 34(21):3755–8. Epub 2018/06/01. https://doi.org/10.1093/bioinformatics/bty419 PMID: 29850778; PubMed Central PMCID: PMCPMC6198858.

53. Pandurangan AP, Ascher DB, Thomas SE, Blundell TL. Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochemical Society transactions. 2017; 45(2):303–11. Epub 2017/04/15. https://doi.org/10.1042/BST20160422 PMID: 28408471; PubMed Central PMCID: PMCPMC5390495.

54. Bloemberg GV, Gagneux S, Böttger EC, Keller PM, Stuckia D, Trauner A, et al. Acquired Resistance to Bedaquiline and Delamanid in Therapy for Tuberculosis. New England Journal of Medicine. 2015; 373(20):1986–8. https://doi.org/10.1056/NEJMc1505196 PMID: 26559594. Language: English. Entry Date: 20151121. Revision Date: 20161125. Publication Type: case study. Journal Subset: Biomedical.

55. Almeida D, Ioerger T, Tyagi S, Li SY, Mdluli K, Andries K, et al. Mutations in pepQ Confer Low-Level Resistance to Bedaquiline and Clofazimine in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2016; 60(8):4590–9. Epub 2016/05/18. https://doi.org/10.1128/AAC.00753-16 PMID: 27185800; PubMed Central PMCID: PMCPMC4958187.

56. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic acids research. 2014; 42(Web Server issue):W320–4. Epub 2014/04/23. https://doi.org/10.1093/nar/gku316 PMID: 24753421; PubMed Central PMCID: PMCPMC4086106.

Supplementary Information

Empirical ways to identify novel Bedaquiline resistance mutations

Malancha Karmakar[1,2,3], Kathryn E. Holt[2], Sarah J. Dunstan[4], Justin Denholm[1,3], David B. Ascher[2,5]

[1]Victorian Tuberculosis Program, Melbourne Health, Victoria, Australia

[2]Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria 3010, Australia

[3]Department of Microbiology and Immunology, University of Melbourne, Melbourne, Victoria, Australia

[4]The Peter Doherty Institute for Infection and Immunity, University of Melbourne, Victoria, Australia

[5]Department of Biochemistry, University of Cambridge, CB2 1GA, UK

Correspondence should be addressed to D.B.A: Tel: +61 3 90354794: david.ascher@unimelb.edu.au

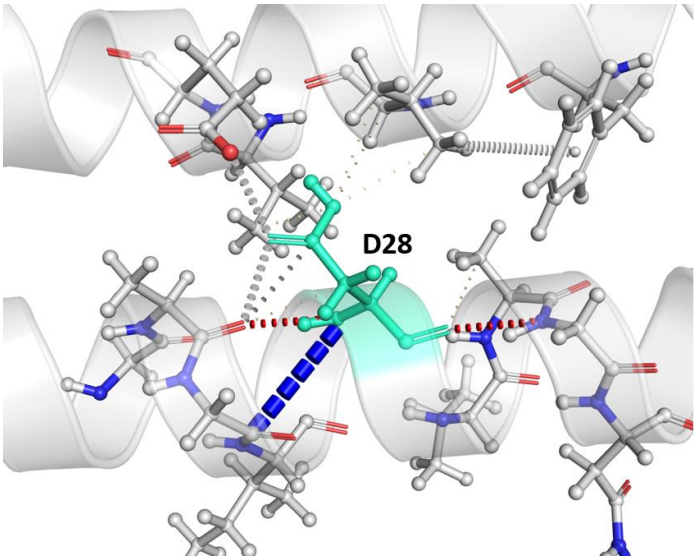
Table S1: The list of different features used to build the empirical model for predicting novel resistance associated mutations in bedaquiline.

Feature	**Effect measured**	**Technique**	**p-value***
mCSM_Stability	Protein stability	Graph based signatures	0.31
SDM	Protein stability	Graph based signatures	0.90
DUET	Protein stability	Graph based signatures	0.65
mCSM_PPI	Protein-protein interaction	Graph based signatures	0.68
DynaMut	Conformational flexibility	Normal mode analysis	0.60
ΔΔG ENCoM	Conformational flexibility	Normal mode analysis	0.66
ΔΔS ENCoM	Changes in Entropy	Normal mode analysis	0.66
mCSM_Lig	Ligand binding affinity	Graph based signatures	0.03
Distance from ligand binding site	Distance of the mutation from the drug (bedaquiline) binding site	Perl script (in-house)	< 2.2e-16
SNAP2	Effect of single nucleotide substitution	Neural Network	0.0002

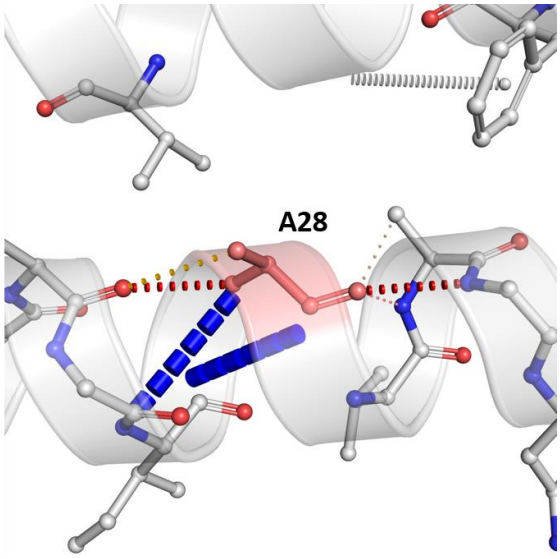
*(Welch two sample t-test)

Figure S1. Images of intermolecular interactions made by the wild-type residue (shown as cyan) and the mutant amino acid (shown as salmon red). Hydrogen bonds are shown in red, halogen bonds in blue, ionic bonds in yellow, hydrophobic bonds in green, π bonds in grey.

1) D28A mutation

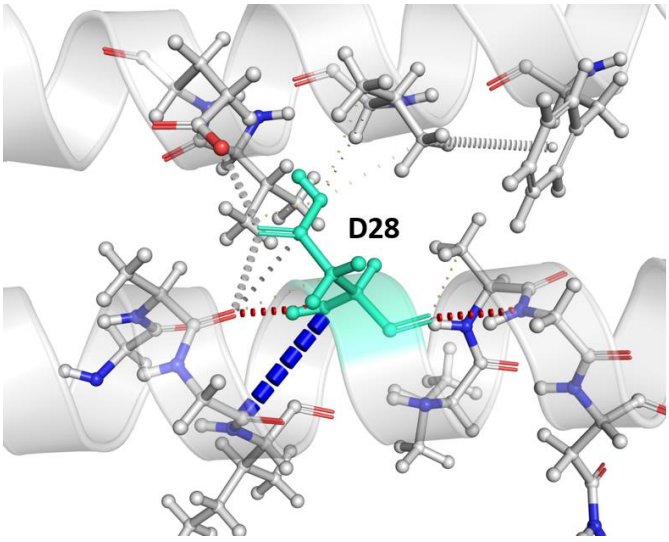


Wild-type D28 interactions

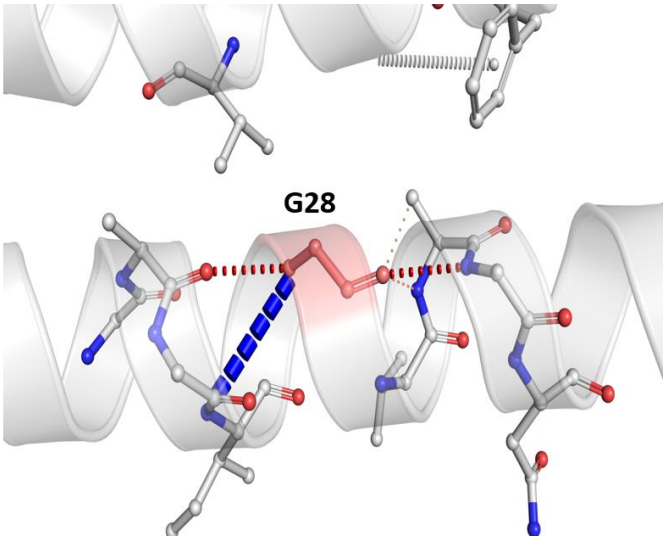


Mutant A28 interactions

2) D28G mutation

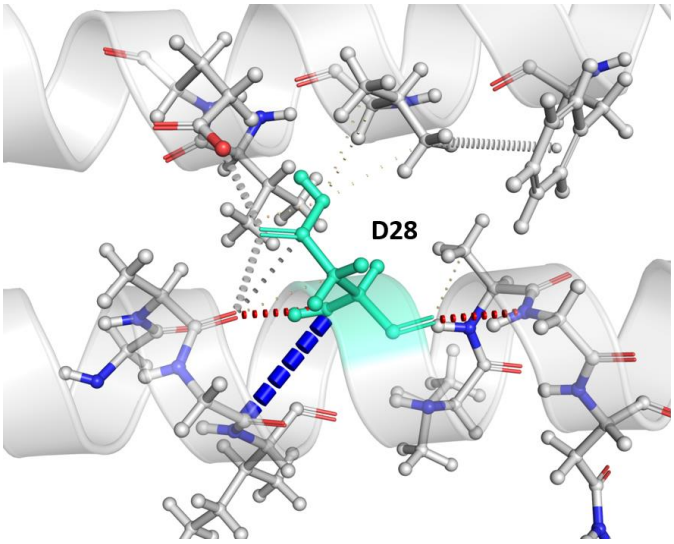


Wild-type D28 interactions

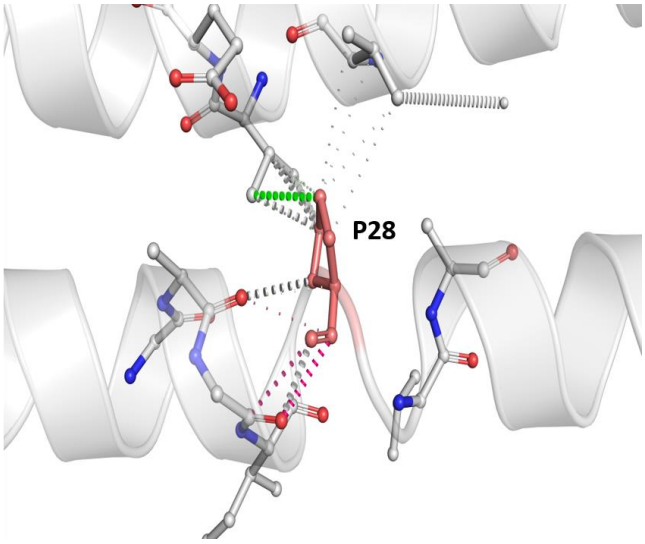


Mutant G28 interactions

3) D28P mutation

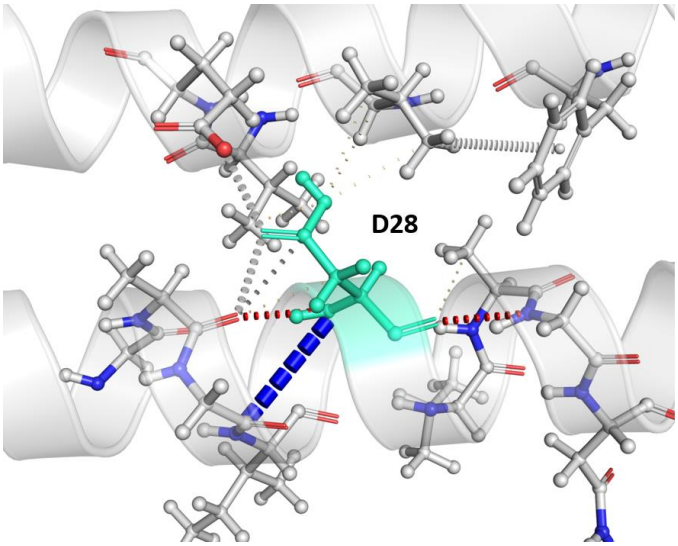


Wild-type D28 interactions

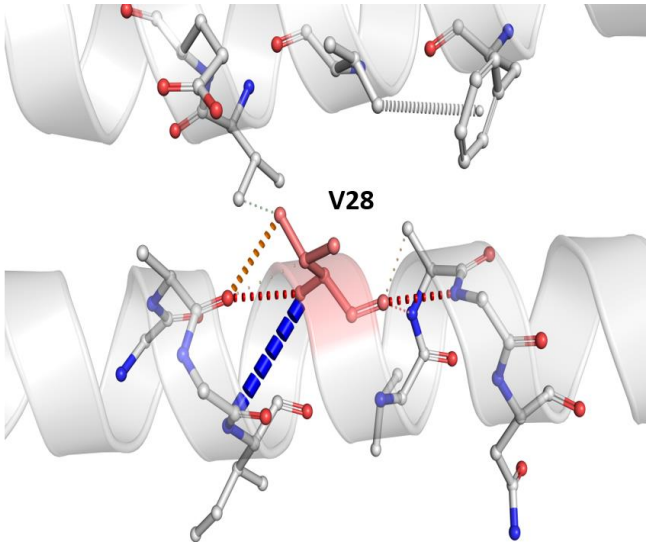


Mutant P28 interactions

5) D28V mutation

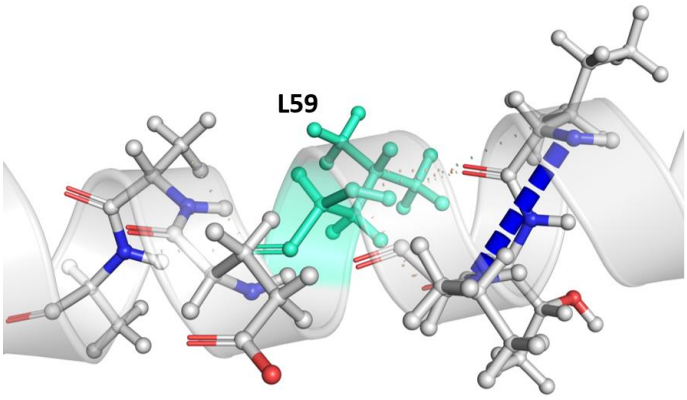


Wild-type D28 interactions

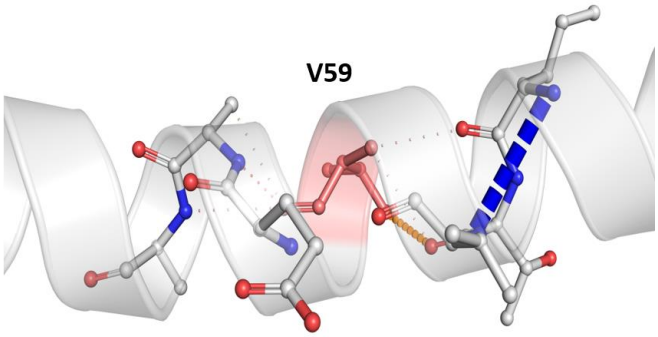


Mutant V28 interactions

5) L59V mutation

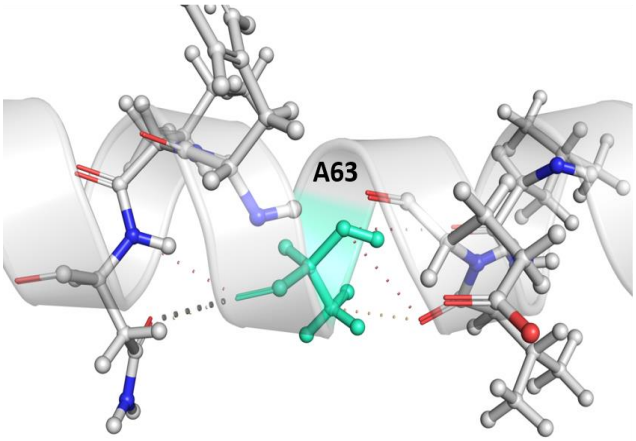


Wild-type L59 interactions

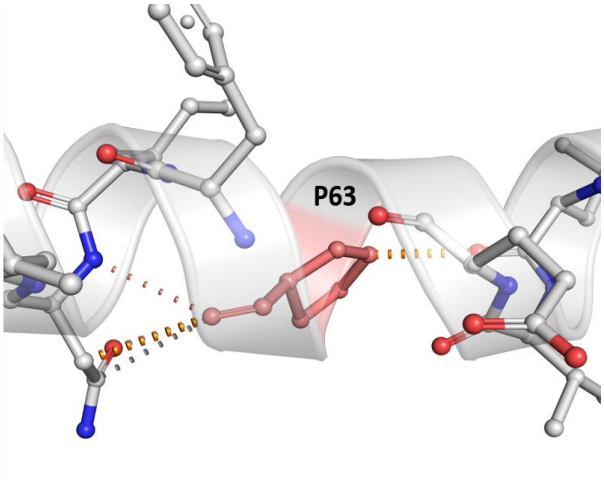


Mutant V59 interactions

6) A63P mutation

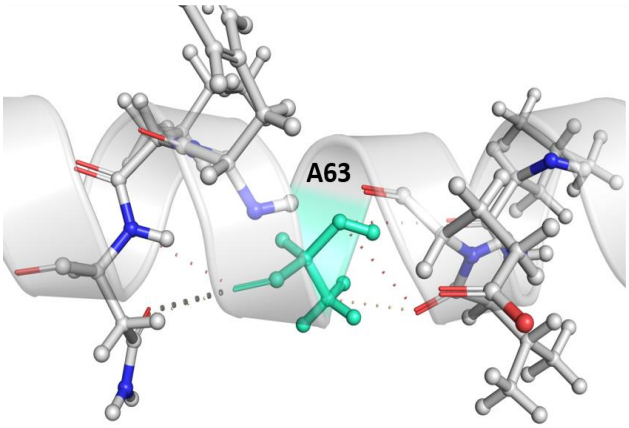


Wild-type A63 interactions

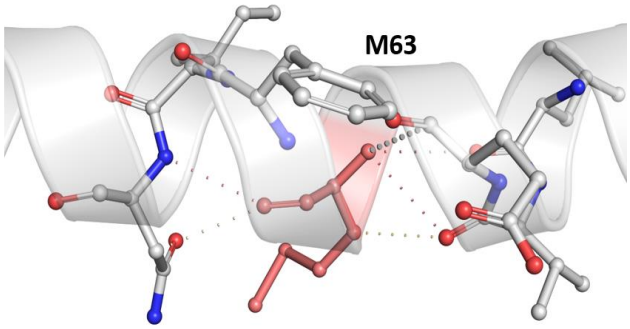


Mutant P63 interactions

7) A63M mutation

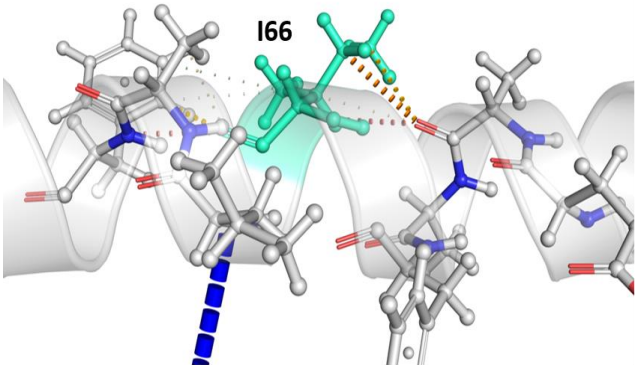


Wild-type A63 interactions

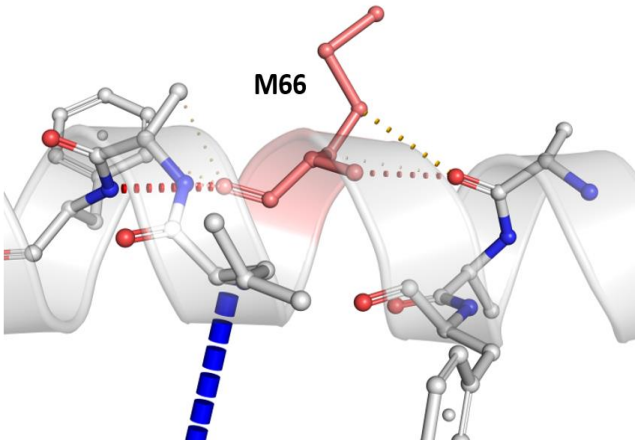


Mutant M63 interactions

8) I66M mutation



Wild-type I66 interactions



Mutant M66 interactions

Figure S2. Detailed molecular interactions between the key proton binding residue E61, and upon its mutation to Asp, with bedaquiline. The wild-type residue is shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation). Hydrogen bonds are shown as orange dashes and ionic bond in yellow

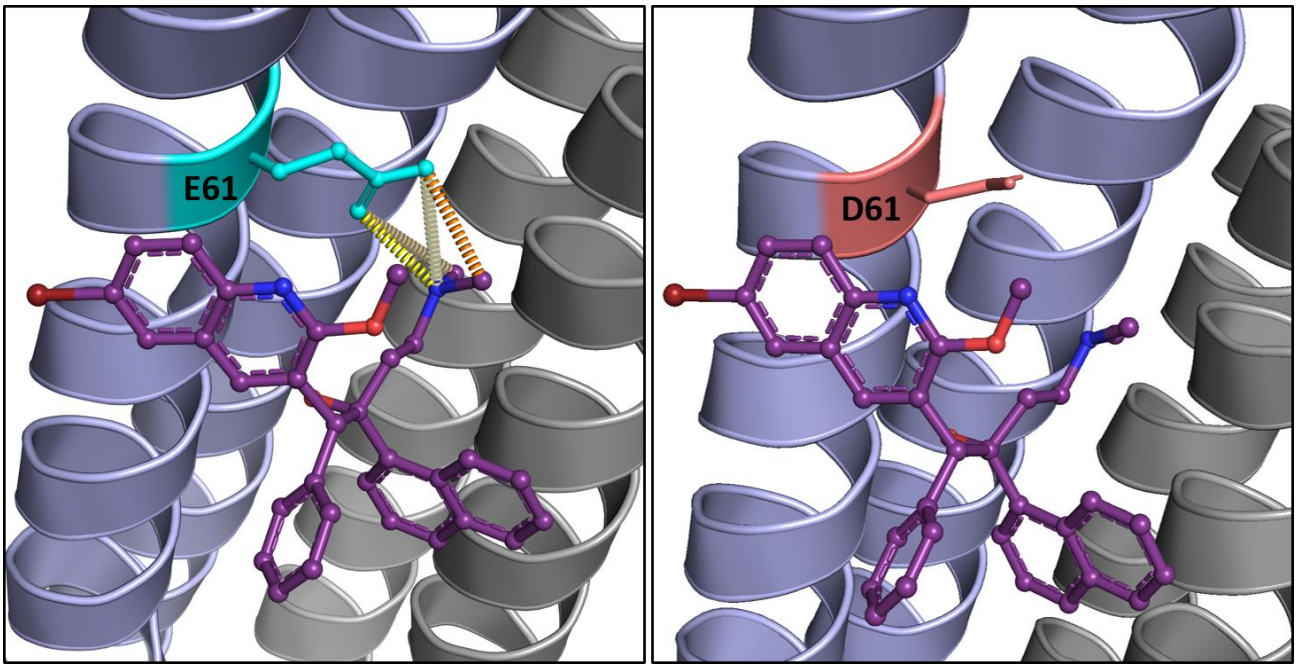


Figure S3. Detailed molecular interactions between two clinically observed bedaquiline resistant variants, with the drug. The wild type residue is shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation). Halogen bonds are represented in blue dashes (amide-amide interaction) and π-bond as grey dashes.

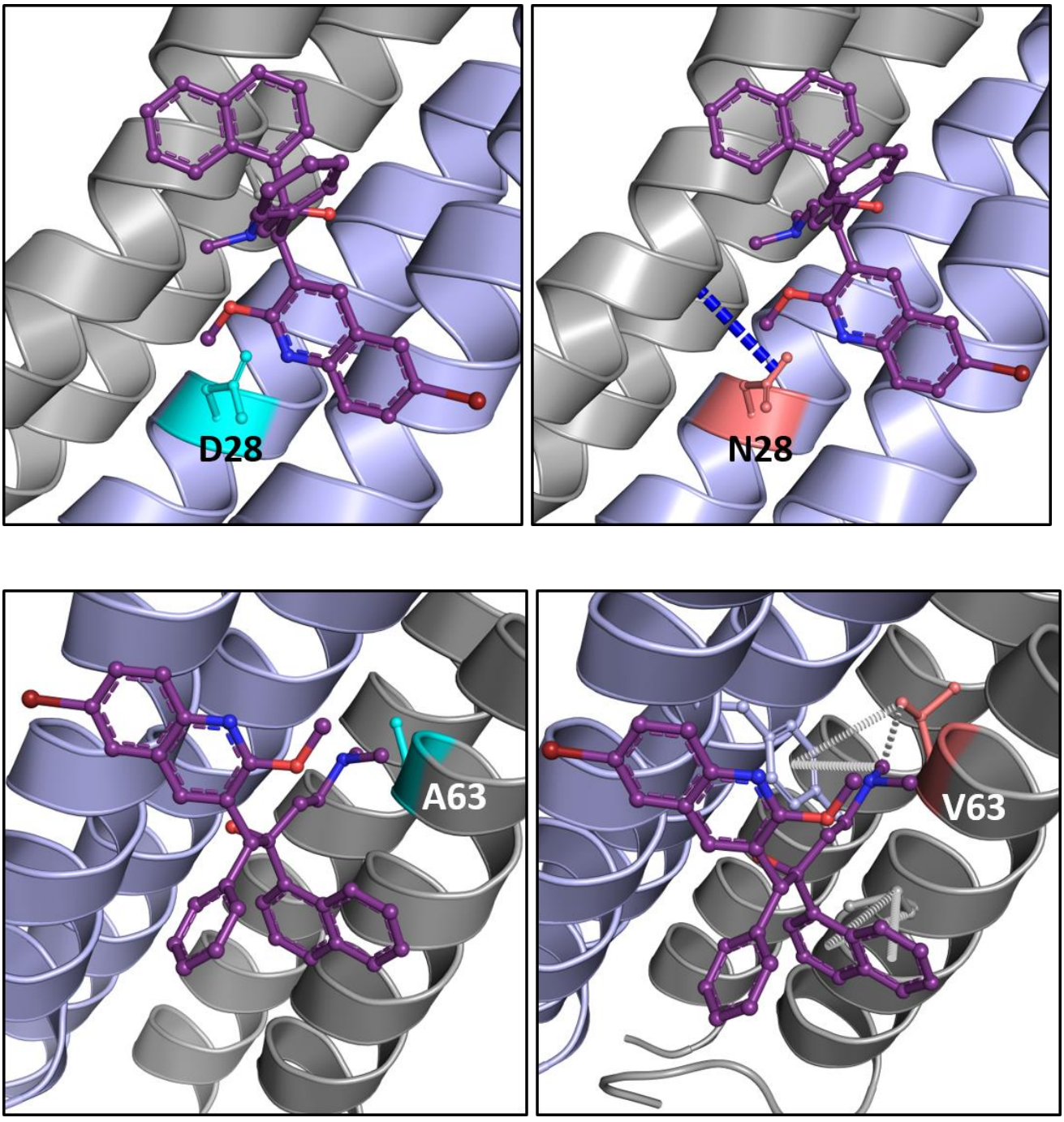
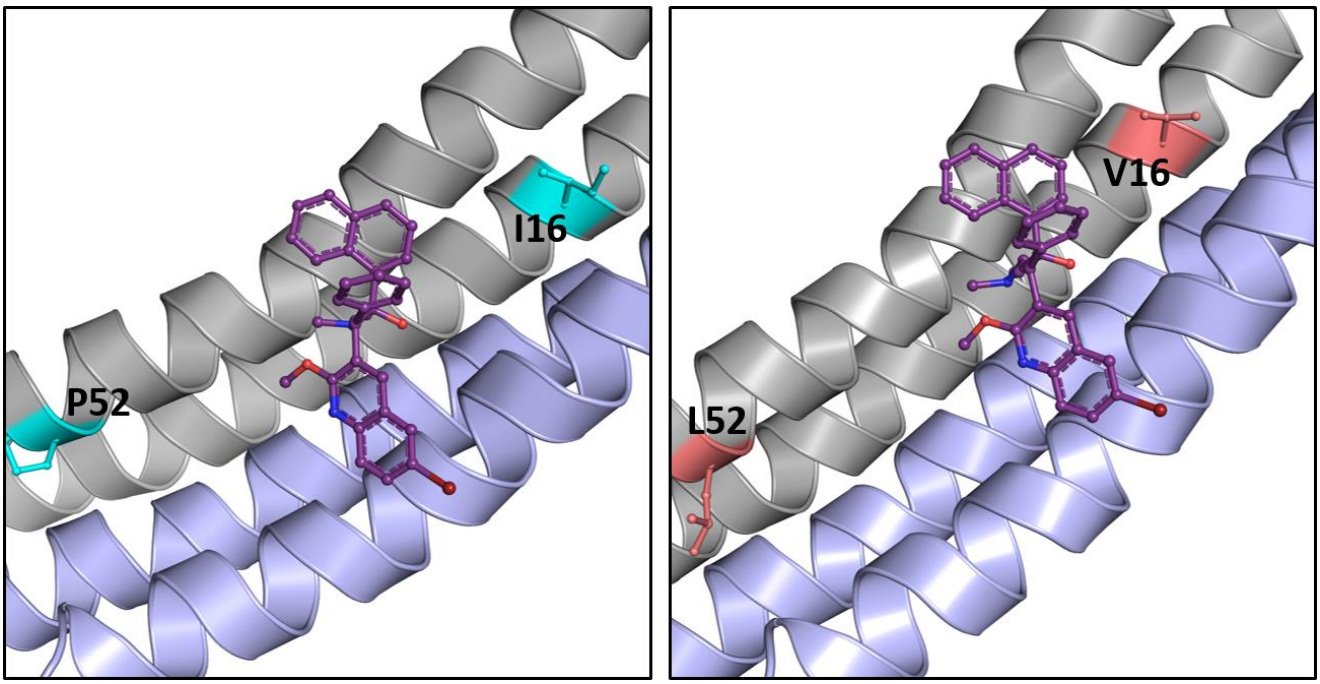


Figure S4. The localization of two circulating *atpE* variants relative to the bedaquiline binding pocket. The wild type residues are shown in cyan and mutant in salmon red in ball and stick representation. Bedaquiline is shown in purple (ball and stick representation).



CHAPTER 5: HYPER TRANSMISSION OF BEIJING LINEAGE *MYCOBACTERIUM TUBERCULOSIS*: SYSTEMATIC REVIEW AND META-ANALYSIS

Summary

Background: The globally distributed "Beijing" lineage of *Mycobacterium tuberculosis* has been associated with outbreaks worldwide. Laboratory based studies have suggested that Beijing lineage may have increased fitness; however, it has not been established whether these differences are of epidemiological significance with regards to transmission.

Objective: Therefore, we undertook a systematic review of epidemiological studies of tuberculosis transmission to compare the transmission dynamics and fitness cost of Beijing lineages versus the non-Beijing lineages.

Methods: We systematically searched Embase and MEDLINE before 31st December 2018, for studies which provided information on the transmission dynamics of the different *M. tuberculosis* lineages. We included articles that conducted population-based cross-sectional or longitudinal molecular epidemiological studies providing information about extent of transmission of different lineages. We then used a random effects model for meta-analysis to produce pooled estimates of transmission ratios for Beijing versus non-Beijing lineage.

Findings: Of 2855 records identified by the search, 42 were included in the review (39,044 patients from 25 countries). Beijing was the most prevalent and highly clustered strain in 76% of the studies. Twenty eligible studies were included in the final meta-analysis. Beijing lineage had a higher likelihood of transmission than non-Beijing lineages (OR 1.51 [0.99; 2.32], $I^2 = 95.0\%$, $\tau^2 = 0.72$).

Interpretation: Beijing lineage *M. tuberculosis* is significantly more likely to be linked to transmission than other lineages.

This chapter has been published in the *Journal of Infection* as a first author publication. *"Hyper transmission of Beijing lineage Mycobacterium tuberculosis: Systematic review and Meta-analysis"*, **Karmakar, M.**, Ascher, D.B., Trauer, M.J., Denholm, J.T. (2019) (doi: 10.1016/j.jinf.2019.09.016)

I thank the external reviewer of the thesis for critically assessing the systematic review and highlighting issues relating to the content of the paper. I have added text in the discussion below which should answer all the questions raised by the reviewer.

Discussion:

In this systematic review I wanted to compare the transmission dynamics of Beijing versus the rest of the lineages. Pooling of lineage 1, 3, 4, 5 and 6 as non-Beijing lineages can be contested because they are not genetically homogenous [220]. The MTBC strains differ in their content of synonymous and non-synonymous SNPs, deletions/ insertions and large duplications. Comas *et al* [122] conducted a WGS on 217 globally distributed clinical strains to calculate the number of pairwise SNPs between strains. They found a difference of 1200 SNPs between two human adapted strains on an average, which corresponds to 0.03% of the genome. Looking into the SNP distance within a lineage, it was seen, Lineage 1 had the highest genetic diversity and with an average of 730 SNPs between any two strains belonging to this lineage; whereas lineage 7 had the lowest corresponding distance with only 230 SNPs. The diversity between lineages for lineage 2, 3 and 4 or the "modern" lineages differed by 970 SNPs on average. Though increasing the number of genomes could influence distances, but it was the first study to indicate genomic distances between and with human adapted lineages [122, 220]. Grouping lineages with

differences in SNPs distances is a limitation in the study and might explain the heterogeneity observed in the study analysis.

Another limitation of the study (as discussed briefly in the published paper) is the issue of artefactual clustering when using molecular genotyping techniques like Spoligotyping, IS*6110*-RFLP and MIRU-VNTR. Ideal molecular typing methods should have desired performance parameters like technical simplicity, robustness, time-efficient, reproducibility and cost effectiveness. They should even accommodate analytical parameters like level of discrimination and stability of the molecular maker being used. A general rule which guides most of the molecular epidemiological investigations is the "discriminatory power" of the different typing methods. MIRU-VNTR has the highest discriminatory power (of the three typing methods mentioned), but when we combine it with Spoligotyping the discriminatory power significantly improves and makes the analysis more reliable [221]. But this is not enough, we even need to factor in other parameters such as the study setting, duration of the study and completeness of sampling [112]. The evolutionary rate is reflected by the stability of the genetic markers over time, often referred to as the molecular clock. MIRU-VNTR is considered to have a slower molecular clock than IS*6110*-RFLP which helps in detecting epidemiologically related cases [222]. Then comes the definition of a "cluster", which has a direct impact on generating diversified genetic patterns. While few studies stick to the strict rule of including isolates with identical genotype [223-225], other studies are lenient on the cluster definition and tolerate a single- or double-band difference in the RFLP profiles or double locus variations in the MIRU-VNTR profiles [226, 227]. Since we do not have an ideal definition of a cluster, therefore the decision to include or exclude an isolate is a matter of arbitrariness. We might have a higher chance of detecting a cluster when we slightly relax the cluster definition, but the consequence is lowering the likelihood of them being epidemiologically related [112]. Therefore, the current solution to all the issues raised above is whole genome sequencing, because currently the existing genotyping methods would not work in diverse settings or populations or be equally good in answering

specific epidemiological questions. Though we cannot deny the fact that molecular typing methods has significantly advanced our knowledge of the transmission and pathogenesis of mycobacteria.

This work helped me understand the genetic diversity in MTBC, which can be attributed to the fact that about two thirds of SNPs in coding regions are non-synonymous (i.e. amino acid changing) [228-230], and hence has an impact on the pathobiological phenotype [231, 232]. In future, if I do get an opportunity, I would like to explore the transmission dynamics of modern lineages (lineage 2, 3 and 4). I would compare Beijing lineage (sub-lineage of lineage 2) with LAM (L4.2) or Haarlem (L4.1.2) lineage (sub-lineages of lineage 4) to see if they are equally widespread globally and what factors determine such observations.

Journal of Infection 79 (2019) 572–581

Contents lists available at ScienceDirect

Journal of Infection

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

Hyper transmission of Beijing lineage *Mycobacterium tuberculosis*: Systematic review and meta-analysis

Malancha Karmakar[a,b,c,d], James M. Trauer[a,e], David B. Ascher[b,d,f], Justin T. Denholm[a,c,*]

[a] Victorian Tuberculosis Program, Melbourne Health, 792 Elizabeth Street, Melbourne, Victorian 3000 Australia
[b] Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria 3010, Australia
[c] Department of Microbiology and Immunology, at the Doherty Institute of Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia
[d] Structural Biology and Bioinformatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
[e] School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[f] Department of Biochemistry, University of Cambridge, CB2 1GA, UK

ARTICLE INFO

Article history:
Accepted 27 September 2019
Available online 1 October 2019

Keywords:
Beijing lineage
Molecular genotyping
Transmission
Tuberculosis

SUMMARY

Objectives: The globally distributed "Beijing" lineage of *Mycobacterium tuberculosis* has been associated with outbreaks worldwide. Laboratory based studies have suggested that Beijing lineage may have increased fitness; however, it has not been established whether these differences are of epidemiological significance with regards to transmission. Therefore, we undertook a systematic review of epidemiological studies of tuberculosis clustering to compare the transmission dynamics of Beijing lineages versus the non-Beijing lineages.

Methods: We systematically searched Embase and MEDLINE before 31st December 2018, for studies which provided information on the transmission dynamics of the different *M. tuberculosis* lineages. We included articles that conducted population-based cross-sectional or longitudinal molecular epidemiological studies reporting information about extent of transmission of different lineages. The protocol for this systematic review was prospectively registered with PROSPERO (CDR42018088579).

Results: Of 2855 records identified by the search, 46 were included in the review, containing 42,700 patients from 27 countries. Beijing lineage was the most prevalent and highly clustered strain in 72.4% of the studies and had a higher likelihood of transmission than non-Beijing lineages (OR 1·81 [95% 1·28–2·57], $I^2 = 94·0\%$, $\tau^2 = 0·59$, $p < 0·01$).

Conclusions: Despite considerable heterogeneity across epidemiological contexts, Beijing lineage appears to be more transmissible than other lineages.

© 2019 The British Infection Association. Published by Elsevier Ltd. All rights reserved.

Introduction

Mycobacterium tuberculosis (Mtb) has a strictly clonal and hierarchical population structure, due to a near complete absence of horizontal gene transfer. The only apparent modes of evolution of modern strains are through single nucleotide substitution, deletion and duplication events.[1] Because of the clonal structure of Mtb, comparative genotypic analyses from diverse geographic populations can provide unique insights into dissemination dynamics and evolutionary genetics of the pathogen.[2]

Genotypic evaluation of strain relatedness is frequently used to complement epidemiological evidence of transmission. Genotyping can be performed using a variety of techniques that interrogate different classes of genetic markers and generate either strain-specific banding patterns (IS*6110* DNA fingerprint), bar code-like signals (spoligotyping), or numerical patterns (24 locus-MIRU-VNTR typing)[3] and most recently next generation whole genome sequencing (WGS) for genome-based epidemiology.[4] Increasing molecular identification in recent decades has raised questions regarding potential strain-specific differences in the clinical outcomes and epidemiological characteristics of Mtb infection.[1] Currently, seven lineages have been defined by unique event polymorphism (single nucleotide polymorphism or deletion). Most of these lineages are highly prevalent in specific geographic areas and are named according to their predominant geographical distribution: Lineage 1 (Indo-Oceanic lineage), Lineage 2 (East Asian; includes sub-lineage "Beijing"), Lineage 3 (CAS/ Delhi), Lineage 4 (Euro-American), Lineage 5 (West African 1) and Lineage 6 (West African 2), Lineage 7 (Ethiopia).[5,6] These phylogeographic distributions of Mtb lineages suggest local adaptation of the pathogen to sympatric human populations.

* Corresponding author at: Victorian Tuberculosis Program, Melbourne Health, 792 Elizabeth Street, Melbourne, Victorian 3000 Australia.
E-mail address: Justin.denholm@mh.org.au (J.T. Denholm).

https://doi.org/10.1016/j.jinf.2019.09.016
0163-4453/© 2019 The British Infection Association. Published by Elsevier Ltd. All rights reserved.

Of particular interest has been a sub-lineage of Lineage 2 ("Beijing" strain), which is globally distributed[7] and has been associated with outbreaks.[8,9] In 2006, The European Concerted Action on New Generation Genetic Marker and Techniques for the Epidemiology and Control of Tuberculosis combined available datasets from all over the world (>29,000 patients from 49 studies in 35 countries) to assess the Beijing genotype's prevalence worldwide, trends over time and with age and its association with drug resistance.[10] Beijing lineage has been reported to be associated with an increased risk of acquired drug resistance, increased clinical severity and lesser protection from BCG vaccination.[7,11]

Laboratory based studies have also suggested that Beijing strains may have increased fitness,[12] although it has not been established whether these differences are of epidemiological significance with regards to transmission. Fitness of a transmissible organism can also be assessed by considering its effectiveness in terms of epidemic potential. Epidemic potential may be quantified by estimating the average number of secondary cases caused by a specific genotype after its introduction into an entirely susceptible population. These estimates rely on epidemiological evidence such as cluster studies, epidemiological investigation and model-based studies in human population rather than microbial behaviour in the laboratory because their precise contribution to the empiric success of an individual in the real world is not clear.[13] Therefore, we conducted a systematic review of epidemiological studies of Mtb transmission to quantify the extent of hyper-transmission of Beijing lineages.

Methods

Search strategy and selection criteria

We conducted a systematic review and meta-analysis of Mtb transmission to compare the epidemiological risk of transmission of Beijing versus non-Beijing lineage. Our search strategy was prospectively developed, recorded with the PROSPERO database (CDR42018088579) and conforms to the Preferred Reporting Items for Systematic reviews and Meta-analysis (PRISMA) guidelines.[14]

We searched two electronic databases for primary studies: MEDLINE and EMBASE until 31st of December 2018. Search terms included "tuberculosis", "*Mycobacterium tuberculosis*", "secondary cases", "secondary infection", "Beijing", "East-African Indian", "Euro-American", "West African 1", "West African 2", "Indo-Oceanic". The search was supplemented with additional search terms such as "fitness", "fitness cost", "strain"/"lineage" combined with terms for each lineage listed above, "transmission" and "transmission dynamics" to find relevant articles potentially missed during primary searching. We also incorporated a snowball sampling approach and hand searched articles identified from cross-references of identified articles and from suggestions of experts in the field. The study design involved observational studies (cross-sectional and longitudinal).

The titles and abstract for each of these citations were screened to capture relevant articles, with the following studies excluded: (1) studies not in English; (2) posters and reviews; (3) studies that lacked genotyping data; (4) studies related to *M. bovis* or *M. africanum* or non-tuberculous mycobacteria (5) studies focusing on immunological comparisons of plasma cytokine levels in peripheral blood mononuclear cells (6) proteomic approaches to understand the hypervirulence of Beijing isolates (7) studies which only involved multidrug resistant (MDR)-TB or extensively drug resistant (XDR)-TB patients (8) studies which focused on the single patient transmission chain and (9) studies limited to a single lineage only. Full text of the remaining citations was obtained and reviewed thoroughly against inclusion criteria. Disagreements between reviewers were resolved by consensus.

For an article to be included in the review, we required that the following information was reported: genotyping information for the patients with TB (pulmonary and/or extrapulmonary) relevant to the study irrespective of smear status, HIV status and age group. To account for recent transmission, a two-year cutoff period was considered ideal because it broadly coincides with the epidemiologically-observed high-risk period for the development of active TB after recent infection.[15–18] Using the 2-year cutoff period, an index case was defined as a pulmonary TB episode with a DNA fingerprint pattern that had not been assigned to another case within the preceding two years.[19] A secondary case was any case with an identical fingerprint pattern to the index case that was diagnosed no more than two years after the index case. We also investigated clustering information as it provides an indication of overall transmission leading to disease during the study period mentioned in each article. Included articles were required to provide information on either the number of secondary cases and index cases or the number of clustered cases, unique isolates and clusters for both Beijing and non-Beijing lineages. In all included studies, "cluster" was defined as ≥2 patients whose case isolates had identical DNA fingerprints. The percentage of recent transmission, which was our primary outcome measure, was calculated by the formula: $(n_c - c)/n$, where n is the total number of isolates, c is the number of clusters, and n_c is the total number of clustered isolates.[20–24] The clustering index was calculated by the total number of clustered isolates in each group divided by the total number of isolates for the group.[25]

Data analysis

The results of the electronic searches were compiled in Microsoft Excel and duplicate citations were removed. A data extraction form was used to extract the following information: authors, title, country of study, DOI, year of study, smear status of patients, number of secondary cases, number of index cases, transmission indices, number of index and secondary cases or number of clustered and non-clustered/unique isolates, HIV co-infection, resistance information for first-line drugs: isoniazid, rifampicin, ethambutol and pyrazinamide, age groups and conclusions.

Data were analysed using the meta-package[26] for the R programming language for statistical computing (version 3.2.3). We calculated pooled estimates of recent transmission, with their associated odds ratio (OR), standard error (95% CI), standard deviation (z) and p values for both fixed and random-effects models. Meta-analysis was done using the Mantel–Haenszel method; Hartung–Knapp adjustment for random effects model and Paule–Mandel estimator for τ^2. Continuity correction of 0.5 in studies with zero cell frequencies was used. Heterogeneity was assessed analytically by I^2 and Cochrane Q test.

Results

Systematic review

2843 articles were identified by the preliminary search strategy, with a further 12 articles identified from snowball sampling and manual review. After duplicate removal, 776 unique citations were identified, of which 504 publications were eligible for full text review and 46 met all eligibility criteria (Fig. 1).

The 46 included articles reported information on 42,700 patients diagnosed with tuberculosis from 27 countries. Various molecular genotyping methods were used, providing information on clustering by genotype. Table 1 presents an overview of the different molecular typing techniques used. We included twelve studies that used IS*6110*-RFLP for typing, twenty-eight studies that used spoligotyping, thirty-two studies that used MIRU-VNTR

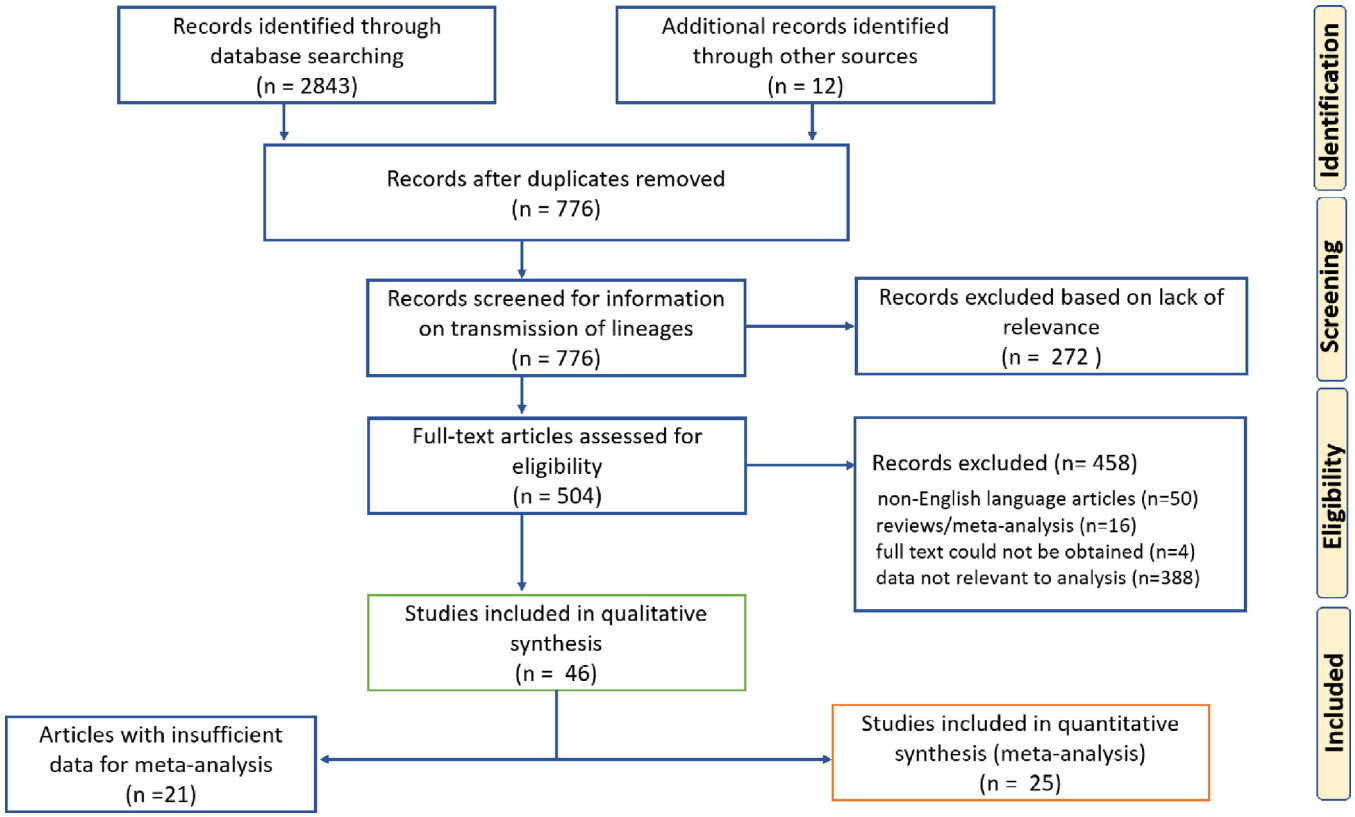


Fig. 1. Flow diagram of the study selection.

Table 1
Characteristics of included studies.

Study	Country	Region	Sample size	Year	Molecular typing	Findings
Anh et al.[27]	Vietnam	Ho Chi Minh City	563	2000	Spoligotyping	Beijing lineage constituted 53.5% of total isolates
Caminero et al.[28]	Spain	Gran Canaria Island	651	2001	IS*6110*-RFLP Spoligotyping	Beijing lineage constituted of the largest cluster (75 cases)
Banu et al.[29]	Bangladesh	Dhaka City	48	2004	Spoligotyping MIRU-VNTR	Beijing lineage constituted 31.3% of total isolates, of which 73.3% were clustered
Cox et al.[30]	Uzbekistan and Turkmenistan	Karakalpakstan, Dashoguz Velayat	382	2005	IS*6110*-RFLP Spoligotyping	Beijing constituted of 50.0% of the total isolates, of which 55.0% were clustered
Drobneiwski et al.[31]	Russia	Samara Region	880	2005	Spoligotyping 12 MIRU-VNTR	Beijing constituted of 63.4% of the total isolates
Hasan et al.[32]	Pakistan	Karachi, Punjab Province, Sindh Province, Northwest Frontier Province and Balouchistan Province	314	2006	Spoligotyping	Beijing constituted of 6.0% of total isolates of which 9.0% were clustered; Lineage 3 constituted of 39.0 of isolates
Dou et al.[33]	Taiwan		208	2008	Spoligotyping, 19 MIRU-VNTR, NTF loci typing and RD deletion number determination	Beijing lineage was the most prevalent, and was present in 40.0% of specimens from the aboriginal population, 72.4% of veterans, and 56.0 of the general population
Cowley et al.[34]	South Africa	Cape Town	291	2008	Spoligotyping	Beijing constituted 23.4% of total isolates
Mokrousov et al.[35]	Russia	Kaliningrad	90	2008	12 MIRU-VNTR	Beijing constituted of 41 of 90 isolates, representing the largest cluster (45.6%)
Van der Spuy et al.[36]	South Africa	Cape Town, Western Cape	1920	2009	IS*6110*-RFLP	Beijing constituted 39.2% of the total isolates of which 81.8% were clustered cases

(*continued on next page*)

Table 1 (*continued*)

Study	Country	Region	Sample size	Year	Molecular typing	Findings
Pardini et al.[37]	Georgia	Abkhazia	311	2009	IS*6110*-RFLP Spoligotyping	Beijing constituted 25.1% of total isolates and was significantly associated with clustering (OR 2.7)
Parwati et al.[38]	Indonesia	Jarkata and Bandung	844	2010	Spoligotyping	Beijing constituted 33.4% of isolates
Hu et al.[39]	China	Deqing County in Zhejiang Province and Guanyun County in Jiangsu Province (eastern China)	399	2010	IS*6110*-RFLP	Beijing constituted 80.1% of all isoniazid-resistant isolates, of which 56.2% were clustered
Shamputa et al.[40]	South Korea		208	2010	24 MIRU-VNTR Spoligotyping	Beijing constituted 97.1% of total isolates, but the clustering rate was low (22.3%)
Gallego et al.[41]	Australia	New South Wales	855	2010	12 MIRU-VNTR Spoligotyping	Beijing constituted 24.0% of total isolates along with the cluster having the highest number of isolates (49)
Wang et al.[42]	China	Heilongjiang Province	200	2011	Spoligotyping, Beijing family specific PCR, 19 MIRU-VNTR	Beijing lineage represented 89.5% of all isolates, of which 16.8% were clustered
Weisenberg et al.[25]	USA	New York City	3911	2012	IS*6110*-RFLP	Beijing constituted 15.1% of total isolates, of which 23.9% were clustered
Buu et al.[43]	Vietnam	Tien Giang Province (Southern Vietnam)	2207	2012	IS6110-RFLP	Beijing constituted 35.6% of total isolates, of which 37.0% were clustered; Lineage 1 constituted of 67.0 of clustered isolates
Aleksic et al.[44]	Kiribati	South Tarawa	74	2013	24 MIRU-VNTR IS*6110*-RFLP Spoligotyping	Beijing constituted 49.0% of total isolates of which 62.8% were clustered
Al-Hajoj et al.[45]	Saudi Arabia		902	2013	Spoligotyping 24 MIRU-VNTR	Beijing constituted 5.8% of all isolates, of which 55.8% were clustered
Langlois-Klassen et al.[46]	Canada	Alberta	1397	2013	IS*6110*-RFLP Spoligotyping	Beijing constituted 19.0% of all isolates, of which 21.0% were clustered
Lu et al.[47]	China	Jiangsu Province	497	2014	Spoligotyping 15 MIRU-VNTR	Beijing constituted 81.1% of all isolates, of which 32.5% were clustered
Liu et al.[48]	China	Gansu Province	426	2014	Spoligotyping 15 MIRU-VNTR	Beijing constituted 87.6% of all isolates and the largest cluster
Liu et al.[20]	China	Jiangsu Province	441	2014	Seven loci MIRU-VNTR (3820, Qub11a, Qub11b, Qub18, Qub26, MIRU26 and Mtub21)	Beijing constituted 89.3% of all isolates, but the clustering rate was low (4.4%)
Chen et al.[47]	Taiwan		177	2014	Spoligotyping and 24 MIRU-VNTR	Beijing constituted 35.2% of all isolates, of which 42.9% were clustered
Gurjav et al.[23]	Australia	Sydney, New South Wales	1128	2014	24 MIRU-VNTR	Beijing constituted 27.6% of all isolates, of which 40.5% were clustered
Zmak et al.[49]	Croatia		1587	2014	15 MIRU-VNTR	Lineage 4 constituted 66.7% and Beijing constituted 0.1% of the total isolates
Yang et al.[50]	China	Five sites	2274	2015	Different sets of MIRU-VNTR, hypervariable VNTR loci (3820, 1982, 3232 and 4120)	Beijing strain were more likely to be clustered (OR 1.67)
Yuan et al.[51]	China	Xinjiang Province	381	2015	24 MIRU-VNTR	Beijing constituted 57.5% of all isolates, of which 11.9% were clustered
Mathema et al.[52]	South Africa	15 mines (Gauteng, North West, and Free State)	1240	2015	IS*6110*-RFLP	Beijing constituted 13.6% of all isolates and most of the large clusters
Barletta et al.[53]	Peru	Lima	844	2015	Spoligotyping 15 MIRU-VNTR	Beijing constituted 16.4% of total isolates of which 59.2% were clustered (Lineage 4 was predominant)
Nebenzahl-Guimaraes et al.[54]	Netherlands		4436	2015	Spoligotyping 24 MIRU-VNTR	Beijing constituted 12.8% of total isolates of which 29.7% were clustered (Lineage 4 was predominant)

(*continued on next page*)

Table 1 (*continued*)

Study	Country	Region	Sample size	Year	Molecular typing	Findings
Globan et al.[55]	Australia	Victoria	2377	2015	15 MIRU-VNTR 24 MIRU-VNTR	Beijing constituted 20.7% of total isolates of which 80.9% were clustered
Hu et al.[56]	China	six rural counties	1222	2016	24 MIRU-VNTR Spoligotyping	Beijing constituted 79.1% of all isolates, of which 22.6% was clustered
Gurjav et al[24]	Australia	New South Wales	1692	2016	24 MIRU-VNTR WGS	Beijing constituted 27.8% of total isolates of which 35.7% were clustered
Liu et al.[57]	China	Beijing	679	2017	Spoligotyping 12 MIRU-VNTR	Beijing constituted 81.7% of total isolates of which 45.2% were clustered
Murase et al.[58]	Japan	37 prefectures	981	2017	28 MIRU-VNTR	Beijing constituted 70.6% of isolates of which 77.0% were clustered
Lalor et al.[59]	England		1646	2017	24 MIRU-VNTR	No increased clustering in the Beijing lineage compared to non-Beijing (increased transmission in Lineage 4 and CAS observed)
Liu et al.[60]	China	Xinjiang	311	2017	15 MIRU-VNTR Spoligotyping	Beijing constituted 72.0% of all isolates, of which 60.3% were clustered
Sharma et al.[61]	India	Ghatampur, Agra	355	2017	Spoligotyping 12 MIRU-VNTR	Beijing constituted 3.9% of all isolates, of which 3.0% were clustered; Lineage 3 was predominant
Riyahi Zaniani et al.[62]	Iran	Isfahan	49	2017	15 MIRU-VNTR	Beijing constituted 24.4% of all isolates, while Lineage 4 constituted 44.9% of isolates; overall low clustering rates
Yamamoto et al.[63]	Japan	Airin area, Osaka City	596	2018	24 MIRU-VNTR	Beijing constituted 80.3% of all isolates, of which 41.8% were clustered
Liu et al.[64]	China	Beijing	1189	2018	Spoligotyping VNTR typing	Beijing constituted 83.3% of isolates and was significantly associated with clustering (22.7%)
Holt et al.[65]	Vietnam	Districts 1, 4, 5, 6 and 8, Tan Binh, Binh Thanh and Phu Nhuan	1635	2018	WGS	Beijing constituted 59.0% of isolates of which 31.5% were clustered
Uddin et al.[66]	Bangladesh	Mymensingh, Netrokona, Kishoreganj, Jamalpur and Tangail districts (northeast part of Bangladesh)	244	2018	Spoligotyping 12 MIRU-VNTR	Beijing constituted of 7.4% of all isolates and Lineage 1 constituted of 27.0%
Bainomugisa et al. [67]	Papua New Guinea		100	2018	WGS	95 out of 100 clinical isolates typed belonged to Beijing stain

Notes: IS*6110*-RFLP – Restriction Fragment length polymorphism targeting the insertion sequence IS*6110*.
MIRU-VNTR – Mycobacterial Interspersed Repetitive Units (MIRU) specific multiple locus Variable Number of Tandem Repeats (VNTR) analysis.
WGS – Whole Genome Sequencing.
PCR – Polymerase Chain Reaction.
NTF – 556 bp of intervening sequence.
RD – Regions of differences.

typing and three studies that used WGS. Twenty-five studies used multiple methods of molecular typing to investigate the genotypic diversity of Mtb isolates. Studies included in the review were from a wide range of geographical settings, including nineteen high and eight low incidence settings.

Beijing lineage constituted the greatest proportion of total isolates in thirty-three of the forty-six studies (71.7%) included in the review (Table 1). Nineteen out of twenty-six studies (73.1%) had a higher clustering index for Beijing than non-Beijing strains (Table 2). Eleven studies had recent transmission rates that were higher for Beijing and three out of four studies which reported the mean number of secondary cases (transmission index) observed higher numbers in Beijing clusters; therefore 77.8% of the studies had a higher primary outcome measure for Beijing (Table 3).

Longitudinal reporting from several countries has found that Beijing strains constituted a growing proportion of total cases [43,53,68]. High rates of ongoing transmission of Beijing were seen in high-incidence settings, including Kiribati,[44] Saudi Arabia,[45] Vietnam,[27,43,65] India,[61,69] Spain,[28] Bangladesh,[29] Taiwan,[33,70] Uzbekistan and Turkmenistan,[30] Russia,[31,35] China,[39,47,50,56,57,64] Japan,[58,63] Georgia,[37] Estonia,[68] Indonesia,[38] South Africa [36] and one low-incidence setting, the Netherlands.[54] Low level transmission was observed in Australia, with clustering analysis revealing that the largest clusters comprised of Beijing lineage.[24,41,55] However, Beijing lineage did not show increased transmissibility compared to non-Beijing lineage in other settings with comprehensive and effective TB prevention and care practices, including the United Kingdom [59] and Canada.[46] In Pakistan it was observed that Beijing was well established in the region and was not a result of recent transmission.[32] Low levels of transmission were also observed in South Korea [40] and in certain rural areas of China.[20,42,51] In South African pediatric [71] and goldmining [52] populations no significant association was found between Beijing lineage and recent transmission.

Beijing and its association with age

Clustering of Beijing lineage in younger age groups is particularly likely to reflect recent transmission. 60.0% of the studies

Table 2
Clustering percentages reported or calculated in different studies for Beijing and non-Beijing strains (* reported in the study).

Study	Year of study	Beijing proportion clustered (%)	Non-Beijing proportion clustered (%)
Anh et al.[27]	2000	53.46	46.53
Cox et al.[30]*	2005	54.73	25.00
Duo et al.[33]	2008	75.63	41.37
Van der Spuy et al.[36]	2009	81.81	59.89
Hu et al.[39]	2010	56.19	15.38
Wang et al.[42]*	2011	16.80	0.00
Buu et al.[43]	2012	37.15	45.32
Weisenberg et al.[25]*	2013	34.80	31.30
Langlois-Klassen et al.[46]	2013	21.31	37.28
Al-Hajoj et al.[45]*	2013	55.76	34.65
Aleksic et al.[44]*	2013	62.79	37.20
Liu et al.[20]	2014	8.07	7.27
Chen et al.[47]	2014	57.14	32.09
Barletta et al.[53]*	2015	59.23	71.71
Nebenzahl-Guimaraes et al.[54]*	2015	32.00	27.50
Globan et al.[55]	2015	17.20	27.63
Yuan et al.[51]*	2015	11.87	24.69
Yang et al.[50]	2015	80.85	71.63
Hu et al.[56]*	2016	22.60	7.80
Liu et al.[57]*	2017	45.21	28.57
Murase et al.[58]	2017	22.68	77.31
Liu et al.[60]*	2017	60.27	25.29
Sharma et al.[61]	2017	2.99	10.96
Yamamato et al.[63]*	2018	41.33	36.75
Liu et al.[64]*	2018	22.70	9.00
Holt et al.[65]*	2018	31.50	14.00

Table 3
Recent transmission proportions reported or calculated in different studies for Beijing and non-Beijing strains (* reported in the study).

Recent transmission			
Study	Year of study	Beijing (%)	Non-Beijing (%)
Duo et al.[33]	2008	52.81	36.60
Van der Spuy et al.[36]*	2009	73.00	45.20
Wang et al.[42]*	2011	10.00	0.00
Weisenberg et al.[25]*	2013	23.90	25.68
Gurjav et al.[23]*	2014	26.90	6.20
Liu et al.[20]*	2014	4.43	3.99
Chen et al.[47]	2014	88.88	32.11
Barletta et al.[53]	2015	53.80	57.33
Yuan et al.[51]	2015	5.47	11.11
Gurjav et al.[24]*	2016	24.30	8.60
Liu et al.[57]*	2017	45.21	28.57
Liu et al. [60]*	2017	45.53	16.09
Liu et al.[64]	2018	20.52	5.52
Yamamato et al.[63]	2018	18.42	30.72
Transmission index (mean number of secondary cases)			
Study	Year of study	Beijing	Non-Beijing
Langlois-Klassen et al.[46]*	2013	0.06	0.14
Globan et al.[55]	2015	7.29	2.96
Nebenzahl-Guimaraes et al.[54]*	2015	1.18	1.02
Lalor et al.[59]*	2017	2.17	1.76

included in the review observed a strong association between clustering and younger age among the Beijing strains. For example, a greater degree of clustering was observed in the 25–44 year age group for studies undertaken in China,[20,50,64] Japan[58] and Indonesia[38]; the cross-sectional study from South Africa done in 15 gold mines across three provinces showed highest clustering among the 45–54 year age group[52]; and in Estonia the majority of clustered cases occurred in individuals aged 30–39 years.[68] A study from Saudi Arabia[45] found that Beijing was distributed equally among all age groups. In Australia[55] and other low incidence countries like Netherlands,[54] Beijing was the most common genotype among young adults (15–29 years old) and in the elderly (<60 years old).[23,24]

Beijing and its association with drug resistance

While the focus of this review is on transmissibility, the presence of drug resistance in isolates may be relevant to risk of secondary infection. We therefore summarise data on drug resistance in the studies identified by our systematic review. Thirty-three out of forty-six included studies reported associations between specific lineages and drug resistance, of which twenty studies (60.6%) showed significantly higher proportion of drug resistance among Beijing lineage. In studies conducted in China, some found higher rates of drug resistance among Beijing strains,[39,50,72] while others found no difference.[42,47] In the Taiwanese aboriginal population, a strong association was found between Beijing and MDR-TB.[47] 39.0% of the Beijing isolates (97.1% of the total isolates) found in South Korea were from MDR-TB or XDR-TB patients.[40] The association between MDR-TB and Beijing genotype in Vietnam was strongly associated with resistance to streptomycin.[43,65] Considerable ongoing transmission of MDR-TB strains of the Beijing lineage was observed in India,[61,69] Bangladesh,[66] Pakistan,[32] Papua New Guinea,[67] Russia,[31,35] Georgia,[37] Uzbekistan and Turkmenistan.[30] In Australia, the number of cases of MDR-TB was small and rates of drug resistance were unchanged since the 2006; however, the Beijing strain was found to be associated with a higher incidence of drug resistance.[23] We also found studies which reported no significant difference in drug resistance distribution between Beijing and non-Beijing lineages.[42,43,47,51,53,56,57,64] In the South African gold miner population, the AH strain (X family) was found to be associated with drug resistance and outbreaks.[52]

Meta-analysis

Following assessment of clustering and transmission indices of the different studies included in the review, we proceeded to our pre-planned meta-analysis. Twenty-five articles had information to conduct the meta-analysis. The odds ratio for the fixed effects model was 1.48 (95% CI 1.38 to 1.58, $z = 11.78$ $p < 0.0001$), while the odds ratio for the random effects model was 1.81 (95% CI 1.28 to 2.57, $z = 3.53$, p 0.0017) (Fig. 2). There was an even contribution

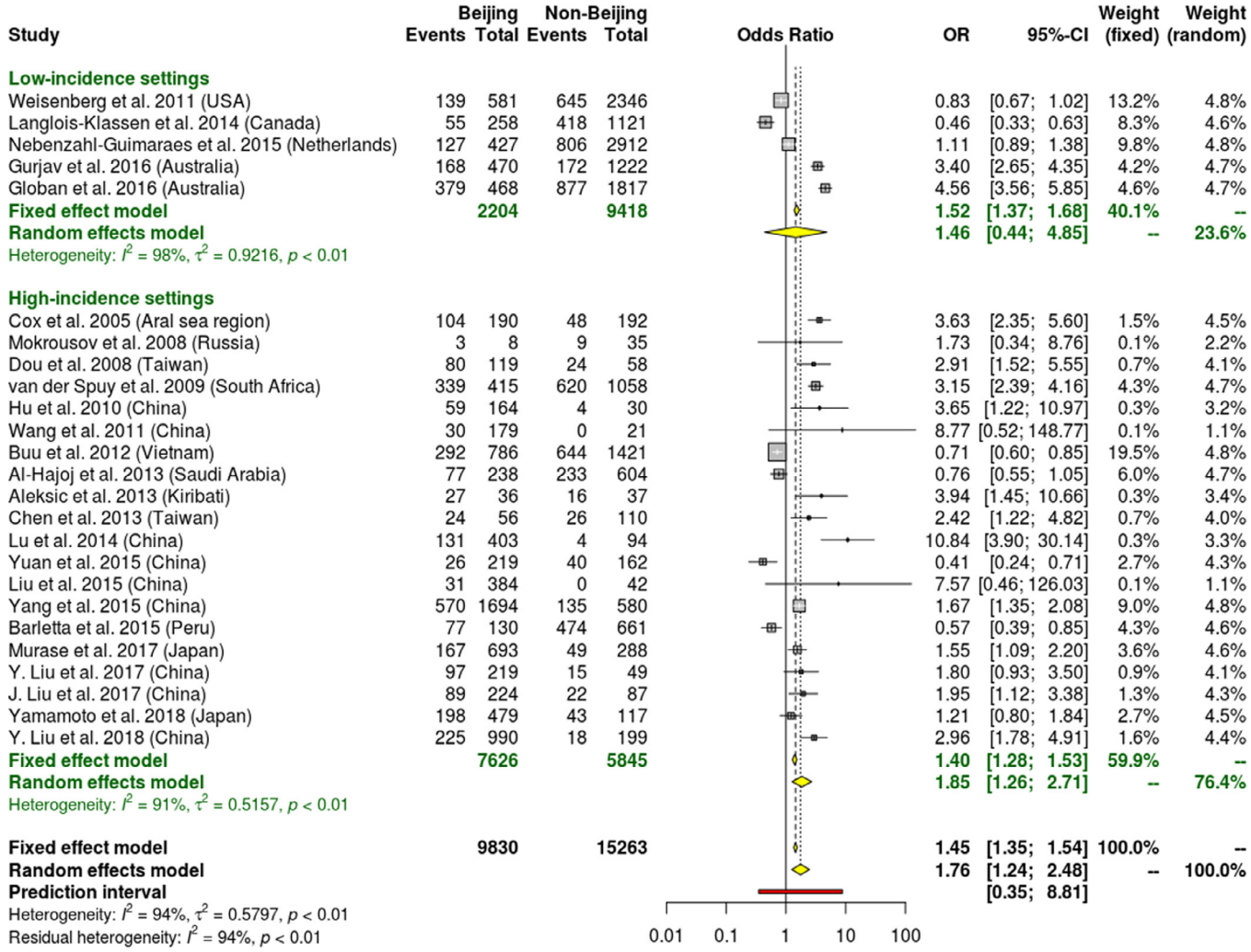


Fig. 2. Forest plot displaying the pooled estimates of transmission for Beijing and non-Beijing strains: The studies have been arranged in chronological order according to their date of publication. The first five studies are from low-incidence setting and the remaining 20 studies are from high-incidence setting.

from each included study (approximately 4% weight for each), but statistical heterogeneity was high.

Discussion

We found that Beijing lineage of Mtb was more likely to be associated with transmission than non-Beijing lineages. The strength of the observed relationship between Beijing lineage and transmission (OR 1.81) was notable and reflects a finding likely to be of epidemiological significance.

While our report has identified a statistically significant association between Beijing lineage TB and transmission, the mechanism for such an effect is inadequately understood. This finding may reflect either the selection of defined sub-lineages in different geographical settings, or the adaptation of strains in a defined Mtb sub-lineage capable of spreading more readily in certain human populations. It seems plausible that evolutionarily modern lineages like Beijing induce weaker immune response than ancient lineages, and this response potentially provides modern lineages with a selective advantage in terms of more rapid disease progression and/or transmission in human populations.[73] However, influence on transmission from microbiological fitness, differential immune response, or other mechanisms also remain plausible explanations. As observed in several studies East-African Indian (EAI) lineage was associated with notably low clustering rates, suggesting they are less likely to be transmitted, raising the possibility of future strain replacement.[54,74] Also, the frequency of transfer between diverse population groups like Vietnam[65] and Eastern Europe[37] supports previous assumptions that the Beijing lineage is a host generalist, capable of moving between ethnically diverse host populations.[9,48]

Our study's strengths include its systematic nature and emphasis on epidemiologic transmissibility, and our findings are limited by the heterogeneity of outcomes and variation in epidemiological and genomic definitions adopted. Classical molecular genotyping has been nearly used for thirty years to define transmission chains / clusters, but it comes with an inherent limitation: overestimation of recent transmission events.[24,42] Spoligotyping has lower discriminatory power compared to MIRU-VNTR; however, a combination of both shows better discriminatory power.[75] Majority of the studies included in the review used both Spoligotyping and MIRU-VNTR as genotyping methods to determine clusters. Studies that only used Spoligotyping were not included in the meta-analysis to avoid overestimation of recent transmission. If these studies were further paired with whole genome sequencing-based approaches the extent of overestimation could be refined further.[76] With the ever-decreasing cost of whole genome sequencing and easier implementation in a variety of settings (especially high-incidence, low-resource settings), it is likely to become

an integral part of the epidemiological approach to track and stop TB.

The high genotypic diversity seen in a low-incidence setting like Australia reflect the large number of overseas-born patients who migrate from all over the world rather than local transmission.[24,41,55] By contrast, studies from Asia and Russia highlighted the high levels of genome homoplasy within the Beijing strain family.[23,40,58] In high endemic settings like India,[61] Taiwan[33] and South Africa,[52] the high genetic diversity of the bacillary load could be explained by a mobile population in combination with reactivation, appearance and disappearance of individual clones and the long incubation period of the disease. Socio-demographic factors like lack of permanent housing, which leads to congregation of people in specific locations and spreading the infection, was observed as a correlate of clustering in Estonia.[68] We also think that it is unlikely a founder effect has a significant role in apparent clustering of Beijing lineage for several reasons. First, historically substantial shifts have been seen in lineage distribution in recent decades, suggesting a dynamic environment where transmission between regions remains relevant. Second, we have included studies where Beijing is both a majority and minority strain, minimising the potential impact of a founder effect. Finally, we have also included a two-year cut off period for defining clustering, which should also be helpful in concentrating the effect seen towards recent transmission.

The definition of fitness includes a microorganism's ability to survive, reproduce and to be transmitted.[13] Mutations leading to drug resistance development may influence the fitness of the microorganism. It has been speculated that low physiological cost of rifampicin resistance and compensatory mutations restoring fitness of Mtb maybe responsible for the widespread dissemination of the Beijing strain.[12] It was also observed that Beijing strains that were MDR were universally resistant to streptomycin.[43,58,65] An association between Beijing lineage and the development of drug resistance could influence clustering of isolates. This is expected to result in a selective advantage for Beijing strain and therefore would lead to higher prevalence of Beijing.[30] The hypervirulence of Beijing strains can be attributed to deletions in *ppe38*, which is responsible for the secretion of a subset of ESX-5 substrates.[77]

This review reinforces the epidemiological significance of Mtb lineages and highlights the importance of combining optimal molecular strain typing with epidemiological data. Further research into the mechanisms of increased transmissibility is required and translating genotypic data into programmatic algorithms. Mathematical models of TB transmission incorporate the process of infection; interventions leading to faster diagnosis and therefore reduced transmission.[78] Effective reproductive number (R_e), which represents the average number of secondary cases arising from a primary case of active TB is commonly used to describe infectiousness. In our current review, we are unable to estimate fully the *Re* from the available data, because our analysis only considers clustered events separated by less than two years and its well-known that late reactivation episodes after this time period are important in sustaining transmission of Mtb. Although a value of one is an important *Re* threshold for disease persistence in a population in general, the relative magnitude of *Re* for two co-circulating strains is of greater relevance to which Mtb strain will be sustained within a population. For outbreaks of a single pathogen, heterogeneous transmission has been shown to favour stochastic extinction as well as explosive outbreaks.[79] Given the high heterogeneity of TB transmission,[80,81] similar principles may apply to a multi-strain competition, in which one strain may replace another more rapidly than predicted by models that assume well-mixed populations. This may explain some of the heterogeneity in our findings.

Expansion of genotyping techniques holds great promise for optimizing public health management of TB. Inclusion of clustering information in routine public health responses is already used for tailoring strategies to reduce Mtb transmission and reactivation. Our results suggest that strategies enhancing contact tracing towards Beijing lineages could be evaluated further, particularly in high incidence settings where they are likely to contribute most to onward transmission and perpetuating the global TB epidemic.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Malancha Karmakar: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft. **James M. Trauer:** Methodology, Investigation, Writing - review & editing. **David B. Ascher:** Supervision, Writing - review & editing. **Justin T. Denholm:** Conceptualization, Data curation, Investigation, Funding acquisition, Project administration, Supervision, Writing - review & editing.

Acknowledgements

M.K was funded by the Melbourne Research Scholarship and supported in part by the Victorian Government's OIS Program. JTD receives funding from the Medical Research Future Fund and the National Health and Medical Research Council.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jinf.2019.09.016.

References

1. Gagneux S, Small PM. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *Lancet Infect Dis* 2007;**7**(5):328–37.
2. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains. *Trends Microbiol* 2002;**10**(1):45–52.
3. Schurch AC, van Soolingen D. DNA fingerprinting of Mycobacterium tuberculosis: from phage typing to whole-genome sequencing. *Infect Genet Evol* 2012;**12**(4):602–9.
4. Cannas A, Mazzarelli A, Di Caro A, Delogu G, Girardi E. Molecular Typing of Mycobacterium Tuberculosis Strains: A Fundamental Tool for Tuberculosis Control and Elimination. *Infect Dis Rep* 2016;**8**(2):6567.
5. Shabbeer A, Cowan LS, Ozcaglar C, Rastogi N, Vandenberg SL, Yener B, et al. TB–Lineage: an online tool for classification and analysis of strains of Mycobacterium tuberculosis complex. *Infect Genet Evol* 2012;**12**(4):789–97.
6. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic co-expansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* 2013;**45**(10):1176–82.
7. Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D. Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review. *Emerg Infect Dis* 2002;**8**(8):843–9.
8. Wiens KE, Woyczynski LP, Ledesma JR, Ross JM, Zenteno-Cuevas R, Goodridge A, et al. Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. *BMC Medicine* 2018;**16**(1):196.
9. Coscolla M, Gagneux S. Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin Immunol* 2014;**26**(6):431–44.
10. Beijing/W genotype Mycobacterium tuberculosis and drug resistance. *Emerg Infect Dis* 2006;**12**(5):736–43.
11. Hanekom M, Gey van Pittius NC, McEvoy C, Victor TC, Van Helden PD, Warren RM. Mycobacterium tuberculosis Beijing genotype: a template for success. *Tuberculosis (Edinb)* 2011;**91**(6):510–23.
12. Toungoussova OS, Caugant DA, Sandven P, Mariandyshev AO, Bjune G. Impact of drug resistance on fitness of Mycobacterium tuberculosis strains of the W-Beijing genotype. *FEMS Immunol Med Microbiol* 2004;**42**(3):281–90.

13. Cohen T, Sommers B, Murray M. The effect of drug resistance on the fitness of Mycobacterium tuberculosis. *Lancet Infect Dis* 2003;**3**(1):13–21.
14. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;**151**(4):264–9 w64.
15. Trauer JM, Moyo N, Tay EL, Dale K, Ragonnet R, McBryde ES, et al. Risk of active tuberculosis in the five years following infection . . . 15%? *Chest* 2016;**149**(2):516–25.
16. Sloot R, van der Loeff MF, Kouw PM, Borgdorff MW. Risk of tuberculosis after recent exposure. A 10-year follow-up study of contacts in Amsterdam. *Am J Respir Crit Care Med* 2014;**190**(9):1044–52.
17. Hart PD, Sutherland I. BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. *Br Med J* 1977;**2**(6082):293–5.
18. Combs DL, O'Brien RJ, Geiter LJ. USPHS tuberculosis short-course chemotherapy trial 21: effectiveness, toxicity, and acceptability. The report of final results. *Ann Intern Med* 1990;**112**(6):397–406.
19. Ferebee SH. Controlled chemoprophylaxis trials in tuberculosis. A general review. *Bibl Tuberc* 1970;**26**:28–106.
20. Liu M, Jiang W, Liu Y, Zhang Y, Wei X, Wang W. Increased genetic diversity of the Mycobacterium tuberculosis W-Beijing genotype that predominates in eastern China. *Infect Genet Evol* 2014;**22**:23–9.
21. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med* 1994;**330**(24):1703–9.
22. Murray M, Alland D. Methodological problems in the molecular epidemiology of tuberculosis. *Am J Epidemiol* 2002;**155**(6):565–71.
23. Gurjav U, Jelfs P, McCallum N, Marais BJ, Sintchenko V. Temporal dynamics of Mycobacterium tuberculosis genotypes in New South Wales, Australia. *BMC Infect Dis* 2014;**14**:455.
24. Gurjav U, Outhred AC, Jelfs P, McCallum N, Wang Q, Hill-Cawthorne GA, et al. Whole Genome Sequencing Demonstrates Limited Transmission within Identified Mycobacterium tuberculosis Clusters in New South Wales, Australia. *PloS One* 2016;**11**(10):e0163612.
25. Weisenberg SA, Gibson AL, Huard RC, Kurepina N, Bang H, Lazzarini LCO, et al. Distinct Clinical and Epidemiological Features of Tuberculosis in New York City Caused by the RD(Rio) Mycobacterium tuberculosis Sublineage. *Infect Genet Evol* 2012;**12**(4):664–70.
26. G S. General Package for Meta-Analysis. *R News* 2007;**7**:40–5.
27. Anh DD, Borgdorff MW, Van LN, Lan NT, van Gorkom T, Kremer K, et al. Mycobacterium tuberculosis Beijing genotype emerging in Vietnam. *Emerg Infect Dis* 2000;**6**(3):302–5.
28. Caminero JA, Pena MJ, Campos-Herrero MI, Rodriguez JC, Garcia I, Cabrera P, et al. Epidemiological evidence of the spread of a Mycobacterium tuberculosis strain of the Beijing genotype on Gran Canaria Island. *Am J Respir Crit Care Med* 2001;**164**(7):1165–70.
29. Banu S, Gordon SV, Palmer S, Islam MR, Ahmed S, Alam KM, et al. Genotypic analysis of Mycobacterium tuberculosis in Bangladesh and prevalence of the Beijing strain. *J Clin Microbiol* 2004;**42**(2):674–82.
30. Cox HS, Kubica T, Doshetov D, Kebede Y, Rusch-Gerdess S, Niemann S. The Beijing genotype and drug resistant tuberculosis in the Aral Sea region of Central Asia. *Respir Res* 2005;**6**:134.
31. Drobniewski F, Balabanova Y, Nikolayevsky V, Ruddy M, Kuznetzov S, Zakharova S, et al. Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. *Jama* 2005;**293**(22):2726–31.
32. Hasan Z, Tanveer M, Kanji A, Hasan Q, Ghebremichael S, Hasan R. Spoligotyping of Mycobacterium tuberculosis isolates from Pakistan reveals predominance of Central Asian Strain 1 and Beijing isolates. *J Clin Microbiol* 2006;**44**(5):1763–8.
33. Dou HY, Tseng FC, Lu JJ, Jou R, Tsai SF, Chang JR, et al. Associations of Mycobacterium tuberculosis genotypes with different ethnic and migratory populations in Taiwan. *Infect Genet Evol* 2008;**8**(3):323–30.
34. Cowley D, Govender D, February B, Wolfe M, Steyn L, Evans J, et al. Recent and rapid emergence of W-Beijing strains of Mycobacterium tuberculosis in Cape Town, South Africa. *Clin Infect Dis: an official publication of the Infectious Diseases Society of America* 2008;**47**(10):1252–9.
35. Mokrousov I, Otten T, Zozio T, Turkin E, Nazemtseva V, Sheremet A, et al. At Baltic crossroads: a molecular snapshot of Mycobacterium tuberculosis population diversity in Kaliningrad, Russia. *FEMS Immunol Med Microbiol* 2009;**55**(1):13–22.
36. van der Spuy GD, Kremer K, Ndabambi SL, Beyers N, Dunbar R, Marais BJ, et al. Changing Mycobacterium tuberculosis population highlights clade-specific pathogenic characteristics. *Tuberculosis (Edinb)* 2009;**89**(2):120–5.
37. Pardini M, Niemann S, Varaine F, Iona E, Meacci F, Orru G, et al. Characteristics of drug-resistant tuberculosis in Abkhazia (Georgia), a high-prevalence area in Eastern Europe. *Tuberculosis (Edinb)* 2009;**89**(4):317–24.
38. Parwati I, Alisjahbana B, Apriani L, Soetikno RD, Ottenhoff TH, van der Zanden AG, et al. Mycobacterium tuberculosis Beijing genotype is an independent risk factor for tuberculosis treatment failure in Indonesia. *J Infect Dis* 2010;**201**(4):553–7.
39. Hu Y, Hoffner S, Jiang W, Wang W, Xu B. Extensive transmission of isoniazid resistant M. tuberculosis and its association with increased multidrug-resistant TB in two rural counties of eastern China: a molecular epidemiological study. *BMC Infect Dis* 2010;**10**:43.
40. Shamputa IC, Lee J, Allix-Beguec C, Cho EJ, Lee JI, Rajan V, et al. Genetic diversity of Mycobacterium tuberculosis isolates from a tertiary care tuberculosis hospital in South Korea. *J Clin Microbiol* 2010;**48**(2):387–394.
41. Gallego B, Sintchenko V, Jelfs P, Coiera E, Gilbert GL. Three-year longitudinal study of genotypes of Mycobacterium tuberculosis in a low prevalence population. *Pathology* 2010;**42**(3):267–72.
42. Wang J, Liu Y, Zhang CL, Ji BY, Zhang LZ, Shao YZ, et al. Genotypes and characteristics of clustering and drug susceptibility of Mycobacterium tuberculosis isolates collected in Heilongjiang Province, China. *J Clin Microbiol* 2011;**49**(4):1354–62.
43. Buu TN, van Soolingen D, Huyen MN, Lan NT, Quy HT, Tiemersma EW, et al. Increased transmission of Mycobacterium tuberculosis Beijing genotype strains associated with resistance to streptomycin: a population-based study. *PloS One* 2012;**7**(8):e42323.
44. Aleksic E, Merker M, Cox H, Reiher B, Sekawi Z, Hearps AC, et al. First molecular epidemiology study of Mycobacterium tuberculosis in Kiribati. *PloS One* 2013;**8**(1):e55423.
45. Al-Hajoj S, Varghese B, Al-Habobe F, Shoukri MM, Mulder A, van Soolingen D. Current trends of Mycobacterium tuberculosis molecular epidemiology in Saudi Arabia and associated demographical factors. *Infect Genet Evol* 2013;**16**:362–8.
46. Langlois-Klassen D, Senthilselvan A, Chui L, Kunimoto D, Saunders LD, Menzies D, et al. Transmission of Mycobacterium tuberculosis Beijing Strains, Alberta, Canada, 1991-2007. *Emerg Infect Dis* 2013;**19**(5):701–11.
47. Lu W, Lu B, Liu Q, Dong H, Shao Y, Jiang Y, et al. Genotypes of Mycobacterium tuberculosis isolates in rural China: using MIRU-VNTR and spoligotyping methods. *Scand J Infect Dis* 2014;**46**(2):98–106.
48. Liu J, Tong C, Liu J, Jiang Y, Zhao X, Zhang Y, et al. First insight into the genotypic diversity of clinical Mycobacterium tuberculosis isolates from Gansu Province, China. *PloS One* 2014;**9**(6):e99357.
49. Zmak L, Obrovac M, Katalinic Jankovic V. First insights into the molecular epidemiology of tuberculosis in Croatia during a three-year period, 2009 to 2011. *Scand J Infect Dis* 2014;**46**(2):123–9.
50. Yang C, Shen X, Peng Y, Lan R, Zhao Y, Long B, et al. Transmission of Mycobacterium tuberculosis in China: a population-based molecular epidemiologic study. *Clin Infect Dis: an official publication of the Infectious Diseases Society of America* 2015;**61**(2):219–27.
51. Yuan L, Mi L, Li Y, Zhang H, Zheng F, Li Z. Genotypic characteristics of Mycobacterium tuberculosis circulating in Xinjiang, China. *Infect Dis (Lond)* 2016;**48**(2):108–15.
52. Mathema B, Lewis JJ, Connors J, Chihota VN, Shashkina E, van der Meulen M, et al. Molecular epidemiology of Mycobacterium tuberculosis among South African gold miners. *Ann Am Thorac Soc* 2015;**12**(1):12–20.
53. Barletta F, Otero L, de Jong BC, Iwamoto T, Arikawa K, Van der Stuyft P, et al. Predominant Mycobacterium tuberculosis Families and High Rates of Recent Transmission among New Cases Are Not Associated with Primary Multidrug Resistance in Lima, Peru. *J Clin Microbiol* 2015;**53**(6):1854–63.
54. Nebenzahl-Guimaraes H, Verhagen LM, Borgdorff MW, van Soolingen D. Transmission and Progression to Disease of Mycobacterium tuberculosis Phylogenetic Lineages in The Netherlands. *J Clin Microbiol* 2015;**53**(10):3264–71.
55. Globan M, Lavender C, Leslie D, Brown L, Denholm J, Raios K, et al. Molecular epidemiology of tuberculosis in Victoria, Australia, reveals low level of transmission. *Int J Tuberc Lung Dis* 2016;**20**(5):652–8.
56. Hu Y, Mathema B, Zhao Q, Zheng X, Li D, Jiang W, et al. Comparison of the socio-demographic and clinical features of pulmonary TB patients infected with sub-lineages within the W-Beijing and non-Beijing Mycobacterium tuberculosis. *Tuberculosis (Edinb)* 2016;**97**:18–25.
57. Liu Y, Jiang X, Li W, Zhang X, Wang W, Li C. The study on the association between Beijing genotype family and drug susceptibility phenotypes of Mycobacterium tuberculosis in Beijing. *Sci Rep* 2017;**7**(1):15076.
58. Murase Y, Izumi K, Ohkado A, Aono A, Chikamatsu K, Yamada H, et al. Prediction of Local Transmission of Mycobacterium tuberculosis Isolates of a Predominantly Beijing Lineage by Use of a Variable-Number Tandem-Repeat Typing Method Incorporating a Consensus Set of Hypervariable Loci. *J Clin Microbiol* 2018;**56**(1):01.
59. Lalor MK, Anderson LF, Hamblion EL, Burkitt A, Davidson JA, Maguire H, et al. Recent household transmission of tuberculosis in England, 2010-2012: retrospective national cohort study combining epidemiological and molecular strain typing data. *BMC Med* 2017;**15**(1):105.
60. Liu J, Li J, Liu J, Zhao X, Lian L, Liu H, et al. Genotypic Diversity of Mycobacterium tuberculosis Clinical Isolates in the Multiethnic Area of the Xinjiang Uygur Autonomous Region in China. *Biomed Res Int* 2017;**2017**:3179535.
61. Sharma P, Katoch K, Chandra S, Chauhan DS, Sharma VD, Couvin D, et al. Comparative study of genotypes of Mycobacterium tuberculosis from a Northern Indian setting with strains reported from other parts of India and neighboring countries. *Tuberculosis (Edinb)* 2017;**105**:60–72.
62. Riyahi Zaniani F, Moghim S, Mirhendi H, Ghasemian Safaei H, Fazeli H, Salehi M, et al. Genetic Lineages of Mycobacterium tuberculosis Isolates in Isfahan, Iran. *Curr Microbiol* 2017;**74**(1):14–21.
63. Yamamoto K, Takeuchi S, Seto J, Shimouchi A, Komukai J, Hase A, et al. Longitudinal genotyping surveillance of Mycobacterium tuberculosis in an area with high tuberculosis incidence shows high transmission rate of the modern Beijing subfamily in Japan. *Infect Genet Evol* 2018.
64. Liu Y, Zhang X, Zhang Y, Sun Y, Yao C, Wang W, et al. Characterization of Mycobacterium tuberculosis strains in Beijing, China: drug susceptibility phenotypes and Beijing genotype family transmission. *BMC Infect Dis* 2018;**18**(1):658.
65. 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* 2018;**50**(6):849–56.

66. Uddin MKM, Ahmed M, Islam MR, Rahman A, Khatun R, Hossain MA, et al. Molecular characterization and drug susceptibility profile of Mycobacterium tuberculosis isolates from Northeast Bangladesh. *Infect Genet Evol: journal of molecular epidemiology and evolutionary genetics in infectious diseases* 2018;**65**:136–43.
67. Bainomugisa A, Lavu E, Hiashiri S, Majumdar S, Honjepari A, Moke R, et al. Multi-clonal evolution of multi-drug-resistant/extensively drug-resistant Mycobacterium tuberculosis in a high-prevalence setting of Papua New Guinea for over three decades. *Microb Genomics* 2018;**4**(2):02.
68. Toit K, Altraja A, Acosta CD, Viiklepp P, Kremer K, Kummik T, et al. A four-year nationwide molecular epidemiological study in Estonia: risk factors for tuberculosis transmission. *Public Health Action* 2014;**4**(Suppl 2):S34–40.
69. Almeida D, Rodrigues C, Ashavaid TF, Lalvani A, Udwadia ZF, Mehta A. High incidence of the Beijing genotype among multidrug-resistant isolates of Mycobacterium tuberculosis in a tertiary care center in Mumbai, India. *Clin Infect Dis: an official publication of the Infectious Diseases Society of America* 2005;**40**(6):881–6.
70. Chen YY, Chang JR, Huang WF, Kuo SC, Yeh JJ, Lee JJ, et al. Molecular epidemiology of Mycobacterium tuberculosis in aboriginal peoples of Taiwan, 2006-2011. *J Infect* 2014;**68**(4):332–7.
71. Marais BJ, Hesseling AC, Schaaf HS, Gie RP, van Helden PD, Warren RM. Mycobacterium tuberculosis transmission is not related to household genotype in a setting of high endemicity. *J Clin Microbiol* 2009;**47**(5):1338–43.
72. Jiao W, Liu Z, Han R, Zhao X, Dong F, Dong H, et al. A country-wide study of spoligotype and drug resistance characteristics of Mycobacterium tuberculosis isolates from children in China. *PloS One* 2013;**8**(12):e84315.
73. 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**(3):e1001307.
74. Albanna AS, Reed MB, Kotar KV, Fallow A, McIntosh FA, Behr MA, et al. Reduced transmissibility of East African Indian strains of Mycobacterium tuberculosis. *PloS One* 2011;**6**(9):e25075.
75. Pitondo-Silva A, Santos AC, Jolley KA, Leite CQ, Darini AL. Comparison of three molecular typing methods to assess genetic diversity for Mycobacterium tuberculosis. *J Microbiol Methods* 2013;**93**(1):42–8.
76. Meehan CJ, Moris P, Kohl TA, Pecerska J, Akter S, Merker M, et al. The relationship between transmission time and clustering methods in Mycobacterium tuberculosis epidemiology. *EBioMedicine* 2018;**37**:410–16.
77. Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, et al. Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. *Nat Microbiol* 2018;**3**(2):181–8.
78. Menzies NA, Cohen T, Lin HH, Murray M, Salomon JA. Population health impact and cost-effectiveness of tuberculosis diagnosis with Xpert MTB/RIF: a dynamic simulation and economic evaluation. *PLoS Med* 2012;**9**(11):e1001347.
79. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. *Nature* 2005;**438**(7066):355–9.
80. Melsew YA, Gambhir M, Cheng AC, McBryde ES, Denholm JT, Tay EL, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. *BMC Infect Dis* 2019;**19**(1):244.
81. Ypma RJ, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. *Epidemiology (Cambridge, Mass)* 2013;**24**(3):395–400.

CHAPTER 6: ESTIMATING THE RISK OF TUBERCULOSIS DRUG RESISTANCE AMPLIFICATION IN HIGH-BURDEN SETTINGS

In this Chapter I wanted to use the knowledge I had gathered from the structural studies to investigate the likelihood of resistance spreading in a population. I built a compartmental epidemiological model and used an adaptive metropolis algorithm to estimate the risk of resistance amplification for isoniazid and rifampicin due to treatment failure. The model provided additional estimates for the relative fitness and transmission rate associated with each drug resistant strain and the case detection rate. We observed rifampicin resistant strains were more likely to be transmitted than acquired through amplification, while both mechanisms of acquisition were important contributors in the case of isoniazid resistance. This finding emphasizes the important of prioritizing testing algorithms for the early detection of isoniazid resistance.

This Chapter has been submitted to *International Journal of Epidemiology* as a first author publication, titled "Estimating the risk of tuberculosis drug resistance amplification in high-burden settings" (2021), **Malancha Karmakar**, Romain Ragonnet, David B. Ascher, James M. Trauer and Justin T. Denholm.

Estimating tuberculosis drug resistance amplification rates in high-burden settings

Malancha Karmakar[1,2,3], Romain Ragonnet[4], David B. Ascher[1,2], James M. Trauer[4], Justin T. Denholm[3]

[1] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[2] Structural Biology and Bioinformatics, Department of Biochemistry, University of Melbourne, Melbourne, Victoria, Australia

[3] Victorian Tuberculosis Program and Department of Microbiology and Immunology, Doherty Institute of Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia

[4] School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

Abstract

Background: Antimicrobial resistance develops following the accrual of mutations in the bacterial genome, and may variably impact organism fitness and hence, transmission risk. Classical representation of tuberculosis (TB) dynamics using a single or two strain (DS/MDR-TB) model typically does not capture elements of this important aspect of TB epidemiology. To understand and estimate the likelihood of resistance spreading in high drug-resistant TB incidence settings, we used epidemiological data to develop a mathematical model of *Mycobacterium tuberculosis* (*Mtb*) transmission.

Methods: A four-strain (drug-susceptible (DS), isoniazid mono-resistant (INH-R), rifampicin mono-resistant (RIF-R) and multidrug-resistant (MDR)) compartmental deterministic *Mtb* transmission model was developed to explore the progression from DS- to MDR-TB in The Philippines and Viet Nam. The models were calibrated using data from national tuberculosis prevalence surveys and drug resistance surveys. We used an adaptive Metropolis algorithm to estimate the risks of drug resistance amplification among unsuccessfully treated individuals and the fitness costs associated with different types of drug resistance.

Results: The estimated proportion of INH-R acquisition among failing treatments was 0.84 (95% CI 0.79 – 0.89) for The Philippines and 0.77 (95% CI 0.71 – 0.84) for Viet Nam. The proportion of RIF-R acquisition among failing treatments was 0.05 (95% CI 0.04 – 0.07) for The Philippines and 0.011 (95% CI 0.010 – 0.012) for Viet Nam. In the Philippines, the estimated proportion of drug resistance resulting from transmission was 50% (95% CI 43 – 70) for INH-R, 52% (95% CI 43 – 70) for RIF-R and 40% (95% CI 28 – 52) for MDR-TB. For Viet Nam, the estimated proportion of drug resistance due to transmission was 67% (95% CI 54 – 73) for INH-R, 63% (95% CI 55 – 71) for RIF-R and 43% (95% CI 34 – 51) for MDR-TB.

Discussion: The risk of resistance amplification due to treatment failure for INH was dramatically higher than RIF. We observed RIF-R strains were more likely to be transmitted than acquired through amplification, while both mechanisms of acquisition were important contributors in the case of INH-R. These findings highlight the complexity of drug resistance dynamics in high-incidence settings, and emphasize the importance of prioritizing testing algorithms which allow for early detection of INH-R.

Introduction

Despite being both a preventable and curable disease, more than 10 million people develop tuberculosis (TB) each year, with 1.4 million deaths in 2019 [1]. Although 63 million lives have been saved through improvements in programmatic TB management this century, the increase in drug-resistant (DR-TB) cases is increasingly concerning [1]. Multidrug-resistant TB (MDR-TB; defined as resistance to both first-line drugs isoniazid (INH) and rifampicin (RIF)) is a particular barrier to TB control efforts [2]. In 2019, 465,000 people were diagnosed with MDR-TB [1]. MDR-TB can be acquired by transmission (primary resistance) or develop *in vivo* through inadequate or incomplete treatment (secondary resistance), and the relative contribution of these mechanisms is likely to vary by context [3]. In all settings, though, careful optimization of both clinical and public health management of MDR-TB is required to ensure good outcomes.

Mathematical modeling is increasingly used to support programmatic optimization for TB [4-6]. Accounting appropriately for DR-TB in mathematical models of disease is critical, as it differs considerably from drug-sensitive TB (DS-TB) in both epidemiological parameters and relevant outcomes. Some variation in disease characteristics is relatively well-understood, including the prolonged treatment duration [7], adverse event rates [8] and diagnostic pathway performance [9, 10]. However, considerable uncertainties persist regarding important characteristics of MDR-TB, including pathogen's fitness, transmissibility and risk of resistance amplification related to treatment [11, 12]. Attempts to better characterize these features of MDR-TB have been challenging, in part due to the diversity of gene mutations which may confer resistance, many of which have limited clinical and epidemiological outcome data to inform model parameterization. Computational biological approaches have recently been used to bridge this gap, providing tools to estimate the fitness and resistance impact of even novel TB mutations [13-15].

Modeling also offers an opportunity to quantify amplification and transmission of drug-resistant TB, by fitting dynamic models to observed data. We therefore aimed to incorporate epidemiological data into an empirically calibrated model, in order to explore parameter estimation for drug resistance amplification and transmission associated with both INH and RIF.

Methods

2.1 Constructing the mathematical model and defining epidemiological parameters

We designed a deterministic compartmental model of *Mtb* transmission to capture five mutually exclusive health states with regards to TB infection and disease - susceptible (S), early latent (L_A), late latent (L_B), infectious (I) and recovered (R). The model included four TB strains: drug-

susceptible (DS-TB, compartment subscript S), isoniazid mono-resistant (INH-R, compartment subscript H), rifampicin mono-resistant (RIF-R, compartment subscript R) and MDR-TB (compartment subscript M). It is to be noted that the strains are not phylogenetically related.

We assumed homogenous mixing in a closed population:

$$N = S + L_{\mathrm{AS}} + L_{\mathrm{AH}} + L_{\mathrm{AR}} + L_{\mathrm{AM}} + L_{\mathrm{BS}} + L_{\mathrm{BH}} + L_{\mathrm{BR}} + L_{\mathrm{BM}} + I_{\mathrm{S}} + I_{\mathrm{H}} + I_{\mathrm{R}} + I_{\mathrm{M}} + R$$

All deaths are replaced as new births (rate π) entering the susceptible compartment. This includes both deaths due to TB disease (μ_i), as well as a universal population-wide death rate (μ).

When individuals in a population are infected with *Mtb*, they transition from the susceptible compartment (S) to the early latent compartment (L_A). The force of infection (λ) associated with each strain is defined as:

$$\lambda_X = r_X \times \beta \times I_X$$

where "x" indexes the drug resistance pattern – S, H, R or M. β is the "effective contact rate" for DS-TB, defined as the product of the average number of contacts between two individuals per unit time and the probability of DS-TB transmission per contact. The relative transmissibility of the different strains is denoted r_X and uses the DS-TB strain's transmissibility as reference ($r_S = 1$). In other words, r_X represents the TB strains' relative fitness.

People entering the early latent compartment (L_A) can either progress directly to the active disease compartment (I) at rate ε, or transition to the late latent compartment (L_B) at rate κ. Progression from L_B to the active disease state occurs at a much slower rate (ν), and is referred to as reactivation. Once individuals have entered the infectious compartment, one of the following six processes can occur: 1) the person may be correctly identified as having active TB and commenced on treatment (rate τ), thence progressing towards cure and transitioning to the recovered (R) compartment; 2) person may be correctly identified to have DS-TB or DR-TB and commenced on treatment but experiences treatment failure without experiencing resistance amplification to other drugs and stay in the same infectious compartment; 3) spontaneous recovery (rate γ) with transition to the recovered compartment (R); 4) TB-related death (μ_i) 5) dying of natural causes or 6) the infecting strain could acquire resistance (α_H and/or α_R) to isoniazid (INH-R), rifampicin (RIF-R) or MDR-TB and move to I_H, I_R and ultimately to I_M compartments. To capture the progressive accrual of resistance with each transition, only one level of additional resistance not already present can be obtained during a disease episode. People who have spontaneously recovered from past TB or successfully completed treatment are both represented as a single compartment (R) on the assumption that prognosis is equivalent regardless of the infecting strain from which each person recovers. Once treatment is complete,

the recovered person can transition back to L_A through reinfection, represented as δ. We define δ as:

$$\delta_X = RR_r * \lambda_X$$

where, RR_r is the "relative risk of re-infection once recovered"

Latently infected people also have a risk of re-infection with the same or other strains represented as θ in the model; and the re-infecting strain would "override" the existing strain. We define θ_x similarly to δ_x as:

$$\theta_X = RR_i * \lambda_X$$

where, RR_i is the "relative risk of re-infection once latently infected"

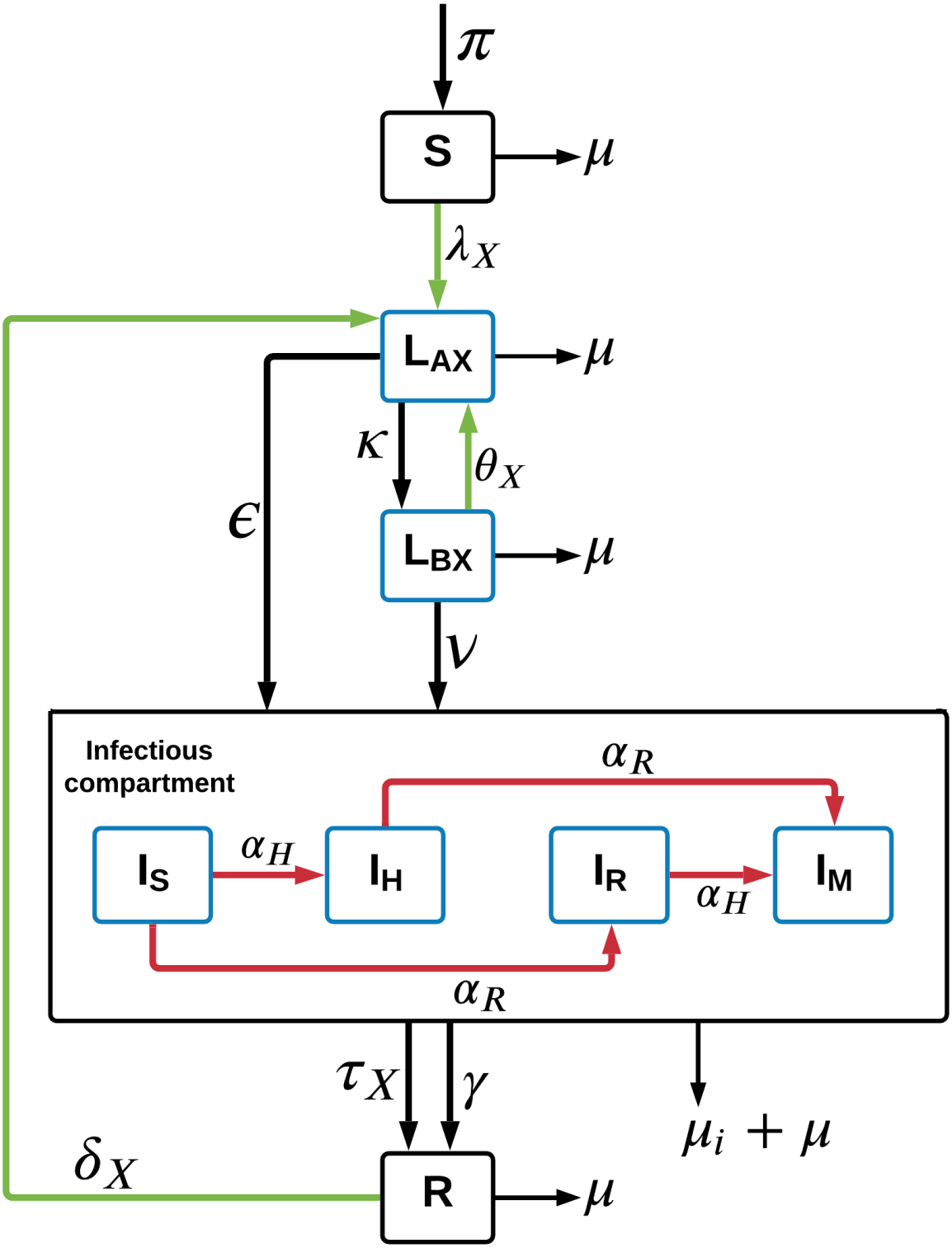

π
S
μ
λ_X
L_{AX}
μ
κ
θ_X
ϵ
L_{BX}
μ
ν
Infectious compartment
α_R
I_S
α_H
I_H
I_R
α_H
I_M
α_R
τ_X
γ
$\mu_i + \mu$
δ_X
R
μ

Figure 1: Structure of four strain Mtb transmission model. *The symbols S, L_A, L_B, I and R represent uninfected/susceptible, early latent, late latent, infected and recovered health states, respectively. The subscript "X" used in L_A and L_B compartments, indexes the drug resistance patterns, with S, H, R and M representing susceptible, isoniazid mono-resistance, rifampicin mono-resistance and multidrug resistance respectively. The infectious compartment is elaborated in the figure to show the amplification flows, parameterized with α_H and α_R (red arrows). The green arrows represent infection/transmission flows, black arrows represent constant progression flows. Compartments stratified according to resistance profiles are shown in blue.*

It is to be noted that the figure does not show individuals who are latently infected with a given strain will have the same strain if they develop active disease. An elaborated diagram in presented in the supplementary sheets where all the compartments modelled have been shown (S1).

Ordinary differential equations used to define the four-strain model

$$\frac{dS}{dt} = \pi - (\lambda_S + \lambda_H + \lambda_R + \lambda_M + \mu)\, S$$

$$\frac{dL_{AS}}{dt} = \lambda_S S - (\epsilon + \kappa + \mu)L_{AS} + \theta_S(L_{BS} + L_{BH} + L_{BR} + L_{BM}) + \delta_S R$$

$$\frac{dL_{AH}}{dt} = \lambda_H S - (\epsilon + \kappa + \mu)L_{AH} + \theta_H(L_{BS} + L_{BH} + L_{BR} + L_{BM}) + \delta_H R$$

$$\frac{dL_{AR}}{dt} = \lambda_R S - (\epsilon + \kappa + \mu)L_{AR} + \theta_R(L_{BS} + L_{BH} + L_{BR} + L_{BM}) + \delta_R R$$

$$\frac{dL_{AM}}{dt} = \lambda_M S - (\epsilon + \kappa + \mu)L_{AM} + \theta_M(L_{BS} + L_{BH} + L_{BR} + L_{BM}) + \delta_M R$$

$$\frac{dL_{BS}}{dt} = \kappa L_{AS} - (\nu + \theta_S + \theta_H + \theta_R + \theta_M + \mu)L_{BS}$$

$$\frac{dL_{BH}}{dt} = \kappa L_{AH} - (\nu + \theta_S + \theta_H + \theta_R + \theta_M + \mu)L_{BH}$$

$$\frac{dL_{BR}}{dt} = \kappa L_{AR} - (\nu + \theta_S + \theta_H + \theta_R + \theta_M + \mu)L_{BR}$$

$$\frac{dL_{BM}}{dt} = \kappa L_{AM} - (\nu + \theta_S + \theta_H + \theta_R + \theta_M + \mu)L_{BM}$$

$$\frac{dI_S}{dt} = \epsilon L_{AS} + \nu L_{BS} - \alpha_H I_S - \alpha_R I_S - (\gamma + \tau_S + \mu_i + \mu)I_S$$

$$\frac{dI_H}{dt} = \epsilon L_{AH} + \nu L_{BH} + \alpha_H I_S - \alpha_R I_H - (\gamma + \tau_H + \mu_i + \mu)I_H$$

$$\frac{dI_R}{dt} = \epsilon L_{AR} + \nu L_{BR} - \alpha_H I_R + \alpha_R I_S - (\gamma + \tau_R + \mu_i + \mu) I_R$$

$$\frac{dI_M}{dt} = \epsilon L_{AM} + \nu L_{BM} + \alpha_H I_R + \alpha_R I_H - (\gamma + \tau_M + \mu_i + \mu) I_M$$

$$\frac{dR}{dt} = (\tau_S + \gamma) I_S + (\tau_H + \gamma) I_H + (\tau_R + \gamma) I_R + (\tau_M + \gamma) I_M - (\delta_S + \delta_H + \delta_R + \delta_M + \mu) R$$

2.2 Parameter Estimation

An adaptive Metropolis algorithm was used to estimate model parameters [16], including drug resistance amplification rates. Parameters can be categorized as universal, country-specific and time-variant parameters, as presented in Table 1.

Universal parameters:

From the literature we gathered information on disease-specific and epidemiological parameters to calibrate the *Mtb* transmission model. We considered these parameters to be universal to all TB settings and so assigned the same values for all strains and settings (Table 1A).

Table 1: Epidemiological parameters used for calibrating the model and their prior distribution ranges.

A) Universal parameters

Parameter	Value	Prior distribution	Source
Early progression (ε) (year^{-1})	0.401775	Uniform [0.1 – 0.8]	[17]
Transition to late latency (κ) (year^{-1})	3.6525	Uniform [1.0 – 7.0]	[17]
Reactivation (ν) (year^{-1})	0.002008875	Uniform [0.0009, 0.006]	[17]
Spontaneous recovery (γ) (year^{-1})	0.2	Gamma [0.16, 0.29], mode = 0.20	[18]
Natural mortality (μ) (year^{-1})	0.0142		
TB-specific mortality (μ_i) (year^{-1})	0.2	Gamma [0.06, 1.06], mode = 0.08	[18]
Relative risk of reinfection once infected	0.21	-	[19]

B) Country-specific and time-variant parameters (used for model calibration)

Parameter	Country		Prior Distribution	Source
	The Philippines	Viet Nam		
Transmission rate (β)	[1 - 35]	[1 - 30]	Uniform	Fitted
Fitness cost of INH-R TB strain	[0.50 – 1.20]		Uniform	[20], [21]
Fitness cost of RIF-R TB strain	[0.50 – 1.20]		Uniform	[22], [12]
Fitness cost of MDR-TB strain	[0.50 – 0.99]		Uniform	[23]
Proportion of failures developing RIF-R TB (ρ_R)	[0.01 – 0.99]		Uniform	Fitted
Proportion of failures developing INH-R TB (ρ_H)	[0.01 – 0.99]		Uniform	Fitted
Relative risk of reinfection once recovered	[0.50 – 1.50]		Uniform	Fitted
CDR start time	[1950 -1970]		Uniform	Fitted
CDR final value	[0.30 – 0.80]		Uniform	Fitted

(*CDR – Case detection rate)

Defining time-variant model processes

To capture the rise of drug resistance over time, we allowed the case detection rate (CDR, a proportion) and the treatment success rate (TSR) to vary over time. People diagnosed with active TB are commenced on treatment upon identification and move from the infectious compartments (I, I_H, I_R and I_M) to the recovered compartment (R). The transition from the infectious to the recovered compartment is represented using the parameter "τ". τ is dependent on the TB detection rate "d" and the treatment success rate (TSR) and is mathematically expressed as -

$$\tau\,(t) = d(t) * \mathrm{TSR}\,(t)$$

where, "d" is calculated by solving the following CDR equation:

$$CDR(t) = \frac{d(t)}{d(t) + \gamma + \mu_i + \mu}$$

which results in:

$$d(t) = \frac{CDR(t)}{(1 - CDR(t)) * (\gamma + \mu_i + \mu)}$$

TSR is the probability of a person being first tested and ultimately put on treatment to be cured, or simply put the probability of treatment success at presentation. This parameter was varied by strain. We used sigmoidal functions to model progressive increases for both the CDR and the TSR between 1950 and 2020. The final value of the TSR was set to the most recent TSR estimate reported by the WHO. In contrast, the final value of the CDR was varied during calibration. This allowed flexibility in simulating the historical dynamics of TB control in the countries considered.

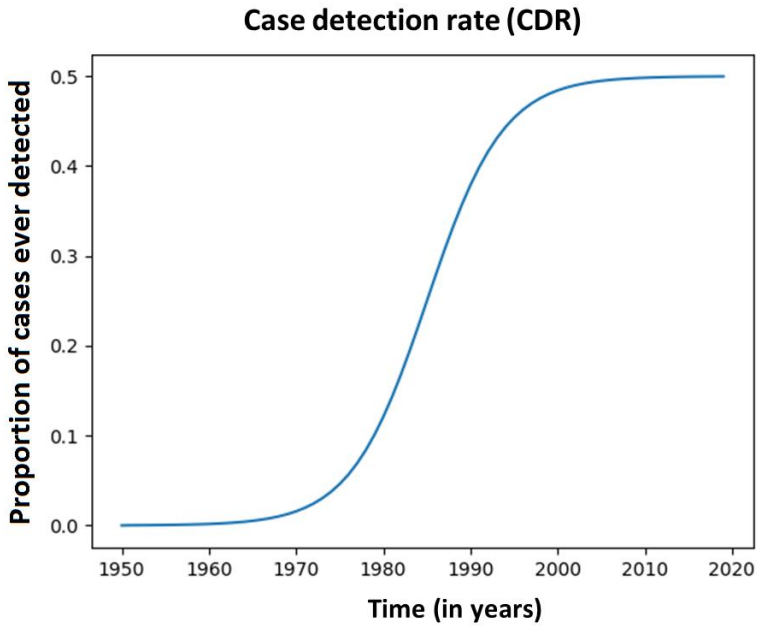


Figure 2: Example of time-variant case detection rate (CDR) (final value=50%).

Defining the amplification rate

Treatment for tuberculosis begins once individuals are detected with TB and the TB strain is correctly identified. Treatment then proceeds and may result in three possible outcomes: death, successful treatment or treatment failure. Treatment failure can further be associated with new acquisition of resistance to one additional drug that was not previously present in the infecting organism. INH and RIF are part of the standard regimen for the treatment of drug-susceptible strains. Gain in resistance to either INH or RIF is represented using amplification rates α_H or α_R respectively in the model. Mathematical representation of INH and RIF mono-resistant amplification is shown as -

$$\text{Rate of amplification } (\alpha_H) = d\,(t) * \big(1 - \text{TSR}\,(t)\big) * \rho_H$$
$$\text{Rate of amplification } (\alpha_R) = d\,(t) * \big(1 - \text{TSR}\,(t)\big) * \rho_R$$

where,
ρ_H = Proportion of previously INH-susceptible individuals that acquire resistance on treatment failure, and
ρ_R = Proportion of previously RIF-susceptible individuals that acquire resistance on treatment failure

2.3 Model calibration to prevalence and notification data

Prevalence data

The model presented above was calibrated to country-specific data. We fitted the models using TB prevalence estimates from national TB prevalence surveys (Viet Nam: 2006-2007 and 2017-2018; The Philippines: 2007 and 2016) and drug-resistance prevalence from national DR-TB surveys (Viet Nam: 2011; The Philippines: 2009, 2016). The detailed estimates are presented in Table 2:

Table 2: Summaries of prevalence survey results and drug resistance survey data for Philippines and Viet Nam.

A) TB Prevalence data

Country	Year	TB prevalence (per 100, 000)	95% CI	Source
Viet Nam	2006-2007	307.2	248.8 – 365.6	[24]
	2017-2018	322	260 – 399	[25]
The Philippines	2007	660	510-810	[26]
	2016	1159	1016-1301	[27]

B) Drug resistance data

Country	Drug resistance	Year	Drug resistance (%)	95% CI	Source
Viet Nam	Isoniazid mono resistance	2011	14.86	12.15 – 17.56	[28]
	Rifampicin mono resistance	2011	0.23	0.1 – 0.35	[28]
	MDR-TB	2011	6.93	4.22 – 9.63	[28]
	Isoniazid mono resistance	2009	9.44	7.95 - 10.92	[29]

The Philippines	Rifampicin mono resistance	2009	1.008	0.71 - 1.304	[29]
	MDR-TB	2009	5.8	4.3 – 7.5	[29]
The Philippines	Isoniazid mono resistance	2016	12.43	11.1 - 13.75	[27]
	Rifampicin mono resistance	2016	0.82	0.44 - 1.19	[27]
	MDR-TB	2016	3.35	2.53 - 4.41	[27]

Notification data: We used WHO-reported TB notifications as a calibration target for both models. For Viet Nam, in 2018, 102,171 cases were notified and for The Philippines 382,543 cases were notified and we calibrated to the per capita notification rates corresponding to these values.

Uncertainty analysis

Once we defined the parameters in our model, we next reviewed literature for information on the prior distributions of uncertain parameters (Table 1B).

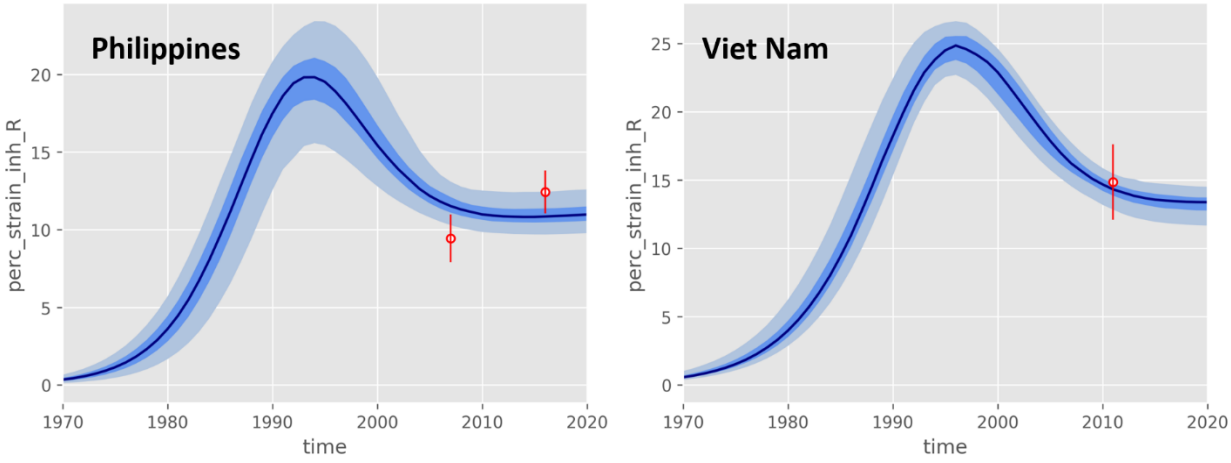
The adaptive Metropolis algorithm [16] was used to generate samples from the posterior distribution of the parameters from 25,000 iterations for each country. The primary estimates are reported as the posterior median value for all parameters of interest such as amplification proportions, CDR, relative fitness of each modelled strain and the relative risk of infection once recovered (δ). The intervals reported are obtained by calculating the 25^{th} and 75^{th} percentile of each parameter's posterior distribution. Programming was done in Python 3.7.3 and all code and associated data are publicly available on GitHub (github.com/malanchak/AuTuMN).

Results

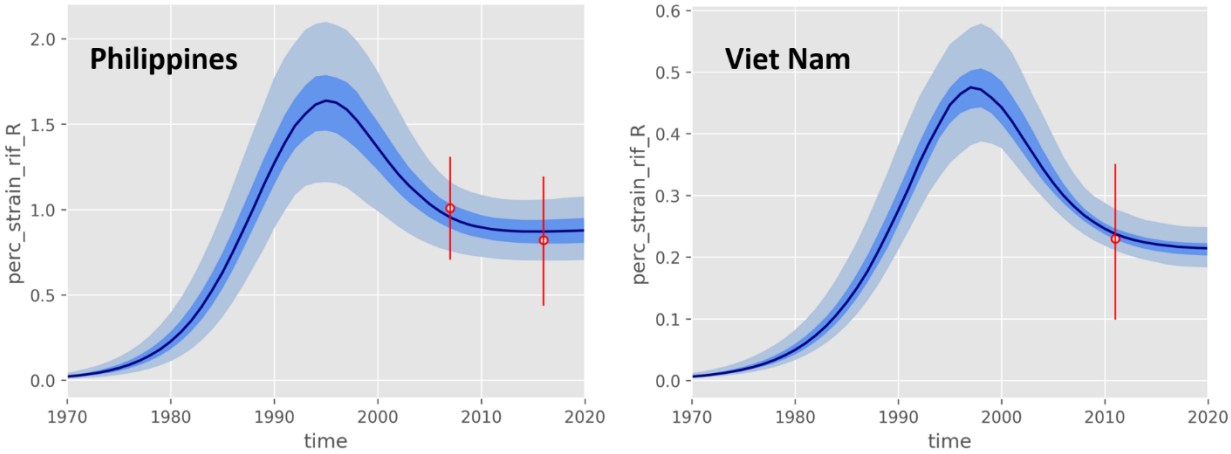
Calibration of the model

Figure 3A, 3B and 3C show the model fits to reported INH-R, RIF-R and MDR levels for the high DR-TB settings, The Philippines and Viet Nam respectively.

A) Calibration fit for isoniazid resistance (%)



B) Calibration fit for rifampicin resistance (%)



C) Calibration fit for MDR (%)

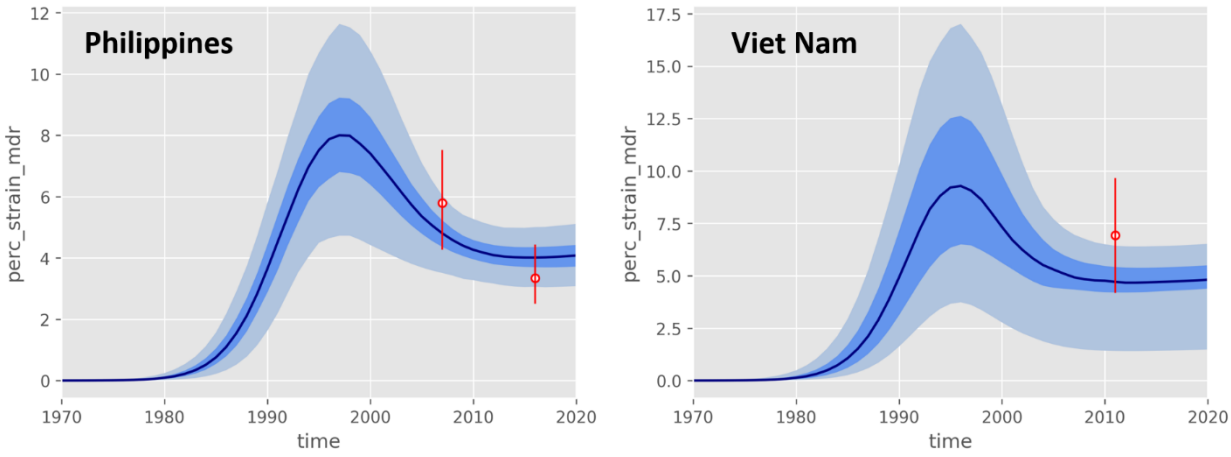


Figure 3: Model calibration*: A) Isoniazid mono resistance B) Rifampicin mono resistance and C) MDR-TB. The red dots with the line represent the empiric data (including intervals) obtained from the drug resistance surveys of the Philippines and Viet Nam. The model predictions are represented in blue solid line as median, interquartile range (dark blue shade) and central 95% credible interval (light blue shade).*

Table 3 shows the posterior distributions of all calibrated parameters.

Table 3: Posterior distribution of parameters obtained using the Bayesian analysis

DR- TB related parameter	**Estimate (median, 50 % CI)**	
	The Philippines	**Viet Nam**
Proportion of previously INH-susceptible individuals that acquire resistance on treatment failure	0.84 (0.79-0.89)	0.77 (0.71 – 0.84)
Proportion of previously RIF-susceptible individuals that acquire resistance on treatment failure	0.05 (0.04 - 0.07)	0.011 (0.010 – 0.012)
Relative fitness of INH-R TB strains	0.87 (0.83 – 0.92)	0.98 (0.95 – 1.00)
Relative fitness of RIF-R TB strains	0.78 (0.74 – 0.84)	0.77 (0.73 – 0.81)
Relative fitness of MDR-TB strains	0.67 (0.58 – 0.71)	0.64 (0.56 – 0.75)
CDR final/maximum value	0.49 (0.47 – 0.51)	0.66 (0.63 – 0.69)

Universal parameters	**Estimate (median, 50 % CI)**	
	The Philippines	**Viet Nam**
Rate of rapid progression (ε) (year-1)	0.33 (0.28 – 0.37)	0.22 (0.19 – 0.29)
Rate of transition towards late latency (κ) (year-1)	5.49 (4.78 – 5.95)	3.62 (3.13 – 4.92)
Rate of re-activation (ν) (year-1)	0.003 (0.002 – 0.004)	0.0017 (0.0016 –0.0018)
Relative risk of re-infection after recovery (δ)	0.74 (0.64 – 0.86)	0.64 (0.56 – 0.81)

Table 4: Estimates obtained for proportions of incident DR-TB due to direct transmission rather than DR amplification

DR-TB	**Estimate (median, 50 % CI)**	
	The Philippines	Viet Nam
INH-R TB	50 (43 – 70)	67 (54 – 73)
RIF-R TB	52 (43 – 70)	63 (55 – 71)
MDR TB	40 (28 – 52)	43 (34 – 51)

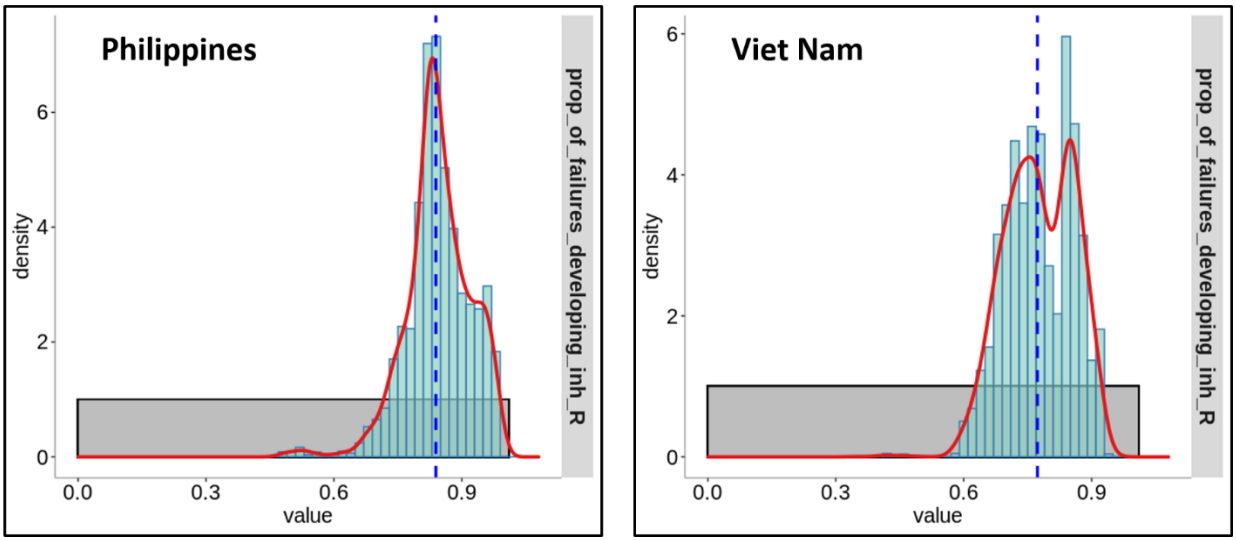
Drug resistance amplification and transmission

We observed higher proportions of drug resistance amplification for INH compared to RIF for both the high DR-TB incidence settings we simulated (Figure 4). The estimated risk of INH-R amplification when treatment fails was 0.84 (95% CI 0.79 – 0.89) for The Philippines and 0.77 (95% CI 0.71 – 0.84) for Viet Nam. The estimated risk of RIF-R acquisition when treatment fails was 0.05 (95% CI 0.04 – 0.07) for The Philippines and 0.011 (95% CI 0.010 – 0.012) for Viet Nam. This meant approximately 84% and 77% of the people who failed treatment in The Philippines and Viet Nam respectively would end up with resistance to INH.

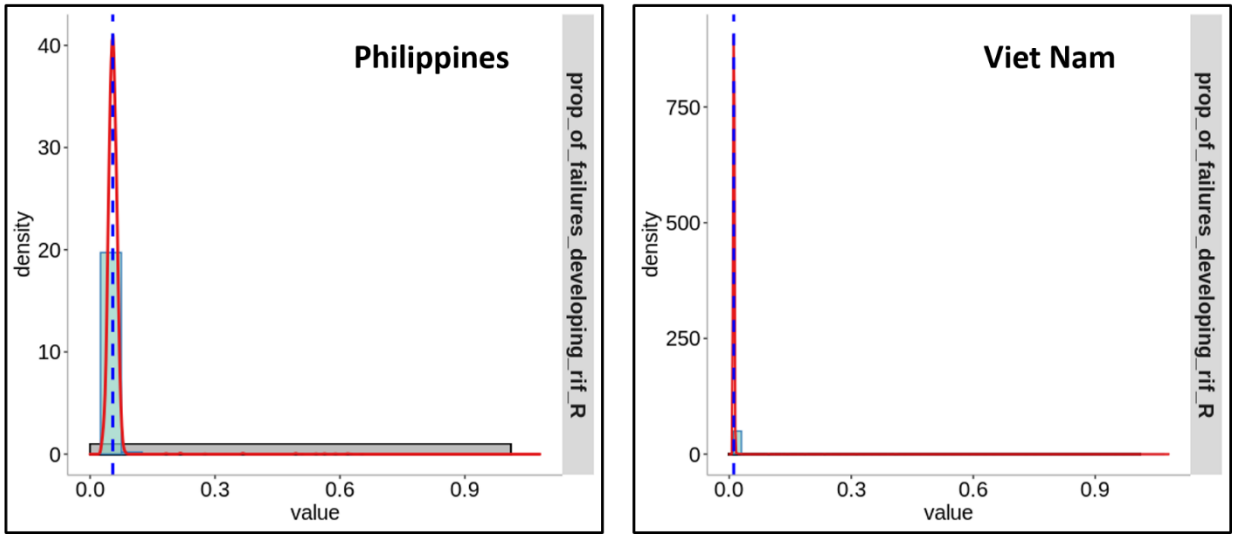
In the Philippines, the proportions of incident INH-R TB due to transmission was 50% (43 – 70), RIF-R TB was 52% (43 – 70) and MDR-TB was 40% (28 – 52). For Viet Nam, the proportions of incident INH-R TB due to transmission was 67% (54 – 73), RIF-R TB was 63% (55 – 71) and MDR-TB was 43% (34 – 51).

In The Philippines, the model estimates for amplification from DS to INH-R was 26 per 100,000 (95% CI 15 – 33) people, followed by 10 per 100,000 (95% CI 6 – 14) people then gain resistance to RIF and moving to the MDR compartment. Comparing this to acquiring RIF resistance first, we see 2 per 100,000 (95% CI 1 – 3) moving from DS to RIF-R, followed by only 0.05 (95% CI 0.02 – 0.08) people gaining resistance to INH to move to the MDR compartment. A similar observation was seen for Viet Nam, the model estimates for amplification from DS to INH-R shows 6 per 100,000 (95% CI 4 – 8) people followed by people 4 per 100,000 (95% CI 3 – 5) gaining resistance to RIF and moving to the MDR compartment. In case of DS to RIF-R transition the estimates were 0.08 per 100,000 (0.06 – 0.11) people, followed by 0.0007 per 100,000 (0.0004 – 0.001) people gaining resistance to INH to move to the MDR compartment.

A)



B)



<u>Figure 4: The estimated risk of INH-R and RIF-R amplification when treatment fails.</u> The probability density function (red line) represents the posterior distribution of the estimates of amplification and the white background represents the prior ranges. The dashed blue line is the median of the estimates. A) Proportion of previously INH-susceptible strains that acquire resistance on treatment failure and B) Proportion of previously RIF-susceptible strains that acquire resistance on treatment failure.

<u>Estimates for relative strain fitness and CDR</u>

The posterior estimates of relative fitness associated with INH-R strains for The Philippines was 0.87 (95% CI 0.83 – 0.92) and 0.98 (95% CI 0.95 – 1.00) for Viet Nam. The relative fitness associated with RIF-R strains for The Philippines was 0.78 (95% CI 0.74 – 0.84) and 0.77 (95% CI 0.73 – 0.81) for Viet Nam. The relative fitness associated with MDR-TB strains in The Philippines was 0.67 (95% CI 0.58 – 0.71) compared to 0.64 (95% CI 0.56 – 0.75) for Viet Nam.

Our study also provided information on estimates of CDR with high precision for both the settings, as inclusion of notification and prevalence of infection data for the analysis helped in constraining the parameter. The estimates obtained for The Philippines was 0.49 (95% CI 0.47 – 0.51) and for Viet Nam was 0.66 (95% CI 0.63 – 0.69).

<u>Discussion</u>

From this modeling study we were able to construct a model which successfully replicated epidemiological dynamics in two higher burden TB settings, incorporating parameters drawn from microbiological fitness data. Using this model to explore the development of drug resistance in these contexts, we found that a much higher proportion of treatment failure resulted in amplification for INH-R rather than for RIF-R. This finding is consistent with observed higher rates of INH-R globally and allows consideration of factors which might be mechanistically important for understanding and planning a programmatic response.

One factor likely to play a significant role in preferential INH-R amplification is current methods of DR-TB detection which prioritize RIF's resistance identification. According to WHO and many country guidelines, TB patients with strains found to be resistant to RIF need to start on a recommended MDR-TB treatment regimen. Longer MDR-TB regimens, and historical second-line therapy regimens, frequently have INH included in them, irrespective of resistance to INH being either undetermined or confirmed. Re-treatment regimens in particular have often incorporated prolonged durations of INH therapy – for example the category II regimen used in the Philippines comprised of 8 months of INH, RIF and ethambutol supplemented by streptomycin for the initial 2 months, and pyrazinamide for the initial 3 months (2SHRZE/HRZE/ 5HRE)[30], and older treatment regimens used in Viet Nam comprised of 8 months of INH and ethambutol supplemented by initial two months of streptomycin, pyrazinamide and rifampicin (2SHRZ/6HE) [28]. These factors may be further amplified by use of isoniazid in the private sector and/or through community pharmacy settings, where worse guideline adherence and increased risk resistance development has been shown[31, 32] but with poorer treatment outcomes compared to NTPs [33, 34].

In addition to programmatic insights, our model provides novel information on parametrizing CDR. This is important, as this parameter cannot be measured directly yet plays a significant role in informing robust mathematical model of TB transmission. As with any mathematical representation our model has certain limitations. Our model was calibrated to TB prevalence and DR surveys to estimate the risk of resistance amplification. But the definition of a TB case may change between surveys, even within the same country. We have adopted a simplified model structure that does not capture factors such as age, comorbidities and other heterogeneity associated with TB epidemics. These factors may affect the risk of resistance amplification. Our model is primarily built for pulmonary tuberculosis and does not include extra-pulmonary TB data, as our primary focus was on transmission. For the same reason, this model has been parametrized from adult TB data given the limited TB transmission from young children to others. In our model we assumed the risk of INH-R amplification is the same starting from I_S, as compared to starting from I_R; the same applies for RIF-R amplification. We even assumed the fitness cost of

MDR-TB is independent of that of INH-R of RIF-R. Therefore, these limitations can potentially influence the estimated risk of resistance amplification.

Historically, diagnosis of MDR-TB has been reliant on culture-based phenotypic testing, which in high-burden settings may be applied selectively, such as after treatment failure. As part of the global policy to control DR-TB, many high burden settings have pledged to deploy the molecular diagnostic assay Xpert MTB/RIF (detects resistance only in RIF), which is a nucleic acid amplification test that can be directly applied to sputum samples [35] [36]. As the presence of RIF resistance is highly predictive of MDR-TB, these policies have led to significant improvements in the appropriate initiation of second-line therapy [37]. However, as our work highlights, these algorithms may also be associated with selecting for and further amplifying INH resistance. Alternative molecular tools, such as the line probe assay MTBDR*plus* [38] or Xpert MTB/RIF Ultra [39] can identify both RIF and INH resistance, and may offer the programmatic advantages of rapid MDR-TB diagnosis while avoiding this secondary effect [40]. Further research into the association between specific INH resistance mutations and differential risk of transmission will be helpful in better defining the public health impact of this effect [41].

While rapid molecular diagnostics will continue to be important for programmatic adoption, it is also important to recognize that the principle of unrecognized resistance amplification demonstrated here can be repeated for any resistance not routinely addressed in diagnostic algorithms. It is therefore essential to incorporate genome sequencing into surveillance programs, to maximize the clinical and public health benefits [42]. With recent developments in next generation sequencing techniques, we have now have high-throughput diagnostic tools for the detection of DR-TB which are both fast and efficient [43]. While such tools are currently in routine use only in high resource settings, the benefits associated with these tools should be prioritized for high burden contexts to support optimal individual and program outcomes [44, 45].

References

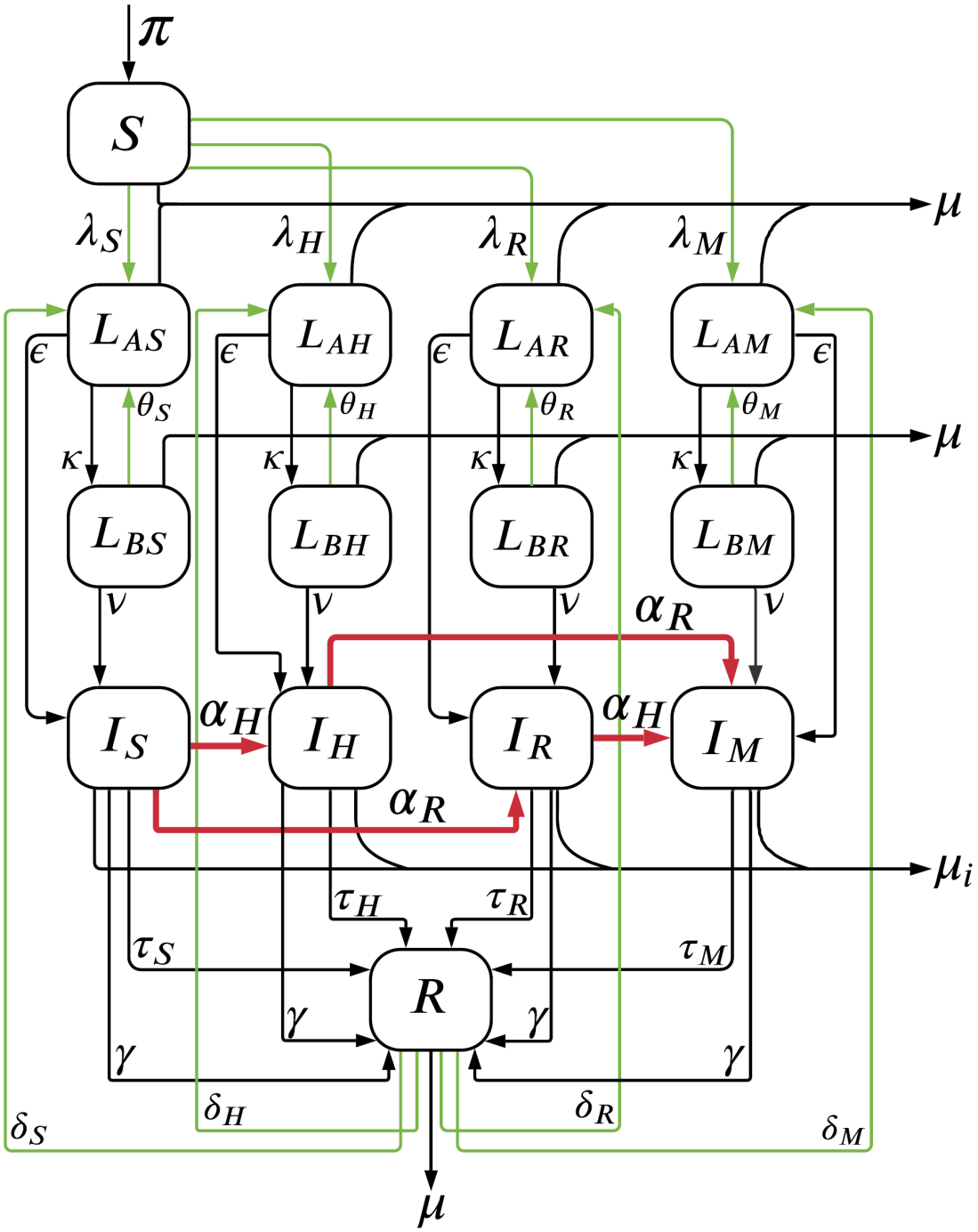
1. WHO, *Global Tuberculosis Report 2020.* https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2020, 2020.
2. WHO, *WHO consolidated guidelines on drug-resistant tuberculosis treatment.* 2019.
3. Ragonnet, R., et al., *High rates of multidrug-resistant and rifampicin-resistant tuberculosis among re-treatment cases: where do they come from?* BMC Infectious Diseases, 2017. **17**(1): p. 36.
4. Trauer, J.M., et al., *Modular programming for tuberculosis control, the "AuTuMN" platform.* BMC Infectious Diseases, 2017. **17**(1): p. 546.
5. Zwerling, A., S. Shrestha, and D.W. Dowdy, *Mathematical Modelling and Tuberculosis: Advances in Diagnostics and Novel Therapies.* Advances in Medicine, 2015. **2015**: p. 907267.
6. Fors, J., et al., *Mathematical model and tool to explore shorter multi-drug therapy options for active pulmonary tuberculosis.* PLoS Comput Biol, 2020. **16**(8): p. e1008107.
7. Pontali, E., M.C. Raviglione, and G.B. Migliori, *Regimens to treat multidrug-resistant tuberculosis: past, present and future perspectives.* European Respiratory Review, 2019. **28**(152): p. 190035.
8. Herrera, M., et al., *Modeling the Spread of Tuberculosis in Semiclosed Communities.* Computational and Mathematical Methods in Medicine, 2013. **2013**: p. 648291.
9. Wikell, A., et al., *Diagnostic pathways and delay among tuberculosis patients in Stockholm, Sweden: a retrospective observational study.* BMC Public Health, 2019. **19**(1): p. 151.
10. Naidoo, P., et al., *Pathways to multidrug-resistant tuberculosis diagnosis and treatment initiation: a qualitative comparison of patients' experiences in the era of rapid molecular diagnostic tests.* BMC health services research, 2015. **15**: p. 488-488.
11. Becerra, M.C., et al., *Transmissibility and potential for disease progression of drug resistant Mycobacterium tuberculosis: prospective cohort study.* BMJ, 2019. **367**: p. l5894.
12. Knight, G.M., et al., *The Distribution of Fitness Costs of Resistance-Conferring Mutations Is a Key Determinant for the Future Burden of Drug-Resistant Tuberculosis: A Model-Based Analysis.* Clin Infect Dis, 2015. **61Suppl 3**(Suppl 3): p. S147-54.
13. Karmakar, M., et al., *Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy.* Am J Respir Crit Care Med, 2018. **198**(4): p. 541-544.
14. Karmakar, M., et al., *Structure guided prediction of Pyrazinamide resistance mutations in pncA.* Sci Rep, 2020. **10**(1): p. 1875.
15. Karmakar, M., et al., *Empirical ways to identify novel Bedaquiline resistance mutations in AtpE.* PLoS One, 2019. **14**(5): p. e0217169.
16. Haario, H., E. Saksman, and J. Tamminen, *An Adaptive Metropolis Algorithm.* Bernoulli, 2001. **7**(2): p. 223-242.
17. Ragonnet, R., et al., *Optimally capturing latency dynamics in models of tuberculosis transmission.* Epidemics, 2017. **21**: p. 39-47.
18. Ragonnet, R., et al., *Revisiting the Natural History of Pulmonary Tuberculosis: a Bayesian Estimation of Natural Recovery and Mortality rates.* Clin Infect Dis, 2020.
19. Trauer, J.M., et al., *Risk of Active Tuberculosis in the Five Years Following Infection . . . 15%?* Chest, 2016. **149**(2): p. 516-525.
20. Cohen, T., B. Sommers, and M. Murray, *The effect of drug resistance on the fitness of Mycobacterium tuberculosis.* Lancet Infect Dis, 2003. **3**(1): p. 13-21.
21. Borrell, S. and S. Gagneux, *Infectiousness, reproductive fitness and evolution of drug-resistant Mycobacterium tuberculosis.* Int J Tuberc Lung Dis, 2009. **13**(12): p. 1456-66.

22. Gagneux, S., *Fitness cost of drug resistance in Mycobacterium tuberculosis.* Clin Microbiol Infect, 2009. **15 Suppl 1**: p. 66-8.
23. Cohen, T. and M. Murray, *Modeling epidemics of multidrug-resistant M. tuberculosis of heterogeneous fitness.* Nat Med, 2004. **10**(10): p. 1117-21.
24. Hoa, N.B., et al., *National survey of tuberculosis prevalence in Viet Nam.* Bull World Health Organ, 2010. **88**(4): p. 273-80.
25. Nguyen, H.V., et al., *The second national tuberculosis prevalence survey in Vietnam.* PLoS One, 2020. **15**(4): p. e0232142.
26. Tupasi, T.E., et al., *Significant decline in the tuberculosis burden in the Philippines ten years after initiating DOTS.* Int J Tuberc Lung Dis, 2009. **13**(10): p. 1224-30.
27. Department of Health, R.o.T.P., *National Tuberculosis Prevalence Survey 2016 Philippines.* 2016. **http://www.ntp.doh.gov.ph/downloads/publications/Philippines_2016%20National%20TB%20Prevalence%20Survey_March2018.pdf**.
28. Nhung, N.V., et al., *The fourth national anti-tuberculosis drug resistance survey in Viet Nam.* Int J Tuberc Lung Dis, 2015. **19**(6): p. 670-5.
29. *Nationwide drug resistance survey of tuberculosis in the Philippines.* Int J Tuberc Lung Dis, 2009. **13**(4): p. 500-7.
30. Chiang, C.-Y. and A. Trébucq, *Tuberculosis re-treatment after exclusion of rifampicin resistance.* European Respiratory Journal, 2018. **51**(2): p. 1702282.
31. Quy, H.T., et al., *Public-private mix for improved TB control in Ho Chi Minh City, Vietnam: an assessment of its impact on case detection.* Int J Tuberc Lung Dis, 2003. **7**(5): p. 464-71.
32. Tupasi, T.E., et al., *Bacillary disease and health seeking behavior among Filipinos with symptoms of tuberculosis: implications for control.* Int J Tuberc Lung Dis, 2000. **4**(12): p. 1126-32.
33. Lönnroth, K., et al., *Private tuberculosis care provision associated with poor treatment outcome: comparative study of a semi-private lung clinic and the NTP in two urban districts in Ho Chi Minh City, Vietnam. National Tuberculosis Programme.* Int J Tuberc Lung Dis, 2003. **7**(2): p. 165-71.
34. Buu, T.N., K. Lönnroth, and H.T. Quy, *Initial defaulting in the National Tuberculosis Programme in Ho Chi Minh City, Vietnam: a survey of extent, reasons and alternative actions taken following default.* Int J Tuberc Lung Dis, 2003. **7**(8): p. 735-41.
35. Tsara, V., E. Serasli, and P. Christaki, *Problems in diagnosis and treatment of tuberculosis infection.* Hippokratia, 2009. **13**(1): p. 20-2.
36. Campbell, E.A., et al., *Structural mechanism for rifampicin inhibition of bacterial rna polymerase.* Cell, 2001. **104**(6): p. 901-12.
37. Dlamini, M.T., et al., *Whole genome sequencing for drug-resistant tuberculosis management in South Africa: What gaps would this address and what are the challenges to implementation?* Journal of clinical tuberculosis and other mycobacterial diseases, 2019. **16**: p. 100115-100115.
38. Nathavitharana, R.R., et al., *Multicenter Noninferiority Evaluation of Hain GenoType MTBDRplus Version 2 and Nipro NTM+MDRTB Line Probe Assays for Detection of Rifampin and Isoniazid Resistance.* J Clin Microbiol, 2016. **54**(6): p. 1624-1630.
39. Chakravorty, S., et al., *Detection of Isoniazid-, Fluoroquinolone-, Amikacin-, and Kanamycin-Resistant Tuberculosis in an Automated, Multiplexed 10-Color Assay Suitable for Point-of-Care Use.* J Clin Microbiol, 2017. **55**(1): p. 183-198.
40. Talbot, E.A. and M. Pai, *Tackling drug-resistant tuberculosis: we need a critical synergy of product and process innovations.* Int J Tuberc Lung Dis, 2019. **23**(7): p. 774-782.
41. Fregonese, F., et al., *Comparison of different treatments for isoniazid-resistant tuberculosis: an individual patient data meta-analysis.* Lancet Respir Med, 2018. **6**(4): p. 265-275.

42. Dunstan, S.J., D.A. Williamson, and J.T. Denholm, *Understanding the global tuberculosis epidemic: moving towards routine whole-genome sequencing.* Int J Tuberc Lung Dis, 2019. **23**(12): p. 1241-1242.
43. Mahomed, S., et al., *Whole genome sequencing for the management of drug-resistant TB in low income high TB burden settings: Challenges and implications.* Tuberculosis (Edinb), 2017. **107**: p. 137-143.
44. Luo, T., et al., *Whole-genome sequencing to detect recent transmission of Mycobacterium tuberculosis in settings with a high burden of tuberculosis.* Tuberculosis (Edinb), 2014. **94**(4): p. 434-40.
45. Sulis, G. and M. Pai, *Isoniazid-resistant tuberculosis: A problem we can no longer ignore.* PLoS Med, 2020. **17**(1): p. e1003023.

Supplementary Data

S1: A detailed representation of the Mtb transmission model.



CHAPTER 7: CONCLUSION

The spread of anti-microbial resistance in TB is a major challenge for clinical care and global public health, and so optimal understanding and control of TB epidemics is of international interest. Current strategies to control TB are primarily aimed at reducing transmission through rapid identification of infectious patients by deploying fast and accurate diagnostic measures, followed by treatment with effective drugs according to resistance detected through these tools [233]. During my doctoral thesis I have demonstrated computational protein structural tools can be useful for predicting drug resistance, optimizing treatment regimens, and informing models of drug resistance emergence. These tools have an important role in strengthening the translational value of TB genomics, allowing better directed and more effective programmatic responses in the face of expanding drug resistance.

Isoniazid, rifampicin, ethambutol and pyrazinamide are the four main first-line drugs used to treat DS-TB While the first three drugs are mainly responsible for killing the activating replicating bacterium, pyrazinamide kills the dormant bacilli. PZA being a sterilizing drug, is generally part of majority of the regimens to treat TB but is associated with dangerous side-effects. PZA is a pro-drug and is converted into its active form pyrazinoic acid with the help of the enzyme PncA (pyrazinamidase). Between 70 – 90% of mutations responsible for resistance in pyrazinamide are harbored in the pncA gene. The WHO recommended phenotypic DST to determine resistance associated with pyrazinamide is the Wayne and Bactec MGIT 960 methods. These methods, however, have poor reproducibility, are labor intensive and need specialised laboratory set up to conduct the tests. Looking closely into the protein structure of PncA and mapping the mutations on to it helped in understanding the discrepancies associated with the DST results. Scores were generated from biophysical and evolutionary analyses to observe differential patterns between susceptible and resistant non-synonymous missense mutations. This information was used to train a supervised machine learning algorithm to develop an empirical classifier that could identify novel

drug resistance mutations with 80% accuracy in pncA. The classifier has been deployed as a freely available user-friendly webserver SUSPECT-PZA. This could be readily implemented in genomic sequencing pipelines as a potential tool to determine resistance associated with pncA in resource limited settings that have higher TB burden. Our tool can be used as an alternative to help in faster and accurate determination of resistance thereby helping in reducing the spread of drug resistance.

As progress was being made developing the novel pipeline to determine novel resistance mutations in pncA, a case was presented from the Royal Melbourne Hospital, Melbourne, Australia. A 42-year old Vietnamese woman was diagnosed with MDR-TB and she was receiving a cocktail of medication which included PZA. Whole genome sequencing identified a novel frameshift mutation in the pncA gene. To determine effectiveness of the drug PZA a real time analysis was carried out to validate whether the novel frameshift mutation lead to loss of function of the protein. The analysis revealed the frameshift mutation lead to a stop codon at the 29th amino acid position of the protein structure. Due to incomplete synthesis, the mutant protein lacked the catalytic binding site which is required to bind the pro-drug and convert it into its active form. Thus, it led to altered patient treatment and was the first reported use of structural information to guide clinical resistance detection.

MDR and XDR-TB involves long, expensive and difficult to manage treatment regimens with unfortunately sub-optimal outcomes in most cohorts [234]. An effective treatment regimen for MDR-TB involves stepwise selection of second line drugs depending on the DST results. It is often seen at programmatic levels that either the second line drugs or the facility to perform the DST is unavailable [235]. Moreover, adherence to treatment due to severe adverse effects of second-line drugs is another other major issue [236]. New drugs have been developed and approved in the past few years to treat MDR/XDR-TB like Bedaquiline and Delamanid. Bedaquiline binds to the c-subunit of ATP synthase (atpE). Clinical and *in vitro* resistance has been detected in atpE and Rv0678, a transcriptional repressor of the MmpS5-MmpL5 efflux pump [237]. Despite this, a standard protocol to determine *in vitro* susceptibility of bedaquiline has not been developed and agreed upon [79]. Being an important drug,

which is effective against both DS and DR-TB, which in future could replace both INH and RIF, it is necessary to urgently develop a protocol to determine its susceptibility. The same structural pipeline was used to develop a predictive tool for the drug target atpE for bedaquiline. Recognising the lack of published data correlating bedaquiline testing with clinical outcomes, an alternative approach was developed for determine the susceptible variants (explained in chapter 5). This work led to establishing the first computational tool which could determine novel drug resistance in bedaquiline and can in slowing down the rise of drug resistance towards bedaquiline

While building these predictive models an important observation was made with respect to lineages and its impact on the susceptibility of a variant. Some mutations had different effects on susceptibility depending on underlying lineage. Beijing lineage, which is a sub-lineage of Lineage 2 (East Asian) has been under intense scrutiny in recent years, as it has been reported to be found globally and associated with higher rates of drug resistance. A thorough systematic review followed by meta-analysis was conducted and Beijing was found to be hyper-transmissible compared to non-Beijing strains. The finding was epidemiologically significant as clustering data from molecular genotyping methods was used to carry out the investigation. The study established the fact that Beijing lineage had a higher propensity to cause disease and transmit within different geographical settings. It further helped in supporting the idea of incorporation of lineage specific information in the empirical tools I build to make them more powerful and accurate. In future, further exploration will help in understanding the underlying causes which helps Beijing to be more transmissible than the other lineages and how this information can be used to stop its spread.

Mathematical models used to study emergence and control of DR-TB at population level are important tools to understand the natural history of the disease, helps in choosing the correct available intervention and highlights areas where additional research is required [168]. The molecular data gathered during the protein structural studies provided information for developing a four-strain epidemiological compartmental model to estimate proportion of drug resistance amplification for two high drug resistant

TB-incidence settings. The model provided posterior estimates of mono drug resistant amplifications and transmission rates for INH and RIF. RIF resistant strains had a higher propensity to be transmitted compared to acquiring resistance through amplification. In contrast, INH resistant strains had similar rates for transmission and amplification. These findings highlight complexity involved with drug resistance dynamics in high TB burden settings. This leads to the next question whether deploying GeneXpert plus or line probe assay (which can detect resistance in both INH and RIF) over GeneXpert MTB/RIF (detects resistance only in RIF) worldwide could help in bringing down the higher rates of drug resistance amplification in INH and better control the MDR-TB epidemics.

In this work, I have explored how computational tools can improve our clinical care, public health response, and understanding of tuberculosis. With the reduction in the cost to perform WGS, individualized therapy could become available to all TB patients including those diagnosed with DR-TB. Though bringing such therapy to resource limited high burden settings is still a challenge; SUSPECT-PZA and SUSPECT-BDQ could aid in preliminary identification and accurate determination of the susceptibility of drugs in these settings and can be included in their national programs. A further improvement which can be introduced to the updated version of the classifiers in the future is inclusion of lineage specific information for each missense mutation. Chapter 5 helped me understand that susceptibility pattern changes with lineages, therefore including this information in our tools will make it more robust and powerful. Another addition that would make the tool more comprehensive and useful is addition of the other genes responsible for resistance. For bedaquiline, majority of resistance observed in the clinical cases is due to mutations in the efflux pump Rv0678. In case of PZA, it is more complicated as various genes contribute to resistance towards the drug. Addition of classifiers for panD, rpsA and clpC1 might make the tool all-inclusive in a clinical setting. Moreover, the structural studies provides us with information which can be used to design better resistance-resistant antibiotic [238]. Similarly, the mathematical model can be further used to study and understand the rise of INH amplification rates.

This work has established and validated a platform for using protein structural information and computational models to identify drug resistance mutations. Using these insights, I have developed empirical tools to predict phenotypic outcomes of mutations in p53 – a cancer causing gene and for a recessive genetic disorder – alkaptonuria. The work on alkaptonuria has been developed into a user-friendly webserver HGDiscovery (http://biosig.unimelb.edu.au/hgdiscovery/). Therefore, I have demonstrated the translational power of the pipeline I developed during my PhD.

This works provides a foundation for more effective drug resistance screening, public health policy decisions, and provide valuable insights for drug development. It highlights the power of combining different scientific approaches to achieve a common goal – tackle the issue to drug resistance in TB.

REFERENCES

1. Pai, M., et al., *Tuberculosis.* Nature Reviews Disease Primers, 2016. **2**: p. 16076.
2. Taylor, G.M., et al., *Koch's bacillus - a look at the first isolate of Mycobacterium tuberculosis from a modern perspective.* Microbiology, 2003. **149**(Pt 11): p. 3213-20.
3. WHO, *Global Tuberculosis Report, 2020.* 2020.
4. Petrini, B. and S. Hoffner, *Drug-resistant and multidrug-resistant tubercle bacilli.* Int J Antimicrob Agents, 1999. **13**(2): p. 93-7.
5. Padda IS, M.R.K., *Antitubercular Medications.* In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing;, 2020.
6. Rozwarski, D.A., et al., *Modification of the NADH of the Isoniazid Target (InhA) from Mycobacterium tuberculosis.* Science, 1998. **279**(5347): p. 98-102.
7. Campbell, E.A., et al., *Structural mechanism for rifampicin inhibition of bacterial rna polymerase.* Cell, 2001. **104**(6): p. 901-12.
8. Goude, R., et al., *The Arabinosyltransferase EmbC Is Inhibited by Ethambutol in Mycobacterium tuberculosis.* Antimicrobial Agents and Chemotherapy, 2009. **53**(10): p. 4138-4146.
9. Aldred, K.J., et al., *Fluoroquinolone interactions with Mycobacterium tuberculosis gyrase: Enhancing drug activity against wild-type and resistant gyrase.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(7): p. E839-E846.
10. Khisimuzi, M. and M. Zhenkun, *Mycobacterium tuberculosis DNA Gyrase as a Target for Drug Discovery.* Infectious Disorders - Drug Targets, 2007. **7**(2): p. 159-168.
11. Kenney, T.J. and G. Churchward, *Cloning and sequence analysis of the rpsL and rpsG genes of Mycobacterium smegmatis and characterization of mutations causing resistance to streptomycin.* J Bacteriol, 1994. **176**(19): p. 6153-6.
12. Alangaden, G.J., et al., *Mechanism of resistance to amikacin and kanamycin in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 1998. **42**(5): p. 1295-7.
13. Suzuki, Y., et al., *Detection of kanamycin-resistant Mycobacterium tuberculosis by identifying mutations in the 16S rRNA gene.* J Clin Microbiol, 1998. **36**(5): p. 1220-5.
14. Stanley, R.E., et al., *The structures of the anti-tuberculosis antibiotics viomycin and capreomycin bound to the 70S ribosome.* Nature Structural & Molecular Biology, 2010. **17**(3): p. 289-293.
15. Deoghare, S., *Bedaquiline: a new drug approved for treatment of multidrug-resistant tuberculosis.* Indian journal of pharmacology, 2013. **45**(5): p. 536-537.
16. Gler, M.T., et al., *Delamanid for multidrug-resistant pulmonary tuberculosis.* N Engl J Med, 2012. **366**(23): p. 2151-60.
17. Manjunatha, U., H.I.M. Boshoff, and C.E. Barry, *The mechanism of action of PA-824.* Communicative & Integrative Biology, 2009. **2**(3): p. 215-218.
18. Ippolito, J.A., et al., *Crystal Structure of the Oxazolidinone Antibiotic Linezolid Bound to the 50S Ribosomal Subunit.* Journal of Medicinal Chemistry, 2008. **51**(12): p. 3353-3356.
19. Davies, P.D., *The role of DOTS in tuberculosis treatment and control.* Am J Respir Med, 2003. **2**(3): p. 203-9.
20. Zhang Y Fau - Shi, W., et al., *Mechanisms of Pyrazinamide Action and Resistance.* (2165-0497 (Electronic)).
21. (WHO), W.H.O., *WHO consolidated guidelines on drug-resistant tuberculosis treatment.* 2019.
22. Organization, G.W.H., *WHO treatment guidelines for drug-resistant tuberculosis (2016 update)*

2016. **(https://apps.who.int/iris/bitstream/handle/10665/250125/9789241549639-eng.**

.

23. Moliva, J.I., J. Turner, and J.B. Torrelles, *Prospects in Mycobacterium bovis Bacille Calmette et Guerin (BCG) vaccine diversity and delivery: why does BCG fail to protect against tuberculosis?* Vaccine, 2015. **33**(39): p. 5035-41.
24. Cars, O., et al., *Meeting the challenge of antibiotic resistance.* Bmj, 2008. **337**: p. a1438.
25. Frost, L.S., et al., *Mobile genetic elements: the agents of open source evolution.* Nat Rev Microbiol, 2005. **3**(9): p. 722-32.
26. Gal-Mor, O. and B.B. Finlay, *Pathogenicity islands: a molecular toolbox for bacterial virulence.* Cell Microbiol, 2006. **8**(11): p. 1707-19.
27. Zainuddin, Z.F. and J.W. Dale, *Does Mycobacterium tuberculosis have plasmids?* (0041-3879 (Print)).
28. Gillespie, S.H., *Evolution of drug resistance in Mycobacterium tuberculosis: clinical and molecular perspective.* Antimicrob Agents Chemother, 2002. **46**(2): p. 267-74.
29. Eldholm, V. and F. Balloux, *Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out.* Trends Microbiol, 2016. **24**(8): p. 637-648.
30. Farhat, M.R., et al., *Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis.* Nat Genet, 2013. **45**(10): p. 1183-9.
31. Wright, G.D., *Aminoglycoside-modifying enzymes.* Curr Opin Microbiol, 1999. **2**(5): p. 499-503.
32. Brennan, P.J., *Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis.* Tuberculosis (Edinb), 2003. **83**(1-3): p. 91-7.
33. Garima, K., et al., *Differential expression of efflux pump genes of Mycobacterium tuberculosis in response to varied subinhibitory concentrations of antituberculosis agents.* Tuberculosis (Edinb), 2015. **95**(2): p. 155-61.
34. Andersson, D.I., *The biological cost of mutational antibiotic resistance: any practical conclusions?* Curr Opin Microbiol, 2006. **9**(5): p. 461-5.
35. Koch, A., V. Mizrahi, and D.F. Warner, *The impact of drug resistance on Mycobacterium tuberculosis physiology: what can we learn from rifampicin?* Emerg Microbes Infect, 2014. **3**(3): p. e17.
36. Maisnier-Patin, S. and D.I. Andersson, *Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution.* Res Microbiol, 2004. **155**(5): p. 360-9.
37. Gagneux, S., et al., *The competitive cost of antibiotic resistance in Mycobacterium tuberculosis.* Science, 2006. **312**(5782): p. 1944-6.
38. Barter, D.M., et al., *Tuberculosis and poverty: the contribution of patient costs in sub-Saharan Africa--a systematic review.* BMC Public Health, 2012. **12**: p. 980.
39. de Welzen, L., et al., *Whole-Transcriptome and -Genome Analysis of Extensively Drug-Resistant Mycobacterium tuberculosis Clinical Isolates Identifies Downregulation of ethA as a Mechanism of Ethionamide Resistance.* Antimicrobial Agents and Chemotherapy, 2017. **61**(12): p. e01461-17.
40. Kim, S.J., *Drug-susceptibility testing in tuberculosis: methods and reliability of results.* European Respiratory Journal, 2005. **25**(3): p. 564-569.
41. Organisation, W.H., *Technical manual for drug susceptibility testing of medicines used in the treatment of tuberculosis.* 2018.
42. Rodwell, T.C., et al., *Predicting extensively drug-resistant Mycobacterium tuberculosis phenotypes with genetic mutations.* J Clin Microbiol, 2014. **52**(3): p. 781-9.
43. Boehme, C.C., et al., *Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study.* Lancet, 2011. **377**(9776): p. 1495-505.
44. Hillemann, D., S. Rusch-Gerdes, and E. Richter, *Evaluation of the GenoType MTBDRplus assay for rifampin and isoniazid susceptibility testing of Mycobacterium tuberculosis strains and clinical specimens.* J Clin Microbiol, 2007. **45**(8): p. 2635-40.
45. *Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis*. 2014, IEEE. p. 618.
46. Albanaz, A.T.S., et al., *Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design.* (1746-045X (Electronic)).

47. Chen, J., et al., *Early detection of multidrug- and pre-extensively drug-resistant tuberculosis from smear-positive sputum by direct sequencing.* BMC Infectious Diseases, 2017. **17**(1): p. 300.
48. Kitzman, J.O., et al., *Massively parallel single-amino-acid mutagenesis.* Nat Methods, 2015. **12**(3): p. 203-6, 4 p following 206.
49. Pires, D.E., et al., *In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity.* Sci Rep, 2016. **6**: p. 19848.
50. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
51. Choi, Y. and A.P. Chan, *PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels.* Bioinformatics, 2015. **31**(16): p. 2745-7.
52. Bromberg, Y. and B. Rost, *SNAP: predict effect of non-synonymous polymorphisms on function.* Nucleic Acids Res, 2007. **35**(11): p. 3823-35.
53. Worth, C.L., R. Preissner, and T.L. Blundell, *SDM--a server for predicting effects of mutations on protein stability and malfunction.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W215-22.
54. Pires, D.E., D.B. Ascher, and T.L. Blundell, *mCSM: predicting the effects of mutations in proteins using graph-based signatures.* Bioinformatics, 2014. **30**(3): p. 335-42.
55. Pires, D.E., D.B. Ascher, and T.L. Blundell, *DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W314-9.
56. Pires, D.E., T.L. Blundell, and D.B. Ascher, *mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance.* Sci Rep, 2016. **6**: p. 29575.
57. Frappier, V., M. Chartier, and R.J. Najmanovich, *ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability.* Nucleic Acids Res, 2015. **43**(W1): p. W395-400.
58. Rodrigues, C.H.M., D.E.V. Pires, and D.B. Ascher, *DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability.* Nucleic Acids Research, 2018. **46**(W1): p. W350-W355.
59. Heifets, L. and P. Lindholm-Levy, *Pyrazinamide sterilizing activity in vitro against semidormant Mycobacterium tuberculosis bacterial populations.* Am Rev Respir Dis, 1992. **145**(5): p. 1223-5.
60. Zumla, A., et al., *Tuberculosis treatment and management--an update on treatment regimens, trials, new drugs, and adjunct therapies.* Lancet Respir Med, 2015. **3**(3): p. 220-34.
61. Steele, M.A. and R.M. Des Prez, *The role of pyrazinamide in tuberculosis chemotherapy.* Chest, 1988. **94**(4): p. 845-50.
62. Zhang, Y., et al., *'Z(S)-MDR-TB' versus 'Z(R)-MDR-TB': improving treatment of MDR-TB by identifying pyrazinamide susceptibility.* Emerg Microbes Infect, 2012. **1**(7): p. e5.
63. Chedore, P., et al., *Potential for erroneous results indicating resistance when using the Bactec MGIT 960 system for testing susceptibility of Mycobacterium tuberculosis to pyrazinamide.* J Clin Microbiol, 2010. **48**(1): p. 300-1.
64. Dillon, N.A., et al., *Anti-tubercular Activity of Pyrazinamide is Independent of trans-Translation and RpsA.* Scientific Reports, 2017. **7**(1): p. 6135.
65. Shi, W., et al., *Pyrazinamide Inhibits Trans-Translation in Mycobacterium tuberculosis.* Science, 2011. **333**(6049): p. 1630-1632.
66. Zimhony, O., et al., *Pyrazinamide inhibits the eukaryotic-like fatty acid synthetase I (FASI) of Mycobacterium tuberculosis.* Nature Medicine, 2000. **6**(9): p. 1043-1047.
67. Gopal, P., et al., *Pyrazinamide triggers degradation of its target aspartate decarboxylase.* Nature Communications, 2020. **11**(1): p. 1661.
68. Shi, W., et al., *Aspartate decarboxylase (PanD) as a new target of pyrazinamide in Mycobacterium tuberculosis.* Emerging Microbes & Infections, 2014. **3**(1): p. 1-8.
69. Zhang, S., et al., *Mutations in panD encoding aspartate decarboxylase are associated with pyrazinamide resistance in Mycobacterium tuberculosis.* Emerging Microbes & Infections, 2013. **2**(1): p. 1-5.
70. Sambandamurthy, V.K., et al., *A pantothenate auxotroph of Mycobacterium tuberculosis is highly attenuated and protects mice against tuberculosis.* Nature Medicine, 2002. **8**(10): p. 1171-1174.
71. Karmakar, M., et al., *Structure guided prediction of Pyrazinamide resistance mutations in pncA.* Scientific Reports, 2020. **10**(1): p. 1875.

72. Karmakar, M., et al., *Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy.* Am J Respir Crit Care Med, 2018. **198**(4): p. 541-544.
73. Koul, A., et al., *Diarylquinolines target subunit c of mycobacterial ATP synthase.* Nat Chem Biol, 2007. **3**(6): p. 323-4.
74. Guo, H., et al., *Structure of mycobacterial ATP synthase bound to the tuberculosis drug bedaquiline.* Nature, 2021. **589**(7840): p. 143-147.
75. Hards, K., et al., *Ionophoric effects of the antitubercular drug bedaquiline.* Proc Natl Acad Sci U S A, 2018. **115**(28): p. 7326-7331.
76. Hards, K., et al., *Bactericidal mode of action of bedaquiline.* J Antimicrob Chemother, 2015. **70**(7): p. 2028-37.
77. Haagsma, A.C., et al., *Selectivity of TMC207 towards mycobacterial ATP synthase compared with that towards the eukaryotic homologue.* Antimicrob Agents Chemother, 2009. **53**(3): p. 1290-2.
78. Field, S.K., *Bedaquiline for the treatment of multidrug-resistant tuberculosis: great promise or disappointment?* Ther Adv Chronic Dis, 2015. **6**(4): p. 170-84.
79. Salfinger, M. and G.B. Migliori, *Bedaquiline: 10 years later, the drug susceptibility testing protocol is still pending.* Eur Respir J, 2015. **45**(2): p. 317-21.
80. Karmakar, M., et al., *Empirical ways to identify novel Bedaquiline resistance mutations in AtpE.* PLoS One, 2019. **14**(5): p. e0217169.
81. WHO, *The Use of Bedaquiline in the Treatment of Multidrug-Resistant Tuberculosis.* 2013.
82. Coll, F., et al., *Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences.* Genome Med, 2015. **7**(1): p. 51.
83. Nguyen, T.V.A., et al., *Bedaquiline Resistance: Its Emergence, Mechanism, and Prevention.* Clin Infect Dis, 2018. **66**(10): p. 1625-1630.
84. Segala, E., et al., *New mutations in the mycobacterial ATP synthase: new insights into the binding of the diarylquinoline TMC207 to the ATP synthase C-ring structure.* Antimicrob Agents Chemother, 2012. **56**(5): p. 2326-34.
85. Ozcaglar, C., et al., *Epidemiological models of Mycobacterium tuberculosis complex infections.* Mathematical biosciences, 2012. **236**(2): p. 77-96.
86. Ragonnet, R., et al., *High rates of multidrug-resistant and rifampicin-resistant tuberculosis among re-treatment cases: where do they come from?* BMC Infect Dis, 2017. **17**(1): p. 36.
87. Orr, H.A., *Fitness and its role in evolutionary genetics.* Nature reviews. Genetics, 2009. **10**(8): p. 531-539.
88. Cohen, T., B. Sommers, and M. Murray, *The effect of drug resistance on the fitness of Mycobacterium tuberculosis.* Lancet Infect Dis, 2003. **3**(1): p. 13-21.
89. Cole, S.T., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.* Nature, 1998. **393**(6685): p. 537-44.
90. Thierry, D., et al., *IS6110, an IS-like element of Mycobacterium tuberculosis complex.* Nucleic acids research, 1990. **18**(1): p. 188-188.
91. Yesilkaya, H., et al., *Natural transposon mutagenesis of clinical isolates of Mycobacterium tuberculosis: how many genes does a pathogen need?* Journal of bacteriology, 2005. **187**(19): p. 6726-6732.
92. van Soolingen, D., et al., *Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis.* J Clin Microbiol, 1991. **29**(11): p. 2578-86.
93. Park, Y.-K., G.-H. Bai, and S.-J. Kim, *Restriction Fragment Length Polymorphism Analysis of Mycobacterium tuberculosis Isolated from Countries in the Western Pacific Region*. Journal of Clinical Microbiology, 2000. **38**(1): p. 191-197.
94. Comas, I., et al., *Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies.* PLoS One, 2009. **4**(11): p. e7815.
95. Heersma, H.F., K. Kremer, and J.D. van Embden, *Computer analysis of IS6110 RFLP patterns of Mycobacterium tuberculosis.* Methods Mol Biol, 1998. **101**: p. 395-422.
96. Hermans, P.W., et al., *Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains.* Infection and immunity, 1991. **59**(8): p. 2695-2705.

97. van Embden, J.D., et al., *Genetic markers for the epidemiology of tuberculosis.* Res Microbiol, 1992. **143**(4): p. 385-91.
98. Kontsevaya, I.S., V.V. Nikolayevsky, and Y.M. Balabanova, *Molecular epidemiology of tuberculosis: Objectives, methods, and prospects.* Molecular Genetics, Microbiology and Virology, 2011. **26**(1): p. 1.
99. Kamerbeek, J., et al., *Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology.* Journal of clinical microbiology, 1997. **35**(4): p. 907-914.
100. Mathema, B., et al., *Molecular epidemiology of tuberculosis: current insights.* Clinical microbiology reviews, 2006. **19**(4): p. 658-685.
101. Brudey, K., et al., *Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology.* BMC Microbiology, 2006. **6**(1): p. 23.
102. Demay, C., et al., *SITVITWEB--a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology.* Infect Genet Evol, 2012. **12**(4): p. 755-66.
103. Mazars, E., et al., *High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology.* Proceedings of the National Academy of Sciences, 2001. **98**(4): p. 1901-1906.
104. Frothingham, R. and W.A. Meeker-O'Connell, *Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats.* Microbiology (Reading), 1998. **144 (Pt 5)**: p. 1189-96.
105. Supply, P., et al., *Identification of novel intergenic repetitive units in a mycobacterial two-component system operon.* Mol Microbiol, 1997. **26**(5): p. 991-1003.
106. Allix-Béguec, C., et al., *Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of Mycobacterium tuberculosis complex isolates.* Journal of clinical microbiology, 2008. **46**(8): p. 2692-2699.
107. Weniger, T., et al., *MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria.* Nucleic acids research, 2010. **38**(Web Server issue): p. W326-W331.
108. Scott, A.N., et al., *Sensitivities and Specificities of Spoligotyping and Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing Methods for Studying Molecular Epidemiology of Tuberculosis.* Journal of Clinical Microbiology, 2005. **43**(1): p. 89-94.
109. Flores, L., et al., *Large Sequence Polymorphisms Classify Mycobacterium tuberculosis Strains with Ancestral Spoligotyping Patterns.* Journal of Clinical Microbiology, 2007. **45**(10): p. 3393-3395.
110. Mokrousov, I., et al., *Mycobacterium tuberculosis Beijing genotype in Russia: in search of informative variable-number tandem-repeat loci.* Journal of clinical microbiology, 2008. **46**(11): p. 3576-3584.
111. Allix-Béguec, C., et al., *Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of Mycobacterium tuberculosis Beijing isolates.* J Clin Microbiol, 2014. **52**(1): p. 164-72.
112. Jagielski, T., et al., *Current Methods in the Molecular Typing of Mycobacterium tuberculosis and Other Mycobacteria.* BioMed Research International, 2014. **2014**: p. 645802.
113. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.
114. Kato-Maeda, M., J.Z. Metcalfe, and L. Flores, *Genotyping of Mycobacterium tuberculosis: application in epidemiologic studies.* Future Microbiol, 2011. **6**(2): p. 203-16.
115. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
116. Niemann, S., et al., *Genomic diversity among drug sensitive and multidrug resistant isolates of Mycobacterium tuberculosis with identical DNA fingerprints.* PLoS One, 2009. **4**(10): p. e7407.
117. Doughty, E.L., et al., *Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer.* PeerJ, 2014. **2**: p. e585.
118. MacLean, D., J.D.G. Jones, and D.J. Studholme, *Application of 'next-generation' sequencing technologies to microbial genetics.* Nature Reviews Microbiology, 2009. **7**(4): p. 96-97.
119. Guthrie, J.L. and J.L. Gardy, *A brief primer on genomic epidemiology: lessons learned from Mycobacterium tuberculosis.* Ann N Y Acad Sci, 2017. **1388**(1): p. 59-77.

120. Schulz zur Wiesch, P., J. Engelstädter, and S. Bonhoeffer, *Compensation of Fitness Costs and Reversibility of Antibiotic Resistance Mutations.* Antimicrobial Agents and Chemotherapy, 2010. **54**(5): p. 2085-2095.
121. Borrell, S. and S. Gagneux, *Strain diversity, epistasis and the evolution of drug resistance in Mycobacterium tuberculosis.* Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 2011. **17**(6): p. 815-820.
122. Comas, I., et al., *Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans.* Nat Genet, 2013. **45**(10): p. 1176-82.
123. Brosch, R., et al., *A new evolutionary scenario for the Mycobacterium tuberculosis complex.* Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3684-9.
124. Supply, P., et al., *Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis.* Nat Genet, 2013. **45**(2): p. 172-9.
125. Brites, D., et al., *A New Phylogenetic Framework for the Animal-Adapted Mycobacterium tuberculosis Complex.* Front Microbiol, 2018. **9**: p. 2820.
126. Smith, N.H., et al., *Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis.* Nat Rev Microbiol, 2009. **7**(7): p. 537-44.
127. Couvin, D., et al., *Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database.* Infect Genet Evol, 2019. **72**: p. 31-43.
128. Stucki, D., et al., *Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages.* Nat Genet, 2016. **48**(12): p. 1535-1543.
129. Merker, M., et al., *Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage.* Nat Genet, 2015. **47**(3): p. 242-9.
130. Rutaihwa, L.K., et al., *Multiple Introductions of Mycobacterium tuberculosis Lineage 2–Beijing Into Africa Over Centuries.* Frontiers in Ecology and Evolution, 2019. **7**(112).
131. Austin, D.J., K.G. Kristinsson, and R.M. Anderson, *The relationship between the volume of antimicrobial consumption in human communities and the frequency of resistance.* Proc Natl Acad Sci U S A, 1999. **96**(3): p. 1152-6.
132. Lipsitch, M. and B.R. Levin, *Population dynamics of tuberculosis treatment: mathematical models of the roles of non-compliance and bacterial heterogeneity in the evolution of drug resistance.* Int J Tuberc Lung Dis, 1998. **2**(3): p. 187-99.
133. Magana-Arachchi, D.N., *Epidemiology of Multidrug Resistant Tuberculosis (MDR-TB), Tuberculosis - Current Issues in Diagnosis and Management.* DOI: 10.5772/54882., 2013(https://www.intechopen.com/books/tuberculosis-current-issues-in-diagnosis-and-management/epidemiology-of-multidrug-resistant-tuberculosis-mdr-tb-).
134. Kendall, E.A., M.O. Fofana, and D.W. Dowdy, *Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis.* Lancet Respir Med, 2015. **3**(12): p. 963-72.
135. Dye, C., et al., *Erasing the world's slow stain: strategies to beat multidrug-resistant tuberculosis.* Science, 2002. **295**(5562): p. 2042-6.
136. Sisson, S.A., Y. Fan, and M.M. Tanaka, *Sequential Monte Carlo without likelihoods.* Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1760-5.
137. Luciani, F., et al., *The epidemiological fitness cost of drug resistance in Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2009. **106**(34): p. 14711-5.
138. Blower, S.M. and T. Chou, *Modeling the emergence of the 'hot zones': tuberculosis and the amplification dynamics of drug resistance.* Nat Med, 2004. **10**(10): p. 1111-6.
139. Cohen, T. and M. Murray, *Modeling epidemics of multidrug-resistant M. tuberculosis of heterogeneous fitness.* Nat Med, 2004. **10**(10): p. 1117-21.
140. Bottger, E.C. and B. Springer, *Tuberculosis: drug resistance, fitness, and strategies for global control.* Eur J Pediatr, 2008. **167**(2): p. 141-8.
141. Mitchison, D.A. and J.B. Selkon, *The bactericidal activities of antituberculous drugs.* Am Rev Tuberc, 1956. **74**(2 Part 2): p. 109-16; discussion, 116-23.
142. Johnsson, K. and P.G. Schultz, *Mechanistic Studies of the Oxidation of Isoniazid by the Catalase Peroxidase from Mycobacterium tuberculosis.* Journal of the American Chemical Society, 1994. **116**(16): p. 7425-7426.

143. Lei, B., C.J. Wei, and S.C. Tu, *Action mechanism of antitubercular isoniazid. Activation by Mycobacterium tuberculosis KatG, isolation, and characterization of inha inhibitor.* J Biol Chem, 2000. **275**(4): p. 2520-6.
144. Marrakchi, H., G. Lanéelle, and A.K. Quémard, *InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II.* Microbiology (Reading), 2000. **146 (Pt 2)**: p. 289-296.
145. Dessen, A., et al., *Crystal structure and function of the isoniazid target of Mycobacterium tuberculosis.* Science, 1995. **267**(5204): p. 1638-41.
146. Rawat, R., A. Whitty, and P.J. Tonge, *The isoniazid-NAD adduct is a slow, tight-binding inhibitor of InhA, the Mycobacterium tuberculosis enoyl reductase: adduct affinity and drug resistance.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 13881-6.
147. Vilchèze, C., et al., *Inactivation of the inhA-encoded fatty acid synthase II (FASII) enoyl-acyl carrier protein reductase induces accumulation of the FASI end products and cell lysis of Mycobacterium smegmatis.* J Bacteriol, 2000. **182**(14): p. 4059-67.
148. Vilchèze, C., et al., *Transfer of a point mutation in Mycobacterium tuberculosis inhA resolves the target of isoniazid.* Nat Med, 2006. **12**(9): p. 1027-9.
149. Zhang, Y., T. Garbe, and D. Young, *Transformation with katG restores isoniazid-sensitivity in Mycobacterium tuberculosis isolates resistant to a range of drug concentrations.* Mol Microbiol, 1993. **8**(3): p. 521-4.
150. Slayden, R.A. and C.E. Barry, 3rd, *The genetics and biochemistry of isoniazid resistance in mycobacterium tuberculosis.* Microbes Infect, 2000. **2**(6): p. 659-69.
151. Vilchèze, C. and W.R. Jacobs, Jr., *Resistance to Isoniazid and Ethionamide in Mycobacterium tuberculosis: Genes, Mutations, and Causalities.* Microbiol Spectr, 2014. **2**(4): p. Mgm2-0014-2013.
152. Wehrli, W., *Rifampin: Mechanisms of Action and Resistance.* Reviews of Infectious Diseases, 1983. **5**: p. S407-S411.
153. Archambault, J. and J.D. Friesen, *Genetics of eukaryotic RNA polymerases I, II, and III.* Microbiol Rev, 1993. **57**(3): p. 703-24.
154. Zaczek, A., et al., *Genetic evaluation of relationship between mutations in rpoB and resistance of Mycobacterium tuberculosis to rifampin.* BMC microbiology, 2009. **9**: p. 10-10.
155. *Xpert MTB/RIF: WHO Policy update and Implementation manual.* https://www.who.int/tb/laboratory/xpert_launchupdate/en/, 2016.
156. Stagg, H.R., et al., *Isoniazid-resistant tuberculosis: a cause for concern?* The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease, 2017. **21**(2): p. 129-139.
157. Jenkins, H.E., M. Zignol, and T. Cohen, *Quantifying the burden and trends of isoniazid resistant tuberculosis, 1994-2009.* PLoS One, 2011. **6**(7): p. e22927.
158. Romanowski, K., et al., *Treatment outcomes for isoniazid-resistant tuberculosis under program conditions in British Columbia, Canada.* BMC Infectious Diseases, 2017. **17**(1): p. 604.
159. Falzon, D. and D. van Cauteren, *Demographic features and trends in tuberculosis cases in the European Region, 1995-2005.* Euro Surveill, 2008. **13**(12).
160. Johnson, A., *Increasing numbers of isoniazid-monoresistant TB in the USA.* Thorax, 2009. **64**(4): p. 338-338.
161. Nagu, T.J., et al., *Effects of isoniazid resistance on TB treatment outcomes under programmatic conditions in a high-TB and -HIV setting: a prospective multicentre study.* Journal of Antimicrobial Chemotherapy, 2016. **72**(3): p. 876-881.
162. Kumar, P., et al., *High degree of multi-drug resistance and hetero-resistance in pulmonary TB patients from Punjab state of India.* Tuberculosis (Edinb), 2014. **94**(1): p. 73-80.
163. Gegia, M., et al., *Outcomes among tuberculosis patients with isoniazid resistance in Georgia, 2007-2009.* Int J Tuberc Lung Dis, 2012. **16**(6): p. 812-6.
164. Nhung, N.V., et al., *The fourth national anti-tuberculosis drug resistance survey in Viet Nam.* Int J Tuberc Lung Dis, 2015. **19**(6): p. 670-5.
165. Hang, N.T., et al., *Primary drug-resistant tuberculosis in Hanoi, Viet Nam: present status and risk factors.* PLoS One, 2013. **8**(8): p. e71867.
166. Cornejo Garcia, J.G., et al., *Treatment outcomes for isoniazid-monoresistant tuberculosis in Peru, 2012-2014.* PLoS One, 2018. **13**(12): p. e0206658.

167. Helb, D., et al., *Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology.* J Clin Microbiol, 2010. **48**(1): p. 229-37.
168. Cohen, T., et al., *Mathematical models of the epidemiology and control of drug-resistant TB.* Expert Rev Respir Med, 2009. **3**(1): p. 67-79.
169. Liao, C.M. and Y.J. Lin, *Assessing the transmission risk of multidrug-resistant Mycobacterium tuberculosis epidemics in regions of Taiwan.* Int J Infect Dis, 2012. **16**(10): p. e739-47.
170. Kuddus, M.A., et al., *Modeling drug-resistant tuberculosis amplification rates and intervention strategies in Bangladesh.* PLoS One, 2020. **15**(7): p. e0236112.
171. Capriotti, E., et al., *Predicting protein stability changes from sequences using support vector machines.* Bioinformatics, 2005. **21 Suppl 2**: p. ii54-8.
172. Cheng, J., A. Randall, and P. Baldi, *Prediction of protein stability changes for single-site mutations using support vector machines.* Proteins, 2006. **62**(4): p. 1125-32.
173. Capriotti, E. and R.B. Altman, *Improving the prediction of disease-related variants using protein three-dimensional structure.* BMC Bioinformatics, 2011. **12 Suppl 4**: p. S3.
174. Tian, J., et al., *Predicting changes in protein thermostability brought about by single- or multi-site mutations.* BMC Bioinformatics, 2010. **11**: p. 370.
175. Topham, C.M., N. Srinivasan, and T.L. Blundell, *Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables.* Protein Eng, 1997. **10**(1): p. 7-21.
176. Bordner, A.J. and R.A. Abagyan, *Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations.* Proteins, 2004. **57**(2): p. 400-13.
177. Kellogg, E.H., A. Leaver-Fay, and D. Baker, *Role of conformational sampling in computing mutation-induced changes in protein structure and stability.* Proteins, 2011. **79**(3): p. 830-8.
178. Cheng, T.M., et al., *Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms.* PLoS Comput Biol, 2008. **4**(7): p. e1000135.
179. Pires, D.E., et al., *Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns.* BMC Genomics, 2011. **12 Suppl 4**: p. S12.
180. Kumar, M.D., et al., *ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.* Nucleic Acids Res, 2006. **34**(Database issue): p. D204-6.
181. Grant, B.J., et al., *Bio3d: an R package for the comparative analysis of protein structures.* Bioinformatics, 2006. **22**(21): p. 2695-2696.
182. Ng, P.C. and S. Henikoff, *Predicting deleterious amino acid substitutions.* Genome Res, 2001. **11**(5): p. 863-74.
183. Krieger, E., S.B. Nabuurs, and G. Vriend, *Homology modeling.* Methods Biochem Anal, 2003. **44**: p. 509-23.
184. Agarwal, S., D. Chadha, and R. Mehrotra, *Molecular modeling and spectroscopic studies of semustine binding with DNA and its comparison with lomustine-DNA adduct formation.* J Biomol Struct Dyn, 2015. **33**(8): p. 1653-68.
185. Lamb, M.L. and W.L. Jorgensen, *Computational approaches to molecular recognition.* Curr Opin Chem Biol, 1997. **1**(4): p. 449-57.
186. Shoichet, B.K., et al., *Lead discovery using molecular docking.* Curr Opin Chem Biol, 2002. **6**(4): p. 439-46.
187. Airey, E., et al., *Identifying Genotype-Phenotype Correlations via Integrative Mutation Analysis.* Methods Mol Biol, 2021. **2190**: p. 1-32.
188. Chernyaeva, E.N., et al., *Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology.* BMC Genomics, 2014. **15**(1): p. 308.
189. Sandgren, A., et al., *Tuberculosis drug resistance mutation database.* PLoS Med, 2009. **6**(2): p. e2.
190. Joshi, K.R., H. Dhiman, and V. Scaria, *tbvar: A comprehensive genome variation resource for Mycobacterium tuberculosis.* Database : the journal of biological databases and curation, 2014. **2014**: p. bat083-bat083.
191. Flandrois, J.-P., G. Lina, and O. Dumitrescu, *MUBII-TB-DB: a database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis.* BMC Bioinformatics, 2014. **15**(1): p. 107.
192. UniProt Consortium, T., *UniProt: the universal protein knowledgebase.* Nucleic Acids Research, 2018. **46**(5): p. 2699-2699.

193. Rose, P.W., et al., *The RCSB protein data bank: integrative view of protein, gene and 3D structural information.* Nucleic Acids Research, 2016. **45**(D1): p. D271-D281.
194. Kotsiantis, S.B., *Supervised Machine Learning: A Review of Classification Techniques*, in *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. 2007, IOS Press. p. 3-24.
195. Hall, M., et al., *The WEKA data mining software: an update.* SIGKDD Explor. Newsl., 2009. **11**(1): p. 10–18.
196. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**(null): p. 2825–2830.
197. Powers, D.M.W., *Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation.* Journal of Machine Learning Technologies, 2011. **2**(1): p. 37-63.
198. Zaki, M.J. and W.M. Jr, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. 2014: Cambridge University Press. 624.
199. Boughorbel, S., F. Jarray, and M. El-Anbari, *Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric.* PloS one, 2017. **12**(6): p. e0177678-e0177678.
200. Dye, C., et al., *Prospects for worldwide tuberculosis control under the WHO DOTS strategy. Directly observed short-course therapy.* Lancet, 1998. **352**(9144): p. 1886-91.
201. Bruchfeld, J., M. Correia-Neves, and G. Källenius, *Tuberculosis and HIV Coinfection.* Cold Spring Harbor perspectives in medicine, 2015. **5**(7): p. a017871-a017871.
202. Restrepo, B.I., *Diabetes and Tuberculosis.* Microbiology spectrum, 2016. **4**(6): p. 10.1128/microbiolspec.TNMI7-0023-2016.
203. Ronacher, K., et al., *Acquired immunodeficiencies and tuberculosis: focus on HIV/AIDS and diabetes mellitus.* Immunol Rev, 2015. **264**(1): p. 121-37.
204. Frieden, T.R., et al., *Tuberculosis.* Lancet, 2003. **362**(9387): p. 887-99.
205. Sutherland, I., E. Svandova, and S. Radhakrishna, *Alternative models for the development of tuberculosis disease following infection with tubercle bacilli.* Bull Int Union Tuberc, 1976. **51**(1): p. 171-9.
206. Manabe, Y.C. and W.R. Bishai, *Latent Mycobacterium tuberculosis-persistence, patience, and winning by waiting.* Nat Med, 2000. **6**(12): p. 1327-9.
207. Blower, S.M., et al., *The intrinsic transmission dynamics of tuberculosis epidemics.* Nature Medicine, 1995. **1**(8): p. 815-821.
208. Kasaie, P., et al., *Timing of tuberculosis transmission and the impact of household contact tracing. An agent-based simulation model.* Am J Respir Crit Care Med, 2014. **189**(7): p. 845-52.
209. Cohen, T., et al., *Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission.* Journal of The Royal Society Interface, 2007. **4**(14): p. 523-531.
210. Waaler, H., A. Geser, and S. Andersen, *The use of mathematical models in the study of the epidemiology of tuberculosis.* Am J Public Health Nations Health, 1962. **52**(6): p. 1002-13.
211. S, F., *An epidemiological model of tuberculosis in the united states.* Bulletin of the National Tuberculosis Association, 1967. **53:4–7**.
212. ReVelle, C.S., W.R. Lynn, and F. Feldmann, *Mathematical models for the economic allocation of tuberculosis control activities in developing nations.* Am Rev Respir Dis, 1967. **96**(5): p. 893-909.
213. Zwerling, A., S. Shrestha, and D.W. Dowdy, *Mathematical Modelling and Tuberculosis: Advances in Diagnostics and Novel Therapies.* Advances in Medicine, 2015. **2015**: p. 907267.
214. Breban, R., R. Vardavas, and S. Blower, *Theory versus data: how to calculate R0?* PLoS One, 2007. **2**(3): p. e282.
215. Ragonnet, R., et al., *Optimally capturing latency dynamics in models of tuberculosis transmission.* Epidemics, 2017. **21**: p. 39-47.
216. Chowell, G., Hyman, J. M., Bettencourt, L. M. A., & Castillo-Chavez, C., *Mathematical and statistical estimation approaches in epidemiology.* Springer Netherlands, 2009. **https://doi.org/10.1007/978-90-481-2313-1**.
217. Becker, N., *The Uses of Epidemic Models.* Biometrics, 1979. **35**(1): p. 295-305.
218. Kyere SN, B.F., Hoggar GF, Jonathan P. , *The Stochastic Model.* Adv Comput Sci., 2018 Jan. **1(1): 105.**
219. Roberts, M., et al., *Nine challenges for deterministic epidemic models.* Epidemics, 2015. **10**: p. 49-53.

220. Coscolla, M. and S. Gagneux, *Consequences of genomic diversity in Mycobacterium tuberculosis.* Seminars in immunology, 2014. **26**(6): p. 431-444.
221. Pitondo-Silva, A., et al., *Comparison of three molecular typing methods to assess genetic diversity for Mycobacterium tuberculosis.* Journal of Microbiological Methods, 2013. **93**(1): p. 42-48.
222. van der Spuy, G.D., et al., *Use of Genetic Distance as a Measure of Ongoing Transmission of Mycobacterium tuberculosis.* Journal of Clinical Microbiology, 2003. **41**(12): p. 5640-5644.
223. Alonso-Rodríguez, N., et al., *Evaluation of the new advanced 15-loci MIRU-VNTR genotyping tool in Mycobacterium tuberculosis molecular epidemiology studies.* BMC Microbiology, 2008. **8**(1): p. 34.
224. Oelemann, M.C., et al., *Assessment of an Optimized Mycobacterial Interspersed Repetitive- Unit-Variable-Number Tandem-Repeat Typing System Combined with Spoligotyping for Population-Based Molecular Epidemiology Studies of Tuberculosis.* Journal of Clinical Microbiology, 2007. **45**(3): p. 691-697.
225. Quitugua, T.N., et al., *Transmission of Drug-Resistant Tuberculosis in Texas and Mexico.* Journal of Clinical Microbiology, 2002. **40**(8): p. 2716-2724.
226. Alonso-Rodriguez, N., et al., *Prospective Universal Application of Mycobacterial Interspersed Repetitive-Unit-Variable-Number Tandem-Repeat Genotyping To Characterize Mycobacterium tuberculosis Isolates for Fast Identification of Clustered and Orphan Cases.* Journal of Clinical Microbiology, 2009. **47**(7): p. 2026-2032.
227. Cave, M.D., et al., *Epidemiologic Import of Tuberculosis Cases Whose Isolates Have Similar but Not Identical IS6110 Restriction Fragment Length Polymorphism Patterns.* Journal of Clinical Microbiology, 2005. **43**(3): p. 1228-1233.
228. Bergval, I., et al., *Combined species identification, genotyping, and drug resistance detection of Mycobacterium tuberculosis cultures by MLPA on a bead-based array.* PLoS One, 2012. **7**(8): p. e43240.
229. Comas, I., et al., *Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved.* Nat Genet, 2010. **42**(6): p. 498-503.
230. Tsolaki, A.G., et al., *Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains.* Proc Natl Acad Sci U S A, 2004. **101**(14): p. 4865-70.
231. Hershberg, R., et al., *High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography.* PLoS Biol, 2008. **6**(12): p. e311.
232. Homolka, S., et al., *High resolution discrimination of clinical Mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms.* PLoS One, 2012. **7**(7): p. e39855.
233. Loddenkemper, R., D. Sagebiel, and A. Brendel, *Strategies against multidrug-resistant tuberculosis.* Eur Respir J Suppl, 2002. **36**: p. 66s-77s.
234. Ahuja, S.D., et al., *Multidrug resistant pulmonary tuberculosis treatment regimens and patient outcomes: an individual patient data meta-analysis of 9,153 patients.* PLoS Med, 2012. **9**(8): p. e1001300.
235. Pontali, E., A. Matteelli, and G.B. Migliori, *Drug-resistant tuberculosis.* Current Opinion in Pulmonary Medicine, 2013. **19**(3): p. 266-272.
236. Seddon, J.A., et al., *Hearing loss in patients on treatment for drug-resistant tuberculosis.* Eur Respir J, 2012. **40**(5): p. 1277-86.
237. Pym, A.S., et al., *Bedaquiline in the treatment of multidrug- and extensively drug-resistant tuberculosis.* Eur Respir J, 2016. **47**(2): p. 564-74.
238. Oldfield, E. and X. Feng, *Resistance-resistant antibiotics.* Trends in Pharmacological Sciences, 2014. **35**(12): p. 664-674.

APPENDIX 1: METHODS USED IN THE SPATIAL ANALYSIS OF TUBERCULOSIS EPIDEMIOLOGY: A SYSTEMATIC REVIEW

Shaweno et al. BMC Medicine (2018) 16:193
https://doi.org/10.1186/s12916-018-1178-4

BMC Medicine

RESEARCH ARTICLE Open Access

Methods used in the spatial analysis of tuberculosis epidemiology: a systematic review

Debebe Shaweno[1,2*], Malancha Karmakar[2,3], Kefyalew Addis Alene[4,5], Romain Ragonnet[1,6], Archie CA Clements[7], James M. Trauer[2,8], Justin T. Denholm[2,3] and Emma S. McBryde[1,9]

Abstract

Background: Tuberculosis (TB) transmission often occurs within a household or community, leading to heterogeneous spatial patterns. However, apparent spatial clustering of TB could reflect ongoing transmission or co-location of risk factors and can vary considerably depending on the type of data available, the analysis methods employed and the dynamics of the underlying population. Thus, we aimed to review methodological approaches used in the spatial analysis of TB burden.

Methods: We conducted a systematic literature search of spatial studies of TB published in English using Medline, Embase, PsycInfo, Scopus and Web of Science databases with no date restriction from inception to 15 February 2017. The protocol for this systematic review was prospectively registered with PROSPERO (CRD42016036655).

Results: We identified 168 eligible studies with spatial methods used to describe the spatial distribution ($n = 154$), spatial clusters ($n = 73$), predictors of spatial patterns ($n = 64$), the role of congregate settings ($n = 3$) and the household ($n = 2$) on TB transmission. Molecular techniques combined with geospatial methods were used by 25 studies to compare the role of transmission to reactivation as a driver of TB spatial distribution, finding that geospatial hotspots are not necessarily areas of recent transmission. Almost all studies used notification data for spatial analysis (161 of 168), although none accounted for undetected cases. The most common data visualisation technique was notification rate mapping, and the use of smoothing techniques was uncommon. Spatial clusters were identified using a range of methods, with the most commonly employed being Kulldorff's spatial scan statistic followed by local Moran's *I* and Getis and Ord's local Gi(d) tests. In the 11 papers that compared two such methods using a single dataset, the clustering patterns identified were often inconsistent. Classical regression models that did not account for spatial dependence were commonly used to predict spatial TB risk. In all included studies, TB showed a heterogeneous spatial pattern at each geographic resolution level examined.

Conclusions: A range of spatial analysis methodologies has been employed in divergent contexts, with all studies demonstrating significant heterogeneity in spatial TB distribution. Future studies are needed to define the optimal method for each context and should account for unreported cases when using notification data where possible. Future studies combining genotypic and geospatial techniques with epidemiologically linked cases have the potential to provide further insights and improve TB control.

Keywords: Spatial analysis, Tuberculosis, Genotypic cluster

* Correspondence: debebesh@gmail.com
[1]Department of Medicine, University of Melbourne, Melbourne, Victoria, Australia
[2]Victorian Tuberculosis Program at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia
Full list of author information is available at the end of the article

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Mycobacterium tuberculosis (*Mtb*) transmission often occurs within a household or small community because prolonged duration of contact is typically required for infection to occur, creating the potential for localised clusters to develop [1]. However, geospatial TB clusters are not always due to ongoing person-to-person transmission but may also result from reactivation of latent infection in a group of people with shared risk factors [1, 2]. Spatial analysis and identification of areas with high TB rates (clusters), followed by characterisation of the drivers of the dynamics in these clusters, have been promoted for targeted TB control and intensified use of existing TB control tools [3, 4].

TB differs from other infectious diseases in several ways that are likely to influence apparent spatial clustering. For example, its long latency and prolonged infectious period allow for significant population mobility between serial cases [5]. Thus, *Mtb* infection acquired in a given location may progress to TB disease in an entirely different region, such that clustering of cases may not necessarily indicate intense transmission but could rather reflect aggregation of population groups at higher risk of disease, such as migrants [6]. Similarly, *Mtb* infection acquired from workplaces and other congregate settings can be wrongly attributed to residential exposure, as only an individual's residence information is typically recorded on TB surveillance documents in many settings [7, 8].

Identifying heterogeneity in the spatial distribution of TB cases and characterising its drivers can help to inform targeted public health responses, making it an attractive approach [9]. However, there are practical challenges in appropriate interpretation of spatial clusters of TB. Of particular importance is that the observed spatial pattern of TB may be affected by factors other than genuine TB transmission or reactivation, including the type and resolution of data and the spatial analysis methods used [10]. For instance, use of incidence data versus notification data could give considerably different spatial pattern [11], as the latter misses a large number of TB cases and could be skewed towards areas with better access to health care in high-burden settings [12, 13]. Thus, spatial analysis using notification data alone in such settings could result in misleading conclusions.

Similarly, the type of model used and the spatial unit of data analysis are important determinants of the patterns identified and their associations [14–16]. That is, different spatial resolutions could lead to markedly different results for the same dataset regardless of the true extent of spatial correlation [15, 17, 18] and the effect observed at a regional level may not hold at the individual level (an effect known as the ecological fallacy) [19]. Therefore, we aimed to review methodological approaches used in the spatial analysis of TB burden. We also considered how common issues in data interpretation were managed, including sparse data, false-positive identification of clustering and undetected cases.

Methods

Data source and search strategy

Our search strategy aimed to identify peer-reviewed studies of the distribution and determinants of TB that employed spatial analysis methods. In this review, studies were considered spatial if they incorporated any spatial approaches (e.g. geocoding, spatial analysis units, cluster detection methods, spatial risk modelling) into the design and analysis of the distribution, determinants and outcomes of TB [20]. We searched Medline, Embase, Web of Science, Scopus and PsycInfo databases from their inception to 15 February 2017 using a combination of keywords and medical subject headings (MeSH) pertaining to our two central concepts: tuberculosis and space. We refined search terms related to the latter concept after reviewing key studies, including a previous systematic review not limited to TB [21]. The full search strategy was adapted to the syntax of the individual database from the following conceptual structure: (tuberculosis OR multidrug-resistant tuberculosis) AND (spatial analysis OR geographic mapping OR spatial regression OR spatiotemporal analysis OR spatial autocorrelation analysis OR geography OR geographic distribution OR geographic information system OR geographically weighted regression OR space-time clustering OR 'spati*' OR 'hotspots' OR cluster analysis) and is provided in the Appendix. Studies targeted to special populations (e.g. homeless, migrants, HIV-infected persons) and that considered the entire population of a region were permitted. Additional papers were also identified through hand searching the bibliographies of retrieved articles and from suggestions from experts in the field.

Eligibility, and inclusion and exclusion criteria

We included peer-reviewed papers that incorporated the spatial analysis approaches described above in the study of TB. After exclusion of duplicates, titles and abstracts were screened by two researchers (DS and MK) to identify potentially eligible studies. Of these papers, articles were excluded hierarchically on the basis of article type, whether the method used could be considered spatial or not and the outcomes assessed. No exclusions were made on the basis of the outcome reported, with studies that considered incidence, prevalence or any TB-related health outcome included. Studies were excluded if the language of the publication was not English, the report was a letter, conference abstract or a review or only reported the temporal (trend) of TB. Spatial studies of

non-tuberculous mycobacteria, non-human diseases and population immunological profiles were also excluded. Full-text articles were excluded if they did not provide sufficient information on the spatial analysis techniques employed. There were no exclusions based on study setting or anatomical site of disease.

Data extraction and synthesis

Three independent reviewers (DS, MK, KAA) performed data extraction using pretested data extraction forms and stored these in a Microsoft Excel 2016 spreadsheet (Microsoft Corporation, Redmond, Washington, USA). Disagreements were resolved by consensus. The following information was extracted from each paper: country, publication year, study aim, data type (notifications or survey), type of TB disease (smear-positive pulmonary, smear-negative pulmonary and extrapulmonary), geographic level, spatial methods (map types, cluster detection methods, statistical regression methods, spatial lag, spatial error, spatial smoothing techniques), time scale and outcomes reported (whether quantification of TB cases or TB-related health outcomes, such as mortality, default from care, disability-adjusted life years (DALYs) and key conclusions). In studies which combined geospatial methods with genotypic clustering methods, we also extracted the genotypic cluster identification methods. Spatial analysis techniques were categorised as either visualisation (mapping), exploration (using statistical tests to identify spatial clusters) or statistical modelling [19, 22]. Counts and proportions were primarily used to summarise study findings. The protocol for this systematic review was prospectively registered with PROSPERO (CRD42016036655). Although we adhered to our original published protocol, here we additionally describe the importance of genotypic methods and the application of spatial methods in informing public health interventions in response to requests during peer review.

Results

Study characteristics

A total of 2350 records were identified from the electronic searches, of which 252 full-text articles were assessed. Of these, 168 articles met all inclusion criteria and were included in the final narrative synthesis (Fig. 1). Using a cutoff of 100 TB cases per 100,000 population in reported incidence in 2016, 111 (66%) of the studies were from low-incidence settings.

All references returned by the search strategy were from the period 1982 to 2017, with 71% published from 2010 onwards (Additional file 1: Figure S1). Earlier studies (predominantly in the 1980s and 1990s) tended to be descriptive visualisations, while studies in the last two decades frequently incorporated cluster detection and risk prediction. More recently, a range of statistical techniques including Bayesian statistical approaches and geographically weighted regression have become increasingly popular.

Key objectives of included studies

Spatial analysis was applied to address a range of objectives (Table 1), with the commonest ones including description of the distribution ($n = 135$), statistical analysis of spatial clustering ($n = 73$) and analysis of risk factors and risk prediction ($n = 64$). Spatial methods were also used to determine the relative importance of transmission by comparison to reactivation as a driver of TB incidence ($n = 25$), the effect of TB interventions ($n = 2$), barriers to TB service uptake ($n = 2$), spatial distribution of TB-related health outcomes (mortality, default, hospitalisation) ($n = 5$), spatial pattern of TB incidence among people living with HIV (PLHIV) ($n = 4$), HIV-related TB mortality ($n = 4$), multidrug-resistant TB (MDR-TB) drivers ($n = 1$), TB outbreak detection ($n = 3$) and drivers of spatial clustering (including the role of congregate settings, such as social drinking venues and schools) ($n = 30$).

Types of TB disease analysed

Spatial analysis was most commonly conducted on data for all types of TB (i.e. without distinction between pulmonary or extrapulmonary; $n = 121$), followed by pulmonary TB only ($n = 28$) and smear-positive pulmonary TB only ($n = 13$). Spatial analysis of multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB) was reported in 15 studies and one study respectively.

Data used and scale of analysis

Nearly all studies used retrospective TB program data (notifications), with the exception of five studies that used prevalence surveys and two prospectively collected data. None of the studies using notification data accounted for undetected/unreported cases. In all included studies, spatial analysis of TB was based on the individual's residence, except for three studies that explored the effect of exposure from social gathering sites.

Spatial analysis was generally done using data aggregated over administrative spatial units ($n = 131$), but the scale of aggregation differed markedly. Common spatial scales included census tract ($n = 20$), district ($n = 15$), postal code ($n = 15$), county ($n = 15$), neighbourhood ($n = 10$), health area ($n = 7$), municipality ($n = 11$), state ($n = 7$), province ($n = 6$), local government area (LGA) ($n = 4$) and ward ($n = 4$). Data were analysed at the individual level in 37 studies, while three studies were reported at a continent and country scale.

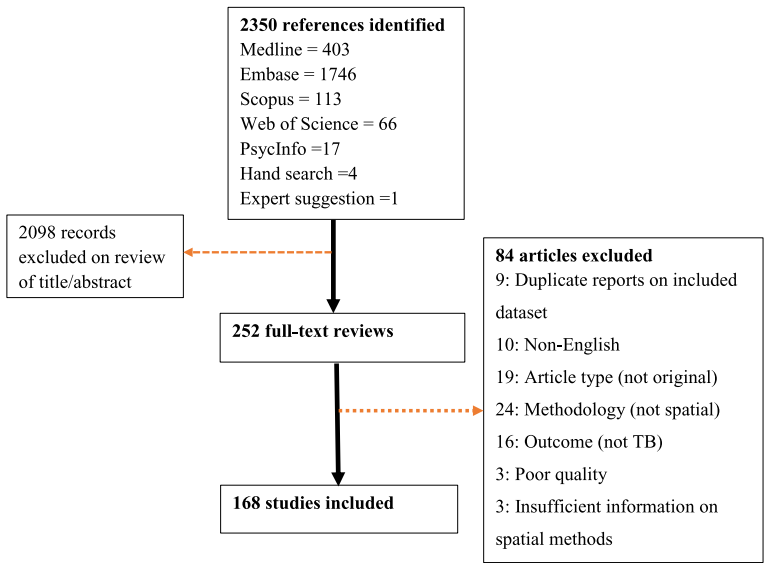


Fig. 1 Study inclusion flow chart

Table 1 Application areas of spatial methods in TB studies

Spatial method application areas	Methods used	References
Spatial TB distribution or spatial clustering	Dot maps, rate maps, thematic maps, Moran's *I*, GetisOrd statistic, NNI Besag and Newel statistic, *k*-functions, spatial scan statistic	[1, 2, 7, 8, 12, 16, 23–41, 44–49, 51–54, 57–72, 75, 93–95, 99, 100, 102–176]
Risk factors	Bayesian CAR models, regression models (with or without including spatial terms), GWR, PCA, mixture models, spatial lag models	[8, 12, 33, 36, 38, 40, 42–44, 46–52, 58, 59, 62, 70, 71, 93, 94, 99–102, 104, 111, 112, 116, 117, 120, 123, 125, 127–129, 131, 136, 137, 141–143, 145, 148, 149, 156, 161, 164, 176–189]
Monitoring spatiotemporal TB trends	Temporal trend maps	[27, 36–39]
Intervention evaluation	Distance map, kernel density map	[73, 74]
Barriers to TB care	Rate map, dot map, travel time map, distance map	[12, 187]
TB program performance	Map (time to detection)	[184]
HIV-related TB incidence	Rate map, dot map, spatial scan statistic	[40, 166, 186, 190]
TB treatment outcomes	Spatial empirical Bayes smoothing, kernel density maps, spatial scan statistic, spatial regression	[152, 155, 179, 183, 191]
Mortality related to TB/HIV coinfection	Rate map, thematic maps, Moran's *I* and spatial regression	[42, 43, 174, 192]
Transmission	Dot maps (congregate settings)	[54, 55, 193]
	Dot maps (cases)	[7, 8]
	Geospatial and genotypic clustering methods	[1, 2, 25, 28, 47, 57, 59–72, 93–95, 169, 194]
Methodological	Spatial scan statistic	[25]
TB outbreak detection	Spatial scan statistic	[1, 25, 28]
Prevalence estimation	Model-based geostatistics	[80]
Drivers of MDR-TB	*k*-function	[35]

NNI nearest neighbourhood index, *CAR models* conditional autoregressive models, *GWR* geographically weighted regression, *PCA* principal component analysis, *HIV* human immunodeficiency virus, *MDR-TB* multidrug-resistant TB

Methods in the spatial analysis of TB

Table 2 shows the range of spatial methods used. Spatial analysis was used to visualise patterns ($n = 154$), explore spatial clusters ($n = 73$) and identify risk factors for clustering ($n = 64$), with risk prediction undertaken by 11 studies. Of the included studies, six did not explicitly report any of these methods but reported statistical results that implied the use of these methods.

Data visualisation

Data visualisation was the most consistently applied technique, with 154 of the studies using at least one data visualisation method to present TB distribution and/or risk factor patterns across space (Table 1). The TB incidence rate was the commonest indicator mapped ($n = 63$), followed by event maps ($n = 37$), which were smoothed using kernel density in seven studies. Data visualisation was based on standardised morbidity ratios (SMR) in 12 studies. Five studies reported maps of trends in TB incidence over time, and thematic maps were used in nine to consider the impact of risk factors on TB incidence by displaying the spatial distribution of other variables. Variables plotted included climate ($n = 1$), socioeconomic factors ($n = 5$), diabetes ($n = 1$) and obesity ($n = 1$).

Approaches used to account for data sparseness

TB is a relatively rare disease at the population level, and burden is typically expressed in terms of cases per 100,000 population. Various approaches were used to account for this sparseness in the number of cases, such as aggregating cases over administrative geographic levels and over time periods (ranging from 1 to 25 years).

An alternative approach was rate smoothing, although this practice was rare, despite the fact that TB rates were the commonest indicators mapped. In the included

Table 2 Spatial methods used in spatial analysis of tuberculosis ($n = 168$)

Method category	Method	Number	References
Visualisation	Rate map	63	[12, 16, 23, 26, 27, 29–34, 37, 41, 44–46, 48, 51, 52, 57, 58, 60, 61, 70, 100, 102, 103, 105, 106, 120, 123–146, 164, 165, 170, 173–176, 195, 196]
	Dot map	37	[2, 7, 8, 35, 40, 47, 53, 54, 59, 66, 67, 72, 73, 75, 95, 107–122, 158, 166, 169, 178, 191, 197]
	SMR map	12	[38, 49, 99, 100, 124, 126, 127, 129, 138, 142, 148, 149]
	Kernel density map	7	[35, 37, 62, 93, 120, 147, 171]
	Case counts maps	3	[108, 167, 172]
	Others*	17	[16, 24, 50, 60, 62, 63, 68, 71, 99, 100, 103, 104, 116, 148, 166, 168, 185, 198]
Spatial cluster analysis	Global Moran's *I*	28	[16, 26, 34, 37, 39, 44, 48, 49, 51, 58, 65, 93, 100, 102, 123, 126, 128, 131, 133, 135, 138, 139, 145, 150, 161, 180, 188, 199]
	Local Moran's *I*	14	[16, 41, 44, 49, 51, 93, 100, 123, 126, 131, 135, 138, 145, 192]
	Kulldorff's spatial scan statistic	43	[1, 2, 23–32, 40, 57, 63, 64, 70, 71, 94, 109–111, 119, 120, 130, 135, 138, 139, 141, 151–160, 163, 164, 166, 191]
	GetisOrd statistic	12	[2, 16, 26, 39, 49, 54, 65, 93, 104, 131, 139, 161]
	k-NN	8	[35, 53, 69, 72, 93, 114, 122, 163]
	k-function	6	[35, 62, 93, 116, 117, 147]
	Besag and Newell statistic	2	[125, 145]
Statistical modelling	Bayesian CAR models	7	[38, 44, 49, 99, 101, 127, 148]
	Geographically weighted regression	6	[16, 50, 93, 102–104]
	Mixture modelling	2	[142, 149]
	Conventional logistic	15	[8, 40, 70, 71, 94, 95, 111, 112, 120, 141, 161, 177, 178, 187, 189]
	Conventional Poisson	5	[46, 125, 136, 145, 156]
	Conventional linear	5	[12, 47, 129, 137, 176]
	Negative binomial	1	[164]
	Factor analysis	6	[50, 103, 117, 143, 146, 170]
	Regression models with spatial terms	9	[42, 48, 51, 58, 100, 116, 128, 131, 188]
	Spatial prediction	11	[38, 42, 43, 62, 80, 99, 101, 127, 131, 148, 181]

SMR standardised morbidity ratio, *k-NN* *k*-nearest neighbourhood test, *CAR* conditional autoregressive
*Includes maps of disability-adjusted life years (DALYs), survival time, factor scores, probability maps, proportion of cases and regression coefficients

studies, smoothed rates were used in six (4%) studies. Similarly, of 12 studies that analysed SMRs, smoothed SMRs were presented in seven. In the included studies, several different data smoothing techniques were used, including fully Bayesian ($n = 8$), empirical Bayes ($n = 4$) and spatial empirical Bayes ($n = 5$). A significant number of visualisation reports ($n = 30$) were not complemented by hypothesis testing, either by exploration methods or modelling approaches. In 12 studies (7%), maps were not presented, but a narrative description of TB burden or a tabular presentation of TB distribution by administrative unit was described.

Spatial cluster (hotspot) identification

Use of at least one spatial cluster identification method was reported in 73 (43%) studies, with Kulldorff's spatial scan statistic used most frequently ($n = 43$), followed by Local Moran test ($n = 14$) and Getis and Ord's local Gi(d) statistic ($n = 12$). Nearest neighbour index (NNI), *k*-function and Besag and Newell methods were reported in eight, six and two studies respectively (Table 1). The presence of overall area-wide heterogeneity was assessed most often using global Moran *I* ($n = 28$). In three studies, no globally significant spatial autocorrelation was seen, although there was spatial clustering locally. Although studies used data aggregated over various spatial scales, only one evaluated the impact of spatial scale on the hotspot detection performance of the spatial scan statistic. Use of individual address-level data improved the sensitivity of the spatial scan statistic compared to data aggregated at the administrative level.

Simultaneous use of two spatial cluster detection methods was reported in 11 studies and showed differences in hotspot identification that ranged from complete disagreement to some degree of similarity (Table 3).

False-positive clustering

Not all spatial clusters are true clusters. False-positive clusters can arise from various sources, including data and methods used, and unmeasured confounding. Given that notification data were by far the most commonly used data source in the spatial analyses reviewed here, it could not be determined if these clusters represented true clusters of tuberculosis incidence or if they were caused by factors such as pockets of improved case detection. The role of differential TB detection has been documented in some studies from low-income settings, where increased spatial TB burden was linked to improved health care access [12].

In addition, rate was the commonest disease indicator used for disease mapping, as well as cluster detection in this study. As described earlier, rates are liable to stochasticity and can lead to false-positive clustering. However, rate smoothing and stability (sensitivity) analysis of clusters identified using rates was done in only a few studies [23, 24]. This remains an important area of consideration in the future spatial analysis of TB.

Table 3 Comparisons of spatial clusters from multiple cluster identification methods

Author, year	Methods	Outcome	Conclusion
Alene, K, 2017 [49]	Local Moran's *I* Getis and Ord	Clustered Clustered	50% similarity (two non-significant clusters identified by LISA)
Álvarez-Hernández, G., et al. 2010 [145]	Local Moran's *I* Besag and Newell	No significant Clustered	Widely conflicting
Dangisso M, et al. 2015 [26]	Getis and Ord Spatial scan statistic	Clustered Clustered	Similar overall pattern, but marked differences by years
Feske, M., et al. 2011 [93, 178]	Getis and Ord GWR residuals	Clustered Heterogeneous	Similar overall pattern, but some local differences
Ge E, et al. 2016 [139]	Getis and Ord Spatial scan statistic	Clustered Clustered	Similar overall pattern, but differences in some locations and across time
Haase I, et al. 2007 [2]	Hotspot analysis SaTScan	Clustered Clustered	Similar overall pattern, but some local differences
Hassarangsee S, et al. 2015 [138]	LISA Spatial scan statistic	Clustered Clustered	Very similar, but not identical
Li L, et al. 2016 [135]	LISA Spatial scan statistic	No significant cluster, Clustered	Widely conflicting
Maceiel ELN, et al. 2010 [131]	LISA, Getis and Ord Model prediction	Clustered Heterogeneous	Widely conflicting
Wubuli A, et al. 2015 [16]	LISA Getis and Ord	Clustered Clustered	Similar overall pattern, but some local differences
Wang T, et al 2016 [102]	Spatial scan statistic Getis and Ord	Clustered Clustered	Similar overall pattern, but some local differences

GWR geographically weighted regression; *LISA* local indicators of spatial association

Spatiotemporal analysis

Temporal scale

In the spatial analysis of TB, the time window is an important dimension that influences the spatial pattern of TB [25]. As TB is relatively a rare disease at the population level and has a long incubation period, detection of apparent spatial clusters requires a longer time scale than for acute infectious diseases that may form spatial clusters within days of the start of outbreak. Because of this, the included studies were based on cases that accumulated over considerable time periods, ranging from 1 to 25 years, with use of data aggregated over 5 years being the most frequent practice (20%).

Approaches

Generally, two approaches were used in the space-time cluster analysis of TB. The first uses classical space-time clustering using algorithms which scan space over a changing time window, such as Kulldorff's spatial scan statistic [23, 25–29]. The second approach is to account for the temporal dimension by repeating the spatial analysis for each time unit [26, 30–35]. In some studies, spatial patterns in temporal trends of TB incidence were determined as increasing or decreasing [27, 36–39].

Spatial statistical modelling

Different statistical modelling approaches were used to describe the relationship between TB and ecological factors in 65 (39%) studies, including nine spatially explicit models using Bayesian approaches. Conditional autoregressive (CAR) models were used in nine models to account for spatial correlation. Classical regression models were used in 33, while non-Bayesian spatial regression models were reported in 12.

Of the regression models that evaluated the effect on model fit of including spatial structure (spatial error or spatial lag), the inclusion of spatial structure improved the performance of the model in seven studies and failed to do so in two (based on deviance information criteria). Spatial lag was explicitly modelled in seven studies and highlighted the significant influence of neighbouring locations on TB distribution.

Traditional models including a Bayesian approach assumed a stationary relationship between TB and its spatial covariates and hence imposed a single (global) regression model on the entire study area. Only six studies used a geographically weighted regression (a local regression model) to accommodate variation in the association between TB and its risk factors from place to place and showed spatially varying (non-stationary) effects ($n = 6$). Other models used included mixture modelling ($n = 2$) and factor analysis using principal component analysis (PCA) ($n = 4$).

Results from spatial analysis

Geographic distribution of TB

The geographic distribution of TB was heterogeneous in all included studies both from low- and high-incidence settings, although no formal hypothesis testing was presented in 55 (33%). An exception was one study from South Africa that reported no significant clustering of cases among HIV patients on ART [40]. Spatial analysis was also used to describe the drivers of drug-resistant tuberculosis, with tighter spatial aggregation of MDR-TB cases compared with non-MDR cases taken as evidence of transmission of MDR-TB [41].

Spatial analyses into both HIV and TB investigated outcomes including HIV-associated TB incidence ($n = 4$) and spatial patterns of TB/HIV-related mortality ($n = 4$). All such studies revealed significant spatial heterogeneity. TB/HIV-related mortality in children was linked to areas with low socio-economic status and maternal deaths [42, 43].

Spatial methods used to study the impact of community-based TB treatment showed marked improvement in access compared to health facility-based treatment approaches ($n = 1$), and similar studies demonstrated travel time and distance to be important barriers to TB control ($n = 2$).

Correlations with social and environmental factors

The observed spatial patterns of TB were consistently linked to areas with poverty ($n = 14$), overcrowding and non-standard housing ($n = 9$), ethnic minority populations ($n = 3$), population density ($n = 2$), low education status ($n = 2$), health care access ($n = 3$) and immigrant populations ($n = 5$). However, a minority of studies have also found conflicting or non-significant associations between TB and poverty [44–46], population density [47–49] and unemployment [45, 47].

Four studies (including three from China) examined the correlation of climatic factors with TB incidence, with conflicting results. Two province-level studies in China using data from different time periods found TB burden to be associated with increasing annual average temperature [33, 50], although correlation with humidity was conflicting. Positive associations were observed with average precipitation [33, 50] and with air pressure [33] in these studies, while inverse associations were observed with sun exposure [50] and with wind speed [33]. In contrast, a county-level study which used average monthly climate data within a single province of China found the reverse, with temperature, precipitation, wind speed and sunshine exposure showing associations in the opposite direction [51]. A study that compared TB incidence between regions with different climatic conditions showed higher incidence at dry regions and low incidence in humid regions [52].

Space-time analysis to detect TB outbreaks

Studies reporting the application of the spatial methods in the early identification of TB outbreak were uncommon. Space-time TB studies using retrospective surveillance data in the USA found that the spatial scan statistic and other methods could effectively detect outbreaks months before local public authorities became aware of the problem [25, 28]. However, as space-time clusters of TB can be due to either ongoing transmission or reactivation, characterising the drivers that resulted in the spatial clustering is essential. Findings from studies which compared the timeliness and accuracy of space-time clusters in identifying TB outbreaks varied with spatial resolution and the background population, with two studies from the USA detecting ongoing outbreaks [25, 28], in contrast to false alarms due to reactivation TB among immigrants in a study from Canada [1].

Spatial analysis of the source of TB infection

Spatial methods were also used to determine the role of households and congregate settings (e.g. social gathering venues, schools) on TB transmission risk (Table 1). The role of the household was determined by cross-referencing child and adolescent TB infection or disease with adult TB in two studies [7, 8]. In these studies, the importance of household exposure declined with the age of the child, such that TB disease or infection was related to residential exposure to adult TB in younger children but not adolescents.

Congregate settings, which pose increased transmission risk, were identified using multiple techniques that included linking TB cases to social gathering places [53] and mapping the distribution of rebreathed air volume (RAV) [54] (including grading these settings based on TB transmission principles [55]). These approaches identified schools and social gathering sites as high-risk areas.

Identifying local drivers

Recent transmission is a critical mechanism driving local TB epidemiology in high-burden settings, while reactivation of remotely acquired infection is thought to predominate in most low-endemic settings [4, 56]. Geospatial clusters may reflect increased disease risk due to geographic proximity, which may correspond to recent transmission, or reactivation of latent TB infection in an aggregate of individuals infected elsewhere or both [57]. In the reviewed studies, spatial methods coupled with other methods were used to identify which of these two mechanisms drives local TB epidemiology in the following three ways.

Combining spatial clusters with cohort clustering: TB clustering can occur from ongoing transmission or from reactivation of latent infection among high-risk subgroups due to shared characteristics such as similar country of birth rather than a shared transmission network, a phenomenon known as cohort clustering. Cohort cluster analysis is used to identify selected high-risk population subgroups for targeted interventions based on the relative TB incidence they bear. The Lorenz curve is a simple visualisation tool that compares the clustering (inequality) in the subgroup of interest across regions and over time. One study, which combined such cohort (birth country) cluster analysis using the Lorenz curve of inequality with spatial cluster analysis [31] revealed colocation of these cluster types, suggesting the presence of both transmission and reactivation. Spatial clusters among foreign-born persons covered too large an area compared to clusters among the locally born to be consistent with direct person-to-person transmission. In addition, spatial modelling was also applied to differentiate the role of transmission from reactivation by assessing spatial dependence. The presence of spatial dependence (autocorrelation) was taken to indicate transmission, while its absence was considered to indicate reactivation [58].

Combining spatial and genotype clustering: Genotypic clustering of TB may be used as a proxy for recent transmission, such that geospatial clusters in which cases are genotypically clustered may be taken as stronger evidence for locations where recent transmission has occurred. These approaches were combined to quantify the role of recent transmission and determine geographical locations of such transmission in 25 studies. This was done either by determining the spatial distribution of genotypic clusters [25, 28, 59–69] or by assessing the genotypic similarity of cases contained within geospatial clusters [2, 57, 65, 70, 71].

The findings from these studies varied considerably by the country and sub-population studied (locally born versus immigrants) (Table 4). Genotypic clusters were spatially clustered in many studies, providing evidence of recent local transmission. In some studies, cases in geospatial clusters were less likely to be dominated by genotypically similar cases (i.e. were dominated by unique strains) than cases outside the geospatial clusters, implying spatial aggregation of reactivation TB [57]. This finding highlights that geospatial hotspots in low TB incidence settings are not necessarily areas of recent transmission and spatial clustering may be primarily mediated by social determinants, such as migration, HIV and drug abuse [57].

Table 4 Overlap between spatial and molecular clustering

Authors	Country	Genotyping methods	Findings
Bishai WR, et al. 1998 [95]	USA	IS6110-RFLP and PGRS	Genotypic clusters with epidemiologic links were spatially clustered but 76% of DNA clustered cases lack epidemiologic links.
Mathema B, et al. 2002 [169]	USA	IS6110-RFLP and spoligotyping	Genotypic clusters showed spatial aggregation
Richardson M, et al. 2002 [72]	South Africa	IS6110-RFLP and spoligotyping	Spatial aggregation of genotypic clusters was limited
Nguyen D, et al. 2003 [69]	Canada	IS6110-RFLP and spoligotyping	Genotypically similar cases were not more spatially clustered than genotypically unique cases
Moonan P, et al. 2004 [61]	USA	IS*6110*-RFLP and spoligotyping	Genotypic clusters were spatially heterogeneous
Jacobson L, et al. 2005 [59]	Mexico	IS*6110*-RFLP and spoligotyping	Spatial patterns were similar for both cases categorised as reactivation or recent transmission
Haase I, et al. 2007 [2]	Canada	IS*6110*-RFLP and spoligotyping	In spatial TB clusters of immigrants, there was significant genotype similarity
Higgs B, et al. 2007 [25]	USA	IS*6110*-RFLP and PGRS	Space-time clusters contained genotypic clusters
Feske ML, et al. 2011 [93, 178]	USA	IS*6110*-RFLP and spoligotyping	Genotypically clustered cases were randomly distributed across space
Evans JT, et al. 2011 [66]	UK	Spoligotyping and MIRU-VNTR	Genotypic clusters showed spatial aggregation
Nava-Aguilera E, et al. 2011 [67]	Mexico	Spoligotyping	Genotypic clusters were not spatially aggregated
Prussing C, et al. 2013 [57]	USA	Spoligotyping and 12- MIRU-VNTR	Cases in geospatial clusters were equally or less likely to share similar genotypes than cases outside geospatial clusters
Tuite AR, et al. 2013 [94]	Canada	Spoligotyping and 24-MIRU-VNTR	The proportion of cases in genotypic clusters was five times that seen in spatial clusters (23% vs 5%)
Kammerer JS, et al. 2013 [28]	USA	Spoligotyping and 12-MIRU-VNTR	Genotypically similar cases were spatially clustered
Verma A, et al. 2014 [1]	Canada	IS*6110*-RFLP and Spoligotyping	Space-time clusters contained few or no genotypically similar cases
Izumi K, et al. 2015 [65]	Japan	IS*6110*-RFLP	Both genotypically similar and unique strains formed spatial hotspots
Chamie G, et al. 2015 [194]	Uganda	Spoligotyping	Genotypic clusters shared social gathering sites (clinic, place of worship, market or bar)
Chan-Yeung M, et al. 2005 [47]	Hong Kong	IS*6110*-RFLP	Spatial locations of genotypic clusters and unique cases did not differ by their sociodemographic characteristics
Gurjav U, et al. 2016 [70]	Australia	24-MIRU-VNTR	Spatial hotspots were characterised by a high proportion of unique strains; less than 4% of cases in spatial clusters were genotypically similar
Ribeiro FK, et al. 2016 [62]	Brazil	IS*6110*-RFLP and Spoligotyping	Genotypic clusters were spatially clustered
Saavedra-Campos M, et al. 2016 [71]	England	24-MIRU-VNTR	10% of cases clustered spatially and genotypically
Seraphin MN, et al. 2016 [64]	USA	Spoligotyping and 24-MIRU-VNTR	22% of cases among USA-born and 5% among foreign-born clustered spatially and genotypically
Yuen CM, et al. 2016 [68]	USA	Spoligotyping and 24-MIRU-VNTR	Genotype clustered cases were spatially heterogeneous
Yeboah-Manu D, et al. 2016 [63]	Ghana	IS6110 and rpoB PCR	Genotypic clusters showed spatial aggregation
Zelner J, et al. 2016 [60]	Peru	24-MIRU-VNTR	Genotypic clusters showed spatial aggregation

PGRS polymorphic GC-rich repetitive sequence

Combinations of multiple methods were typically used for genotyping, with the commonest being IS6110 restriction fragment length polymorphism (IS6110-RFLP) and spoligotyping ($n = 9$), followed by mycobacterial interspersed repetitive unit variable number tandem repeat (MIRU-VNTR) and spoligotyping ($n = 5$), although use of a single method was reported in six studies (Table 4). No identified studies reported use of whole genome sequencing.

Temporal distribution of genotypically clustered cases

The temporal pattern of genotypic clustering could provide insights to distinguish between transmission and reactivation. In some studies, the temporal distribution of genotypically clustered cases indicated periods of 1 to more than 8 years between the genotypically clustered cases [1, 72], implying reactivation TB could also show genotypic similarity.

Use of spatial methods to inform public health interventions

In addition to their use in characterising the spatial distribution and determinants of TB, spatial methods have been used to inform TB-related public health interventions. In these studies, spatial analysis methods have proved to be attractive in guiding public health interventions, although their application to TB care beyond research is not well documented. For instance, spatial analysis techniques have been used to identify locations with a high density of TB cases (termed hotspots, although this definition was not based on spatial statistical tests). Community screening was then conducted in these areas, and its yield was compared to that from routine service provision. This GIS-guided screening was found to considerably improve the detection of individuals with latent TB infection and other infectious diseases [73]. Similarly, a study from South Africa highlighted the potential for using GIS to promote community-based DOTS by locating and geographically linking TB patients to their nearest supervision sites, although programmatic implementation of this approach was not reported [74].

The potential for spatial methods to be used for the early detection of TB outbreaks has also been described, although the findings widely varied based on the background population [1, 28]. Spatial cluster analysis using data at higher geographic resolutions improves the method's performance in cluster detection [25].

Discussion

While a range of methodologies has been employed in divergent contexts, we found that essentially all geospatial studies of TB have demonstrated significant heterogeneity in spatial distribution. Spatial analysis was applied to improve understanding of a range of TB-related issues, including the distribution and determinants of TB, the mechanisms driving the local TB epidemiology, the effect of interventions and the barriers to TB service uptake. Recently, geospatial methods have been combined with genotypic clustering techniques to understand the drivers of local TB epidemiology, although most such studies remain limited to low-endemic settings.

In almost all reviewed studies, retrospective program data (notifications) were used. Notification data, especially from resource-scarce settings, suffer from the often large proportion of undetected cases and are heavily dependent on the availability of diagnostic facilities [12]. None of the spatial studies of TB that used notification data accounted for undetected cases, such that the patterns in the spatial distribution and clustering could be heavily influenced by case detection performance [11]. Hence, distinguishing the true incidence pattern from the detection pattern has rarely been undertaken, despite its importance in interpretation.

The problems of undetected cases could be compounded in the spatial analysis of drug-resistant forms of TB, especially in resource-scarce settings where testing for drug-resistant TB is often additionally conditional on the individual's risk factors for drug resistance [75]. However, recently, there have been some attempts to account for under-detection in the spatial analysis of TB. A Bayesian geospatial modelling approach presented a framework to estimate TB incidence and case detection rate for any spatial unit and identified previously unreported spatial areas of high burden [11]. Another approach is to estimate incidence using methods such as capture-recapture [76, 77] and mathematical modelling [78]. If case detection rate is truly known for a defined region, incidence can be calculated as notifications divided by case detection rate, although this is rarely if ever the case. Spatial analysis using prevalence data could also be considered in areas where such data are available.

In relation to the data problems outlined above, spatial analysis of TB could benefit from the use of model-based geostatistics, which is commonly used in other infectious diseases [79], although there are few studies that consider *Mtb* [80]. In particular, measurement of TB prevalence is impractical to perform at multiple locations due to logistic reasons. Therefore, model-based geostatistics can be used to predict disease prevalence in areas that have not been sampled from prevalence values at nearby locations at low or no cost, producing smooth continuous surface estimates.

Mapping of notification rates was the most commonly used data visualisation technique, in which TB cases were categorised at a particular administrative spatial level. This approach has the advantage of easy interpretability, although it can introduce bias because the size of the regions and the locations of their boundaries typically reflect administrative requirements, which may not reflect the spatial distribution of epidemiological factors [19, 22]. In addition, patterns observed across regions may depend on the spatial scale chosen, an effect known as the modifiable areal unit problem (MAUP) [17]. Because the choice of spatial scale mainly depends on the limitations of available data [81], only one study was able to provide a systematic evaluation of the effect of scale on spatial patterns, demonstrating improved performance of Kulldorff's spatial scan statistic method at a high geographic resolution [25]. Different spatial resolutions could lead to markedly different results for the same dataset regardless of the true extent of correlation, due to averaging (aggregation effect) or other spatial processes operating at different scales

[15, 17, 18]. Assessing the presence of this effect should be a priority for future studies using aggregated data in spatial TB studies.

Bayesian smoothing techniques can mitigate the problems of stochastically unstable rates from areas with small population [81], although such techniques were not widely used in the included studies and so false spatial clustering remains an important consideration. The less frequent use of rate smoothing techniques in the spatial analysis of TB could have various explanations, including lack of software packages that are easily accessible to the wider user (although GeoDa spatial software currently provides an accessible platform to people with limited statistical or mathematical backgrounds [82]). It may also be that most spatial analyses of TB are based on data aggregated over larger geographic areas from several years, such that the problem of statistical stochasticity may not be a major problem, although this was not explicitly discussed in the included studies.

In all studies that applied spatial cluster identification tools, TB cases were clustered irrespective of whether the setting was low or high endemic. However, in studies that incorporated more than one cluster identification method, areas identified as hotspots were not identical, with the extent of agreement between the alternative methods highly variable. This could be partly attributable to different methods testing separate hypotheses, such that these results may correctly support one hypothesis while refuting another. However, there is no consensus on how to interpret these findings appropriately and consistently [82, 83], and method selection did not typically appear to be based on such considerations [84, 85]. Thus, caution is required when considering interventions assessing clusters with one method only, as is frequently undertaken in TB spatial analysis [22].

Use of multiple cluster detection methods and requiring their overlap to represent a truly high-risk area is increasingly recommended [82, 84, 86]. However, this approach could also increase the risk of false-positive spatial clustering when different methods are used serially until significant clusters are observed [85]. Sensitivity analysis of spatial clustering [87, 88] and cluster validation using geostatistical simulations [23, 89, 90] can help identify robust clusters. While methods that adjust for confounding are generally preferred [91], further investigative strategies including data collection and cluster surveillance are required to validate an observed spatial cluster before introducing interventions [84, 85]. Although the focus of this study is TB, several methodological considerations outlined here would remain true for many infectious diseases.

In several studies, presence of spatial clustering or spatial autocorrelation in TB distribution was considered to reflect ongoing TB transmission, while its absence was taken to indicate reactivation [58]. Recently, molecular techniques have been combined with geospatial methods to understand the drivers of local TB epidemiology, although findings from these studies vary by country and the subset of the population studied. While spatial clustering of genotypically related cases was reported in several studies and likely reflected intense local TB transmission [61, 65], spatial clusters were dominated by genotypically unique strains in some studies, implying that reactivation was the dominant process [47, 72]. Hence, the combination of genotypic and geospatial techniques can improve understanding of the relative contribution of reactivation and transmission and other local contributors to burden.

Notwithstanding the general principles outlined above, not all spatial clusters of genotypically related cases will necessarily result from recent transmission, as simultaneous reactivation of remotely acquired infection and limited genetic variation in the pathogen population can also lead to genotypic similarity of spatially clustered cases [2, 92]. In some studies, the time between the first and last diagnosis of the cases in the genetic cluster ranged from 1 to more than 8 years [1, 72], suggesting that genotypic clustering could occur from spatially clustered reactivation. Similarly, limited spatial aggregation of genotypically clustered cases [72, 93, 94] and lack of epidemiological links between genotypically clustered cases in some studies may reflect migration of the human population over the extended time scale over which TB clusters occur [95], although casual transmission creating spatially diffuse clusters is an alternative explanation.

The extent of genotypic similarity between cases also depends on the discriminatory power of the genotyping method and the diversity of the pathogen population. Compared to whole genome sequencing, standard molecular genotyping (spoligotyping, MIRU-VNTR and IS6110) methods generally overestimate TB transmission with a false-positive clustering rate of 25 to 75% based on strain prevalence in the background population [92, 96]. The accuracy of these tests in distinguishing ongoing transmission from genetically closely related strains is very low among immigrants from high TB incidence settings with limited pathogen diversity [92, 97]. Thus, care should be taken when interpreting the genotypic similarity of cases among immigrant groups, as independent importation of closely related strains is possible. The frequent finding of more extensive genotypic than spatial clusters [71, 94] may reflect overestimation by the genotypic methods [98]. On the other hand, TB transmission might not result in apparent spatial clustering due to reasons that include population movement, poor surveillance and unmeasured confounding.

Regression models used for spatial analysis of TB were either conventional regression models or models that incorporated spatial effects. Although the former was more commonly employed, the majority of models incorporating spatial effects confirmed that accounting for spatial correlation improved model fit [11, 33, 44, 58, 99–101]. Conventional regression models assume spatial independence of model residuals and so ignore the potential presence of spatial autocorrelation, such that non-spatial models may lead to false conclusions regarding covariate effects.

The use of the conventional regression models described above may be appropriate for spatial analysis and spatial prediction, in the case that spatial dependence in residuals has been ruled out. Under this approach, the standard procedure is to start with classical ordinary least squares (OLS) regression models and then look for spatial dependence in the residuals, which implies the need for a spatially explicit regression model [82]. Several of the models reviewed here did not appear to adopt this approach, and so, caution is required when interpreting the findings from such analyses.

Most regression models treat the association between TB rates and ecological factors as global and are unable to capture local variation in the estimates of the association. However, geographically weighted regression (GWR) estimates coefficients for all spatial units included [22] and has often found the effect of risk factors on TB incidence to be spatially variable [16, 102–104], implying that global models may be inadequate to consider locally appropriate interventions. Few studies were able to perform explicit Bayesian spatial modelling incorporating information from nearby locations, thereby producing stable and robust estimates for areas with small populations and robust estimates of the effects of covariates [91].

While our review focused on methodological issues, several consistent observations were noted. Most importantly, all studies included in this review demonstrated that TB displayed a heterogeneous spatial pattern across various geographic resolutions. This reflects the underlying tendency for spatial dependence that can be caused by person-to-person transmission, socio-economic aggregation [49] and environmental effects [58, 93]. However, in nearly all included studies, spatial analyses of TB were based on the individual's residence, although considerable TB infection is acquired from workplaces and other social gathering sites [8, 54]. Such studies could wrongly attribute TB acquired from such sites to residential exposure, leading to resource misallocation.

Several models have shown significant associations between TB rates and demographic, socioeconomic and risk-factor variables, although it is difficult to rule out publication bias favouring studies with positive findings. However, associations observed between TB rates and different factors such as population density, unemployment and poverty at the population level varied across studies. These were recognised as important individual-level risk factors, highlighting the potential for ecological fallacy.

We did not perform individual study level analysis of bias in this review. Analyses in the reviewed studies involved counts and proportions across different spatial distributions, rather than comparisons across different treatment/exposure groups. Standard tools of bias analysis predominantly focus on different treatment groups within cohorts (absent from our included studies) and hence are not applicable to this review. We have however discussed many potential sources of bias in the studies included in our review.

Most of the reviewed studies were from high-income settings, which may either reflect publication bias or a focus of research efforts on such settings. In high-incidence settings, the more limited use of spatial analysis methods could reflect a lack of access to resources (e.g. georeferenced data and spatial software packages) or insufficient expertise in these settings. However, it is these high-transmission settings which stand to gain the most from an improved understanding of TB spatial patterns and also these settings in which geospatial clustering may be most important epidemiologically.

Conclusions

A range of spatial analysis methodologies have been employed in divergent contexts, with virtually all studies demonstrating significant heterogeneity in spatial TB distribution regardless of geographic resolution. Various spatial cluster detection methods are available, although there is no consensus on how to interpret the considerable inconsistencies in the outputs of these methods applied to the same dataset. Further studies are needed to determine the optimal method for each context and research question and should also account for unreported cases when using notifications as input data where possible. Combining genotypic and geospatial techniques with epidemiologically linkage of cases has the potential to improve understanding of TB transmission.

Appendix

Search strings

Search terms used in Embase, Medline, PsycInfo, Scopus and Web of Science

The exp refers to explode which means include all subheadings underneath spatial analysis. When exploded, it contains geographic mapping, spatial regression and spatiotemporal analysis.

Brackets () denote subject headings (MeSH in Medline and Emtree in Embase) terms highlighted by the database.

Medline and PsycInfo

1. (exp spatial analysis) OR (Geographic information systems) OR (Space-time clustering) OR geographic* analys*.mp OR spati*regres*.mp OR spat*temp*.mp OR spat* analys*.mp OR spat* temp* analys*.mp OR spat* temp* pattern*.mp OR geography* distribut*.mp OR spat* temp* distribut*.mp OR heterogen* distribut.mp OR spacetime cluster*mp OR space-time cluster*mp OR hotspot.mp Or hot spots. mp OR GIS OR spati*
2. (tuberculosis) OR (tuberculosis, multidrug resistant) OR TB.mp
3. 1 AND 2

Embase

1. (spatial analysis) OR (geographic mapping) OR (spatial regression) OR (Spatiotemporal analysis OR (spatial autocorrelation analysis) OR (geography) OR (geographic distribution) OR (geographically weighted regression) OR (geographic information systems) OR (cluster analysis) OR geographic* analys*.mp OR spati*regres*.mp OR spat*temp*.mp OR spat* analys*.mp OR spat* temp* analys*.mp OR spat* temp* pattern*.mp OR geography* distribut*.mp OR spat* temp* distribut*.mp OR heterogen* distribut.mp OR spacetime cluster*mp OR space-time cluster*mp OR hotspot.mp Or hot spots. mp OR GIS OR spati*
2. (tuberculosis) OR (multidrug resistant tuberculosis) OR TB.mp
3. 1 AND 2

Scopus

("Spatial analysis" OR
"Spatio-temporal analysis" OR
"Geographic Information System" OR
"Geographic Mapping" OR
"geographic distribution" OR
"spatial regression" OR
"spatial autocorrelation analysis" OR
"Spatiotemporal analysis" OR
hotspot OR
"hot spot" AND tuberculosis/TB

Web of science

[(Spatial analysis) OR
(Spatio-temporal analysis) OR
(Geographic Information System) OR
(Geographic Mapping) OR
(geographic distribution) OR
(spatial regression) OR
(spatial autocorrelation analysis) OR
(Spatiotemporal analysis) OR
(hotspot) OR
(hot spot)] AND (Tuberculosis)

Additional file

Additional file 1: Figure S1. Trends in the spatial analysis of TB (note—the study included publications up to February 15, 2017). (DOCX 17 kb)

Abbreviations

CAR models: Conditional autoregressive models; GIS: Geographic information system; GWR: Geographically weighted regression; HIV: Human immunodeficiency virus; LISA: Local indicators of spatial association; NNI: Nearest neighbourhood index; PCA: Principal component analysis; TB: Tuberculosis

Acknowledgements

The authors are grateful to the University of Melbourne librarians for their extensive assistance in sourcing articles.

Funding

We did not receive funding for this study. Debebe Shaweno is the recipient of the Melbourne International Research Scholarship and Melbourne International Fee Remission Scholarship. James Trauer is a recipient of an Early Career Fellowship from the NHMRC (APP1142638).

Availability of data and materials

A list of included studies has been made available. The study protocol can be accessed on PROSPERO (CRD42016036655).

Authors' contributions

DS and EM conceived the study, which was refined by JD and JT. DS developed data extraction checklist, and DS, MK and KA extracted the data. DS drafted the manuscript, and all authors provided input into revisions and approved the final draft for submission.

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

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Medicine, University of Melbourne, Melbourne, Victoria, Australia. [2]Victorian Tuberculosis Program at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia. [3]Department of Microbiology and Immunology, University of Melbourne, Melbourne, Victoria, Australia. [4]Research School of Population Health, College of Health and Medicine, The Australian National University, Canberra, Australia. [5]Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia. [6]Burnet Institute, Melbourne, Australia. [7]Curtin University, Bentley, Western Australia, Australia. [8]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia. [9]Australian Institute of Tropical Health and Medicine, James Cook University, Townsville, Queensland, Australia.

Received: 15 May 2018 Accepted: 20 September 2018
Published online: 18 October 2018

References

1. Verma A, Schwartzman K, Behr MA, Zwerling A, Allard R, Rochefort CM, Buckeridge DL. Accuracy of prospective space-time surveillance in detecting tuberculosis transmission. Spatial Spatio-Temp Epidemiol. 2014;8:47–54.
2. Haase I, Olson S, Behr MA, Wanyeki I, Thibert L, Scott A, Zwerling A, Ross N, Brassard P, Menzies D, et al. Use of geographic and genotyping tools to characterise tuberculosis transmission in Montreal. Int J Tuber Lung Dis. 2007;11(6):632–8.
3. Theron G, Jenkins HE, Cobelens F, Abubakar I, Khan AJ, Cohen T, Dowdy DW. Data for action: collection and use of local data to end tuberculosis. Lancet. 2015;386(10010):2324–33.
4. Yates TA, Khan PY, Knight GM, Taylor JG, McHugh TD, Lipman M, White RG, Cohen T, Cobelens FG, Wood R, et al. The transmission of Mycobacterium tuberculosis in high burden settings. Lancet Infect Dis. 2016;16(2):227–38.
5. Dye C, Loyd K. Tuberculosis. In: Jamison DTBJ, Measham AR, editors. Disease control priorities in developing countries. 2nd ed. Washington DC: WorldBank; 2006.
6. McBryde ES, Denholm JT. Risk of active tuberculosis in immigrants: effects of age, region of origin and time since arrival in a low-exposure setting. Med J Aust. 2012;197(8):458–61.
7. Middelkoop K, Bekker LG, Morrow C, Zwane E, Wood R. Childhood tuberculosis infection and disease: a spatial and temporal transmission analysis in a South African township. Samj South Afr Med J. 2009;99(10):738–43.
8. Middelkoop K, Bekker LG, Morrow C, Lee N, Wood R. Decreasing household contribution to TB transmission with age: a retrospective geographic analysis of young people in a South African township. BMC Infect Dis. 2014;14:221.
9. Keshavjee S, Dowdy D, Swaminathan S. Stopping the body count: a comprehensive approach to move towards zero tuberculosis deaths. Lancet. 2015;386(10010):e46–7.
10. Sasson C, Cudnik MT, Nassel A, Semple H, Magid DJ, Sayre M, Keseg D, Haukoos JS, Warden CR. Identifying high-risk geographic areas for cardiac arrest using three methods for cluster analysis. Acad Emerg Med. 2012;19(2):139–46.
11. Shaweno D, Trauer JM, Denholm JT, McBryde ES. A novel Bayesian geospatial method for estimating tuberculosis incidence reveals many missed TB cases in Ethiopia. BMC Infect Dis. 2017;17(1):662.
12. Dangisso MH, Datiko DG, Lindtjorn B. Accessibility to tuberculosis control services and tuberculosis programme performance in southern Ethiopia. Glob Health Action. 2015;8:29443.
13. World Health Organization. Global tuberculosis report 2016: World Health Organization; 2016.
14. Clements ACA, Lwambo NJS, Blair L, Nyandindi U, Kaatano G, Kinung'hi S, Webster JP, Fenwick A, Brooker S. Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. Tropical Med Int Health. 2006;11(4):490–503.
15. Lai P-C, So F-M, Chan K-W. Spatial epidemiological approaches in disease mapping and analysis: CRC Press; 2008.
16. Wubuli A, Xue F, Jiang D, Yao X, Upur H, Wushouer Q. Socio-demographic predictors and distribution of pulmonary tuberculosis (TB) in Xinjiang, China: a spatial analysis. PLoS One. 2015:**10**(12).
17. Manley D, Flowerdew R, Steel D. Scales, levels and processes: studying spatial patterns of British census variables. Comput Environ Urban Syst. 2006;30(2):143–60.
18. Cressie N. Statistics for spatial data. Terra Nova. 1992;4(5):613–7.
19. Pfeiffer D. Spatial analysis in epidemiology. Oxford ; New York: Oxford University Press; 2008.
20. Kirby RS, Delmelle E, Eberth JM. Advances in spatial epidemiology and geographic information systems. Ann Epidemiol. 2017;27(1):1–9.
21. Smith CM, Le Comber SC, Fry H, Bull M, Leach S, Hayward AC. Spatial methods for infectious disease outbreak investigations: systematic literature review. Eurosurveillance. 2015;20(39):1–21.
22. Durr PA, Gatrell AC. GIS and spatial analysis in veterinary science: Cabi; 2004.
23. Nunes C. Tuberculosis incidence in Portugal: spatiotemporal clustering. Int J Health Geogr [Electronic Resource]. 2007;6:30.
24. Bhatt V, Tiwari N. A spatial scan statistic for survival data based on Weibull distribution. Stat Med. 2014;33(11):1867–76.
25. Higgs BW, Mohtashemi M, Grinsdale J, Kawamura LM. Early detection of tuberculosis outbreaks among the San Francisco homeless: trade-offs between spatial resolution and temporal scale. PLoS One [Electronic Resource]. 2007;2(12):e1284.
26. Dangisso MH, Datiko DG, Lindtjorn B. Spatio-temporal analysis of smear-positive tuberculosis in the Sidama Zone, Southern Ethiopia. PLoS One. 2015;10(6).
27. Areias C, Briz T, Nunes C. Pulmonary tuberculosis space-time clustering and spatial variation in temporal trends in Portugal, 2000-2010: an updated analysis. Epidemiol Infect. 2015;143(15):3211–9.
28. Kammerer JS, Shang N, Althomsons SP, Haddad MB, Grant J, Navin TR. Using statistical methods and genotyping to detect tuberculosis outbreaks. Int J Health Geogr. 2013;12:15.
29. Wang T, Xue F, Chen Y, Ma Y, Liu Y. The spatial epidemiology of tuberculosis in Linyi City, China, 2005-2010. BMC Public Health. 2012;12(1).
30. Silva AP, Souza WV, Albuquerque Mde F. Two decades of tuberculosis in a city in Northeastern Brazil: advances and challenges in time and space. Rev Soc Bras Med Trop. 2016;49(2):211–21.
31. Roth D, Otterstatter M, Wong J, Cook V, Johnston J, Mak S. Identification of spatial and cohort clustering of tuberculosis using surveillance data from British Columbia, Canada, 1990–2013. Soc Sci Med. 2016;168:214–22.
32. Gurjav U, Burneebaatar B, Narmandakh E, Tumenbayar O, Ochirbat B, Hill-Cawthorne GA, Marais BJ, Sintchenko V. Spatiotemporal evidence for cross-border spread of MDR-TB along the Trans-Siberian Railway line. Int J Tuber Lung Dis. 2015;19(11):1376–82.
33. Cao K, Yang K, Wang C, Guo J, Tao LX, Liu QR, Gehendra M, Zhang YJ, Guo XH. Spatial-temporal epidemiology of tuberculosis in Mainland China: an analysis based on Bayesian theory. Int J Environ Res Public Health. 2016;13(5).
34. de Queiroga RP, de Sa LD, Nogueira Jde A, de Lima ER, Silva AC, Pinheiro PG, Braga JU. Spatial distribution of tuberculosis and relationship with living conditions in an urban area of Campina Grande--2004 to 2007. Rev Bras Epidemiol. 2012;15(1):222–32.
35. Lin H, Shin S, Blaya JA, Zhang Z, Cegielski P, Contreras C, Asencios L, Bonilla C, Bayona J, Paciorek CJ, et al. Assessing spatiotemporal patterns of multidrug-resistant and drug-sensitive tuberculosis in a South American setting. Epidemiol Infect. 2011;139(11):1784–93.
36. Davidow AL, Marmor M, Alcabes P. Geographic diversity in tuberculosis trends and directly observed therapy, New York City, 1991 to 1994. Am J Respir Crit Care Med. 1997;156(5):1495–500.
37. Venâncio TS, Tuan TS, Nascimento LFC. Indidence of tuberculosis in children in the state of São Paulo, Brazil, under spatial approach. Cien Saude Colet. 2015;20(5):1541–7.
38. Jafari-Koshki T, Arsang-Jang S, Raei M. Applying spatiotemporal models to study risk of smear-positive tuberculosis in Iran, 2001-2012. Int J Tuber Lung Dis. 2015;19(4):469–74.
39. Jia ZW, Jia XW, Liu YX, Dye C, Chen F, Chen CS, Zhang WY, Li XW, Cao WC, Liu HL, et al. Spatial analysis of tuberculosis cases in migrants and permanent residents, Beijing, 2000-2006. Emerg Infect Dis. 2008;14(9):1413–20.
40. Houlihan CF, Mutevedzi PC, Lessells RJ, Cooke GS, Tanser FC, Newell ML. The tuberculosis challenge in a rural South African HIV programme. BMC Infect Dis. 2009;10 (no pagination)(23).
41. Jenkins HE, Plesca V, Ciobanu A, Crudu V, Galusca I, Soltan V, Serbulenco A, Zignol M, Dadu A, Dara M, et al. Assessing spatial heterogeneity of multidrug-resistant tuberculosis in a high-burden country. Eur Respir J. 2013;42(5):1291–301.
42. Musenge E, Vounatsou P, Collinson M, Tollman S, Kahn K. The contribution of spatial analysis to understanding HIV/TB mortality in children: a structural equation modelling approach. Glob Health Action. 2013;6:19266.
43. Musenge E, Vounatsou P, Kahn K. Space-time confounding adjusted determinants of child HIV/TB mortality for large zero-inflated data in rural South Africa. Spatial Spatio-Temporal Epidemiol. 2011;2(4):205–17.
44. Harling G, Castro MC. A spatial analysis of social and economic determinants of tuberculosis in Brazil. Health Place. 2014;25:56–67.
45. De Castro DB, Pinto RC, De Albuquerque BC, Sadahiro M, Braga JU. The socioeconomic factors and the indigenous component of tuberculosis in amazonas. PLoS One. 2016;11(6) (no pagination)(e0158574).
46. Wong MK, Yadav R-P, Nishikiori N, Eang MT. The association between household poverty rates and tuberculosis case notification rates in Cambodia, 2010. Western Pacific Surveill Response J. 2013;4(1):25–33.

47. Chan-Yeung M, Yeh AGO, Tam CM, Kam KM, Leung CC, Yew WW, Lam CW. Socio-demographic and geographic indicators and distribution of tuberculosis in Hong Kong: a spatial analysis. Int J Tuber Lung Dis. 2005; 9(12):1320–6.
48. Shaweno D, Shaweno T, Trauer JM, Denholm JT, McBryde ES. Heterogeneity of distribution of tuberculosis in Sheka Zone, Ethiopia: drivers and temporal trends. Int J Tuber Lung Dis. 2017;21(1):79–85 and i.
49. Alene KA, Viney K, McBryde ES, Clements ACA. Spatial patterns of multidrug resistant tuberculosis and relationships to socioeconomic, demographic and household factors in northwest Ethiopia. PLoS One. 2017;12(2) (no pagination)(e0171800).
50. Li XX, Wang LX, Zhang J, Liu YX, Zhang H, Jiang SW, Chen JX, Zhou XN. Exploration of ecological factors related to the spatial heterogeneity of tuberculosis prevalence in P. China. Glob Health Action. 2014;7:23620.
51. Rao HX, Zhang X, Zhao L, Yu J, Ren W, Zhang XL, Ma YC, Shi Y, Ma BZ, Wang X, et al. Spatial transmission and meteorological determinants of tuberculosis incidence in Qinghai Province, China: a spatial clustering panel analysis. Infect Dis Pov. 2016;5(1) (no pagination)(45).
52. Beiranvand R, Karimi A, Delpisheh A, Sayehmiri K, Soleimani S, Ghalavandi S. Correlation assessment of climate and geographic distribution of tuberculosis using geographical information system (GIS). Iran J Public Health. 2016;45(1):86–93.
53. Munch Z, Van Lill S, Booysen C, Zietsman H, Enarson D, Beyers N. Tuberculosis transmission patterns in a high-incidence area: a spatial analysis. Int J Tuber Lung Dis. 2003;7(3):271–7.
54. Patterson B, Morrow CD, Kohls D, Deignan C, Ginsburg S, Wood R. Mapping sites of high TB transmission risk: integrating the shared air and social behaviour of TB cases and adolescents in a South African township. Sci Total Environ. 2017;05.
55. Murray EJ, Marais BJ, Mans G, Beyers N, Ayles H, Godfrey-Faussett P, Wallman S, Bond V. A multidisciplinary method to map potential tuberculosis transmission 'hot spots' in high-burden communities. Int J Tuber Lung Dis. 2009;13(6):767–74.
56. Ricks PM, Cain KP, Oeltmann JE, Kammerer JS, Moonan PK. Estimating the burden of tuberculosis among foreign-born persons acquired prior to entering the U.S., 2005–2009. PLOS One. 2011;6(11):e27405.
57. Prussing C, Castillo-Salgado C, Baruch N, Cronin WA. Geo-epidemiologic and molecular characterization to identify social, cultural, and economic factors where targeted tuberculosis control activities can reduce incidence in Maryland, 2004-2010. Public Health Rep. 2013;128(Suppl 3):104–14.
58. Ng IC, Wen TH, Wang JY, Fang CT. Spatial dependency of tuberculosis incidence in Taiwan. PLoS One. 2012;7(11).
59. Jacobson LM, Garcia-Garcia Ma DL, Hernandez-Avila JE, Cano-Arellano B, Small PM, Sifuentes-Osornio J, Ponce-De-Leon A. Changes in the geographical distribution of tuberculosis patients in Veracruz, Mexico, after reinforcement of a tuberculosis control programme. Trop Med Int Health. 2005;10(4):305–11.
60. Zelner JL, Murray MB, Becerra MC, Galea J, Lecca L, Calderon R, Yataco R, Contreras C, Zhang ZB, Manjourides J, et al. Identifying hotspots of multidrug-resistant tuberculosis transmission using spatial and molecular genetic data. J Infect Dis. 2016;213(2):287–94.
61. Moonan PK, Bayona M, Quitugua TN, Oppong J, Dunbar D, Jost KC, Burgess G, Singh KP, Weis SE. Using GIS technology to identify areas of tuberculosis transmission and incidence. Int J Health Geogr. 2004;3(1):23.
62. Ribeiro FK, Pan W, Bertolde A, Vinhas SA, Peres RL, Riley L, Palaci M, Maciel EL. Genotypic and spatial analysis of Mycobacterium tuberculosis transmission in a high-incidence urban setting. Clin Infect Dis. 2015;61(5):758–66.
63. Yeboah-Manu D, Asare P, Asante-Poku A, Otchere ID, Osei-Wusu S, Danso E, Forson A, Koram KA, Gagneux S. Spatio-temporal distribution of Mycobacterium tuberculosis complex strains in Ghana. PLoS One. 2016;11(8) (no pagination)(e0161892).
64. Seraphin MN, Lauzardo M, Doggett RT, Zabala J, Morris JG Jr, Blackburn JK. Spatiotemporal clustering of Mycobacterium tuberculosis complex genotypes in Florida: genetic diversity segregated by country of birth. PLoS One [Electronic Resource]. 2016;11(4):e0153575.
65. Izumi K, Ohkado A, Uchimura K, Murase Y, Tatsumi Y, Kayebeta A, Watanabe Y, Ishikawa N. Detection of tuberculosis infection hotspots using activity spaces based spatial approach in an urban Tokyo, from 2003 to 2011. PLoS One. 2015;10(9).
66. Evans JT, Wani RL, Anderson L, Gibson AL, Smith EG, Wood A, Olowokure B, Abubakar I, Mann JS, Gardiner S, et al. A geographically-restricted but prevalent Mycobacterium tuberculosis strain identified in the West Midlands region of the UK between 1995 and 2008. PLoS One. 2011;6(3) (no pagination)(e17930).
67. Nava-Aguilera E, Lopez-Vidal Y, Harris E, Morales-Perez A, Mitchell S, Flores-Moreno M, Villegas-Arrizon A, Legorreta-Soberanis J, Ledogar R, Andersson N. Clustering of Mycobacterium tuberculosis cases in Acapulco: spoligotyping and risk factors. Clin Dev Immunol. 2011;2011:408375.
68. Yuen CM, Kammerer JS, Marks K, Navin TR, France AM. Recent transmission of tuberculosis—United States, 2011–2014. PLoS One. 2016;11(4):e0153728.
69. Nguyen D, Brassard P, Westley J, Thibert L, Proulx M, Henry K, Schwartzman K, Menzies D, Behr MA. Widespread pyrazinamide-resistant Mycobacterium tuberculosis family in a low-incidence setting. J Clin Microbiol. 2003;41(7):2878–83.
70. Gurjav U, Jelfs P, Hill-Cawthorne GA, Marais BJ, Sintchenko V. Genotype heterogeneity of Mycobacterium tuberculosis within geospatial hotspots suggests foci of imported infection in Sydney, Australia. Infect Genet Evol. 2016;40:346–51.
71. Saavedra-Campos M, Welfare W, Cleary P, Sails A, Burkitt A, Hungerford D, Okereke E, Acheson P, Petrovic M. Identifying areas and risk groups with localised Mycobacterium tuberculosis transmission in northern England from 2010 to 2012: spatiotemporal analysis incorporating highly discriminatory genotyping data. Thorax. 2016;71(8):742–8.
72. Richardson M, van Lill SW, van der Spuy GD, Munch Z, Booysen CN, Beyers N, van Helden PD, Warren RM. Historic and recent events contribute to the disease dynamics of Beijing-like Mycobacterium tuberculosis isolates in a high incidence region. Int J Tuber Lung Dis. 2002;6(11):1001–11.
73. Goswami ND, Hecker EJ, Vickery C, Ahearn MA, Cox GM, Holland DP, Naggie S, Piedrahita C, Mosher A, Torres Y, et al. Geographic information system-based screening for TB, HIV, and syphilis (GIS-THIS): a cross-sectional study. PLoS One. 2012;7 (10) (no pagination)(e46029).
74. Tanser F, Wilkinson D. Spatial implications of the tuberculosis DOTS strategy in rural South Africa: a novel application of geographical information system and global positioning system technologies. Trop Med Int Health. 1999;4(10):634–8.
75. Manjourides J, Lin HH, Shin S, Jeffery C, Contreras C, Cruz JS, Jave O, Yagui M, Asencios L, Pagano M, et al. Identifying multidrug resistant tuberculosis transmission hotspots using routinely collected data. Tuberculosis. 2012; 92(3):273–9.
76. Stephen C. Capture-recapture methods in epidemiological studies. Infect Control Hospital Epidemiol. 1996;17(4):262–6.
77. Guernier V, Guégan J-F, Deparis X. An evaluation of the actual incidence of tuberculosis in French Guiana using a capture-recapture model. Microbes Infect. 2006;8(3):721–7.
78. WHO. Technical appendix - methods used to estimate the global burden of disease caused by TB, vol. 2015; 2014.
79. Clements AC, Firth S, Dembelé R, Garba A, Touré S, Sacko M, Landouré A, Bosqué-Oliva E, Barnett AG, Brooker S. Use of Bayesian geostatistical prediction to estimate local variations in Schistosoma haematobium infection in western Africa. Bull World Health Organ. 2009;87(12):921–9.
80. Li XX, Wang LX, Zhang H, Jiang SW, Fang Q, Chen JX, Zhou XN. Spatial variations of pulmonary tuberculosis prevalence co-impacted by socio-economic and geographic factors in People's Republic of China, 2010. BMC Public Health. 2014;14:257.
81. Rytkönen MJ. Not all maps are equal: GIS and spatial analysis in epidemiology. Int J Circumpolar Health. 2004;63(1):9–24.
82. Nassel AF, Root ED, Haukoos JS, McVaney K, Colwell C, Robinson J, Eigel B, Magid DJ, Sasson C. Multiple cluster analysis for the identification of high-risk census tracts for out-of-hospital cardiac arrest (OHCA) in Denver, Colorado. Resuscitation. 2014;85(12):1667–73.
83. Wheeler DC. A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996–2003. Int J Health Geogr. 2007;6(1):13.
84. Wartenberg D, Greenberg M. Solving the cluster puzzle: clues to follow and pitfalls to avoid. Stat Med. 1993;12(19–20):1763–70.
85. Wartenberg D. Investigating disease clusters: why, when and how? J Royal Stat Soc. 2001;164(1):13–22.
86. Burra T, Jerrett M, Burnett RT, Anderson M. Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario. Can Geographer/Le Géographe Canadien. 2002;46(2):160–71.
87. Anselin L. Exploring spatial data with GeoDaTM: a workbook. Urbana. 2004; 51:61801.
88. Anselin L, Syabri I, Kho Y. GeoDa: an introduction to spatial data analysis. In: Handbook of applied spatial analysis; 2010. p. 73–89.

89. Goovaerts P, Jacquez GM. Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. Int J Health Geogr. 2004;3(1):14.
90. Goovaerts P, Jacquez GM. Detection of temporal changes in the spatial distribution of cancer rates using local Moran's I and geostatistically simulated spatial neutral models. J Geogr Syst. 2005;7(1):137–59.
91. Aamodt G, Samuelsen SO, Skrondal A. A simulation study of three methods for detecting disease clusters. Int J Health Geogr. 2006;5(1):15.
92. Stucki D, Ballif M, Egger M, Furrer H, Altpeter E, Battegay M, Droz S, Bruderer T, Coscolla M, Borrell S. Standard genotyping overestimates transmission of Mycobacterium tuberculosis among immigrants in a low-incidence country. J Clin Microbiol. 2016;54(7):1862–70.
93. Feske ML, Teeter LD, Musser JM, Graviss EA. Including the third dimension: a spatial analysis of TB cases in Houston Harris County. Tuberculosis. 2011; 91(SUPPL. 1):S24–33.
94. Tuite AR, Guthrie JL, Alexander DC, Whelan MS, Lee B, Lam K, Ma J, Fisman DN, Jamieson FB. Epidemiological evaluation of spatiotemporal and genotypic clustering of Mycobacterium tuberculosis in Ontario, Canada. Int J Tuber Lung Dis. 2013;17(10):1322–7.
95. Bishai WR, Graham NM, Harrington S, Pope DS, Hooper N, Astemborski J, Sheely L, Vlahov D, Glass GE, Chaisson RE. Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy. JAMA. 1998;280(19):1679–84.
96. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med. 2013; 10(2):e1001387.
97. Wampande EM, Mupere E, Debanne SM, Asiimwe BB, Nsereko M, Mayanja H, Eisenach K, Kaplan G, Boom HW, Gagneux S, et al. Long-term dominance of Mycobacterium tuberculosis Uganda family in peri-urban Kampala-Uganda is not associated with cavitary disease. BMC Infect Dis. 2013;13:484.
98. Streicher EM, Warren RM, Kewley C, Simpson J, Rastogi N, Sola C, van der Spuy GD, van Helden PD, Victor TC. Genotypic and phenotypic characterization of drug-resistant Mycobacterium tuberculosis isolates from rural districts of the Western Cape Province of South Africa. J Clin Microbiol. 2004;42(2):891–4.
99. Souza WV, Carvalho MS, Albuquerque MDFPM, Barcellos CC, Ximenes RAA. Tuberculosis in intra-urban settings: a Bayesian approach. Trop Med Int Health. 2007;12(3):323–30.
100. Erazo C, Pereira SM, Da Conceição N. Costa M, Evangelista-Filho D, Braga JU, Barreto ML. Tuberculosis and living conditions in Salvador, Brazil: a spatial analysis. Rev Panamericana de Salud Publica/Pan American Journal of Public Health. 2014;36(1):24–30.
101. da Roza DL, Caccia-Bava Mdo C, Martinez EZ. Spatio-temporal patterns of tuberculosis incidence in Ribeirao Preto, state of Sao Paulo, southeast Brazil, and their relationship with social vulnerability: a Bayesian analysis. Rev Soc Bras Med Trop. 2012;45(5):607–15.
102. Wang W, Jin YY, Yan C, Ahan A, Cao MQ: Local spatial variations analysis of smear-positive tuberculosis in Xinjiang using geographically weighted regression model. BMC Public Health 2016, 16.
103. Sun W, Gong J, Zhou J, Zhao Y, Tan J, Ibrahim AN, Zhou Y. A spatial, social and environmental study of tuberculosis in China using statistical and GIS technology. Int J Environ Res Public Health [Electronic Resource]. 2015;12(2):1425–48.
104. Liu Y, Jiang S, Liu Y, Wang R, Li X, Yuan Z, Wang L, Xue F. Spatial epidemiology and spatial ecology study of worldwide drug-resistant tuberculosis. Int J Health Geogr. 2011;10.
105. Jenkins HE, Gegia M, Furin J, Kalandadze I, Nanava U, Chakhaia T, Cohen T. Geographical heterogeneity of multidrug-resistant tuberculosis in Georgia, January 2009 to June 2011. Eurosurveillance. 2014;19(11).
106. Gaudette LA, Ellis E. Tuberculosis in Canada: a focal disease requiring distinct control strategies for different risk groups. Tubercle Lung Dis. 1993; 74(4):244–53.
107. Froggatt K. Tuberculosis: spatial and demographic incidence in Bradford, 1980-2. J Epidemiol Community Health. 1985;39(1):20–6.
108. Zorzenon dos Santos RM, Amador A, de Souza WV, de Albuquerque MF, Ponce Dawson S, Ruffino-Netto A, Zarate-Blades CR, Silva CL. A dynamic analysis of tuberculosis dissemination to improve control and surveillance. PLoS One [Electronic Resource]. 2010;5(11):e14140.
109. Touray K, Adetifa IM, Jallow A, Rigby J, Jeffries D, Cheung YB, Donkor S, Adegbola RA, Hill PC. Spatial analysis of tuberculosis in an urban west African setting: is there evidence of clustering? Tropical Med Int Health. 2010;15(6):664–72.
110. Tadesse T, Demissie M, Berhane Y, Kebede Y, Abebe M. The clustering of smear-positive tuberculosis in Dabat, Ethiopia: a population based cross sectional study. PLoS One [Electronic Resource]. 2013;8(5):e65022.
111. Shah L, Choi HW, Berrang-Ford L, Henostroza G, Krapp F, Zamudio C, Heymann SJ, Kaufman JS, Ciampi A, Seas C, et al. Geographic predictors of primary multidrug-resistant tuberculosis cases in an endemic area of Lima, Peru. Int J Tuber Lung Dis. 2014;18(11):1307–14.
112. Lin HH, Shin SS, Contreras C, Asencios L, Paciorek CJ, Cohen T. Use of spatial information to predict multidrug resistance in tuberculosis patients, Peru. Emerg Infect Dis. 2012;18(5):811–3.
113. Lai PC, Low CT, Tse WS, Tsui CK, Lee H, Hui PK. Risk of tuberculosis in high-rise and high density dwellings: an exploratory spatial analysis. Environ Pollution (Barking, Essex : 1987). 2013;183:40–5.
114. Kolifarhood G, Khorasani-Zavareh D, Salarilak S, Shoghli A, Khosravi N. Spatial and non-spatial determinants of successful tuberculosis treatment outcomes: an implication of geographical information systems in health policy-making in a developing country. J Epidemiol Glob Health. 2015;5(3):221–30.
115. Hino P, Villa TC, Sassaki CM, Nogueira Jde A, dos Santos CB. Geoprocessing in health area. Rev Latino-Am Enfermagem. 2006;14(6):939–43.
116. Ge E, Lai PC, Zhang X, Yang X, Li X, Wang H, Wei X. Regional transport and its association with tuberculosis in the Shandong province of China, 2009-2011. J Transp Geogr. 2015;46:232–43.
117. Dragioevio S, Schuurman N, Fitzgerald J. The utility of exploratory spatial data analysis in the study of tuberculosis incidences in an urban Canadian population. Cartographica. 2004;39(2):29–39.
118. Dominkovics P, Granell C, Pérez-Navarro A, Casals M, Orcau À, Caylà JA. Development of spatial density maps based on geoprocessing web services: application to tuberculosis incidence in Barcelona, Spain. Int J Health Geogr. 2011;10.
119. Dogba JB, Cadmus SI, Olugasa BO. Mapping of Mycobacterium tuberculosis cases in post-conflict Liberia, 2008-2012: a descriptive and categorical analysis of age, gender and seasonal pattern. Afr J Med Med Sci. 2014;(43 Suppl):117–24.
120. De Abreu E Silva M, Di Lorenzo Oliveira C, Teixeira Neto RG, Camargos PA. Spatial distribution of tuberculosis from 2002 to 2012 in a midsize city in Brazil. BMC Public Health. 2016;16(1).
121. Cegielski JP, Griffith DE, McGaha PK, Wolfgang M, Robinson CB, Clark PA, Hassell WL, Robison VA, Walker KP Jr, Wallace C. Eliminating tuberculosis one neighborhood at a time.[Reprint in Rev Panam Salud Publica. 2013 Oct; 34(4):284-94 Note: Original is in English and republished one is in Spanish.; PMID: 24301742], [Reprint in Am J Public Health. 2014 Apr;104 Suppl 2:S214-33; PMID: 24899457]. Am J Public Health. 2013;103(7):1292–300.
122. Cadmus SI, Akingbogun AA, Adesokan HK. Using geographical information system to model the spread of tuberculosis in the University of Ibadan, Nigeria. Afr J Med Med Sci. 2010;39(Suppl):193–9.
123. Zhou H, Yang X, Zhao S, Pan X, Xu J. Spatial epidemiology and risk factors of pulmonary tuberculosis morbidity in Wenchuan earthquake-stricken area. J Evid-Based Med. 2016;9(2):69–76.
124. Yeh YP, Chang HJ, Yang J, Chang SH, Suo J, Chen THH. Incidence of tuberculosis in mountain areas and surrounding townships: dose-response relationship by geographic analysis. Ann Epidemiol. 2005;15(7):526–32.
125. Yang X, Liu Q, Zhang R. Epidemiology of pulmonary tuberculosis in Wenchuan earthquake stricken area: population-based study. J Evid-Based Med. 2013;6(3):149–56.
126. Uthman OA. Spatial and temporal variations in incidence of tuberculosis in Africa, 1991 to 2005. World Health Popul. 2008;10(2):5–15.
127. Randremanana RV, Richard V, Rakotomanana F, Sabatier P, Bicout DJ. Bayesian mapping of pulmonary tuberculosis in Antananarivo, Madagascar. BMC Infect Dis. 2010;10 (no pagination)(21).
128. Pereira AG, Medronho Rde A, Escosteguy CC, Valencia LI, Magalhaes Mde A. Spatial distribution and socioeconomic context of tuberculosis in Rio de Janeiro, Brazil. Rev Saude Publica. 2015;49:48.
129. Pang PTT, Leung CC, Lee SS. Neighbourhood risk factors for tuberculosis in Hong Kong. Int J Tuber Lung Dis. 2010;14(5):585–92.
130. Nana Yakam A, Noeske J, Dambach P, Bowong S, Fono LA, Ngatchou-Wandji J. Spatial analysis of tuberculosis in Douala, Cameroon: clustering and links with socio-economic status. Int J Tuber Lung Dis. 2014;18(3):292–7.
131. Maciel ELN, Pan W, Dietze R, Peres RL, Vinhas SA, Ribeiro FK, Palaci M, Rodrigues RR, Zandonade E, Golub JE. Spatial patterns of pulmonary

tuberculosis incidence and their relationship to socio-economic status in Vitoria, Brazil. Int J Tuber Lung Dis. 2010;14(11):1395–402.
132. Lopez De Fede A, Stewart JE, Harris MJ, Mayfield-Smith K. Tuberculosis in socio-economically deprived neighborhoods: missed opportunities for prevention. Int J Tuber Lung Dis. 2008;12(12):1425–30.
133. Liu Y, Li X, Wang W, Li Z, Hou M, He Y, Wu W, Wang H, Liang H, Guo X. Investigation of space-time clusters and geospatial hot spots for the occurrence of tuberculosis in Beijing. Int J Tuber Lung Dis. 2012;16(4): 486–91.
134. Lim JR, Gandhi NR, Mthiyane T, Mlisana K, Moodley J, Jaglal P, Ramdin N, Brust JCM, Ismail N, Rustomjee R, et al. Incidence and geographic distribution of extensively drug-resistant tuberculosis in KwaZulu-Natal Province, South Africa. PLoS One. 2015;10(7).
135. Li L, Xi YL, Ren F. Spatio-temporal distribution characteristics and trajectory similarity analysis of tuberculosis in Beijing, China. Int J Environ Res Public Health. 2016;13(3).
136. Kistemann T, Munzinger A, Dangendorf F. Spatial patterns of tuberculosis incidence in Cologne (Germany). Soc Sci Med. 2002;55(1):7–19.
137. Kakchapati S, Choonpradub C, Lim A. Spatial and temporal variations in tuberculosis incidence, Nepal. Southeast Asian J Trop Med Public Health. 2014;45(1):95.
138. Hassarangsee S, Tripathi NK, Souris M. Spatial pattern detection of tuberculosis: a case study of Si Sa Ket province, Thailand. Int J Environ Res Public Health. 2015;12(12):16005–18.
139. Ge E, Zhang X, Wang X, Wei X. Spatial and temporal analysis of tuberculosis in Zhejiang Province, China, 2009–2012. Infect Dis Poverty. 2016;5(1) (no pagination)(11).
140. Fluegge KR. Using spatial disease patterns and patient-level characteristics to describe prevalence elastic behavior in treatment for latent tuberculosis infection (LTBI). Public Health Nurs. 2015;32(5):517–31.
141. Couceiro L, Santana P, Nunes C. Pulmonary tuberculosis and risk factors in Portugal: a spatial analysis. Int J Tuber Lung Dis. 2011;15(11):1445–54.
142. Chandrasekaran SK, Arivarignan G. Disease mapping using mixture distribution. Indian J Med Res. 2006;123(6):788–98.
143. Burgess L. Tuberculosis and urban ecological structure: the Derby case, 1979–83. East Midland Geogr. 1986;9(1–2):9–20.
144. Beyers N, Gie RP, Zietsman HL, Kunneke M, Hauman J, Tatley M, Donald PR. The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. South Afr Med J Suid-Afrikaanse Tydskrif Vir Geneeskunde. 1996;86(1):40–1 44.
145. Alvarez-Hernandez G, Lara-Valencia F, Reyes-Castro PA, Rascon-Pacheco RA. An analysis of spatial and socio-economic determinants of tuberculosis in Hermosillo, Mexico, 2000-2006. Int J Tuber Lung Dis. 2010;14(6):708–13.
146. Acevedo-Garcia D. Zip code-level risk factors for tuberculosis: neighborhood environment and residential segregation in New Jersey, 1985-1992. Am J Public Health. 2001;91(5):734–41.
147. Pinto ML, da Silva TC, Gomes LCF, Bertolozzi MR, Villavicencio LMM, Azevedo KMFA, de Figueiredo TMRM. Occurrence of tuberculosis cases in Crato, Ceará, from 2002 to 2011: a spatial analisys of specific standards. Rev Brasil Epidemiol. 2015;18(2):313–25.
148. Srinivasan R, Venkatesan P. Bayesian spatio-temporal model for tuberculosis in India. Indian J Med Res. 2015;142(April):478–80.
149. Schlattmann P, Dietz E, Bohning D. Covariate adjusted mixture models and disease mapping with the program DismapWin. Stat Med. 1996;15(7–9): 919–29.
150. Zhao F, Cheng S, He G, Huang F, Zhang H, Xu B, Murimwa TC, Cheng J, Hu D, Wang L. Space-time clustering characteristics of tuberculosis in China, 2005-2011. PLoS One. 2013, 8(12).
151. Zaragoza Bastida A, Hernandez Tellez M, Bustamante Montes LP, Medina Torres I, Jaramillo Paniagua JN, Mendoza Martinez GD, Ramirez Duran N. Spatial and temporal distribution of tuberculosis in the State of Mexico, Mexico. Thescientificworldjournal. 2012;2012:570278.
152. Yamamura M, de Freitas IM, Santo Neto M, Chiaravalloti Neto F, Popolin MA, Arroyo LH, Rodrigues LB, Crispim JA, Arcencio RA. Spatial analysis of avoidable hospitalizations due to tuberculosis in Ribeirao Preto, SP, Brazil (2006-2012). Rev Saude Publica. 2016(50):20.
153. Tiwari N, Kandpal V, Tewari A, Rao KRM, Tolia VS. Investigation of tuberculosis clusters in Dehradun city of India. Asian Pac J Trop Med. 2010; 3(6):486–90.
154. Tiwari N, Adhikari CM, Tewari A, Kandpal V. Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic. Int J Health Geogr [Electronic Resource]. 2006;5:33.
155. Santos Neto M, Yamamura M, Garcia MC, Popolin MP, Rodrigues LB, Chiaravalloti Neto F, Fronteira I, Arcencio RA. Pulmonary tuberculosis in Sao Luis, State of Maranhao, Brazil: space and space-time risk clusters for death (2008-2012). Rev Soc Bras Med Trop. 2015;48(1):69–76.
156. Randremanana RV, Sabatier P, Rakotomanana F, Randriamanantena A, Richard V. Spatial clustering of pulmonary tuberculosis and impact of the care factors in Antananarivo City. Tropical Med Int Health. 2009;14(4):429–37.
157. Rakotosamimanana S, Mandrosovololona V, Rakotonirina J, Ramamonjisoa J, Ranjalahy JR, Randremanana RV, Rakotomanana F. Spatial analysis of pulmonary tuberculosis in Antananarivo Madagascar: tuberculosis-related knowledge, attitude and practice. PLoS One. 2014;9(11).
158. Onozuka D, Hagihara A. Geographic prediction of tuberculosis clusters in Fukuoka, Japan, using the space-time scan statistic. BMC Infect Dis. 2007;7.
159. Olfatifar M, Karami M, Hosseini SM, Parvin M. Clustering of pulmonary tuberculosis in Hamadan province, west of Iran: a population based cross sectional study (2005-2013). J Res Health Sci. 2016;16(3):166–9.
160. Gomez-Barroso D, Rodriguez-Valin E, Ramis R, Cano R. Spatio-temporal analysis of tuberculosis in Spain, 2008-2010. Int J Tuber Lung Dis. 2013;17(6): 745–51.
161. Tsai PJ, Lin ML, Chu CM, Perng CH. Spatial autocorrelation analysis of health care hotspots in Taiwan in 2006. BMC Public Health. 2009;9.
162. Mokrousov I. Genetic geography of Mycobacterium tuberculosis Beijing genotype: a multifacet mirror of human history? Infect Genet Evol. 2008;8(6): 777–85.
163. Brassard P, Henry KA, Schwartzman K, Jomphe M, Olson SH. Geography and genealogy of the human host harbouring a distinctive drug-resistant strain of tuberculosis. Infect Genet Evol. 2008;8(3):247–57.
164. Li T, He XX, Chang ZR, Ren YH, Zhou JY, Ju LR, Jia ZW. Impact of new migrant populations on the spatial distribution of tuberculosis in Beijing. Int J Tuber Lung Dis. 2011;15(2):163–8.
165. Wallace D. The resurgence of tuberculosis in New York City: a mixed hierarchically and spatially diffused epidemic. Am J Public Health. 1994; 84(6):1000–2.
166. Wei W, Wei-Sheng Z, Ahan A, Ci Y, Wei-Wen Z, Ming-Qin C. The characteristics of TB epidemic and TB/HIV co-infection epidemic: a 2007–2013 retrospective study in Urumqi, Xinjiang Province, China. PloS One. 2016;11(10):e0164947.
167. Egunjobi L. Spatial distribution of mortality from leading notifiable diseases in Nigeria. Soc Sci Med. 1993;36(10):1267–72.
168. Marlow MA, Maciel EL, Sales CM, Gomes T, Snyder RE, Daumas RP, Riley LW. Tuberculosis DALY-gap: spatial and quantitative comparison of disease burden across urban slum and non-slum census tracts. J Urban Health. 2015;92(4):622–34.
169. Mathema B, Bifani PJ, Driscoll J, Steinlein L, Kurepina N, Moghazeh SL, Shashkina E, Marras SA, Campbell S, Mangura B, et al. Identification and evolution of an IS6110 low-copy-number Mycobacterium tuberculosis cluster. J Infect Dis. 2002;185(5):641–9.
170. Souza WV, Ximenes R, Albuquerque MFM, Lapa TM, Portugal JL, Lima MLC, Martelli CMT. The use of socioeconomic factors in mapping tuberculosis risk areas in a city of northeastern Brazil. Rev Panamericana de Salud Publica/ Pan American Journal of Public Health. 2000;8(6):403–10.
171. Yamamura M, Santos-Neto M, dos Santos RA, Garcia MC, Nogueira JA, Arcencio RA. Epidemiological characteristics of cases of death from tuberculosis and vulnerable territories. Rev Latino-Am Enfermagem. 2015; 23(5):910–8.
172. Perri BR, Proops D, Moonan PK, Munsiff SS, Kreiswirth BN, Kurepina N, Goranson C, Ahuja SD. Mycobacterium tuberculosis cluster with developing drug resistance, New York, New York, USA, 2003-2009. Emerg Infect Dis. 2011;17(3):372–8.
173. Terlikbayeva A, Hermosilla S, Galea S, Schluger N, Yegeubayeva S, Abildayev T, Muminov T, Akiyanova F, Bartkowiak L, Zhumadilov Z, et al. Tuberculosis in Kazakhstan: analysis of risk determinants in national surveillance data. BMC Infect Dis. 2012;12 (no pagination)(262).
174. Lima MD, Martins-Melo FR, Heukelbach J, Alencar CH, Boigny RN, Ramos AN. Mortality related to tuberculosis-HIV/AIDS co-infection in Brazil, 2000-2011: epidemiological patterns and time trends. Cadernos Saude Publica. 2016;32(10):e00026715.
175. Santos Neto M, Yamamura M, Garcia MCC, Popolin MP, Rodrigues LBB, Chiaravalloti Neto F, Fronteira I, Arcêncio RA. Pulmonary tuberculosis in São

Luis, State of Maranhão, Brazil: space and space-time risk clusters for death (2008-2012). Rev Soc Bras Med Trop. 2015;48(1):69–76.
176. Sousa P, Oliveira A, Gomes M, Gaio AR, Duarte R. Longitudinal clustering of tuberculosis incidence and predictors for the time profiles: the impact of HIV. Int J Tuber Lung Dis. 2016;20(8):1027–32.
177. Crisan A, Wong HY, Johnston JC, Tang P, Colijn C, Otterstatter M, Hiscoe L, Parker R, Pollock SL, Gardy JL. Spatio-temporal analysis of tuberculous infection risk among clients of a homeless shelter during an outbreak. Int J Tuber Lung Dis. 2015;19(9):1033–8.
178. Feske ML, Teeter LD, Musser JM, Graviss EA. Giving TB wheels: public transportation as a risk factor for tuberculosis transmission. Tuberculosis. 2011;91(Suppl 1):S16–23.
179. Herrero MB, Arrossi S, Ramos S, Braga JU. Spatial analysis of the tuberculosis treatment dropout, Buenos Aires, Argentina. Rev Saude Publica. 2015;49.
180. Jacob BJ, Krapp F, Ponce M, Gottuzzo E, Griffith DA, Novak RJ. Accounting for autocorrelation in multi-drug resistant tuberculosis predictors using a set of parsimonious orthogonal eigenvectors aggregated in geographic space. Geospat Health. 2010;4(2):201–17.
181. Kang J, Zhang N, Shi R. A Bayesian nonparametric model for spatially distributed multivariate binary data with application to a multidrug-resistant tuberculosis (MDR-TB) study. Biometrics. 2014;70(4):981–92.
182. Leung CC, Yew WW, Tam CM, Chan CK, Chang KC, Law WS, Wong MY, Au KF. Socio-economic factors and tuberculosis: a district-based ecological analysis in Hong Kong. Int J Tuber Lung Dis. 2004;8(8):958–64.
183. Nunes C, Duarte R, Veiga AM, Taylor B. Who are the patients that default tuberculosis treatment? - space matters! Epidemiol Infect. 2017:1–5.
184. Nunes C, Taylor BM. Modelling the time to detection of urban tuberculosis in two big cities in Portugal: a spatial survival analysis. Int J Tuber Lung Dis. 2016;20(9):1219–25.
185. Obasanya J, Abdurrahman ST, Oladimeji O, Lawson L, Dacombe R, Chukwueme N, Abiola T, Mustapha G, Sola C, Dominguez J, et al. Tuberculosis case detection in Nigeria, the unfinished agenda. Tropical Med Int Health. 2015;20(10):1396–402.
186. Rodrigues AL Jr, Ruffino-Netto A, de Castilho EA. Spatial distribution of M. tuberculosis-HIV coinfection in Sao Paulo State, Brazil, 1991-2001. [Portuguese, English]. Rev Saude Publica. 2006;40(2):265–70.
187. Ross JM, Cattamanchi A, Miller CR, Tatem AJ, Katamba A, Haguma P, Handley MA, Davis JL. Investigating barriers to tuberculosis evaluation in Uganda using geographic information systems. Am J Trop Med Hyg. 2015;93(4):733–8.
188. Tipayamongkholgul M, Podang J, Siri S. Spatial analysis of social determinants for tuberculosis in Thailand. J Med Assoc Thailand = Chotmaihet Thangphaet. 2013;96(Suppl 5):S116–21.
189. Wanyeki I, Olson S, Brassard P, Menzies D, Ross N, Behr M, Schwartzman K. Dwellings, crowding, and tuberculosis in Montreal. Soc Sci Med. 2006;63(2):501–11.
190. Rodrigues-Junior AL, Ruffino-Netto A, de Castilho EA. Spatial distribution of the human development index, HIV infection and AIDS-tuberculosis comorbidity: Brazil, 1982-2007. Rev Brasil Epidemiol. 2014;17(Suppl 2):204–15.
191. Santos-Neto M, Yamamura M, Garcia MC, Popolin MP, Silveira TR, Arcencio RA. Spatial analysis of deaths from pulmonary tuberculosis in the city of Sao Luis, Brazil. J Bras Pneumol. 2014;40(5):543–51.
192. Uthman OA, Yahaya I, Ashfaq K, Uthman MB. A trend analysis and sub-regional distribution in number of people living with HIV and dying with TB in Africa, 1991 to 2006. Int J Health Geogr [Electronic Resource]. 2009;8:65.
193. Carter A, Zwerling A, Olson S, Tannenbaum T-N, Schwartzman K. Tuberculosis and the city. Health Place. 2009;15(3):807–13.
194. Chamie G, Wandera B, Marquez C, Kato-Maeda M, Kamya MR, Havlir DV, Charlebois ED. Identifying locations of recent TB transmission in rural Uganda: a multidisciplinary approach. Trop Med Int Health. 2015;20(4):537–45.
195. McGuigan MA, Yamada J. The geographic distribution of tuberculosis and pyridoxine supply in Ontario. Can J Hospital Pharm. 1995;48(6):348–51.
196. Strauss R, Fulop G, Pfeifer C. Tuberculosis in Austria 1995-99: geographical distribution and trends. Euro surveillance. 2003;8(1):19–26.
197. Smith CM, Hayward AC. DotMapper: an open source tool for creating interactive disease point maps. BMC Infect Dis. 2016;16:145.
198. MRC Tuberculosis and chest Diseases Unit. The geographical distribution of tuberculosis notifications in a national survey of England and Wales (1978–79). Report from the Medical Research Council Tuberculosis and Chest Diseases Unit. Tubercle. 1982;63(2):75–88.
199. Bai J, Zou G, Mu S, Ma Y. Using spatial analysis to identify tuberculosis transmission and surveillance. In: Lecture Notes in Electrical Engineering, vol. 277: LNEE; 2014. p. 337–44.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



APPENDIX 2: EXPLORING PROTEIN SUPERSECONDARY STRUCTURE THROUGH CHANGES IN PROTEIN FOLDING, STABILITY, AND FLEXIBILITY



Chapter 9

Exploring Protein Supersecondary Structure Through Changes in Protein Folding, Stability, and Flexibility

Douglas E. V. Pires, Carlos H. M. Rodrigues, Amanda T. S. Albanaz, Malancha Karmakar, Yoochan Myung, Joicymara Xavier, Eleni-Maria Michanetzi, Stephanie Portelli, and David B. Ascher

Abstract

The ability to predict how mutations affect protein structure, folding, and flexibility can elucidate the molecular mechanisms leading to disruption of supersecondary structures, the emergence of phenotypes, as well guiding rational protein engineering. The advent of fast and accurate computational tools has enabled us to comprehensively explore the landscape of mutation effects on protein structures, prioritizing mutations for rational experimental validation.

Here we describe the use of two complementary web-based in silico methods, DUET and DynaMut, developed to infer the effects of mutations on folding, stability, and flexibility and how they can be used to explore and interpret these effects on protein supersecondary structures.

Key words Missense mutations, Protein stability and folding, Machine learning, Normal mode analysis, Graph-based signatures, DUET, DynaMut

1 Introduction

Proteins are marginally stable, versatile macromolecules involved in a large variety of biochemical processes which are strictly linked and regulated by their native conformation. Mutations leading to changes in protein folding, stability, and conformation can have large phenotypic consequences, responsible for the development of many genetic disorders [1–14], including cancers, and even responsible for changes in drug susceptibility [15–27]. While these effects are commonly thought about in terms of reduced protein stability, mutations leading to increased stability and rigidification of the molecule can be equally deleterious. Maintaining, or enhancing, protein stability, and the identification of mutations that do not negatively affect protein stability, also remains one of the most difficult and important challenges in protein engineering.

Alexander E. Kister (ed.), *Protein Supersecondary Structures: Methods and Protocols*, Methods in Molecular Biology, vol. 1958, https://doi.org/10.1007/978-1-4939-9161-7_9, © Springer Science+Business Media, LLC, part of Springer Nature 2019

While experimental validation of protein thermodynamic parameters remains a laborious task, the development of novel robust and scalable computational methods (Table 1) has allowed for the evaluation of the complete landscape of structural effects of mutations in a protein system and their effects on protein stability and flexibility within minutes, enabling rapid mutation prioritization.

Using the concept of graph-based signatures, we have developed robust methods for quantitatively analyzing effects of single missense mutations on protein stability, flexibility, and interactions [9, 28–37]. DUET [37] (http://biosig.unimelb.edu.au/duet) is a machine learning-based approach that integrates and optimizes two complementary methods in an optimized predictor (mCSM-Stability [36] and SDM [38]) using support vector machines. This method enables the accurate assessment of the effects of mutations on protein folding and stability. DynaMut [28] (http://biosig.unimelb.edu.au/dynamut) is a novel method that takes into account molecular motions and, by combining the graph-based signatures with coarse-grained normal mode analysis, generates a consensus prediction of effects of mutations on the protein conformational repertoire. These methods together compose a powerful platform that allows users to navigate the landscape of mutations effects on folding, stability, and flexibility.

2 Materials

DUET and DynaMut are structure-based methods for assessing effects of single-point missense mutations on protein stability/folding and protein flexibility/conformation, respectively. For both methods, users are required to provide:

1. Wild-type protein structure in PDB format: For both methods, a wild-type structure of the protein of interest in the Protein Data Bank [39] format (.pdb) must be provided to perform the predictions. This can be either (a) an experimentally solved structure, with previously solved structures available in the Protein Data Bank, or (b) a model, for instance, obtained via comparative homology modeling (*see* **Note 1** on how to deal with oligomeric structures). We have previously shown that using homology models built using templates down to 25% sequence identity does not significantly reduce predictive performance of either method (*see* **Note 2**). Users have the option to either upload the structure file or provide the PDB accession code when they wish to use an experimental structure previously deposited into the PDB (http://www.rcsb.org or http://www.ebi.ac.uk/pdbe/) (*see* **Note 3**).
2. Mutation information: The user also needs to supply information on the mutation or mutations they wish to analyze,

Table 1
List of freely available webservers and software for predicting effects of single-point mutations on protein folding, thermostability, and flexibility

	Method	Technique	Data set	Correlation	DOI	Publication year
Folding	mCSM-Stability	Structural signatures	ProTherm—351 mutations	0.73	https://doi.org/10.1093/bioinformatics/btt691	2014
	SDM2	Environment-specific substitution tables	ProTherm—351 mutations	0.61	https://doi.org/10.1093/nar/gkx439	2017
	DUET	Integrated approach	ProTherm—351 mutations	0.71	https://doi.org/10.1093/nar/gku411	2014
	Eris	Physical force field with atomic modeling	ProTherm—351 mutations	0.35	https://doi.org/10.1038/nmeth0607-466	2007
	I-Mutant 2.0	Neighboring residue composition	ProTherm—351 mutations	0.29	https://doi.org/10.1093/nar/gki375	2005
	Auto-Mute	Delaunay tessellation	ProTherm—351 mutations	0.46	https://doi.org/10.1155/2014/278385	2014
	CUPSAT	Atom potentials and torsion angle potentials	ProTherm—351 mutations	0.37	https://doi.org/10.1093/nar/gkl190	2006
	MAESTRO	Statistical scoring functions	ProTherm—351 mutations	0.70	https://doi.org/10.1186/s12859-015-0548-6	2015
	FoldX	Empirical full-atom force field	ProTherm—351 mutations	0.35	https://doi.org/10.1093/nar/gki387	2005
	PoPMuSiC	Statistical potentials and neural networks	ProTherm—351 mutations	0.67	https://doi.org/10.1186/1471-2105-12-151	2011
	NeEMO	Residue interaction networks	ProTherm—351 mutations	0.67	https://doi.org/10.1186/1471-2164-15-S4-S7	2014
Thermal stability	HoTMuSiC	Statistical potentials	ProTherm—1626 mutations	0.59	https://doi.org/10.1038/srep23257	2015
	FireProt	Structural and evolutionary information	ProTherm—1152 mutations	87% precision	https://doi.org/10.1093/nar/gkx285	2017
Flexibility	DynaMut	Structural signatures and NMA	ProTherm (2004)—351 mutations	0.69	https://doi.org/10.1093/nar/gky300	2018

including (1) the chain identifier (one-letter code of the chain, which corresponds to the 22nd column of the coordinate section in the PDB file where the mutation occurs) (*see* **Note 1**) and (2) the mutation code, which consists of the one-letter amino acid residue code of the wild-type residue, the residue number position as in the PDB file (columns 23–26 of the coordinate section), and the one-letter code of the mutated residue (e.g., R282W denotes a mutation from arginine to tryptophan at residue position 282).

3 Methods

3.1 Predicting and Analyzing Effects of Mutation on Protein Stability and Folding with DUET

1. DUET is freely available as a user-friendly web interface and is compatible with most operating systems and browsers. Open up the prediction server, http://biosig.unimelb.edu.au/duet/stability, on a web browser of your preference.
2. Provide the wild-type protein structure of interest by either uploading a PDB file or supplying a valid four-letter PDB accession code (Fig. 1a).
3. DUET offers users the option of two prediction modes, (a) assessing stability effects of a single mutation or (b) systematically evaluating all possible mutations at a given residue position. For a single mutation, users need to provide the mutation information and the mutation chain. For systematic evaluation, the one-letter code of the mutated residue is omitted.

3.2 DUET Prediction Output

1. If a single mutation is provided, after processing, the results page is shown (Fig. 1b), which includes information about the mutation and the predicted effects on stability for DUET and for the individual methods (mCSM-Stability and SDM). An interactive molecular visualization is also shown, allowing users to inspect the wild-type residue environment.
2. For systematic evaluation of a given residue, the predicted effects on protein stability for all 19 possible mutations are shown in tabular format (Fig. 1c).
3. Predicted effects are given as the change in Gibbs Free Energy, ΔΔG (kcal/mol), with negative values denoting destabilizing mutations and positive values, stabilizing ones. While users should interpret the values in the context of the protein system being studied, previous studies have used a rule of thumb that highly destabilizing/stabilizing mutations are those with a predicted $|\Delta\Delta G| > 1.0$ kcal/mol; and moderately destabilizing/stabilizing mutations are those with a predicted |ΔΔG| between 0.5 and 1.0. *See* **Notes 4** and **5** for further information on how to interpret results.

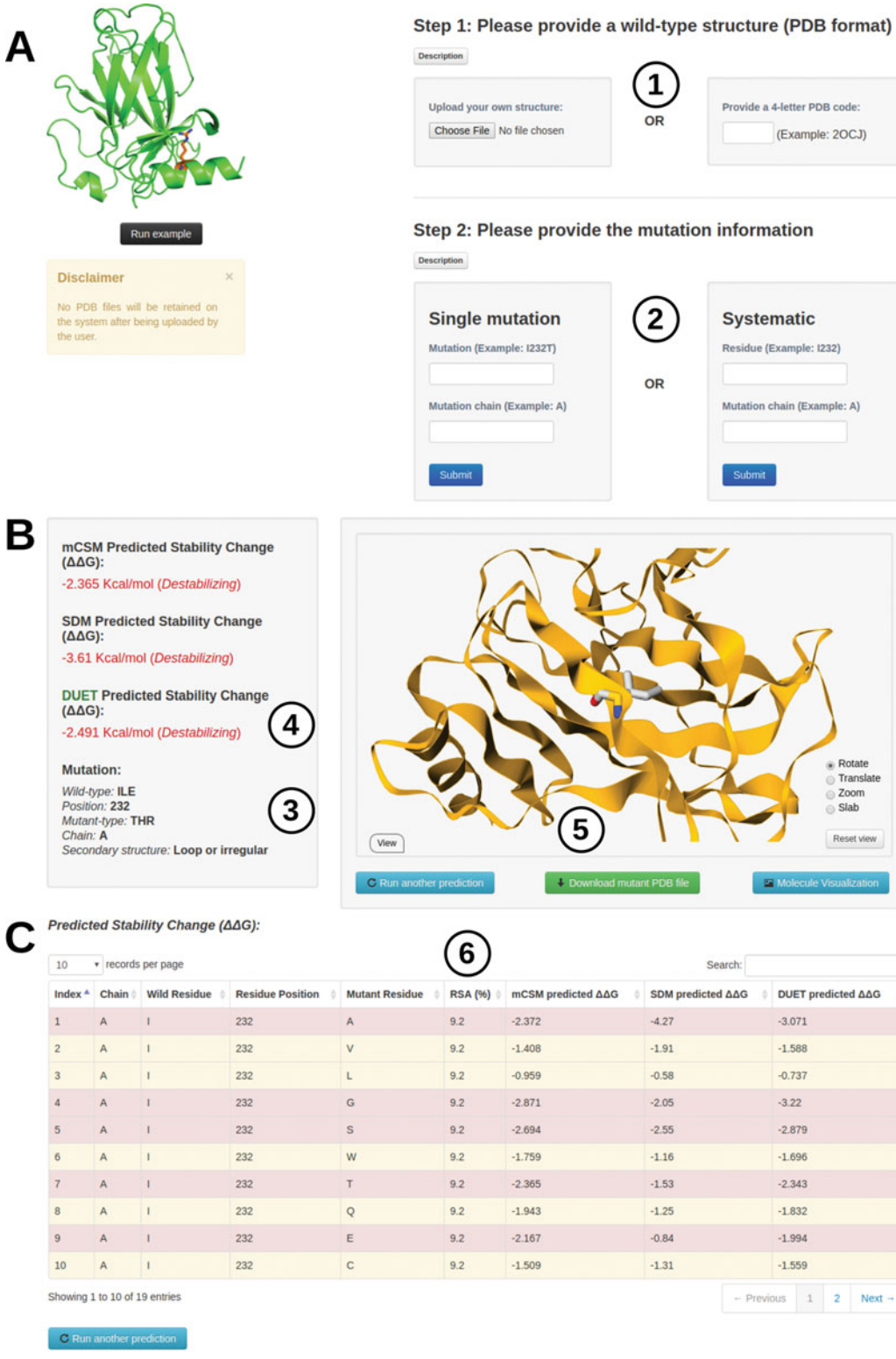


Fig. 1 DUET submission and results web interface. (**a**) The submission page allows users to either provide its own PDB file or inform an accession code of a protein of interest (*1*). Users have the option to analyze a

3.3 Predicting and Analyzing Effects of Mutations on Protein Flexibility and Conformation with DynaMut

1. As with DUET, DynaMut predicted changes upon mutation in protein stability are presented as a change in the Gibbs Free Energy of folding and stability ($\Delta\Delta G$ in kcal/mol), calculated as the difference between the wild-type and mutant proteins: $\Delta\Delta G = \Delta G_{wt} - \Delta G_{mt}$. A positive value denotes a stabilizing mutation, while a negative value denotes a destabilizing one. The DynaMut consensus prediction uses both normal mode analysis and graph-based signatures to more accurately identify stabilizing mutations, a limitation of other published approaches (Fig. 2b).
2. DynaMut is also freely available for use freely as a user-friendly web interface. In order to run a prediction, open up the DynaMut prediction page at http://biosig.unimelb.edu.au/dynamut/prediction on a web browser of your preference (the web server is compatible with the most common operating systems and browsers).
3. Users have the option to either evaluate a single mutation or provide a text file with a list of mutations to be evaluated in the same format discussed above to run DUET (Fig. 2a). There are no limits on the number of mutations that can be analyzed.
4. For both predictions modes, users are required to provide the wild-type protein structure of interest by either uploading a PDB file or supplying a valid four-letter code PDB accession code of a deposited experimental structure (Fig. 2a).

3.4 DynaMut Prediction Output

1. Prediction results: DynaMut will present the results under three main separate tabulated headings: (1) variation of Gibbs Free Energy predictions, (2) interatomic interactions, and (3) deformation/fluctuation analysis. *See* **Notes 4** and **5** for further information on how to interpret results.
2. DynaMut also graphically displays the resulting change in vibrational energy between the wild-type and mutant structures (Fig. 2b). This highlights regions predicted to be more flexible (red) or less flexible (blue) upon mutation. All calculations and representations can be downloaded through links located at the bottom of the results page.

Fig. 1 (continued) specific mutation or perform a systematic analysis of all mutations for a given residue (*2*). (**b**) For single-mutation prediction, the mutation identification (*3*) and the predicted effects on stability are shown (*4*), as well as an interactive molecular visualization (*5*). (**c**) For systematic evaluation of mutation on a given residue, the results are shown in tabular format

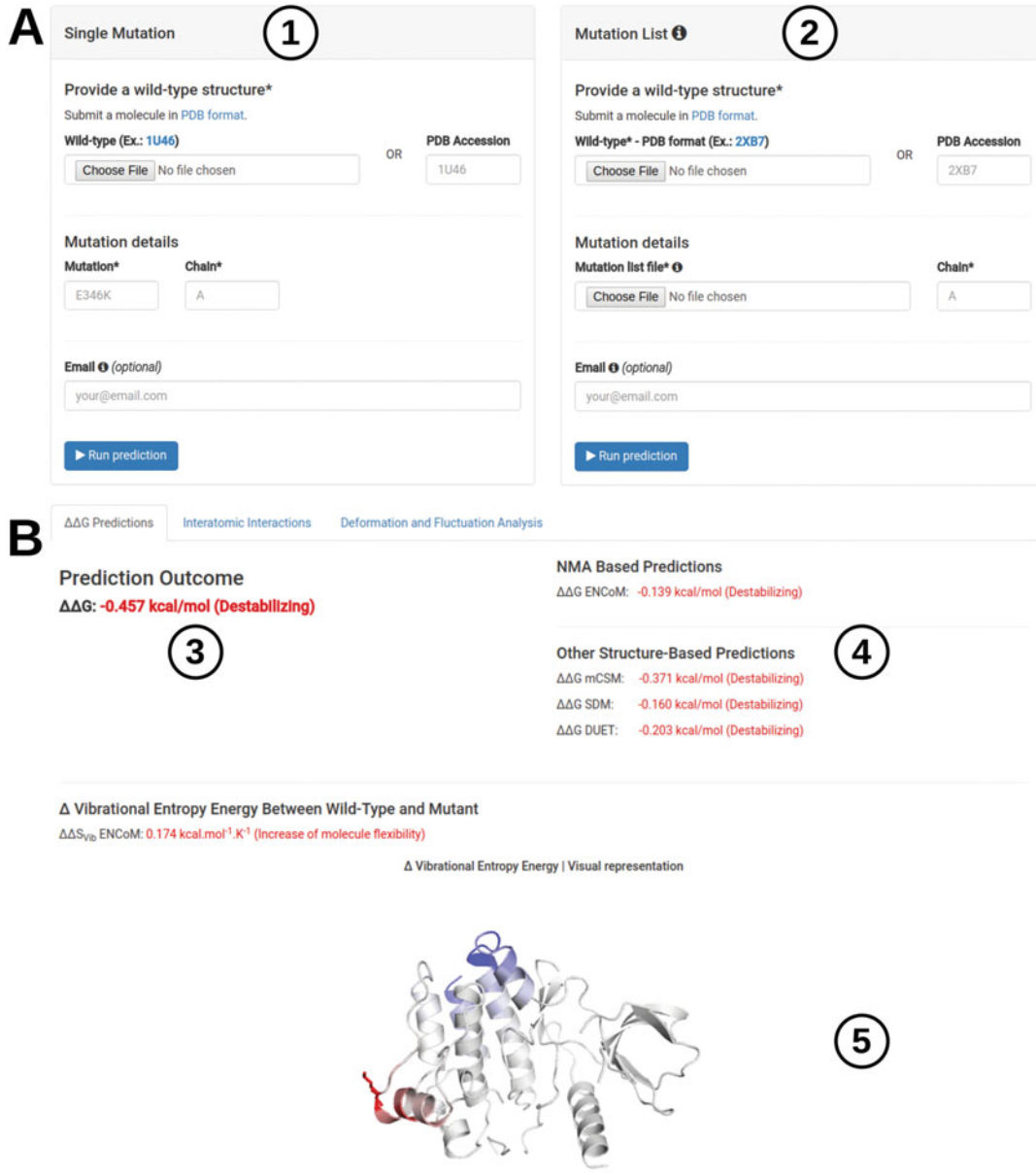


Fig. 2 DynaMut submission and results web interface. (**a**) The submission page allows for the analysis of a single-point mutation (*1*) or a list of mutations (*2*). The main results page (**b**) depicts the predicted effect of mutation by DynaMut (*3*) as well as predicted effects by its individual components (*4*). A depiction of the calculated different in vibration entropy (5) is also shown

3. When multiple mutations are analyzed, these results are presented in a tabulated format, where users are able to open up and analyze each mutation within the single-mutation analysis result interface.

3.5 Visualizing Effects of Mutations on Protein Structure

1. DynaMut also enables visualization of the effects of a mutation within the wild-type and mutant protein structure (Fig. 3).
2. The interatomic interactions made by the wild-type and mutant residues, calculated using Arpeggio [30] (http://biosig.unimelb.edu.au/arpeggioweb/), are visually shown. This enables the user to identify how the mutation will affect the local interaction network—important for maintaining protein stability (Fig. 3a).
3. The normal mode analysis predictions are also shown, highlighting changes in vibrational energy between the wild-type and mutant structures (Fig. 3b).
4. All these representations are downloadable as Pymol session files from links at the bottom of the results page.

4 Notes

1. It is important to notice that both methods, DUET and DynaMut, were conceived to analyze monomer structures. In case of analysis of oligomers, users are advised to filter their PDB files prior to submission, filtering chains of interest (for instance, using the PDBest software [40]). The servers will consider all chains submitted; however, a warning message is exhibited. When considering the effects of mutations on oligomeric structures, it is also important to consider the effects of the mutations on the affinity of the monomers to form the oligomer. This can be assessed using mCSM-PPI (http://biosig.unimelb.edu.au/mcsm/protein_protein).
2. The chain ID for the provided PDB file is a mandatory field, and blank characters are not allowed. Some homology modeling tools do not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. There are several tools available to perform this task.[1]
3. Another source of error comes from structures with multiple models. It is an important practice to filter NMR structures, selecting a single model.
4. Special cases: Mutations to and from prolines. Prolines are the only amino acid whose amino group is connected to the side chain, which in the context of the peptide bond greatly limits torsional angles. The nature of this residue, therefore, needs to be taken into account while analyzing mutation effects. For instance, (1) mutations to prolines in the middle of alpha-helices can introduce kinks, affecting local structure, and

[1] http://www.canoz.com/sdh/renamepdbchain.pl

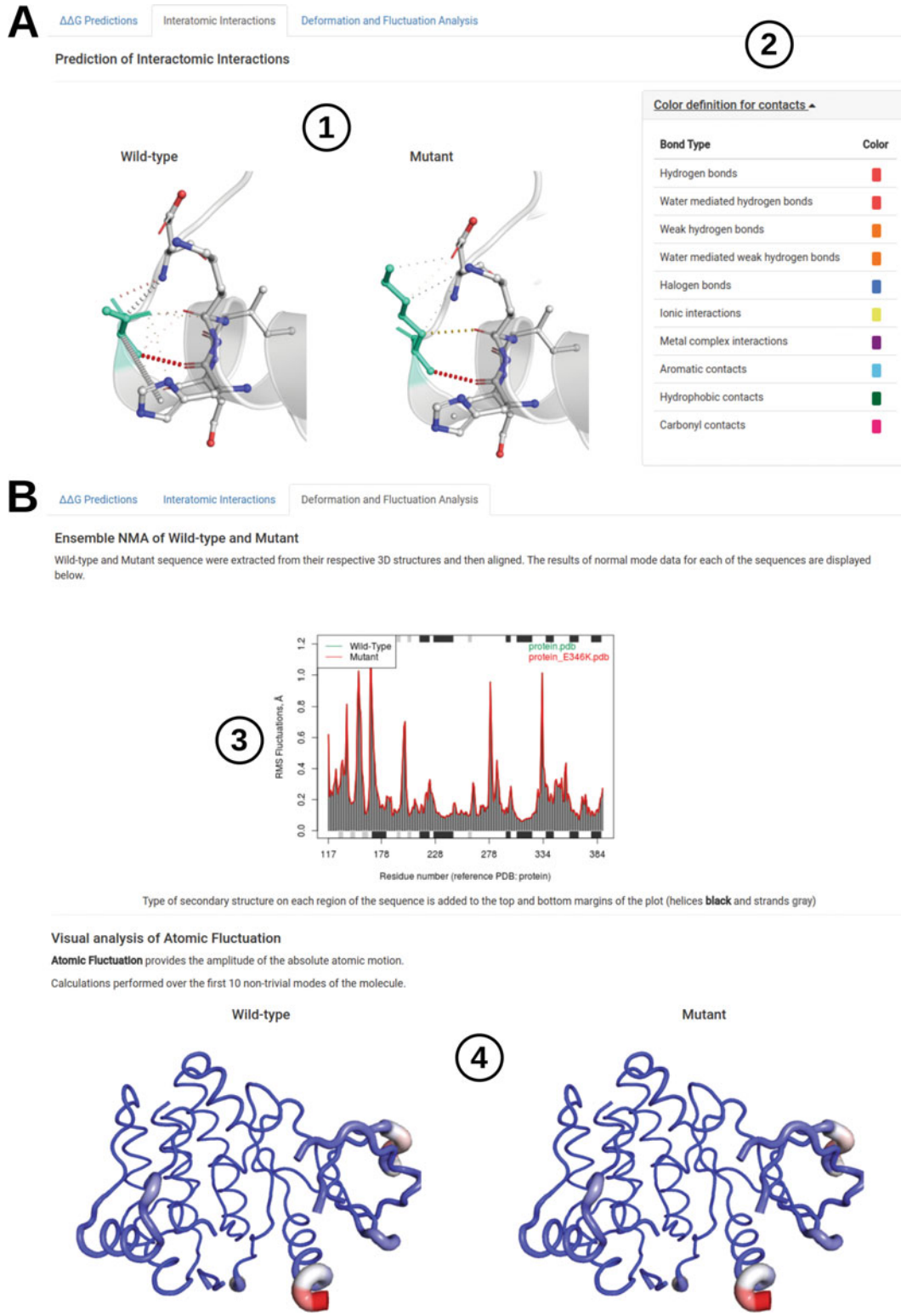


Fig. 3 DynaMut secondary results web interface. (**a**) A depiction of the calculated interatomic interactions (*1*) for wild-type and mutant proteins is shown, with interactions identified by color (*2*). (**b**) Depicts visualizations of the deformation and fluctuation analysis as fluctuation plot per residue (*3*) and atomic fluctuation in the context of the structures (*4*). Figure and individual files (pymol files for molecular visualization) are available for download

(2) since prolines are commonly found in turns and loops, their substitution might interfere with the formation of supersecondary structures such as hairpin loops.

5. Special cases: mutations of positive-phi glycines. Similarly to prolines, positive-phi glycines, while rare in experimental structures, should also be given special consideration due to its torsional angles. Glycines are the only residues capable of adopting positive-phi angles. These glycines are usually conserved across evolution, meaning that mutations of positive-phi glycines tend to be destabilizing.

Acknowledgments

This work was supported by the Australian Government Research Training Program Scholarship [to Y.M., M.K., C.H.M.R. and S.P.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; the Victorian Life Sciences Computation Initiative (VLSCI), an initiative of the Victorian Government, Australia, on its Facility hosted at the University of Melbourne [UOM0017]; the Instituto René Rachou (IRR/FIOCRUZ Minas), Brazil, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [to D.E.V.P.]; and the Department of Biochemistry and Molecular Biology, University of Melbourne [to D.B.A.].

References

1. Andrews KA, Ascher DB, Pires DEV, Barnes DR, Vialard L, Casey RT, Bradshaw N, Adlard J, Aylwin S, Brennan P, Brewer C, Cole T, Cook JA, Davidson R, Donaldson A, Fryer A, Greenhalgh L, Hodgson SV, Irving R, Lalloo F, McConachie M, McConnell VPM, Morrison PJ, Murday V, Park SM, Simpson HL, Snape K, Stewart S, Tomkins SE, Wallis Y, Izatt L, Goudie D, Lindsay RS, Perry CG, Woodward ER, Antoniou AC, Maher ER (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55(6):384–394. https://doi.org/10.1136/jmedgenet-2017-105127
2. Trezza A, Bernini A, Langella A, Ascher DB, Pires DEV, Sodi A, Passerini I, Pelo E, Rizzo S, Niccolai N, Spiga O (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci 58 (12):5320–5328. https://doi.org/10.1167/iovs.17-22158
3. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, Balding DJ, Petrovski S (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27(10):1715–1729. https://doi.org/10.1101/gr.226589.117
4. Soardi FC, Machado-Silva A, Linhares ND, Zheng G, Qu Q, Pena HB, Martins TMM, Vieira HGS, Pereira NB, Melo-Minardi RC, Gomes CC, Gomez RS, Gomes DA, Pires DEV, Ascher DB, Yu H, Pena SDJ (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med

2:7. https://doi.org/10.1038/s41525-017-0009-4

5. Ramdzan YM, Trubetskov MM, Ormsby AR, Newcombe EA, Sui X, Tobin MJ, Bongiovanni MN, Gras SL, Dewson G, Miller JML, Finkbeiner S, Moily NS, Niclis J, Parish CL, Purcell AW, Baker MJ, Wilce JA, Waris S, Stojanovski D, Bocking T, Ang CS, Ascher DB, Reid GE, Hatters DM (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. Cell Rep 19(5):919–927. https://doi.org/10.1016/j.celrep.2017.04.029
6. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montano B, Blundell TL, Ascher DB (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002
7. Chirgadze DY, Ascher DB, Blundell TL, Sibanda BL (2017) DNA-PKcs, allostery, and DNA double-strand break repair: defining the structure and setting the stage. Methods Enzymol 592:145–157. https://doi.org/10.1016/bs.mie.2017.04.001
8. Casey RT, Ascher DB, Rattenberry E, Izatt L, Andrews KA, Simpson HL, Challis B, Park SM, Bulusu VR, Lalloo F, Pires DEV, West H, Clark GR, Smith PS, Whitworth J, Papathomas TG, Taniere P, Savisaar R, Hurst LD, Woodward ER, Maher ER (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5(3):237–250. https://doi.org/10.1002/mgg3.279
9. Pires DE, Chen J, Blundell TL, Ascher DB (2016) In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. https://doi.org/10.1038/srep19848
10. Nemethova M, Radvanszky J, Kadasi L, Ascher DB, Pires DE, Blundell TL, Porfirio B, Mannoni A, Santucci A, Milucci L, Sestini S, Biolcati G, Sorge F, Aurizi C, Aquaron R, Alsbou M, Lourenco CM, Ramadevi K, Ranganath LR, Gallagher JA, van Kan C, Hall AK, Olsson B, Sireau N, Ayoob H, Timmis OG, Sang KH, Genovese F, Imrich R, Rovensky J, Srinivasaraghavan R, Bharadwaj SK, Spiegel R, Zatkova A (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24(1):66–72. https://doi.org/10.1038/ejhg.2015.60
11. Usher JL, Ascher DB, Pires DE, Milan AM, Blundell TL, Ranganath LR (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep 24:3–11. https://doi.org/10.1007/8904_2014_380
12. Jafri M, Wake NC, Ascher DB, Pires DE, Gentle D, Morris MR, Rattenberry E, Simpson MA, Trembath RC, Weber A, Woodward ER, Donaldson A, Blundell TL, Latif F, Maher ER (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov 5(7):723–729. https://doi.org/10.1158/2159-8290.CD-14-1096
13. Hnizda A, Fabry M, Moriyama T, Pachl P, Kugler M, Brinsa V, Ascher DB, Carroll WL, Novak P, Zaliova M, Trka J, Rezacova P, Yang JJ, Veverka V (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia. https://doi.org/10.1038/s41375-018-0073-5
14. Sibanda BL, Chirgadze DY, Ascher DB, Blundell TL (2017) DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science 355 (6324):520–524. https://doi.org/10.1126/science.aak9654
15. Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, Arumugam S, Rajan L, Ebenezer M, Ascher DB, Arnold E, Blundell TL (2018) Structural implications of mutations conferring rifampin resistance in mycobacterium leprae. Sci Rep 8(1):5016. https://doi.org/10.1038/s41598-018-23423-1
16. Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, Denholm JT, Ascher DB (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med. https://doi.org/10.1164/rccm.201712-2572LE
17. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTMH, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN, Thwaites G, Edwards DJ, Nath AP, Pham K, Ascher DB, Farrar J, Khor CC, Teo YY, Inouye M, Caws M, Dunstan SJ (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for EsxW Beijing variant in Vietnam. Nat Genet 50:849–856
18. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, Keri G, Ascher DB, Mondesert G, Vocat A, Lupien A, Sommer R, Vermet H, Lagrange S, Buechler J, Warner DF, McKinney JD, Pato J, Cole ST, Blundell TL, Rizzi M, Mizrahi V (2017) The inosine monophosphate dehydrogenase, GuaB2, is a

vulnerable new bactericidal drug target for tuberculosis. ACS Infect Dis 3(1):5–17. https://doi.org/10.1021/acsinfecdis.6b00102

19. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, Arora K, Ioerger TR, Sacchettini J, Rizzi M, Donini S, Blundell TL, Ascher DB, Rhee K, Breda A, Zhou N, Dartois V, Jonnala SR, Via LE, Mizrahi V, Epemolu O, Stojanovski L, Simeons F, Osuna-Cabello M, Ellis L, MacKenzie CJ, Smith AR, Davis SH, Murugesan D, Buchanan KI, Turner PA, Huggett M, Zuccotto F, Rebollo-Lopez MJ, Lafuente-Monasterio MJ, Sanz O, Diaz GS, Lelievre J, Ballell L, Selenski C, Axtman M, Ghidelli-Disse S, Pflaumer H, Bosche M, Drewes G, Freiberg GM, Kurnick MD, Srikumaran M, Kempf DJ, Green SR, Ray PC, Read K, Wyatt P, Barry CE 3rd, Boshoff HI (2017) Essential but not vulnerable: indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against mycobacterium tuberculosis. ACS Infect Dis 3(1):18–33. https://doi.org/10.1021/acsinfecdis.6b00103
20. Pandurangan AP, Ascher DB, Thomas SE, Blundell TL (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45(2):303–311. https://doi.org/10.1042/BST20160422
21. Albanaz ATS, Rodrigues CHM, Pires DEV, Ascher DB (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12(6):553–563. https://doi.org/10.1080/17460441.2017.1322579
22. White RR, Ponsford AH, Weekes MP, Rodrigues RB, Ascher DB, Mol M, Selkirk ME, Gygi SP, Sanderson CM, Artavanis-Tsakonas K (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog 12(11):e1005977. https://doi.org/10.1371/journal.ppat.1005977
23. Silvino AC, Costa GL, Araujo FC, Ascher DB, Pires DE, Fontes CJ, Carvalho LH, Brito CF, Sousa TN (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. PLoS One 11(7):e0160172. https://doi.org/10.1371/journal.pone.0160172
24. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, Coeck N, Hill-Cawthorne GA, Nair MB, Mallard K, Ramsay A, Campino S, Hibberd ML, Pain A, Rigouts L, Clark TG (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14:31. https://doi.org/10.1186/s12916-016-0575-9
25. Kano FS, Souza-Silva FA, Torres LM, Lima BA, Sousa TN, Alves JR, Rocha RS, Fontes CJ, Sanchez BA, Adams JH, Brito CF, Pires DE, Ascher DB, Sell AM, Carvalho LH (2016) The presence, persistence and functional properties of Plasmodium vivax duffy binding protein II antibodies are influenced by HLA class II allelic variants. PLoS Negl Trop Dis 10(12):e0005177. https://doi.org/10.1371/journal.pntd.0005177
26. Ascher DB, Wielens J, Nero TL, Doughty L, Morton CJ, Parker MW (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. https://doi.org/10.1038/srep04765
27. Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, Spelman DW, Padiglione A, Peleg AY, Holt KE (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom. https://doi.org/10.1099/mgen.0.000165
28. Rodrigues CHM, Pires DEV, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. https://doi.org/10.1093/nar/gky300
29. Pires DE, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45:W241–W246. https://doi.org/10.1093/nar/gkx236
30. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
31. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. https://doi.org/10.1038/srep29575
32. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44(W1):W557–W561. https://doi.org/10.1093/nar/gkw390
33. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res 44(W1):

W469–W473. https://doi.org/10.1093/nar/gkw458

34. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43(Database issue):D387–D391. https://doi.org/10.1093/nar/gku966
35. Pires DE, Blundell TL, Ascher DB (2015) pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem 58 (9):4066–4072. https://doi.org/10.1021/acs.jmedchem.5b00104
36. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691
37. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42. (Web Server issue: W314–W319. https://doi.org/10.1093/nar/gku411
38. Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45:W229–W235. https://doi.org/10.1093/nar/gkx439
39. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
40. Goncalves WR, Goncalves-Almeida VM, Arruda AL, Meira W Jr, da Silveira CH, Pires DE, de Melo-Minardi RC (2015) PDBest: a user-friendly platform for manipulating and enhancing protein structures. Bioinformatics 31(17):2894–2896. https://doi.org/10.1093/bioinformatics/btv223

APPENDIX 3: THERMOMUTDB: A THERMODYNAMIC DATABASE FOR MISSENSE MUTATIONS

Published online 23 October 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D475–D479***
doi: 10.1093/nar/gkaa925

ThermoMutDB: a thermodynamic database for missense mutations

Joicymara S. Xavier[1,2], Thanh-Binh Nguyen[3], Malancha Karmarkar[3,4], Stephanie Portelli[3,4], Pâmela M. Rezende[2], João P.L. Velloso[2], David B. Ascher[3,4,5,*] and Douglas E.V. Pires[3,4,6,*]

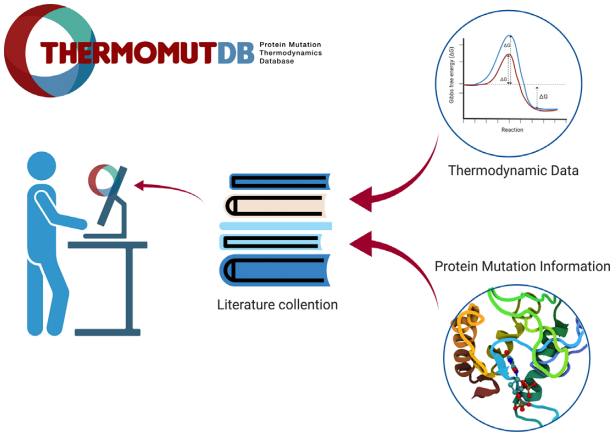
[1]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri, [2]Instituto René Rachou, Fundação Oswaldo Cruz, [3]Bio 21 Institute, University of Melbourne, [4]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, [5]Department of Biochemistry, University of Cambridge and [6]School of Computing and Information Systems, University of Melbourne

Received August 15, 2020; Revised September 21, 2020; Editorial Decision October 05, 2020; Accepted October 12, 2020

ABSTRACT

Proteins are intricate, dynamic structures, and small changes in their amino acid sequences can lead to large effects on their folding, stability and dynamics. To facilitate the further development and evaluation of methods to predict these changes, we have developed ThermoMutDB, a manually curated database containing >14,669 experimental data of thermodynamic parameters for wild type and mutant proteins. This represents an increase of 83% in unique mutations over previous databases and includes thermodynamic information on 204 new proteins. During manual curation we have also corrected annotation errors in previously curated entries. Associated with each entry, we have included information on the unfolding Gibbs free energy and melting temperature change, and have associated entries with available experimental structural information. ThermoMutDB supports users to contribute to new data points and programmatic access to the database via a RESTful API. ThermoMutDB is freely available at: http://biosig.unimelb.edu.au/thermomutdb.

GRAPHICAL ABSTRACT



INTRODUCTION

Protein thermodynamic stability is a fundamental property of proteins that significantly influences their structure, function, expression, and solubility. Changes in protein stability have been shown to be a main driving molecular mechanism of genetic diseases (1–8) and even drug resistance (9–18). Small changes in the protein sequence can have significant consequences on their intricate structures, reflected in changes in their stability and ability to correctly fold (19). This is often a significant consideration whenever considering a new mutation, whether in the context of protein engineering or variant characterisation (20,21).

The accurate prediction of the effects of mutations on protein stability remains a complex and challenging problem. The development of computational approaches to tackle this have required large mutational datasets, however in turn have been limited by the quantity and quality of data available.

*To whom correspondence should be addressed. Tel: +61 3 8344 8185; Email: douglas.pires@unimelb.edu.au
Correspondence may also be addressed to David B. Ascher. Email: david.ascher@unimelb.edu.au

© The Author(s) 2020. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/article/49/D1/D475/5937085 by guest on 11 January 2021

One of the first databases to collect information on the effects of mutations on protein stability, ProTherm (22), led to the exploration and rapid development of new computational approaches (23–28). However, this database has not been updated for 7 years and many errors have been identified previously (29,30), limiting both previous methods and future developments.

To overcome this, we have developed a new comprehensive and user-friendly resource for thermodynamic data from protein mutations, ThemoMutDB. Figure 1 depicts the database development workflow, which is divided into three main stages: (i) data acquisition and curation, (ii) mutation annotation and (iii) web-server development. By using a rigorous and careful data curation approach, ThemoMutDB represents a significant improvement in both the quantity and quality of data. This will not only enable the development of a new generation of methods but also an unbiased assessment of previously proposed ones.

MATERIALS AND METHODS

Data acquisition and curation

Data acquisition for ThermoMutDB was divided into two steps: manual checking of previously mined data available in other resources (Figure 1A) and manual literature curation of new thermodynamic data (Figure 1B). Within ThermoMutDB we captured thermodynamic information, experimental conditions, and literature citations. We also standardized measurements and calculations across the data entries, including temperature in Kelvin, energy in kcal/mol, and Gibbs free energy ($\Delta\Delta G$) as in the formula:

$$\Delta\Delta G = \Delta G\,(\text{wild-type}) - \Delta G\,(\text{mutant})$$

where negative $\Delta\Delta G$ values indicate that the mutation has destabilized the protein and positive $\Delta\Delta G$ values that the mutant protein is more stable.

On the first data acquisition stage, all 1,902 references in ProTherm were manually checked and validated. References that did not contain data about missense mutations were removed, leaving 829 papers. During this process, errors in data fields were corrected, duplicate entries were removed, and 329 new data-points not previously captured, but present in the original papers, were included.

New data were identified through manual literature curation. Optimized search terms (Supplementary Figure S1) were used to identify an initial pool of over 34,000 manuscripts available on PubMed. These were further narrowed down to those that contained experimental thermodynamic results for missense mutations. In total, 393 papers were analyzed and 5,654 new data points obtained, which were confirmed by at least two independent curators. Supplementary Figure S2 shows the distribution of unique mutations collected per year.

Mutation annotation

Collected mutations were mapped to protein structures available at the Protein Data Bank using (31). Different characteristics of the wild-type residue environment were calculated, including secondary structure, torsional angles, relative solvent accessibility (32) and residue depth (33). Additional residue-level information used to annotate the mutations included different substitution matrix scores. Mutation annotations were calculated using the Biopython (34). Mutation effects are also depicted via pharmacophore modeling (23). Pharmacophore modeling has been introduced in the context of mutation analysis in a previous work (23) to characterise the effect of mutations based on the differences in atom counts per pharmacophore type. Mutations that do not map to any available experimental structures are still listed but without any structure-based features calculated.

Database and web interface implementation

The database architecture was developed using SQLAlchemy, a database toolkit for Python (version 2.7). All data is stored in an SQLite database and available to download at http://biosig.unimelb.edu.au/thermomutdb/downloads. The backend system was developed using the Flask Python module (version 1.0.2) and the REStful API uses RestX extension for Flask (version 0.2.0). The web interface was implemented using the Bootstrap (version 4) framework. It also uses HTML5, CSS, JavaScript, and JQuery. JINJA2 templating language for Python was used to dynamically generate HTML templates.

RESULTS

Web interface and usage

ThermoMutDB contains information of the protein, mutational information, experimental methods and conditions, thermodynamic parameters, derived data, and literature information (details are available in Supplementary Table S1 and Figure S2). The database provides a user-friendly web interface that contains five modules: *Explore and Browse*, *Contribute*, *Downloads*, *API* and a detailed tutorial.

Explore and browse. In order to access the data, a search can be performed. This can be done either by selecting the 'Browse' page from the navigation bar or by writing the desired words on search input available on the 'Home' page. In both cases, users can use different filter combinations (Figure 2A), include or exclude columns, and download selected results in several formats (JSON, XML, CSV, TXT, SQL, MS-Excel and PDF).

The search results are shown in an interactive table, with columns providing experimental information recovered from literature and also derived properties (Figure 2B). Aiming to improve user experience, it is possible to visualize a summary for each entry by clicking on the '+' icon. This option can lead to a 'Details' page that shows all information about the mutation and provides related files to download (Supplementary Figure S3).

User contributions. To facilitate a continuous database update, we have implemented a user's contribution section (Supplementary Figure S4), which allows the scientific community to share new data or identify potential errors that will be manually checked by our team. To submit contributions it is just required to fill the form with mutation and

Downloaded from https://academic.oup.com/nar/article/49/D1/D475/5937085 by guest on 11 January 2021

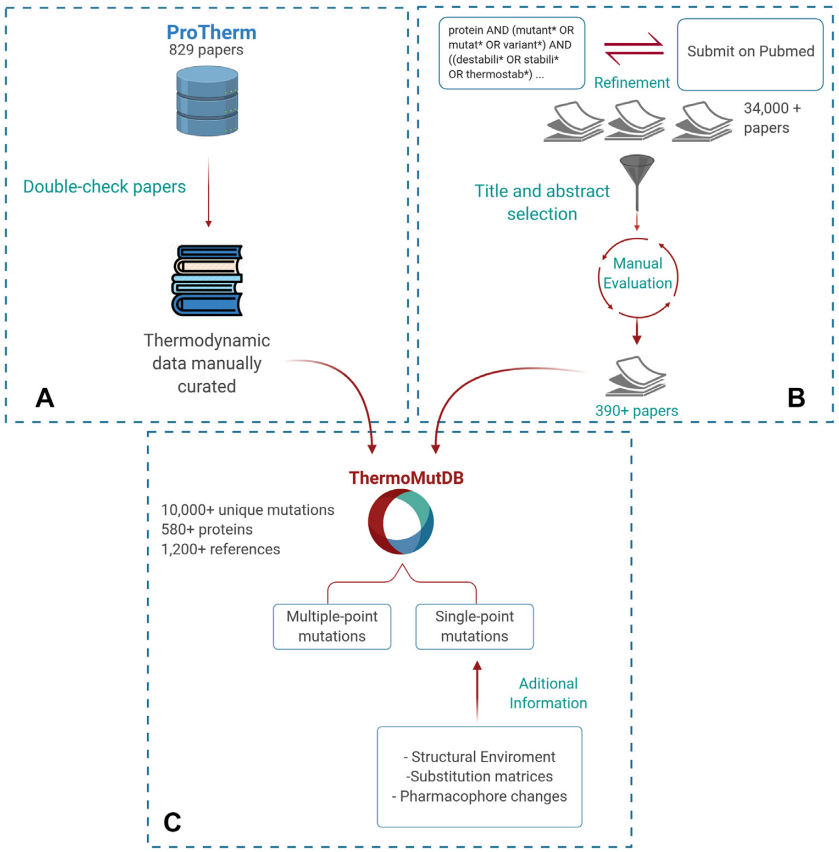


Figure 1. ThermoMutDB workflow for data acquisition and processing. The development workflow is divided into three steps: (**A**) verification of previously available mutation thermodynamics information (**B**) collection and manual curation of new data and (**C**) data aggregation and mutation annotation.

thermodynamics data, to inform a contact email and a reference (paper published, accepted, or pre-print). Although significant effort has been devoted to ensure high quality data curation, users have the option to report any issues with the data to our team. These are important efforts to further expand and improve the database.

Downloads. All data in the database can be downloaded from the 'Download' page in CSV or JSON formats. It is also possible to download the protein structure files related to data available.

Programmatic access via an API. ThermoMutDB supports programmatic access via a RESTful API to allow other services to harness our data easily. The 'API' page provides documentation of all endpoints available and allows users to execute queries using provided fields. Other queries can be performed by passing parameters through the URL (Supplementary Figure S5).

Data statistics

Examining the distribution of mutations in the ThermoMutDB reveals a number of natural biases that need to be taken into consideration when developing, or evaluating, new predictive tools. ThermoMutDB contains thermodynamic information on 14,669 mutations across 588 proteins. This represents a significant increase over ProTherm, with a 83% increase in unique mutations and over 300 new proteins. Supplementary Figure S6 shows the distribution of unique mutations collected per year. The majority of these are single-point mutations (82.8%), with mutations to alanine being over-represented (Figure 3D). This becomes evident when we look at the distribution of wild-type and mutant amino acid residues within the database (Supplementary Figure S7). The most frequent mutations were from Leucine and Valine to Alanine, while 10 mutations were not present in the dataset, including W→G, W→P and C→K among others, which seem to denote large changes in residue physicochemical properties.

As would be expected by chance, two thirds of mutations within the database are destabilising (Supplementary Figure S8). This natural bias creates an extra challenge for computational methods built using this information, in particular those based on machine learning approaches, regarding the prediction of stabilising mutations, which are less well represented. It is important to note, however, that the data on ThermoMutDB represents an increase of over 100% in stabilising mutations in comparison with previous resources. No apparent correlation was identified between the mutation effects and their location within protein structures, with mutations leading to increased and decreased stability similarly distributed across protein structures when looking at residue depth (Supplementary Figure S9). Muta-

Downloaded from https://academic.oup.com/nar/article/49/D1/D475/5937085 by guest on 11 January 2021

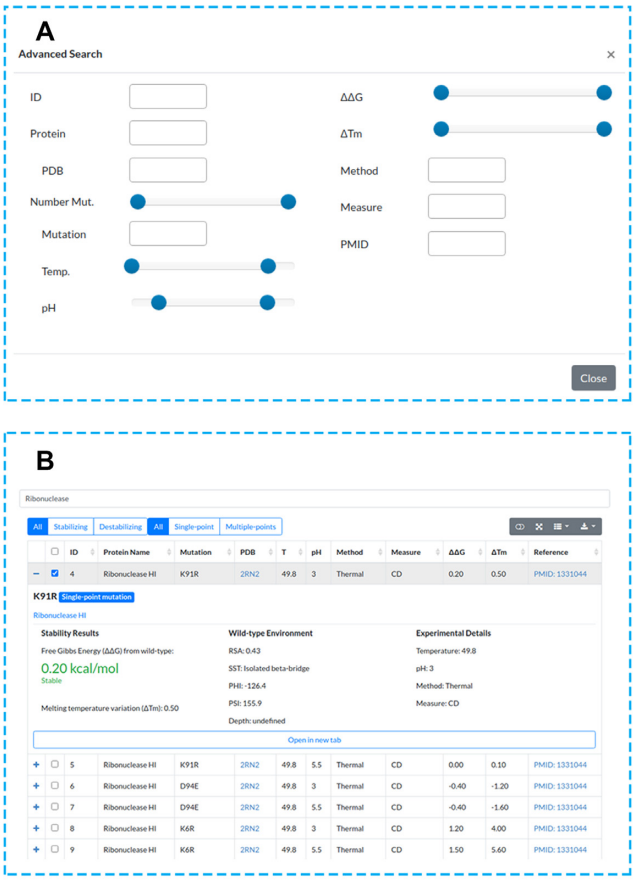


Figure 2. ThermoMutDB web interface search and results pages. (**A**) ThermoMutDB offers 12 query modes, with detailed information available about each query type through the 'Help' page at the top navigation bar and through on-page help in the form of question mark tooltips. (**B**) The general layout of the result page, showing a summary of information for each entry as well as detailed view.



Figure 3. Composition of ThermoMutDB entries. (**A**) depicts the distribution of phylogenetic kingdoms of proteins in the database. (**B**) highlights the distribution of thermodynamic effects of mutation in the database, given as the variation in Gibbs Free Energy ($\Delta\Delta G$). (**C**) Experimental variability of mutation assessed under different conditions and groups. (**D**) Distribution of mutations in ThermoMutDB based on type (mutation to alanine/non-alanine), their location and residue environment.

tions in ThermoMutDB are spread across different protein classes (Supplementary Figure S10) and diverse in terms of secondary structure (Supplementary Figure S11).

Within ThermoMutDB, we identified mutations that had been experimentally measured at least twice and, by comparing the variance between these replicate results (Figure 3C), we identified a Pearson's correlation of 0.9. This provides a measure of the intrinsic noise in the data, and suggests a theoretical maximum performance that should be expected for predictive stability tools built using this data.

DISCUSSION

ThermoMutDB represents a significant increase in availability, reliability and diversity of thermodynamics data linking effects of mutations to protein stability. We believe this resource will have a significant impact on understanding the effects of mutations on protein structure and stability. It will enable experimental scientists to identify previously characterised mutations in proteins of interest, and provide computational scientists with a comprehensive and refined set of experimental data to query the relationship between changes in protein sequence and stability, facilitating the development of new computational tools to analyse these relationships and develop prediction algorithms.

New mutation thermodynamics data collected and compiled in ThermoMutDB will also allow for more robust, comprehensive and independent validation of currently available computational predictors. The database will be continuously maintained and updated, enabling submission of user contributions and data access through an intuitive web-based interface (http://biosig.unimelb.edu.au/thermomutdb) as well as programmatic access through an API.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

Downloaded from https://academic.oup.com/nar/article/49/D1/D475/5937085 by guest on 11 January 2021

FUNDING

D.B.A. and D.E.V.P. were funded by a Newton Fund RCUK-CONFAP Grant awarded by the Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1]; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); Jack Brockhoff Foundation [JBF 4186, 2016]; Wellcome Trust [200814/Z/16/Z]; Investigator Grant from the National Health and Medical Research Council (NHMRC) of Australia [GNT1174405]. Funding for open access charge: Wellcome Trust.
Conflict of interest statement. None declared.

REFERENCES

1. Jafri,M., Wake,N.C., Ascher,D.B., Pires,D.E., Gentle,D., Morris,M.R., Rattenberry,E., Simpson,M.A., Trembath,R.C., Weber,A. *et al.* (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. *Cancer Discov.*, **5**, 723–729.
2. Ramdzan,Y.M., Trubetskov,M.M., Ormsby,A.R., Newcombe,E.A., Sui,X., Tobin,M.J., Bongiovanni,M.N., Gras,S.L., Dewson,G., Miller,J.M.L. *et al.* (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. *Cell Rep.*, **19**, 919–927.
3. Soardi,F.C., Machado-Silva,A., Linhares,N.D., Zheng,G., Qu,Q., Pena,H.B., Martins,T.M.M., Vieira,H.G.S., Pereira,N.B., Melo-Minardi,R.C. *et al.* (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. *NPJ Genom Med*, **2**, 7.
4. Andrews,K.A., Ascher,D.B., Pires,D.E.V., Barnes,D.R., Vialard,L., Casey,R.T., Bradshaw,N., Adlard,J., Aylwin,S., Brennan,P. *et al.* (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. *J. Med. Genet.*, **55**, 384–394.
5. Ascher,D.B., Spiga,O., Sekelska,M., Pires,D.E.V., Bernini,A., Tiezzi,M., Kralovicova,J., Borovska,I., Soltysova,A., Olsson,B. *et al.* (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. *Eur. J. Hum. Genet.*, **27**, 888–902.
6. Hildebrand,J.M., Kauppi,M., Majewski,I.J., Liu,Z., Cox,A.J., Miyake,S., Petrie,E.J., Silk,M.A., Li,Z., Tanzer,M.C. *et al.* (2020) A missense mutation in the MLKL brace region promotes lethal neonatal inflammation and hematopoietic dysfunction. *Nat. Commun.*, **11**, 3150.
7. Pires,D.E.V., Rodrigues,C.H.M. and Ascher,D.B. (2020) mCSM-membrane: predicting the effects of mutations on transmembrane proteins. *Nucleic Acids Res.*, **48**, W147–W153.
8. Trezza,A., Bernini,A., Langella,A., Ascher,D.B., Pires,D.E.V., Sodi,A., Passerini,I., Pelo,E., Rizzo,S., Niccolai,N. *et al.* (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. *Invest. Ophthalmol. Vis. Sci.*, **58**, 5320–5328.
9. Ascher,D.B., Wielens,J., Nero,T.L., Doughty,L., Morton,C.J. and Parker,M.W. (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. *Sci. Rep.*, **4**, 4765.
10. Phelan,J., Coll,F., McNerney,R., Ascher,D.B., Pires,D.E., Furnham,N., Coeck,N., Hill-Cawthorne,G.A., Nair,M.B., Mallard,K. *et al.* (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.*, **14**, 31.
11. Karmakar,M., Globan,M., Fyfe,J.A.M., Stinear,T.P., Johnson,P.D.R., Holmes,N.E., Denholm,J.T. and Ascher,D.B. (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. *Am. J. Respir. Crit. Care Med.*, **198**, 541–544.
12. Portelli,S., Phelan,J.E., Ascher,D.B., Clark,T.G. and Furnham,N. (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Sci. Rep.*, **8**, 15356.
13. Karmakar,M., Rodrigues,C.H.M., Holt,K.E., Dunstan,S.J., Denholm,J. and Ascher,D.B. (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLoS One*, **14**, e0217169.
14. Hawkey,J., Ascher,D.B., Judd,L.M., Wick,R.R., Kostoulias,X., Cleland,H., Spelman,D.W., Padiglione,A., Peleg,A.Y. and Holt,K.E. (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. *Microbial Genomics*, **4**, e000165.
15. Holt,K.E., McAdam,P., Thai,P.V.K., Thuong,N.T.T., Ha,D.T.M., Lan,N.N., Lan,N.H., Nhu,N.T.Q., Hai,H.T., Ha,V.T.N. *et al.* (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.*, **50**, 849–856.
16. Vedithi,S.C., Malhotra,S., Das,M., Daniel,S., Kishore,N., George,A., Arumugam,S., Rajan,L., Ebenezer,M., Ascher,D.B. *et al.* (2018) Structural implications of mutations conferring rifampin resistance in Mycobacterium leprae. *Sci. Rep.*, **8**, 5016.
17. Karmakar,M., Rodrigues,C.H.M., Horan,K., Denholm,J.T. and Ascher,D.B. (2020) Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Sci. Rep.*,**10**, 1875.
18. Portelli,S., Olshansky,M., Rodrigues,C.H.M., D'Souza,E.N., Myung,Y., Silk,M., Alavi,A., Pires,D.E.V. and Ascher,D.B. (2020) Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. *Nat. Genet.*, **52**, 999–1001.
19. Pandurangan,A.P., Ascher,D.B., Thomas,S.E. and Blundell,T.L. (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem. Soc. Trans.*, **45**, 303–311.
20. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
21. Pires,D.E., Chen,J., Blundell,T.L. and Ascher,D.B. (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. *Sci. Rep.*, **6**, 19848.
22. Gromiha,M.M., An,J., Kono,H., Oobatake,M., Uedaira,H. and Sarai,A. (1999) ProTherm: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **27**, 286–288.
23. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
24. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**, W314–W319.
25. Pandurangan,A.P., Ochoa-Montano,B., Ascher,D.B. and Blundell,T.L. (2017) SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.*, **45**, W229–W235.
26. Rodrigues,C.H., Pires,D.E. and Ascher,D.B. (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.*, **46**, W350–W355.
27. Laimer,J., Hiebl-Flach,J., Lengauer,D. and Lackner,P. (2016) MAESTROweb: a web server for structure-based protein stability prediction. *Bioinformatics*, **32**, 1414–1416.
28. Quan,L., Lv,Q. and Zhang,Y. (2016) STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*, **32**, 2936–2946.
29. Yang,Y., Urolagin,S., Niroula,A., Ding,X., Shen,B. and Vihinen,M. (2018) PON-tstab: protein variant stability predictor. Importance of training data quality. *Int. J. Mol. Sci.*, **19**, 1009.
30. Fang,J. (2020) A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief. Bioinform.*, **21**, 1285–1292.
31. Martin,A.C. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics*, **21**, 4297–4301.
32. Gilis,D. and Rooman,M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.*, **272**, 276–290.
33. Chakravarty,S. and Varadarajan,R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*, **7**, 723–732.
34. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Downloaded from https://academic.oup.com/nar/article/49/D1/D475/5937085 by guest on 11 January 2021

APPENDIX 4: A COMPREHENSIVE COMPUTATIONAL PLATFORM TO GUIDE DRUG DEVELOPMENT USING GRAPH-BASED SIGNATURE METHODS



Chapter 7

A Comprehensive Computational Platform to Guide Drug Development Using Graph-Based Signature Methods

Douglas E. V. Pires, Stephanie Portelli, Pâmela M. Rezende, Wandré N. P. Veloso, Joicymara S. Xavier, Malancha Karmakar, Yoochan Myung, João P. V. Linhares, Carlos H. M. Rodrigues, Michael Silk, and David B. Ascher

Abstract

High-throughput computational techniques have become invaluable tools to help increase the overall success, process efficiency, and associated costs of drug development. By designing ligands tailored to specific protein structures in a disease of interest, an understanding of molecular interactions and ways to optimize them can be achieved prior to chemical synthesis. This understanding can help direct crucial chemical and biological experiments by maximizing available resources on higher quality leads. Moreover, predicting molecular binding affinity within specific biological contexts, as well as ligand pharmacokinetics and toxicities, can aid in filtering out redundant leads early on within the process. We describe a set of computational tools which can aid in drug discovery at different stages, from hit identification (EasyVS) to lead optimization and candidate selection (CSM-lig, mCSM-lig, Arpeggio, pkCSM). Incorporating these tools along the drug development process can help ensure that candidate leads are chemically and biologically feasible to become successful and tractable drugs.

Key words Graph-based signatures, mCSM, Mutation, Protein-ligand, Interatomic interactions, Docking, Drug development

1 Introduction

Structure-guided drug development uses knowledge of the three-dimensional structure of the biological target to more efficiently guide the design of small molecule binders. While it has become an integral strategy for both lead generation and optimization, the application of computational tools to take advantage of the explosion in structural information has often required specialist knowledge and resources and in some cases has been limited to commercial software.

Zoltán Gáspári (ed.), *Structural Bioinformatics: Methods and Protocols*, Methods in Molecular Biology, vol. 2112,
https://doi.org/10.1007/978-1-0716-0270-6_7, © Springer Science+Business Media, LLC, part of Springer Nature 2020

Using the concept of graph-based signatures, we have developed a robust, user-friendly, and freely accessible platform to analyze protein structures and interactions [1–12] and guide disease characterization [13–28] and drug development [29–32]. These include methods to perform virtual screening (EasyVS), score protein-small molecule docking solutions (CSM-lig [3]), look at all the molecular interactions being made (Arpeggio [7]), identify mutations that are likely to affect compound binding (mCSM-lig [5]), and characterize the pharmacokinetic and toxicity properties of the proposed molecules (pkCSM [33, 34]). These have been successfully employed in a number of drug development projects [30–32, 35–37] and together comprise a powerful platform that allows users to enhance their structure-guided drug development efforts (Fig. 1). Here we discuss how this platform can be leveraged to guide drug development.

2 Materials

Here we present four structure-based tools to help guide drug development. For each method, users are required to provide:

1. **Wild-type protein structure in PDB format**: For all methods, a wild-type structure in the Protein Data Bank [38] format must be provided to perform the analysis. This can be an experimentally solved structure previously deposited into the Protein Data Bank (www.rcsb.org or http://www.ebi.ac.uk/pdbe/) or a model, for instance, obtained by comparative homology modeling. We have previously shown that homology models built using templates down to 25% sequence identity do not significantly affect the accuracy of the methods [9, 10]. For Arpeggio, CSM-lig, and mCSM-lig, the protein structure file needs to include the ligand of interest, either already present in the experimental structure or computationally docked into the binding site. PDB structures are required to have a valid chain identifier (*see* **Note 1**), a single conformation (multiple occupancies need to be filtered out; *see* **Note 2**), and a single model, in case of NMR structures (*see* **Note 3**).
2. **Three-letter code of the ligand of interest**: When a structure of a protein-ligand complex is provided to the predictive web servers (CSM-lig and mCSM-lig), users will be asked to provide a three-letter code that identifies the residue ID for that ligand within the PDB file, according to the PDB format standards. In addition to the three-letter code, CSM-lig also requires the canonical SMILES of the compound of interest for additional property calculations. Several tools are available to aid users to convert between small molecule formats. These include stand-alone packages such as OpenBabel [39] and Avogadro [40].

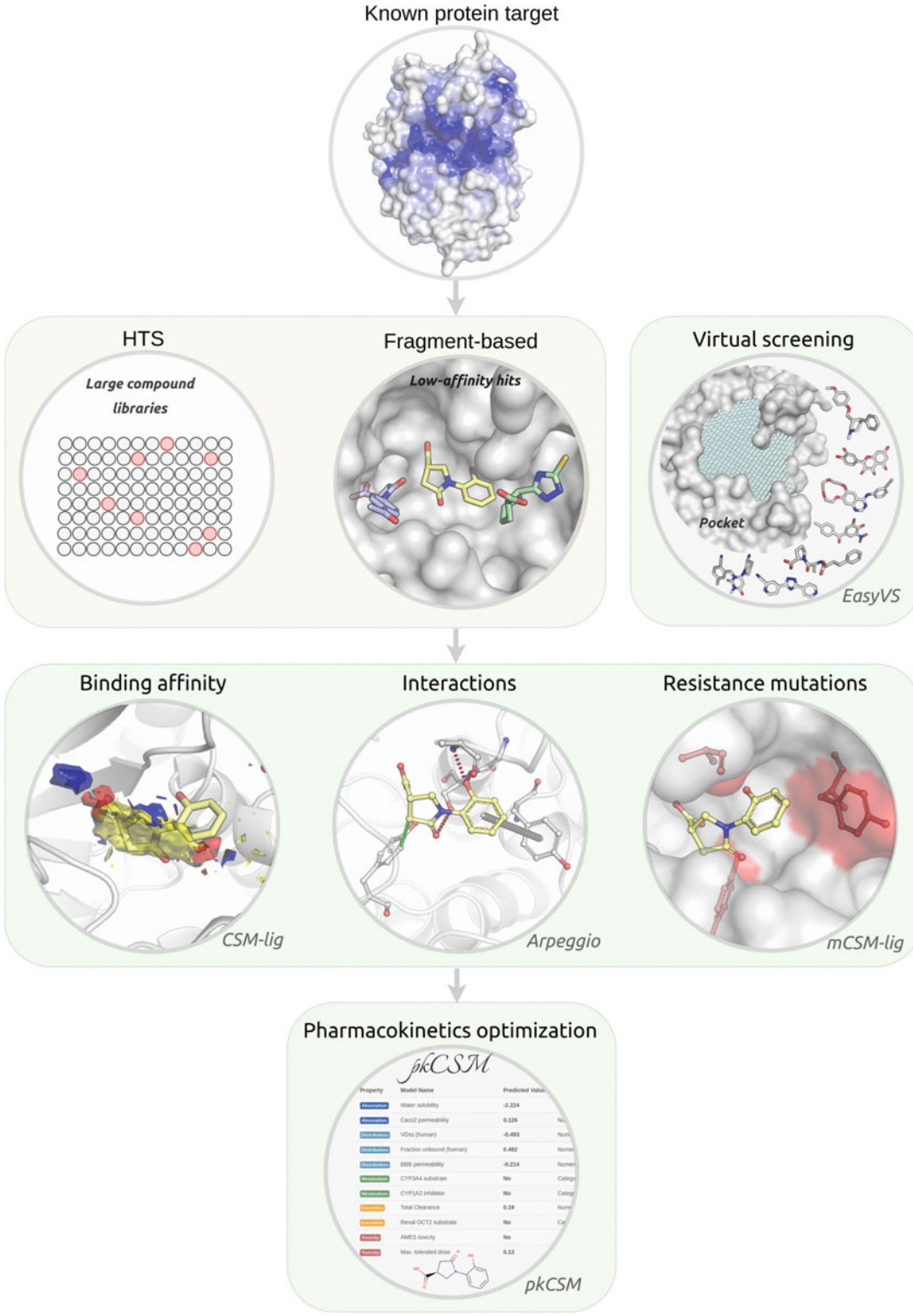


Fig. 1 A structure-based computational platform to guide drug development. To complement and support traditional experimental approaches, including high-throughput screening (HTS) and fragment-based drug discovery, this in silico platform supports hit identification via virtual screening, methods to better understand protein–small molecule interactions, affinity and effects of mutations, as well as the optimization of pharmacokinetic properties

3 Methods

3.1 Performing Automated Docking with EasyVS

1. Virtual screening is a powerful, high-throughput technique for computationally screening large libraries of small molecules (often in the order of millions) in order to identify those ligands which are most likely to bind to a drug target protein. When compared to traditional screening methods, this leads to significantly higher hit rates that can proceed to lead optimization [41, 42]. It can, however, be computationally intensive and usually requires specialist knowledge. EasyVS provides an easy-to-use web interface at http://biosig.unimelb.edu.au/easyvs/, allowing users to rapidly set up and analyze their virtual screening results.
2. Users can upload the structure of the protein target of interest as either a PDB file or by providing the PDB ID of a previously solved experimental structure. Any ligands, ions, or water molecules already bound to the provided structure will be disregarded.
3. On the following step, the provided PDB file or identifier will be processed, and pockets will be automatically detected using Ghecom [43] (Fig. 2a-1). Users can either select one of the identified pockets to determine the docking grid (the three-dimensional space where the ligands will be docked into) or provide specific grid coordinates and size (Fig. 2a-2).
4. Users then need to select the ligand library they want to screen, which includes libraries of purchasable compounds, natural products, or FDA-approved drugs (Fig. 2b). These can be further filtered based upon their molecular properties (e.g., Lipinski's rule of five [44] or the rule of three) or grouped by similarity.
5. The selected molecules will then be docked into the selected docking grid (Fig. 2c-1), and the top 20 poses per ligand can be downloaded. The server also provides an interactive visualization tool to compare ligand docking poses (Fig. 2c-2). The example on this figure shows the docking poses for ligands docked to the Ribosome-Inactivating Protein Ricin A (PDB ID: 1BR5). While poses are sorted by predicted affinity (kcal/mol) using autodock's scoring function, users can evaluate docking poses with alternative approaches, such as CSM-lig [3].

3.2 Predicting Protein-Small Molecule Affinity with CSM-lig

1. Following virtual screening or docking, the affinity of the top docked ligand poses can be quantified using CSM-lig. This is a machine learning-based tool which acts as a scoring function and enables the numerical affinity comparison between poses. It is implemented via an easy-to-use web interface at http://biosig.unimelb.edu.au/csm_lig, which is compatible with most operating systems and browsers.

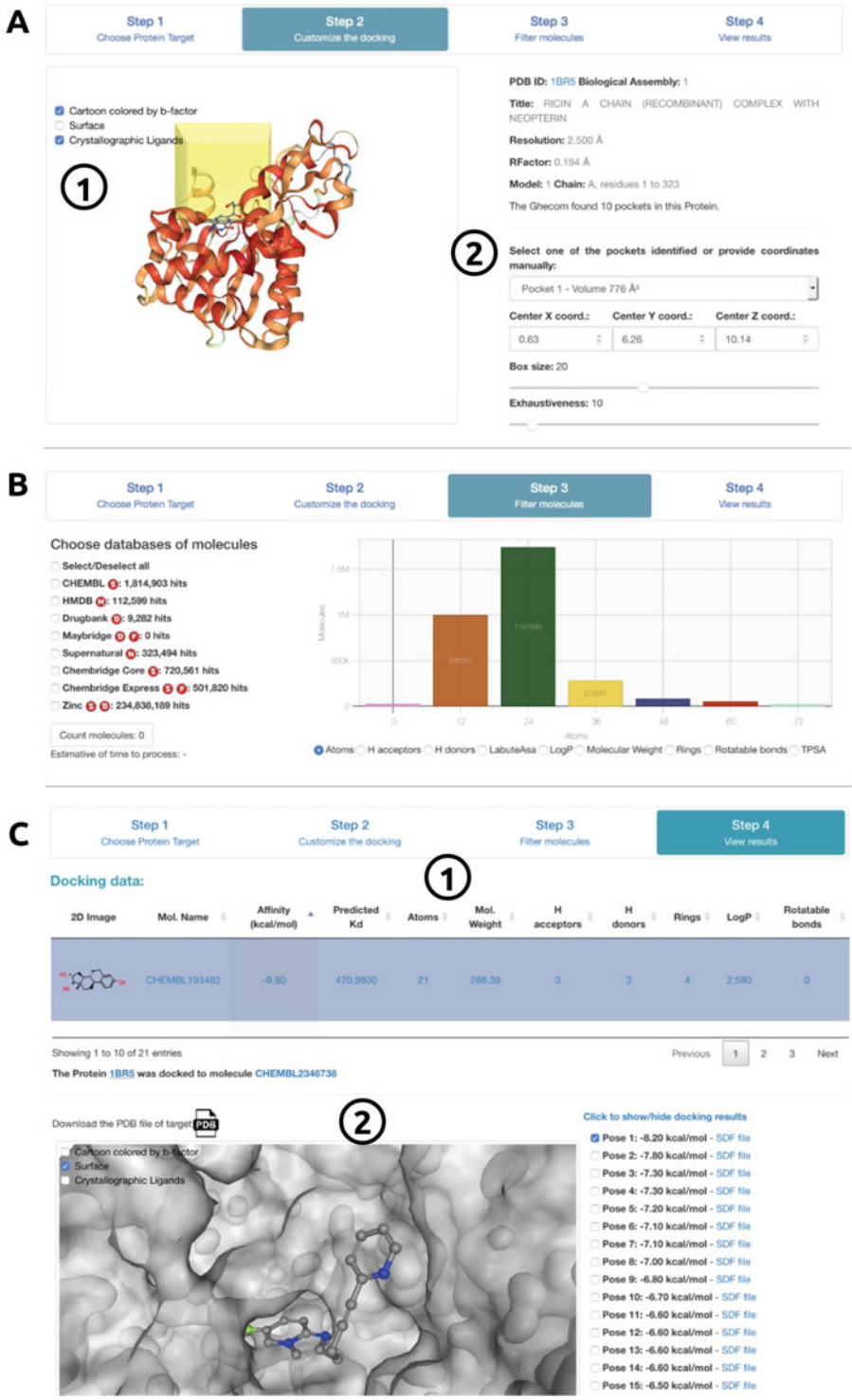


Fig. 2 Automated docking with EasyVS. After choosing a target of interest, EasyVS will automatically identify pockets (**a-1**) and allow user to further customize the docking protocol (**a-2**). A range of ligand libraries can be selected for docking (**b**), including FDA-approved drugs, purchasable compounds, and natural products, which can be further filtered based on physicochemical properties. Docking results are shown in tabular format (**c-1**), depicting ligands, their properties, and docking scores. An interactive viewer allows users to inspect the best poses for each ligand (**c-2**)

2. By selecting the "Predict" tab, users are presented with two job options, "Single Structure" and "Multiple Structures."
3. For "Single Structure" prediction, provide (Fig. 3a-1) the protein-small molecule complex you would like to evaluate the pose of in PDB format (Fig. 3a-2), the three-letter code for the small molecule (as in the provided PDB file) and (Fig. 3a-3) and the SMILES string of the small molecule.
4. Alternatively, for "Multiple Structures," provide two files. The first file (Fig. 3a-4) is a compressed zip file with all protein-small molecule PDB files you would like to evaluate. These could be, for instance, different poses or conformations for a given protein-ligand complex or multiple different complexes. The second (Fig. 3a-5) is a tab-separated file with the following information for each uploaded complex in the .zip file: (a) structure file name (file in PDB format), (b) three-letter code for the small molecule (as in the structure file), and (c) canonical SMILES for the small molecule.
5. The output prediction page for the "Single Structure" jobs depicted in Fig. 1b presents (Fig. 3b-1) the predicted affinity (as $-\log_{10}(\textit{affinity})$ in molar, meaning a compound with an affinity predicted as 1 nM would have a predicted value of 9). The example presented in the figure and the web server shows the affinity prediction for the ligand Zanamivir bound to human sialidase-2 (PDB ID: 2F0Z). For this complex, CSM-lig generates a score of 12.6, denoting very high affinity (larger numbers denote higher affinity). A depiction figure of the small molecule is shown, together with calculated properties, including molecular weight (in Da) and partition coefficient (log *P*), among others (Fig. 3b-2). An interactive visualization of the protein-small molecule complex is also exhibited (Fig. 3b-3). The interatomic non-covalent interactions between protein and small molecule are also calculated and are available as a downloadable Pymol [45] session (Fig. 3b-4). Pharmacokinetics and toxicity predictions by pkCSM for the provided small molecule are also available by clicking on the red button at the bottom-left corner of the results page.
6. The output for "Multiple Structures" jobs are shown in tabular format (Fig. 3c-1), depicting predicted affinity values, SMILES identifying the molecules and their calculated molecular properties. These results are available as a tabular file and can be downloaded (Fig. 3c-2).

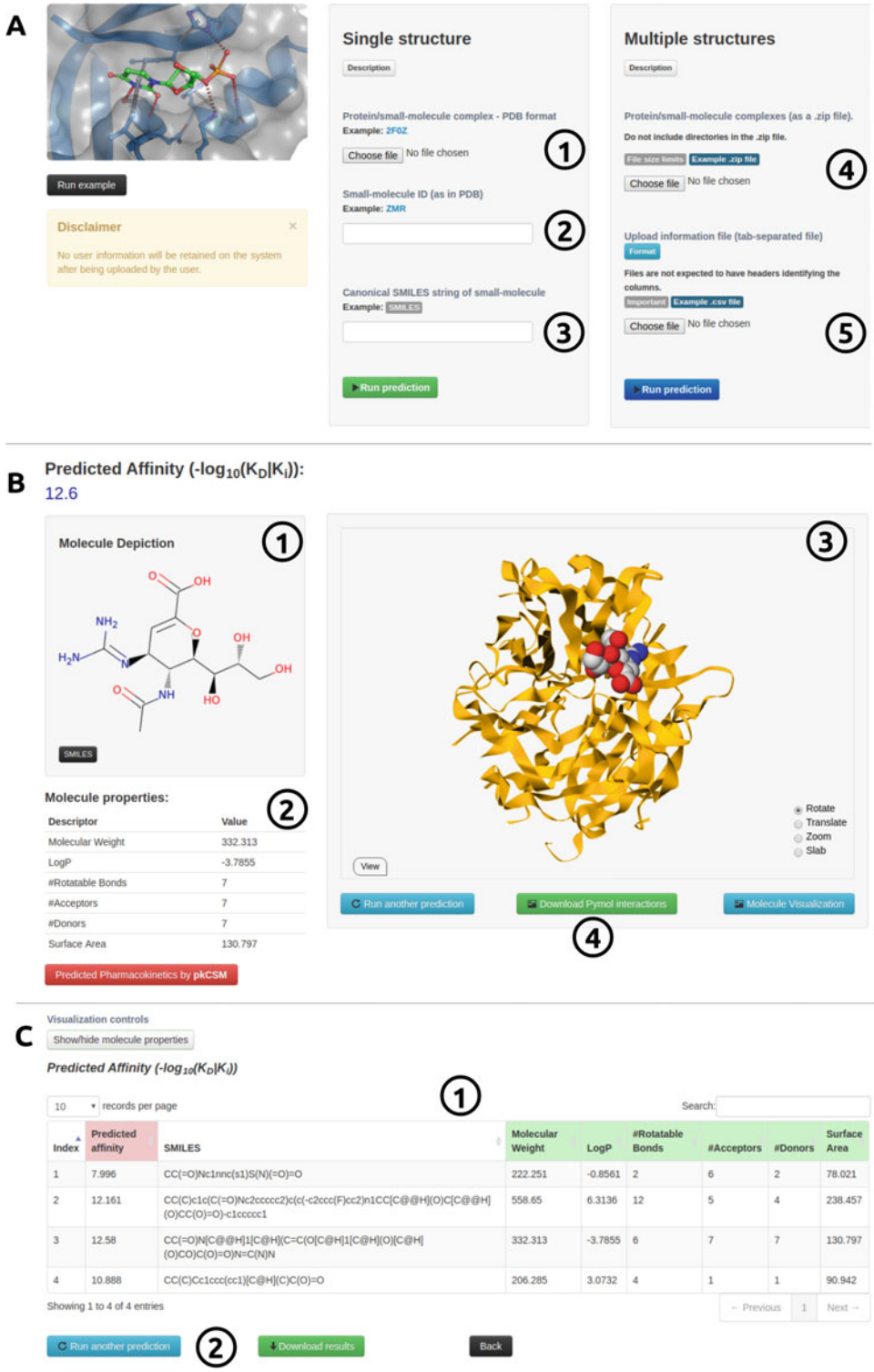


Fig. 3 CSM-lig submission and results web interface. The submission page (**a**) allows users to provide either single or multiple protein-ligand complexes for evaluation. The results page for single complex/pose assessment (**b**) provides the calculated affinity, ligand properties and depiction, as well as an interactive visualization of the complex. For multiple poses, CSM-lig provides the predicted affinities in a downloadable tabular format, together with ligand properties (**c**)

3.3 Depicting and Analyzing Protein-Small Molecule Interactions with Arpeggio

1. Once a structure of the target protein with the candidate molecule is available, either through experimental determination or docking or other alternative approach (for instance, those combining blind docking with molecular dynamics like the Wrap 'n' Shake method [46]), Arpeggio enables the visualization of intermolecular interactions occurring between the lead and its target. During lead optimization, Arpeggio can therefore be used to understand the mechanism of binding and guide medicinal chemistry efforts.
2. Arpeggio is freely available as a user-friendly web interface and is compatible with multiple operating systems and browsers. Open up the prediction server, http://biosig.unimelb.edu.au/arpeggioweb/, on a web browser of your preference.
3. Provide the complexed protein structure of interest by either uploading it as a PDB file or providing the PDB ID of the experimentally solved structure in complex with the ligand of interest (Fig. 4a-1).
4. Select the ligand or ligands of interest under the "Heteroatom" selection heading to calculate all molecular interactions being made by that ligand (Fig. 4b-1; *see* **Note 4**).
5. The results page will show an interactive image of all the molecular interactions made by the ligand(s) selected (Fig. 5a) and a table with a count of the total number of specific molecular interactions being made, including hydrophobic interactions, hydrogen bonds, pi-interactions, and ionic interactions (Fig. 4c).
6. A Pymol session file (PSE file) containing the submitted PDB file and all of the calculated interactions can be downloaded and opened in Pymol to enable visualization of the interaction network in 3D and to facilitate high-quality image generation for manuscripts (Fig. 5b).

3.4 Predicting the Effects of Mutations on Small Molecule Affinity with mCSM-lig

1. During lead optimization, it is important to consider how genetic diversity might affect the binding of candidate molecules and, in particular, if resistance is likely to arise. mCSM-lig uses graph-based signatures to calculate the change upon mutation in small molecule binding affinity. In order to run a prediction, open up the mCSM-lig server at http://biosig.unimelb.edu.au/mcsm_lig/ on a web browser of your preference (the web server is compatible with the most common operating systems and browsers).
2. Users are required to provide the protein structure in complex with the ligand of interest by either uploading a PDB file or supplying a valid four-letter code PDB accession code of a deposited experimental structure (Fig. 6a-1). Users also need to provide the mutation information, the mutation chain, the

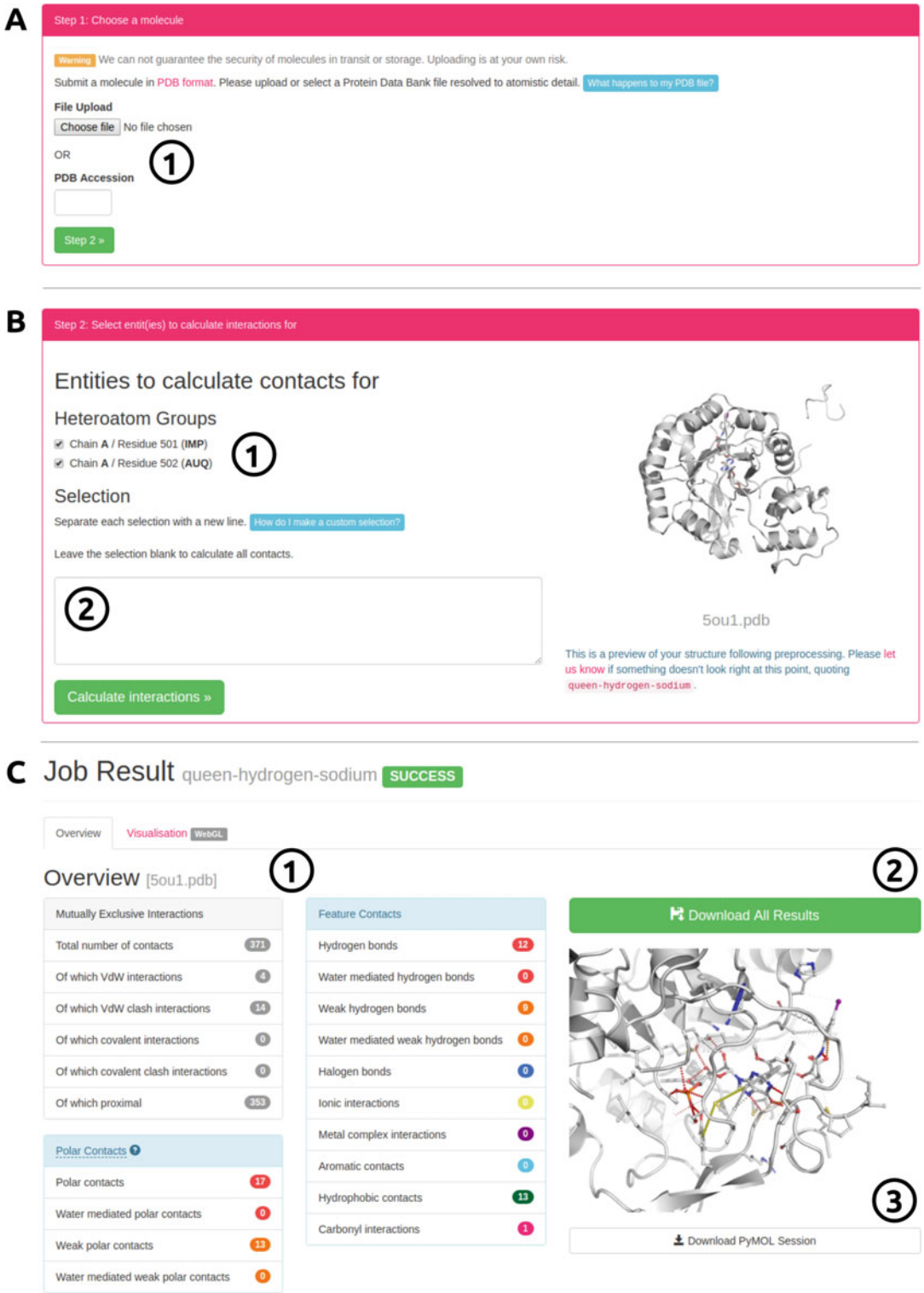


Fig. 4 Arpeggio submission and results web interface. (**a**) The submission page allows users to either provide their own PDB file or an accession code of a deposited experimental structure of the protein of interest. By selecting the molecule of interest (**b**), all molecular interactions will be calculated and displayed (**c**)

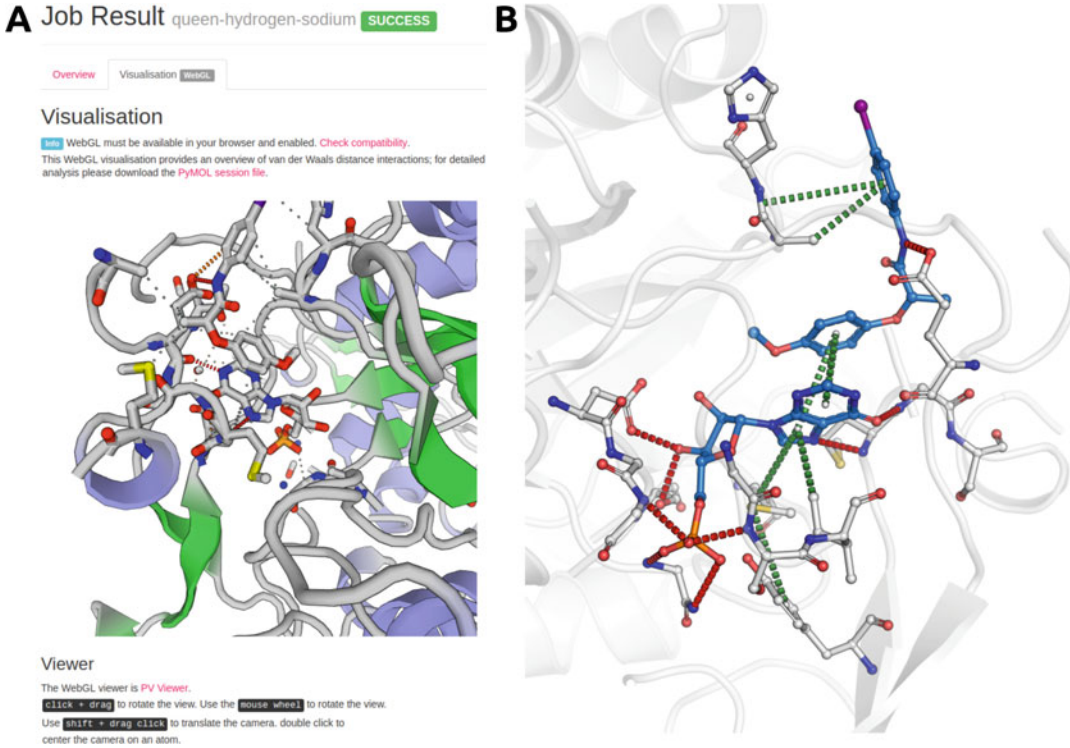


Fig. 5 Molecular interaction visualization using Arpeggio. The molecular interactions calculated by Arpeggio can be visualized either online (**a**) or by downloading the PSE file for visualization in Pymol (**b**)

three-letter code of the ligand of interest in the PDB file, and the approximate binding affinity (in nM) (Fig. 6a-2). If the binding affinity is not available, this can be approximated using CSM-lig. The mCSM-lig values do not vary significantly across most biologically relevant binding affinities.

3. After processing, the results page is shown (Fig. 6b-1), which includes information about the mutation and the predicted effects on the ligand binding affinity. An interactive molecular visualization is shown, allowing users to inspect the wild-type residue environment (Fig. 6b-2).
4. Predicted effects are outputted as the log fold change in binding affinity, in which negative values denote destabilizing mutations and positive values, stabilizing ones. The example shown in Fig. 6 and the web server depicts the prediction for a mutation on the HIV-1 protease bound to an inhibitor. Mutation from Aspartic Acid to Asparagine on residue position 30 is predicted to considerably reduce protein-ligand affinity. While users should interpret the values in the context of the protein system being studied, for competitive binding inhibitors, it is often important to consider the relative effect of a mutation on not only inhibitor binding but also the competitive ligand. This

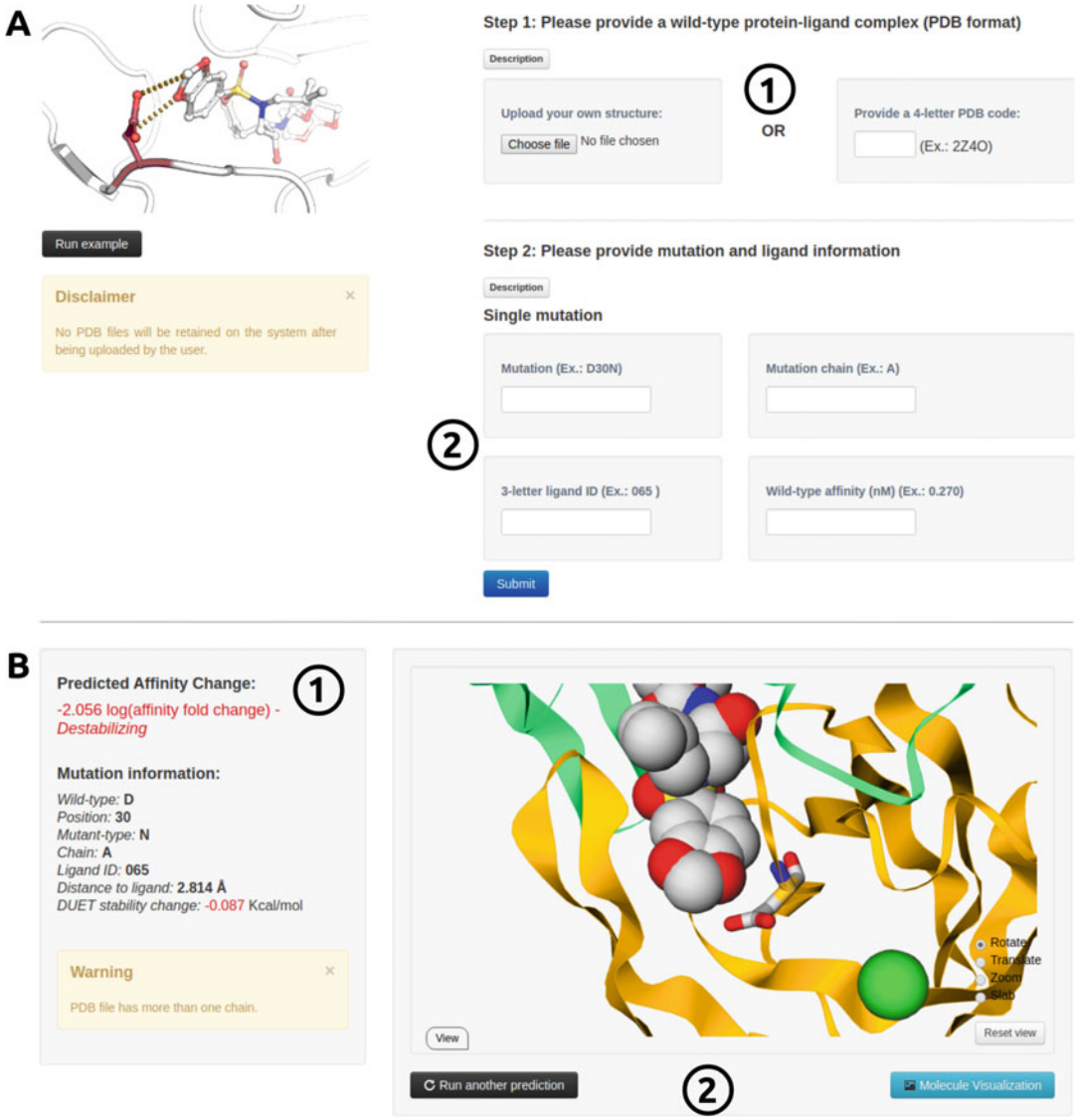


Fig. 6 mCSM-lig submission and results web interface. To predict the effects of a mutation on protein-ligand affinity, users need to provide a protein-ligand structure of interest (**a-1**) as well as mutation and ligand information (**a-2**). Once the calculations have finished, the results page will show the predicted change in ligand binding affinity (**b-1**) as well as an interactive visualization of the mutated residue within its molecular environment (**b-2**)

can be done by submitting a structure of the protein containing the ligand. Resistance mutations are more likely to affect, or have a larger effect, on inhibitor binding affinity than the natural ligand. This has been used to successfully preemptively guide detection of likely resistance variants [29–31, 47–53].

4 Notes

1. The chain ID for the provided PDB file is a mandatory field for CSM-Lig and mCSM-Lig, and blank characters are not allowed. It is possible that homology modeling tools might not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. There are several tools available to perform this task.
2. Another source of error comes from multiple occupancies, common in high-resolution experimental X-ray crystal structures. Multiple occupancies should first be filtered out, with the highest occupancy conformation normally selected.
3. NMR experimental structures often contain multiple models. It is an important practice to filter NMR structures, selecting a single model. The predictive tool will show a warning message in case multiple models are identified.
4. Arpeggio will sometimes fail if the PDB file contains an element with upper and lower case letters (e.g., Fe as opposed to FE). These can be altered using a text editor.

Acknowledgments

This work was supported by the Australian Government Research Training Program Scholarships [to S.P., M.K., Y.M., C.H.M.R.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.]; the Instituto René Rachou (IRR/FIOCRUZ Minas), Brazil, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [to D.E.V.P., P.M.R.]; and the Department of Biochemistry and Molecular Biology, University of Melbourne [to D.B.A.].

References

1. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691
2. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42 (Web Server issue):W314–W319. https://doi.org/10.1093/nar/gku411
3. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44 (W1):W557–W561. https://doi.org/10.1093/nar/gkw390
4. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based

signatures. Nucleic Acids Res 44(W1): W469–W473. https://doi.org/10.1093/nar/gkw458

5. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. https://doi.org/10.1038/srep29575
6. Pires DE, Chen J, Blundell TL, Ascher DB (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. https://doi.org/10.1038/srep19848
7. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
8. Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45(W1): W229–W235. https://doi.org/10.1093/nar/gkx439
9. Rodrigues CH, Ascher DB, Pires DE (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res 46(W1): W127–W132. https://doi.org/10.1093/nar/gky375
10. Rodrigues CH, Pires DE, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46(W1):W350–W355. https://doi.org/10.1093/nar/gky300
11. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43(Database issue):D387–D391. https://doi.org/10.1093/nar/gku966
12. Pires DEV, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45 (W1):W241–W246. https://doi.org/10.1093/nar/gkx236
13. Jafri M, Wake NC, Ascher DB, Pires DE, Gentle D, Morris MR, Rattenberry E, Simpson MA, Trembath RC, Weber A, Woodward ER, Donaldson A, Blundell TL, Latif F, Maher ER (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov 5(7):723–729. https://doi.org/10.1158/2159-8290.CD-14-1096
14. Jubb H, Blundell TL, Ascher DB (2015) Flexibility and small pockets at protein-protein interfaces: new insights into druggability. Prog Biophys Mol Biol 119(1):2–9. https://doi.org/10.1016/j.pbiomolbio.2015.01.009
15. Usher JL, Ascher DB, Pires DE, Milan AM, Blundell TL, Ranganath LR (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep 24:3–11. https://doi.org/10.1007/8904_2014_380
16. Coelho MB, Ascher DB, Gooding C, Lang E, Maude H, Turner D, Llorian M, Pires DE, Attig J, Smith CW (2016) Functional interactions between polypyrimidine tract binding protein and PRI peptide ligand containing proteins. Biochem Soc Trans 44(4):1058–1065. https://doi.org/10.1042/BST20160080
17. Kano FS, Souza-Silva FA, Torres LM, Lima BA, Sousa TN, Alves JR, Rocha RS, Fontes CJ, Sanchez BA, Adams JH, Brito CF, Pires DE, Ascher DB, Sell AM, Carvalho LH (2016) The presence, persistence and functional properties of Plasmodium vivax Duffy binding protein II antibodies are influenced by HLA class II allelic variants. PLoS Negl Trop Dis 10(12):e0005177. https://doi.org/10.1371/journal.pntd.0005177
18. Nemethova M, Radvanszky J, Kadasi L, Ascher DB, Pires DE, Blundell TL, Porfirio B, Mannoni A, Santucci A, Milucci L, Sestini S, Biolcati G, Sorge F, Aurizi C, Aquaron R, Alsbou M, Lourenco CM, Ramadevi K, Ranganath LR, Gallagher JA, van Kan C, Hall AK, Olsson B, Sireau N, Ayoob H, Timmis OG, Sang KH, Genovese F, Imrich R, Rovensky J, Srinivasaraghavan R, Bharadwaj SK, Spiegel R, Zatkova A (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24(1):66–72. https://doi.org/10.1038/ejhg.2015.60
19. Silvino AC, Costa GL, Araujo FC, Ascher DB, Pires DE, Fontes CJ, Carvalho LH, Brito CF, Sousa TN (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. PLoS One 11(7):e0160172. https://doi.org/10.1371/journal.pone.0160172
20. White RR, Ponsford AH, Weekes MP, Rodrigues RB, Ascher DB, Mol M, Selkirk ME, Gygi SP, Sanderson CM, Artavanis-Tsakonas K (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog 12(11):

e1005977. https://doi.org/10.1371/journal.ppat.1005977

21. Casey RT, Ascher DB, Rattenberry E, Izatt L, Andrews KA, Simpson HL, Challis B, Park SM, Bulusu VR, Lalloo F, Pires DEV, West H, Clark GR, Smith PS, Whitworth J, Papathomas TG, Taniere P, Savisaar R, Hurst LD, Woodward ER, Maher ER (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5(3):237–250. https://doi.org/10.1002/mgg3.279
22. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montano B, Blundell TL, Ascher DB (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002
23. Ramdzan YM, Trubetskov MM, Ormsby AR, Newcombe EA, Sui X, Tobin MJ, Bongiovanni MN, Gras SL, Dewson G, Miller JML, Finkbeiner S, Moily NS, Niclis J, Parish CL, Purcell AW, Baker MJ, Wilce JA, Waris S, Stojanovski D, Bocking T, Ang CS, Ascher DB, Reid GE, Hatters DM (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. Cell Rep 19(5):919–927. https://doi.org/10.1016/j.celrep.2017.04.029
24. Soardi FC, Machado-Silva A, Linhares ND, Zheng G, Qu Q, Pena HB, Martins TMM, Vieira HGS, Pereira NB, Melo-Minardi RC, Gomes CC, Gomez RS, Gomes DA, Pires DEV, Ascher DB, Yu H, Pena SDJ (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med 2(1):7. https://doi.org/10.1038/s41525-017-0009-4
25. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, Balding DJ, Petrovski S (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27(10):1715–1729. https://doi.org/10.1101/gr.226589.117
26. Trezza A, Bernini A, Langella A, Ascher DB, Pires DEV, Sodi A, Passerini I, Pelo E, Rizzo S, Niccolai N, Spiga O (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci 58(12):5320–5328. https://doi.org/10.1167/iovs.17-22158
27. Andrews KA, Ascher DB, Pires DEV, Barnes DR, Vialard L, Casey RT, Bradshaw N, Adlard J, Aylwin S, Brennan P, Brewer C, Cole T, Cook JA, Davidson R, Donaldson A, Fryer A, Greenhalgh L, Hodgson SV, Irving R, Lalloo F, McConachie M, McConnell VPM, Morrison PJ, Murday V, Park SM, Simpson HL, Snape K, Stewart S, Tomkins SE, Wallis Y, Izatt L, Goudie D, Lindsay RS, Perry CG, Woodward ER, Antoniou AC, Maher ER (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55(6):384–394. https://doi.org/10.1136/jmedgenet-2017-105127
28. Hnizda A, Fabry M, Moriyama T, Pachl P, Kugler M, Brinsa V, Ascher DB, Carroll WL, Novak P, Zaliova M, Trka J, Rezacova P, Yang JJ, Veverka V (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia 32(6):1393–1403. https://doi.org/10.1038/s41375-018-0073-5
29. Albanaz ATS, Rodrigues CHM, Pires DEV, Ascher DB (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12(6):553–563. https://doi.org/10.1080/17460441.2017.1322579
30. Park Y, Pacitto A, Bayliss T, Cleghorn LA, Wang Z, Hartman T, Arora K, Ioerger TR, Sacchettini J, Rizzi M, Donini S, Blundell TL, Ascher DB, Rhee K, Breda A, Zhou N, Dartois V, Jonnala SR, Via LE, Mizrahi V, Epemolu O, Stojanovski L, Simeons F, Osuna-Cabello M, Ellis L, MacKenzie CJ, Smith AR, Davis SH, Murugesan D, Buchanan KI, Turner PA, Huggett M, Zuccotto F, Rebollo-Lopez MJ, Lafuente-Monasterio MJ, Sanz O, Diaz GS, Lelievre J, Ballell L, Selenski C, Axtman M, Ghidelli-Disse S, Pflaumer H, Bosche M, Drewes G, Freiberg GM, Kurnick MD, Srikumaran M, Kempf DJ, Green SR, Ray PC, Read K, Wyatt P, Barry CE 3rd, Boshoff HI (2017) Essential but not vulnerable: indazole sulfonamides targeting inosine monophosphate dehydrogenase as potential leads against Mycobacterium tuberculosis. ACS Infect Dis 3(1):18–33. https://doi.org/10.1021/acsinfecdis.6b00103
31. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, Keri G, Ascher DB, Mondesert G, Vocat A, Lupien A, Sommer R, Vermet H, Lagrange S, Buechler J, Warner DF, McKinney JD, Pato J, Cole ST, Blundell TL, Rizzi M, Mizrahi V (2017) The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. ACS Infect Dis 3(1):5–17. https://doi.org/10.1021/acsinfecdis.6b00102

32. Trapero A, Pacitto A, Singh V, Sabbah M, Coyne AG, Mizrahi V, Blundell TL, Ascher DB, Abell C (2018) Fragment-based approach to targeting inosine-5′-monophosphate dehydrogenase (IMPDH) from Mycobacterium tuberculosis. J Med Chem 61(7):2806–2822. https://doi.org/10.1021/acs.jmedchem.7b01622
33. Pires DE, Blundell TL, Ascher DB (2015) pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem 58 (9):4066–4072. https://doi.org/10.1021/acs.jmedchem.5b00104
34. Pires DEV, Kaminskas LM, Ascher DB (2018) Prediction and optimization of pharmacokinetic and toxicity properties of the ligand. Methods Mol Biol 1762:271–284. https://doi.org/10.1007/978-1-4939-7756-7_14
35. Sigurdardottir AG, Winter A, Sobkowicz A, Fragai M, Chirgadze D, Ascher DB, Blundell TL, Gherardi E (2015) Exploring the chemical space of the lysine-binding pocket of the first kringle domain of hepatocyte growth factor/scatter factor (HGF/SF) yields a new class of inhibitors of HGF/SF-MET binding. Chem Sci 6(11):6147–6157. https://doi.org/10.1039/c5sc02155c
36. Ascher DB, Jubb HC, Pires DE, Ochi T, Higueruelo A, Blundell TL (2015) Protein-protein interactions: structures and druggability. In: Scapin G, Patel D, Arnold E (eds) Multifaceted roles of crystallography in modern drug discovery. NATO science for peace and security series a: chemistry and biology. Springer, Netherlands, pp 141–163. https://doi.org/10.1007/978-94-017-9719-1_12
37. Pandurangan AP, Ascher DB, Thomas SE, Blundell TL (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45(2):303–311. https://doi.org/10.1042/BST20160422
38. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
39. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. J Cheminform 3:33. https://doi.org/10.1186/1758-2946-3-33
40. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform 4(1):17. https://doi.org/10.1186/1758-2946-4-17
41. Ascher DB, Crespi GA, Ng HL, Morton CJ, Parker MW (2008) Novel therapeutic approaches to treat Alzheimer's disease and memory disorders. J Proteomics Bioinform 1:464–476
42. Chai SY, Yeatman HR, Parker MW, Ascher DB, Thompson PE, Mulvey HT, Albiston AL (2008) Development of cognitive enhancers based on inhibition of insulin-regulated aminopeptidase. BMC Neurosci 9(Suppl 2):S14. https://doi.org/10.1186/1471-2202-9-S2-S14
43. Kawabata T (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins 78(5):1195–1211. https://doi.org/10.1002/prot.22639
44. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46 (1–3):3–26
45. Schrodinger, LLC (2015) The PyMOL molecular graphics system, version 1.8
46. Balint M, Jeszenoi N, Horvath I, van der Spoel D, Hetenyi C (2017) Systematic exploration of multiple drug binding sites. J Cheminform 9(1):65. https://doi.org/10.1186/s13321-017-0255-6
47. Ascher DB, Wielens J, Nero TL, Doughty L, Morton CJ, Parker MW (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. https://doi.org/10.1038/srep04765
48. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, Coeck N, Hill-Cawthorne GA, Nair MB, Mallard K, Ramsay A, Campino S, Hibberd ML, Pain A, Rigouts L, Clark TG (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14(1):31. https://doi.org/10.1186/s12916-016-0575-9
49. Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, Spelman DW, Padiglione A, Peleg AY, Holt KE (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom 4. https://doi.org/10.1099/mgen.0.000165
50. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN, Thwaites G, Edwards DJ, Nath AP, Pham K, Ascher DB, Farrar J, Khor CC, Teo YY, Inouye M, Caws M, Dunstan SJ (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage

and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet 50 (6):849–856. https://doi.org/10.1038/s41588-018-0117-9

51. Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, Denholm JT, Ascher DB (2018) Analysis of a novel pncA mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med 198 (4):541–544. https://doi.org/10.1164/rccm.201712-2572LE
52. Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Sci Rep 8 (1):15356. https://doi.org/10.1038/s41598-018-33370-6
53. Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, Arumugam S, Rajan L, Ebenezer M, Ascher DB, Arnold E, Blundell TL (2018) Structural implications of mutations conferring rifampin resistance in Mycobacterium leprae. Sci Rep 8(1):5016. https://doi.org/10.1038/s41598-018-23423-1

APPENDIX 5: IDENTIFYING GENOTYPE-PHENOTYPE CORRELATIONS VIA INTEGRATIVE MUTATION ANALYSIS



Chapter 1

Identifying Genotype–Phenotype Correlations via Integrative Mutation Analysis

Edward Airey, Stephanie Portelli, Joicymara S. Xavier, Yoo Chan Myung, Michael Silk, Malancha Karmakar, João P. L. Velloso, Carlos H. M. Rodrigues, Hardik H. Parate, Anjali Garg, Raghad Al-Jarf, Lucy Barr, Juliana A. Geraldo, Pâmela M. Rezende, Douglas E. V. Pires, and David B. Ascher

Abstract

Mutations in protein-coding regions can lead to large biological changes and are associated with genetic conditions, including cancers and Mendelian diseases, as well as drug resistance. Although whole genome and exome sequencing help to elucidate potential genotype–phenotype correlations, there is a large gap between the identification of new variants and deciphering their molecular consequences. A comprehensive understanding of these mechanistic consequences is crucial to better understand and treat diseases in a more personalized and effective way. This is particularly relevant considering estimates that over 80% of mutations associated with a disease are incorrectly assumed to be causative. A thorough analysis of potential effects of mutations is required to correctly identify the molecular mechanisms of disease and enable the distinction between disease-causing and non–disease-causing variation within a gene. Here we present an overview of our integrative mutation analysis platform, which focuses on refining the current genotype–phenotype correlation methods by using the wealth of protein structural information.

Key words Genotype–phenotype correlations, Graph-based signatures, mCSM, Mutation, Protein structure, Protein interactions

1 Introduction

Proteins are versatile molecules, responsible for orchestrating a wide range of biological processes. They comprise a single polypeptide chain of amino acids, which folds in 3D space into dynamic structures. How a protein folds is important for determining its functions, including activities and interactions with other molecules. These structures are highly coordinated and conserved across evolution, and small perturbations in the amino acid sequence can disrupt these shapes, functions, and interactions [1, 2]. While

Hugh Cartwright (ed.), *Artificial Neural Networks*, Methods in Molecular Biology, vol. 2190,
https://doi.org/10.1007/978-1-0716-0826-5_1, © Springer Science+Business Media, LLC, part of Springer Nature 2021

missense mutations, causing a change to a single amino acid, are generally less structurally disruptive than nonsense mutations, their effects are highly variable and can be wide-ranging, making their molecular consequences harder to determine. Despite their subtle effects, missense substitutions are related with many different genetic conditions, including cancer, Mendelian diseases, and the emergence of drug resistance.

The introduction of a missense mutation can have many molecular effects, including altering how the protein folds, its dynamics, posttranslational modifications, half-life, localization, activity, and molecular interactions [3]. When analyzing a new mutation, an integrative approach is therefore important to consider the effects it might have on all of these aspects. This enables the identification of specific functional, and structural changes imparted by the mutations, which is essential for a molecular understanding. It can also explain why mutations in the same protein might lead to different diseases, why mutations might cluster in 3D space and how those genetic changes present phenotypically.

Although many assume that an unfavorable phenotype (e.g., pathogenic, drug-resistant) is the result of large, overall destabilizing mutations, mutations with milder effects are often more prevalent in a population, as they are generally under less selective pressure [4, 5]. For example, by assessing mutations in three different tuberculosis proteins that lead to resistance, we have shown that the most frequent resistant mutations were more likely to be associated with overall mild functional effects, and associated reduced fitness cost, allowing for increased prevalence within the bacterial population [4].

Experimentally elucidating the biophysical effects of mutations is an expensive and time-consuming task, usually limited to a few variants in proteins with amenable assays. Over the years, the accumulation of information of experimentally characterized mutations has enabled the development and improvement of computational mutational analysis tools [6]. These computational platforms have shown to be invaluable assets to decipher genotype–phenotype correlations in cancer [7–19], Mendelian diseases [20–26], and detection of antimicrobial resistance [4, 15, 27–35], guiding clinical decisions and driving further research. Here, we introduce a general computational pipeline that uses in silico biophysical predictions and machine learning approaches to harness the wealth of available biological and protein structural information and give insights into genotype–phenotype correlation for clinical use [10].

The mutation cutoff scanning matrix (mCSM) platform is the only comprehensive collection of in silico tools for quantitatively predicting the effects of missense mutations on protein folding, structure, dynamics, and interactions. It includes tools which calculate all possible molecular interactions (Arpeggio [36]), account for changes in protein stability (mCSM-Stability [37], SDM [38],

DUET [39], mCSM-membrane [40], dynamics (DynaMut [41]), protein interactions with other proteins (mCSM-PPI [37], mCSM-PPI2 [42], mCSM-AB [43], mCSM-AB2 [44], mmCSM-AB [45], nucleic acids (mCSM-DNA [37], mCSM-NA [46]), and small molecule ligands (mCSM-lig [47], CSM-lig [48]).

These tools were built using the concept of graph-based signatures [49, 50], which represent the geometry and physicochemical properties of the wild-type protein structure environment as a network or graph, composed of a series of nodes, describing the local mutation environment, and edges, describing the distances between interacting "layers" of surrounding residues. Information on the mutation is captured using the pharmacophore change between the wild-type and the mutant residue, including whether charges or hydrogen donors/acceptors have been gained or lost [37].

This platform allows for accurate biophysical predictions, which, when complemented with other protein analytical tools, can provide a detailed landscape on the specific mutational effects on a protein. We have implemented these within an analytical and supervised machine learning predictive pipeline (Fig. 1), to enable easy and fast characterization of novel mutations and their likely

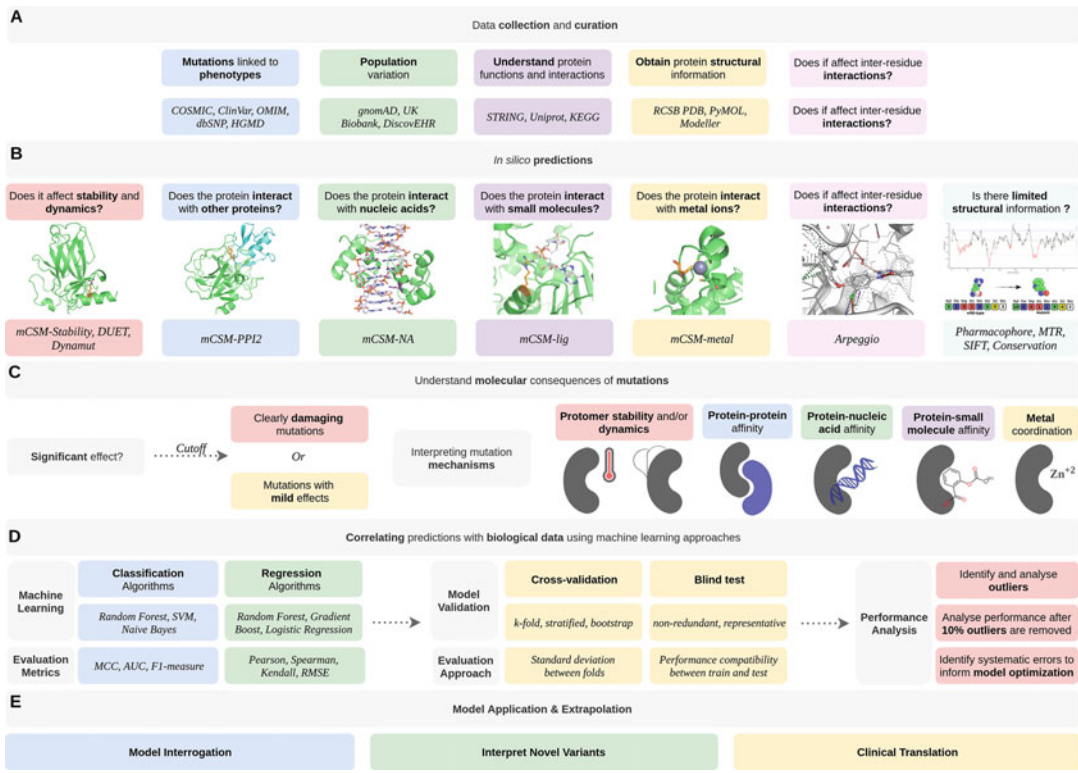


Fig. 1 An overview of the mechanistic characterization of mutations and their biological consequences, to guide the development of tools to predict phenotypic outcomes

clinical phenotypes. This approach has been shown to have big implications in diagnostic and personalized medicine in the post-genomic era.

2 Materials

2.1 Data Curation

2.1.1 Mutation Curation

The foremost requirement for training a machine learning model is appropriate high-quality experimental/clinical data, with suitable representation of the classes under comparison. For human disease, a wealth of freely accessible collections of curated data exist. Previously reported mutations through publications and functional studies are available from dbSNP [51], the largest freely available repository of genetic variation. Variants with evidence of pathogenicity can be viewed from the Human Gene Mutation Database (HGMD) [52] and ClinVar [53], and from disease-specific datasets such as the Catalogue of Somatic Mutations in Cancer (COSMIC). Standing variation is available from genomic sequencing efforts of healthy populations, including over 140,000 healthy humans in gnomAD [54] and 50,000 whole exomes currently available in UK Biobank [55].

When combining data from multiple sources, it is important that all datapoints are comparable. If using genetic coordinates, they should be found on the same assembly of the genome (e.g., GRCh38 vs GRCh37). The mutations themselves (whether reported as genetic or amino acid changes) must be reported on the same transcript, as most genes have multiple reported coding sequences.

2.1.2 Protein Structure Curation

The sequence and functional information for a specific protein can be obtained from Uniprot (https://www.uniprot.org/) [56]. To run the mCSM tools we need crystallographic structures, which can be downloaded from the Protein Data Bank (PDB;http://www.rcsb.org/) [57] or generated via homology modeling or molecular docking (to run mCSM-PPI, mCSM-Lig, or mCSM-NA). Once we have the variant information collected from the resources in Subheading 2.1.1, we map these variants on to the identified protein structures to help visualize the spread and identify potential hotspots, which is easily done using visualization software such as PyMol, as it enables selection of residues being mutated in a 3D manner.

2.2 An Overview of Computational Tools to Analyze Missense Mutations

Over the past two decades there has been an unprecedented growth in both computational power and the amount of biological data available. This has facilitated the development of numerous sequence (Table 1) and structural (Table 2) based computational tools to guide mutation characterization.

Table 1
Available sequence-based predictive tools for mutation analysis

Protein stability and dynamics	
Method	**Corr.[a]**
I-Mutant 2.0	0.62
Auto-Mute	0.64[a]
MUpro	0.75
DynaMine	0.63
DDGun	0.49
INPS-MD/3D	0.58
iStable	0.56[b]
iPTREEE - STAB	0.70
ProMaya	0.79

[a]Pearson's correlation
[b]MCC

Table 2
Available structure-based predictive tools for mutation analysis

Protein stability and dynamics		Protein–protein affinity		Protein–nucleic acid affinity		Protein–small molecule affinity	
Method	**Corr.[a]**	**Method**	**Corr.[b]**	**Method**	**Corr.[c]**	**Method**	**Corr.[d]**
mCSM-Stability	0.69	mCSM-PPI	0.16	mCSM-NA	0.70	mCSM-lig	0.63
DUET	0.68	mCSM-PPI2	0.42				
DynaMut	0.70	BeAtMuSiC	0.28				
SDM2	0.61	MutaBind	0.41				
STRUM	0.79	FoldX	0.12				
PopMuSiC 2.1	0.63	MMPBSA	0.19				
CUPSAT	0.78						
Eris	0.75						
INPS-MD/3D	0.72						

[a]Pearson's correlation when evaluated on blind-test sets derived from the ProTherm database
[b]Kendall rank correlation coefficient on 1007 single-point mutations from CAPRI (T55)
[c]Pearson's correlation on 331 single-point mutations from 38 protein–nucleic acid complexes
[d]Pearson's correlation on 763 single-point mutations from 200 protein–ligand complexes

The mCSM platform is the only available approach to consider all possible molecular effects and has therefore formed the central component of our mutational analysis pipeline. All mCSM

Platform tools are available freely as websites compatible with most web-browsers, but Google Chrome is recommended. A summary of these methods and links to access them is described in Table 3.

Table 3
Computational tools available in the mCSM platform

mCSM tool	Type	Function
Arpeggio[a]	Protein interaction	Calculates 13 different types of interactions between atoms including hydrogen bonds, halogen bonds, carbonyl interactions, and others.
MTR-Viewer[b]	Missense tolerance	A measure of a gene's regional tolerance to missense variation.
mCSM-Stability[c]	Stability	Predict the effects of a mutation on the overall protein stability
SDM2[d]	Stability	Predicts the change in protein stability due to a single mutation using conformationally constrained environment-dependent amino acid substitution tables.
DUET[e]	Stability	Uses mCSM-Stability and SDM2 in order to create a consensus prediction the effects of a mutation on protein stability
DynaMut[f]	Flexibility	Looks to predict the effects of a mutation on protein stability, flexibility, and dynamics
mCSM-PPI[g]	Protein interaction	Predicts the effects of a mutation within a specified protein on its impact with overall protein–protein interactions.
mCSM-PPI2[h]	Protein interaction	Creates a similar prediction to PPI but incorporates the effects of mutations on interresidue noncovalent interaction network using graph kernels, evolutionary information, complex network metrics, and energetic terms.
mCSM-DNA[i]	Protein interaction	Predicts the impact of mutations on the protein interaction with DNA.
mCSM-NA[j]	Protein interaction	Predicts the impact of mutations on the protein interaction with nucleic acids, and uses pharmacophore and information about nucleic acid properties.
mCSM-Lig[k]	Protein interaction	Predicts the effects of single-point mutations on the stability of a protein–ligand complex.

[a]http://biosig.unimelb.edu.au/arpeggioweb/
[b]http://biosig.unimelb.edu.au/mtr-viewer/
[c]http://biosig.unimelb.edu.au/mcsm/stability
[d]http://marid.bioc.cam.ac.uk/sdm2
[e]http://biosig.unimelb.edu.au/duet/
[f]http://biosig.unimelb.edu.au/dynamut/
[g]http://biosig.unimelb.edu.au/mcsm/protein_protein
[h]http://biosig.unimelb.edu.au/mcsm_ppi2/
[i]http://biosig.unimelb.edu.au/mcsm/protein_dna
[j]http://biosig.unimelb.edu.au/mcsm_na/
[k]http://biosig.unimelb.edu.au/mcsm_lig/

3 Methods

3.1 Predicting and Analyzing Structural and Biophysical Effects of Mutations Using the mCSM Platform

The mCSM methods can be categorized by purpose. As shown in Fig. 1, methods are chosen depending on interactions made, and what structural information is available. Below we discuss how each type of predictor can be used and interpreted.

- The user should choose the appropriate tools based on what information is available on their protein of interest (Fig. 1).
- In general, each mCSM tool requires a wild-type protein file, in the PDB format, and the single-point mutation or a list of mutations. Some tools may require additional specific information; Table 4 shows the inputs required for each tool. **Notes 1** and **2** highlight some common issues with the submission inputs.

3.2 mCSM Platform Output

The results of Arpeggio are shown in Fig. 2.

3.2.1 Arpeggio

- After submitting a job, an overview of the type and number of atomic interactions within the protein is shown (Fig. 2a). Arpeggio calculates all types of molecular interactions (Table 5), which are displayed and downloadable along with a visual representation of the atomic contacts overlaid on the protein structure (Fig. 2b).
- The number of each interaction/contact and PyMOL session files can be downloaded for a more detailed analysis.

3.2.2 MTR-Viewer

Gene Viewer

- The MTR gene viewer [5] results page (Fig. 3) shows predicted MTR scores in an interactive line graph with a control panel which allows users to adjust the window size and the ethnicity for MTR estimates. A line graph (Fig. 3a) displays regions that have high variation, low-MTR scored; those in red are most likely to be pathogenic. Any ethnicity-specific MTR scores are shown in blue on the line graph.
- The first lollipop plot (Fig. 3b) shows observed missense (yellow) and synonymous (green) variations based on gnomeAD.
- If the gene of interest is a ClinVar pathogenic gene, their pathogenic (red) and benign (blue) missense variants are displayed under the gnomeAD lollipop plot (Fig. 3c).
- Users can browse results of alternative-transcript (Fig. 3d) of the given query if available.

Variant Query

- The variant query result page (Fig. 4) shows MTR scores for each user-supplied missense variant, providing the estimated regional intolerance. Low MTR scores indicate stronger purifying selection within the population. Users can also press "view" next to a variant to show its position within its gene transcript.

Table 4
Information required to run each mCSM program

		Inputs	
mCSM tool	**Task**	**Step 1**	**Step 2**
Arpeggio	Calculate	Molecule in PDB format or PDB accession code.	Select desired interaction calculation. You can select any (including multiple) part of the PDB file using the syntax: /1/2/3 Where: 1. Chain ID. 2. Residue number. 3. Atom name.
MTR-Viewer	Gene Viewer Variant Queries	Gene, ensembl ID, or Refseq ID Variants as GrCh37 genomic coordinates.	Select window size and overlay sub-population
mCSM-Stability, mCSM-PPI, mCSM-DNA	Prediction	Wild-type protein file in PDB format. For mCSM-PPI and mCSM-DNA, the structure of the complex in PDB format is required.	Single mutation (code and mutation chain), file with a list of mutations and its respective chains or code of residue and the mutation chain.
SDM2	Prediction	Wild-type protein structure in a PDB format or PDB accession code.	Single mutation (code and mutation chain) or residue/position code and the mutation chain.
DUET	Prediction	Wild-type protein structure in a PDB format or PDB accession code.	Single mutation (code and mutation chain)
DynaMut	Analysis Prediction	Wild-type protein structure in a PDB format or PDB accession code. Wild-type protein structure in a PDB format or PDB accession code.	The selection of a Force Field and email (optional field). Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and email (optional field).
mCSM-PPI2	Prediction Analysis	The structure of the complex in PDB format or corresponding PDB accession code. The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and email (optional field). Mutation details (alanine scanning or saturation mutagenesis) and email (optional field).
mCSM-NA	Prediction	The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and the selection of the Nucleic Acid Type.
mCSM-Lig	Prediction	The structure of the complex in PDB format or corresponding PDB accession code.	Single mutation (code and mutation chain) and ligand information (three-letter ligand ID and estimated wild-type affinity).

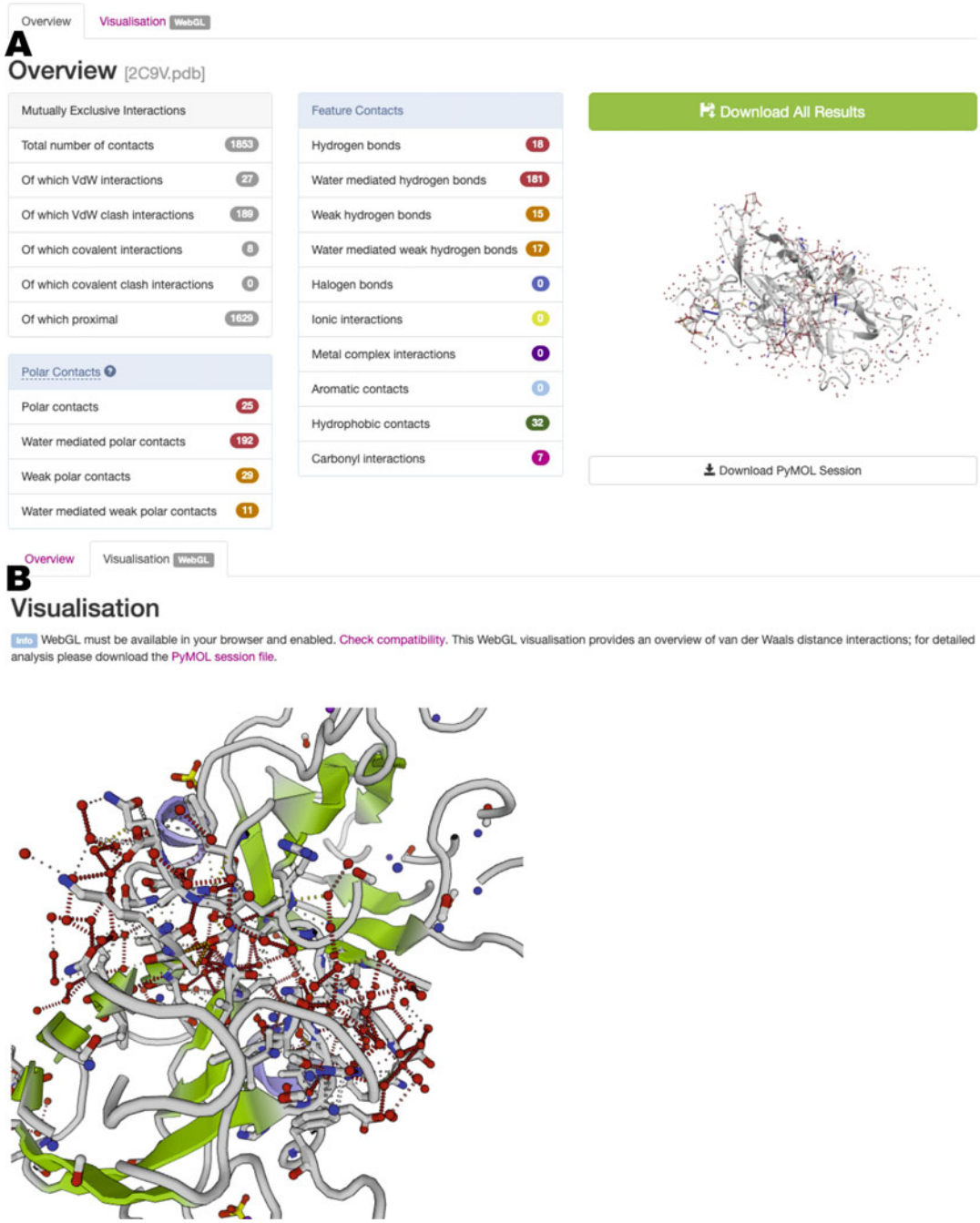


Fig. 2 Output of the Arpeggio tool. (**a**) Overview of the output for the inputted protein including the different types of interactions. (**b**) Visualization of the interactions shown on a protein structure

Table 5
Atomic interactions calculated by Arpeggio

Atomic interaction	Description	Arpeggio class	Bond energy (kJ/mol)
Van der Waals (dipole)	Permanent, induced and instantaneous dipoles	VWD	1–9
Hydrophobic	Between aliphatic and aromatic atoms	Hydrophobic	4–12
Hydrogen bond	Between carboxyl, amide, imidazole, guanidine, amino, hydroxyl and phenolic groups	Hydrogen bonds, weak hydrogen bond, polar contacts, halogen bonds, carbonyl interactions	8–40
Pi interactions	From/to rings	Aromatic contacts	6–70
Electrostatic	Between carboxyl and amino groups	Ionic interactions, metal complex	42–84

3.2.3 mCSM-Stability/PPI/DNA

The impact of mutations on protein stability, protein–protein binding affinity, and protein–DNA affinity can be predicted by mCSM-Stability, mCSM-PPI, mCSM-DNA with three types of prediction; single, multiple and systematic mutation.

Single Mutation

- If the single mutation option is selected in one of the tools within the mCSM platform, it will be shown on a results page after processing. This information includes the predicted value changes (protein stability, protein–protein interaction, protein–DNA interaction) as measured by the change in Gibbs Free Energy $\Delta\Delta G$ kcal/mol (Fig. 5), which is classified as highly destabilizing ($\Delta\Delta G \leq -2$ kcal/mol), destabilizing (-2 kcal/mol $< \Delta\Delta G < 0$ kcal/mol), stabilizing (0 kcal/mol $\leq \Delta\Delta G < 2$ kcal/mol), or highly stabilizing ($\Delta\Delta G \geq 2$ kcal/mol).
- If the structure of a complex is submitted to mCSM-Stability, it will calculate the predicted change in stability of the entire complex. It is therefore often advisable to also run predictions on a PDB file containing the protomer chain alone.
- For mCSM-PPI and mCSM-DNA, for mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there are fewer mutations located further away than 12 Å in the datasets used to train the methods.
- Also shown is an interactive 3D visual representation of the uploaded PDB file (Fig. 5a, right).

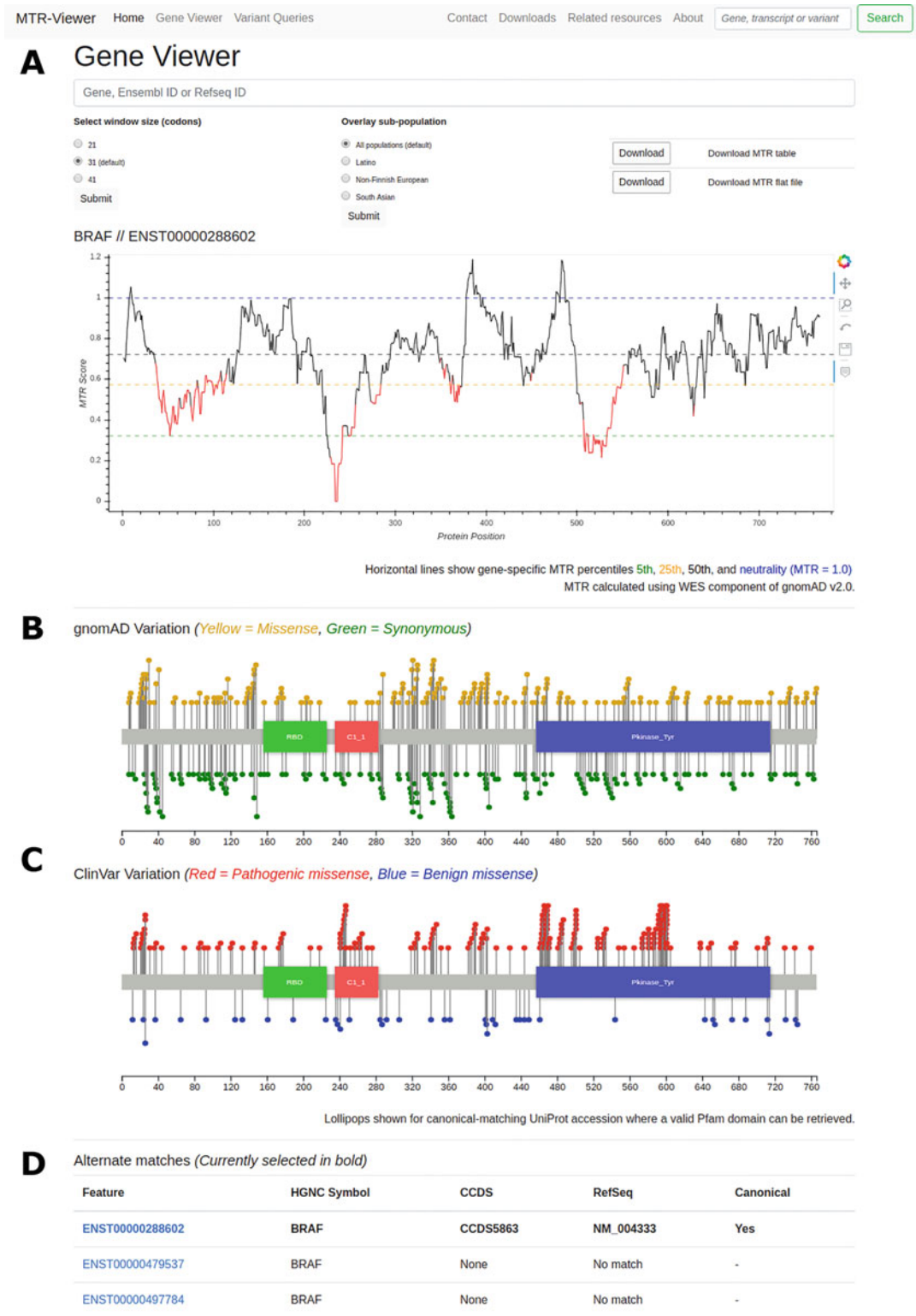
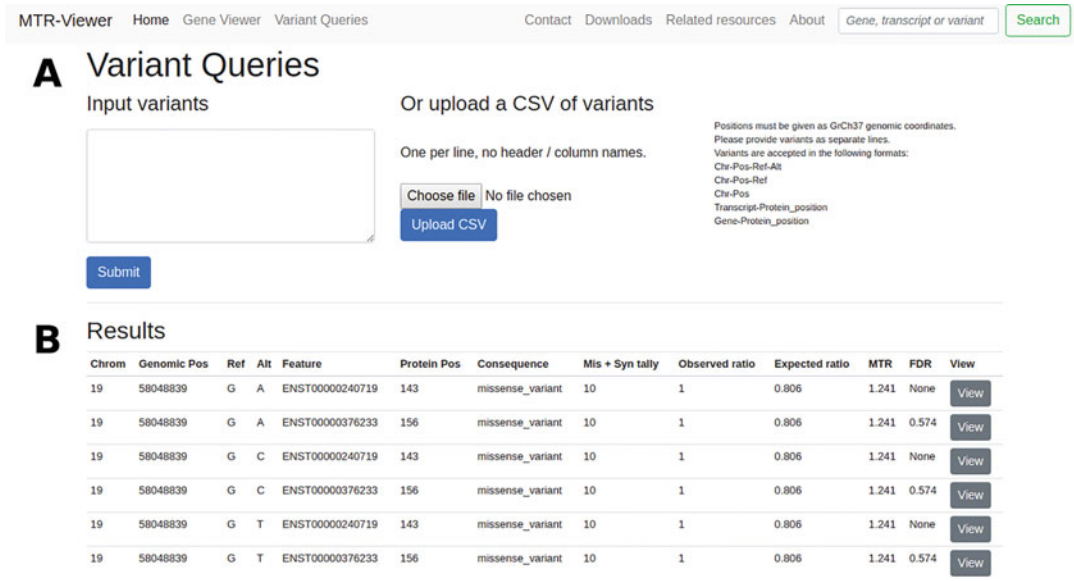


Fig. 3 The MTR Gene Viewer result page. (**a**) The line graph shows MTR scores in red for variations distant from neutrality across the transcript according to selected window size (codons) and subpopulation option. (**b**) The lollipop plot shows observed gnomeAD variation in yellow and green for missense and synonymous variation. (**c**) The second lollipop plot displays pathogenic (red) and benign (blue) missense variants based on ClinVar annotation. (**d**) The alternate transcripts can be shown in a table with RefSeq ID



Chrom	Genomic Pos	Ref	Alt	Feature	Protein Pos	Consequence	Mis + Syn tally	Observed ratio	Expected ratio	MTR	FDR	View
19	58048839	G	A	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	A	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View
19	58048839	G	C	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	C	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View
19	58048839	G	T	ENST00000240719	143	missense_variant	10	1	0.806	1.241	None	View
19	58048839	G	T	ENST00000376233	156	missense_variant	10	1	0.806	1.241	0.574	View

Fig. 4 MTR Variant Queries result page. Calculated results and information for the given input variants (or a CSV). User can check the details through MTR Gene Viewer by clicking on the view button

Multiple or Systematic

- If the option for inputting a list of mutations or systematic was used to analyze the PDB file, then after processing, results will be shown in tabulated form (Fig. 5b), including mutation specific information such as the residue solvent accessibility (RSA), as well as the predicted ΔΔG.
- Each result is also classified, using the predicted ΔΔG value, as highly destabilizing, destabilizing, stabilizing, or highly stabilizing.
- Users can search the result table or download results into a tab-separated text file.

3.2.4 SDM

SDM uses environment-specific amino acid substitution tables [38] and structural features including residue depth [15] and packing density to predict the impact of mutations on protein stability. The result page of single and list mutation is as follows.

Single Mutation

- The single mutation result page (Fig. 6a) provides predicted protein stability changes (ΔΔG), in addition to structural information implemented in SDM including secondary structure, RSA, residue depth and residue occluded packing density (OSP), sidechain–sidechain hydrogen bond (HBOND_SS), sidechain–main chain amide hydrogen bond (HBOND_SN), and sidechain–main chain carbonyl hydrogen bond (HBOND_SO). The integrated 3D viewer also shows the

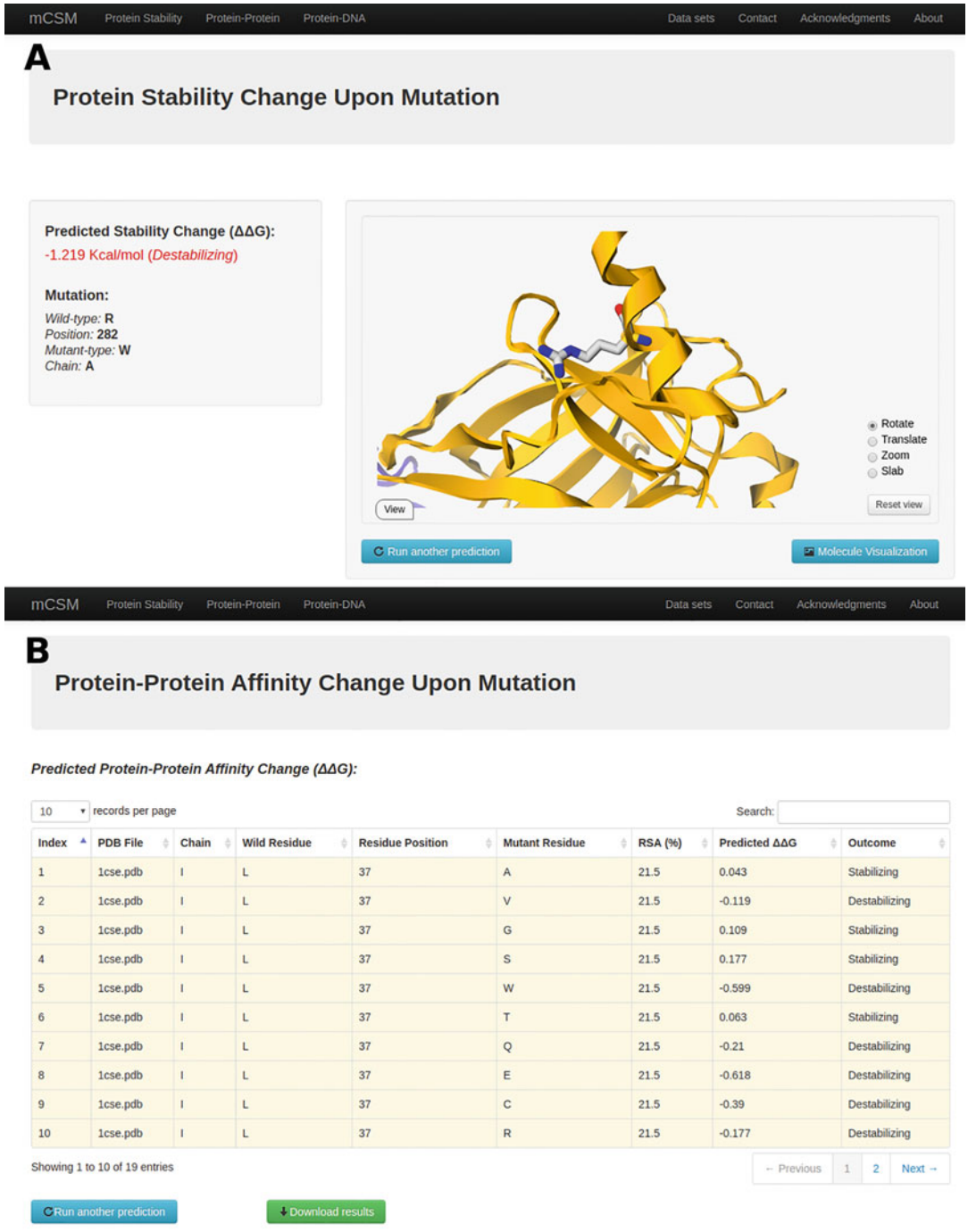


Fig. 5 Result pages for mCSM-Stability, mCSM-PPI and mCSM-DNA. (**a**) mCSM-Stability (single mutation) and (**b**) mCSM-PPI (multiple/systematic mutation). (**a**) The single prediction for example mCSM-Stability page supports 3D interactive viewer for structural analysis. (**b**) The results and information from multiple/systematic prediction for example mCSM-PPI are shown in a table

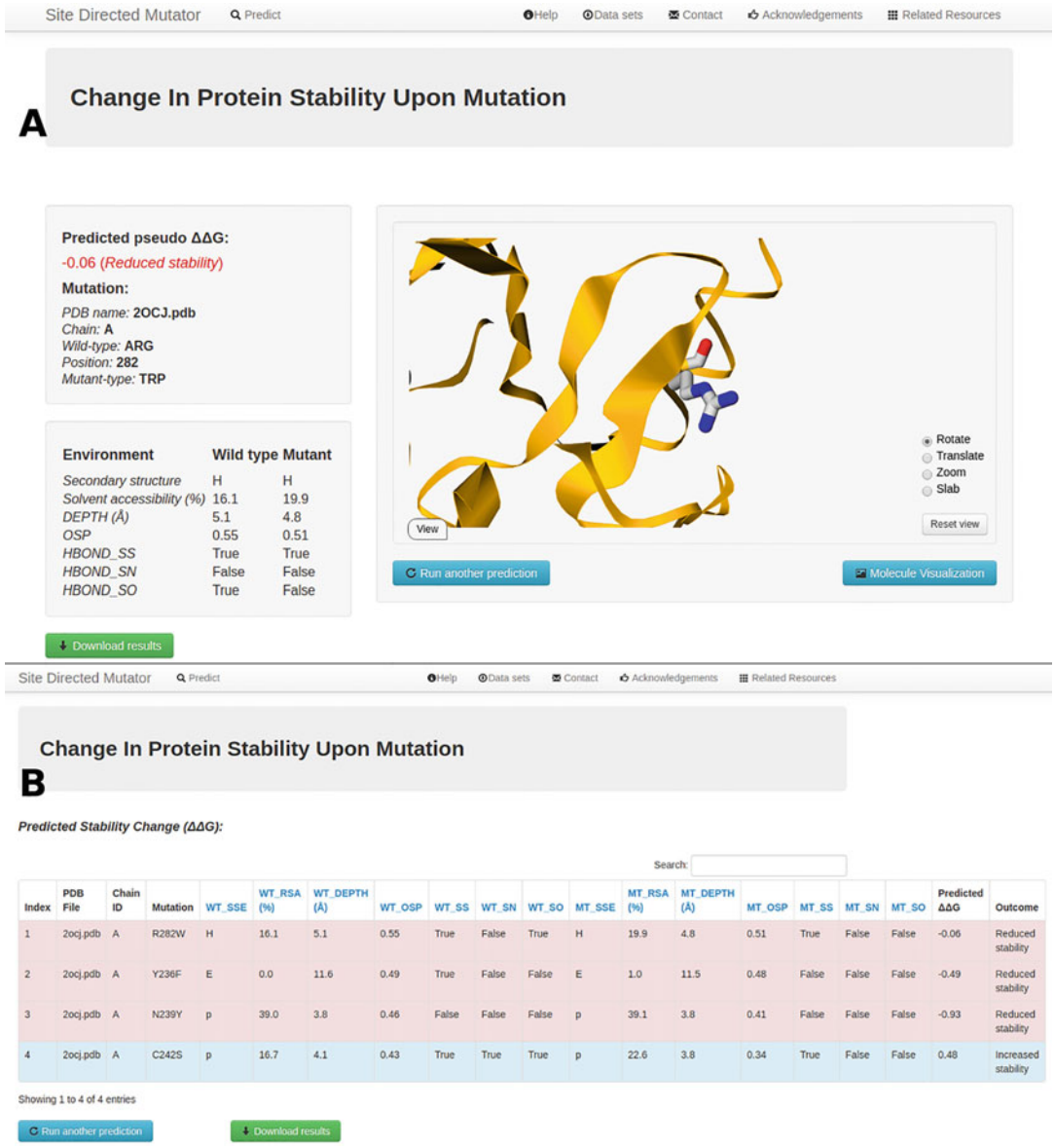


Fig. 6 SDM prediction results for single and list prediction. (**a**) The single prediction displays the predicted ΔΔG with information used on the left panel and 3D structure in a ribbon (protein) and a stick (wild-type amino acid) representation. (**b**) The list prediction gives detailed structural information and predicted ΔΔG in a tabulated form highlighted according to stabilizing (blue) and destabilizing (red) mutation

structure and its wild-type amino acids in ribbon and stick representation.

- Stability changes (ΔΔG) are shown in red with a negative sign if the mutation is predicted to be destabilizing, and in blue with a positive sign if the mutation is predicted to be stabilizing.

Multiple Mutations

- The predicted SDM ΔΔG for a given mutation list is displayed in a tabulated format (Fig. 6b) with their structural features. Users can download all mutant PDB structures and their predicted values in individual files.

3.2.5 DUET

Single Mutation

- The DUET result page (Fig. 7a) provides the predicted stability changes (ΔΔG) with integrated features such as secondary structure and stability changes from mCSM and SDM. While DUET refers to both mCSM and SDM scores, the prediction result can vary between the two methods.
- In the structure viewer (Fig. 7a right), the wild-type amino acid is shown in stick form and users can download the corresponding mutant structure file in PDB format.

Systematic Mutations

- With the systematic prediction (Fig. 7b), users can examine the predicted changes in protein stability using DUET, mCSM, and SDM for all nineteen possible mutations at a given residue position.
- The predictions and the structural information used to calculate the DUET scores are displayed in a downloadable table.

3.2.6 DynaMut

Users can use DynaMut to assess the impact of mutations on protein dynamics and stability with single and list mutation prediction.

Single Mutation

- The results of mutational effects on protein dynamics and stability are shown in Fig. 8a: ΔΔG predictions, interatomic interactions, deformation and fluctuation analysis.
- The ΔΔG prediction page provides predicted values from normal mode analysis (NMA)-based prediction (ΔΔG ENCoM), vibrational entropy energy changes ($\Delta\Delta S_{Vib}$ ENCoM), and other structure-based stability predictions (ΔΔG mCSM, ΔΔG SDM, ΔΔG DUET). Users can visually assess mutational effects on protein flexibility which is colored on the protein structure by vibrational entropy (Fig. 8b) for the region gaining (red) or losing (blue) flexibility. This 3D representation can be downloaded into a Pymol session, high resolution image and CSV file.
- Through the interatomic interactions tab, users can compare molecular interactions between wild-type and mutant structures. The PDB structure with interatomic interactions can be retrieved as a Pymol session file.
- The mutational effects on protein dynamics are shown in the deformation and fluctuation tab. Users can evaluate changes in the amount of local flexibility and atomic fluctuation upon mutation in 3D visual representation; results are downloadable as a CSV file and a Pymol session file.

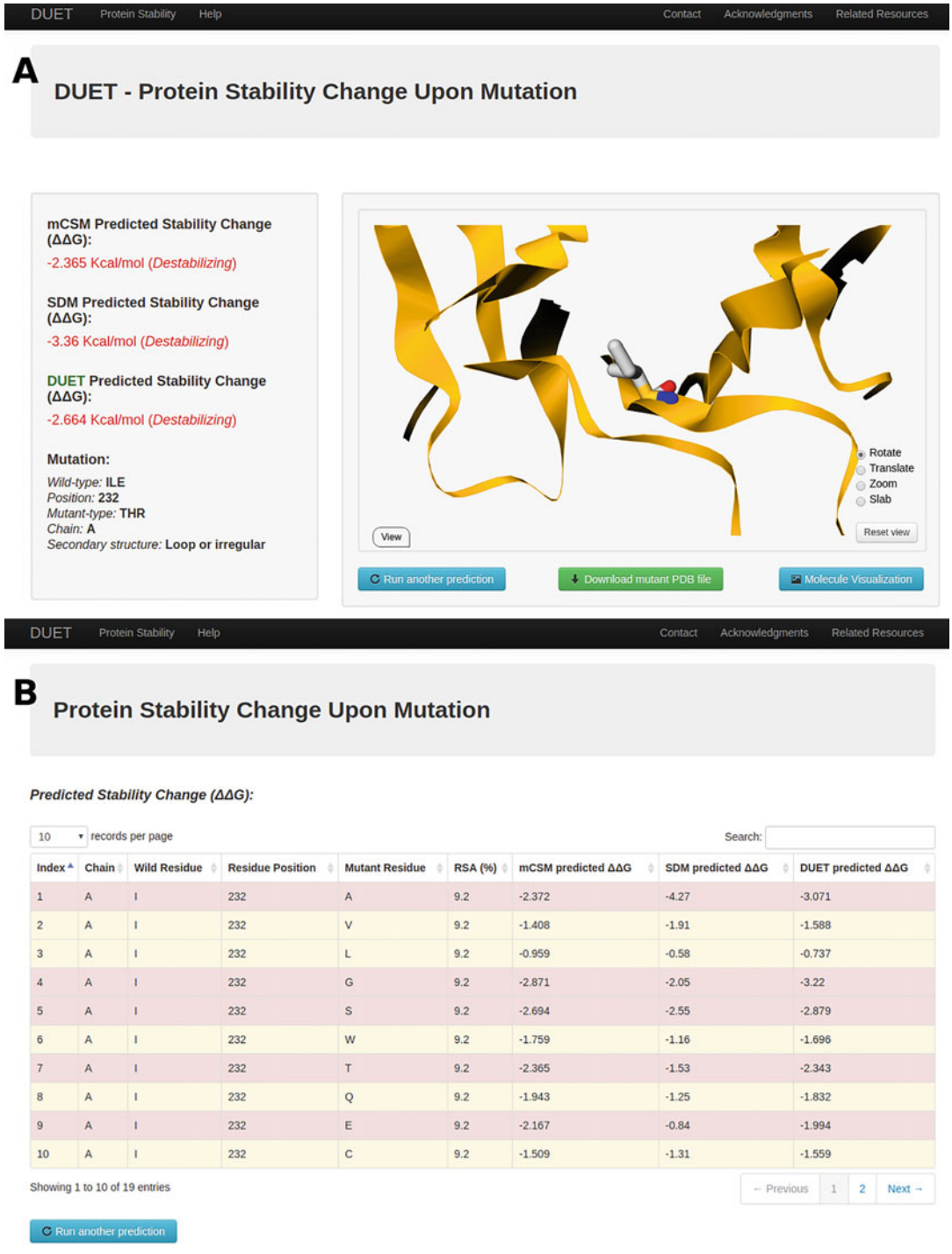


Fig. 7 DUET result pages for single and systematic prediction. (**a**) The single prediction result of DUET shows predicted ΔΔG across SDM and mCSM-Stability with mutation details. (**b**) Systematic prediction results including ΔΔG from DUET, SDM and mCSM-Stability and relative solvent accessible area of wild-type structure

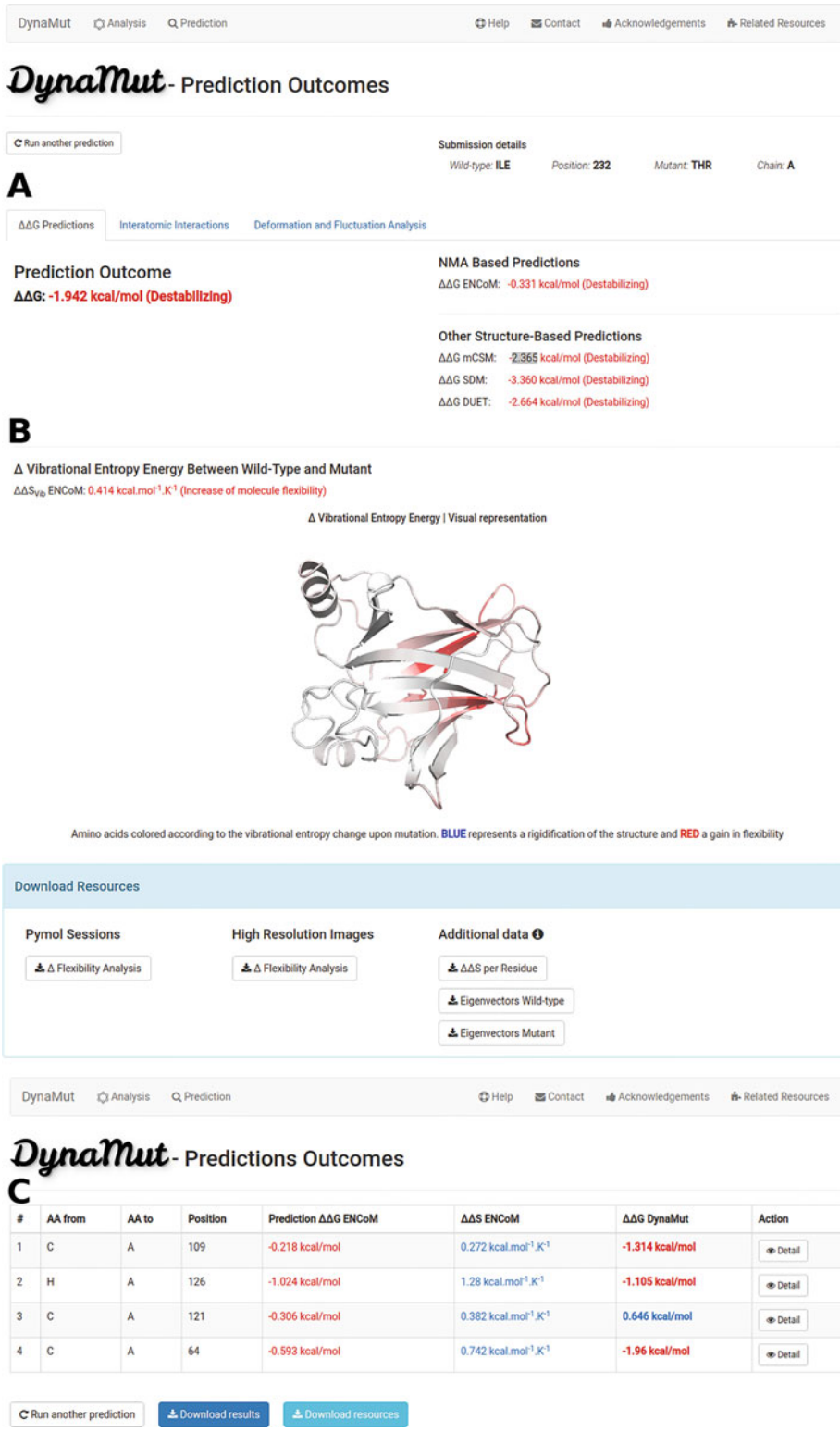


Fig. 8 DynaMut result pages. The single prediction shows predicted DynaMut ΔΔG (**a**, left) and predicted protein stability (ΔΔG) from mCSM-Stability, SDM and DUET and flexibility changes (ΔΔG ENCoM). Users can check vibrational energy changes upon mutation in the panel B. For a multiple mutation, (**b**) list prediction result page shows predicted DynaMut ΔΔG and links to access the corresponding single prediction in table

Multiple Mutations

- For a given mutation list, DynaMut gives all predicted values, including $\Delta\Delta G_{Stability}$ENCoM, $\Delta\Delta S_{Vib}$ ENCoM, and $\Delta\Delta G_{Stability}$ DynaMut, in table format (Fig. 8c). A more detailed analysis is available through the single prediction page of each mutation by clicking on the "Detail" button.

3.2.7 *mCSM-PPI2*

mCSM-PPI2 supports two types of protein–protein affinity prediction: mutation prediction and binding analysis. Mutation prediction gives predicted protein–protein affinity changes based on a given protein–protein complex and the mutation information. Binding analysis considers interface residues within 5 Å from different chains in the complex structure for alanine scanning and saturation mutagenesis.

Single Mutation

- mCSM-PPI2 displays predicted binding affinity changes (ΔΔG) upon mutation in two classes, destabilizing and stabilizing. Mutation details such as the distance to the interface from the given mutation position are also shown (Fig. 9).
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- Users can assess the mutational impact in atomic/residue level through a 3D interactive viewer and a 2D graph. The molecular viewer provides Arpeggio inter/intra interactions for wild-type and mutant structures and the interaction changes between wild-type and mutant allows for investigation of the relationship between nonbonded interaction and protein–protein affinity. For residue-level analysis, the 2D graph can be used to study interresidue interactions of wild-type and mutant in a simple and user-friendly representation.

List Mutation

- For multiple mutation analysis, the result page tabulates predicted ΔΔG with mutation details. Users can access detailed results of each mutation through the single mutation result page and download all entries as a CSV file.

Alanine Scanning

- To identify residues with a greater contribution to the energy of binding (hot-spot) at the interface of interaction, alanine scanning can be used by predicting protein–protein binding affinity changes upon mutations to alanine across all identified interface residues. The predicted ΔΔG values are displayed in table, bar chart, and 3D viewer (Fig. 10a).
- Users can assess the effects of alanine mutation on the interface residues through a bar graph and 3D viewer colored in red and blue for destabilizing and stabilizing mutations, respectively.



Fig. 9 mCSM-PPI2 single prediction result page. The predicted ΔΔG is shown along with two interaction viewers: 3D interactive molecule viewer for atomic interaction analysis and 2D diagram for residue-level interaction analysis



Fig. 10 mCSM-PPI2 interface scanning result pages. The result pages of (**a**) alanine scanning and (**b**) saturation mutagenesis provide a bar chart and a heatmap colored by predicted ΔΔG and average predicted ΔΔG from the nineteen possible mutations, respectively

Saturation Mutagenesis

- The saturation mutagenesis provides the most exhaustive prediction, showing predicted ΔΔG for all identified interface residues when they are changed into nineteen different amino acids. The results are shown in table, heatmap, and 3D molecule viewer, and the interface residues of the 3D viewer are colored by the average ΔΔG of all mutations for each residue.

3.2.8 mCSM-NA

Single Prediction

- The predicted protein–nucleic acid affinity changes on a given structure are shown (Fig. 11a) with other properties such as the type of nucleic acid, solvent accessibility of wild-type protein, and predicted mutational effects from mCSM-Stability.
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- The molecule visualization panel shows the protein–nucleic acid complex with the wild-type amino acid, and the mutation as a stick representation. mCSM-NA allows users to further investigate inter/intraresidue interactions by downloading Pymol session file.

List Mutation

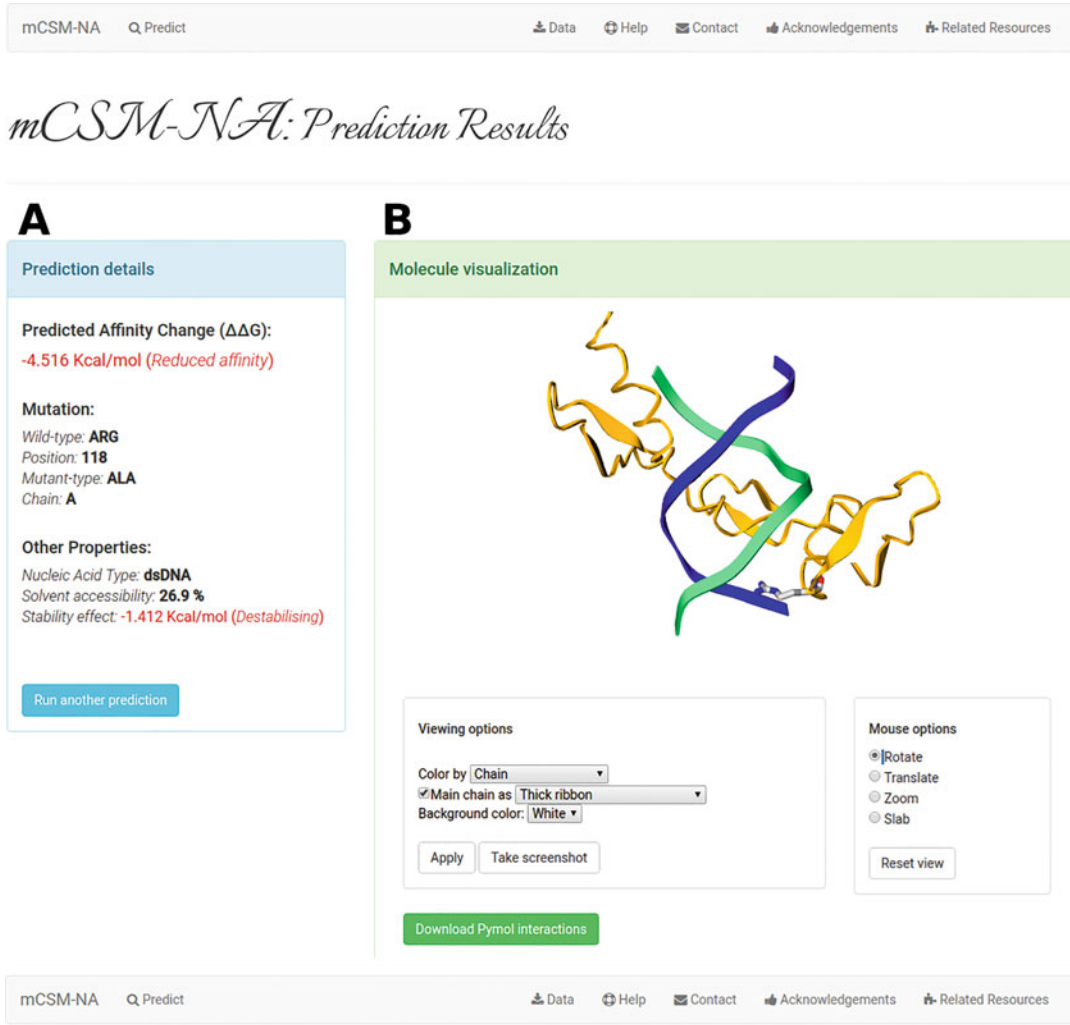
- mCSN-NA provides predicted protein–nucleic acid affinity changes, wild-type RSA, and mutation information for a given list of mutations in a table which is also downloadable in TSV format.

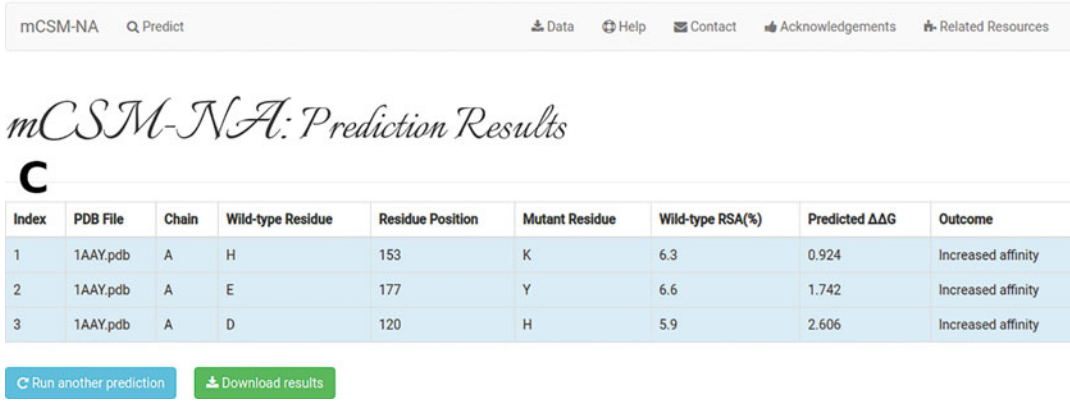
3.2.9 mCSM-lig

- mCSM-lig predicts affinity changes (log affinity fold) between a protein and its ligand upon mutation (Fig. 12a) using additional information such as the closest distance between wild-type residue and ligand and the protein stability change (Kcal/mol) from DUET. The stabilizing and destabilizing mutations are shown in positive and negative values respectively.
- For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
- The wild-type amino acid and ligand are shown in stick and sphere representations in 3D molecule viewer, respectively.

3.3 Identification of Driving Molecular Consequences

The outputs of the predictive tools described above provide the basis for an initial heuristic examination. When trying to interpret the molecular consequences of a specific variant, it is important to remember that phenotypic outcomes are often the result of the





Index	PDB File	Chain	Wild-type Residue	Residue Position	Mutant Residue	Wild-type RSA(%)	Predicted ΔΔG	Outcome
1	1AAY.pdb	A	H	153	K	6.3	0.924	Increased affinity
2	1AAY.pdb	A	E	177	Y	6.6	1.742	Increased affinity
3	1AAY.pdb	A	D	120	H	5.9	2.606	Increased affinity

Fig. 11 mCSM-NA result pages for single and list mutation prediction. In the single prediction result page, predicted protein–DNA affinity changes and mutation information are displayed in the prediction details (**a**) and the 3D viewer shows protein–DNA complex and wild-type amino acid in a ribbon and stick representation (**b**). The results of list prediction are shown in a tabulated form (**c**) and users can save the results in a TSV format

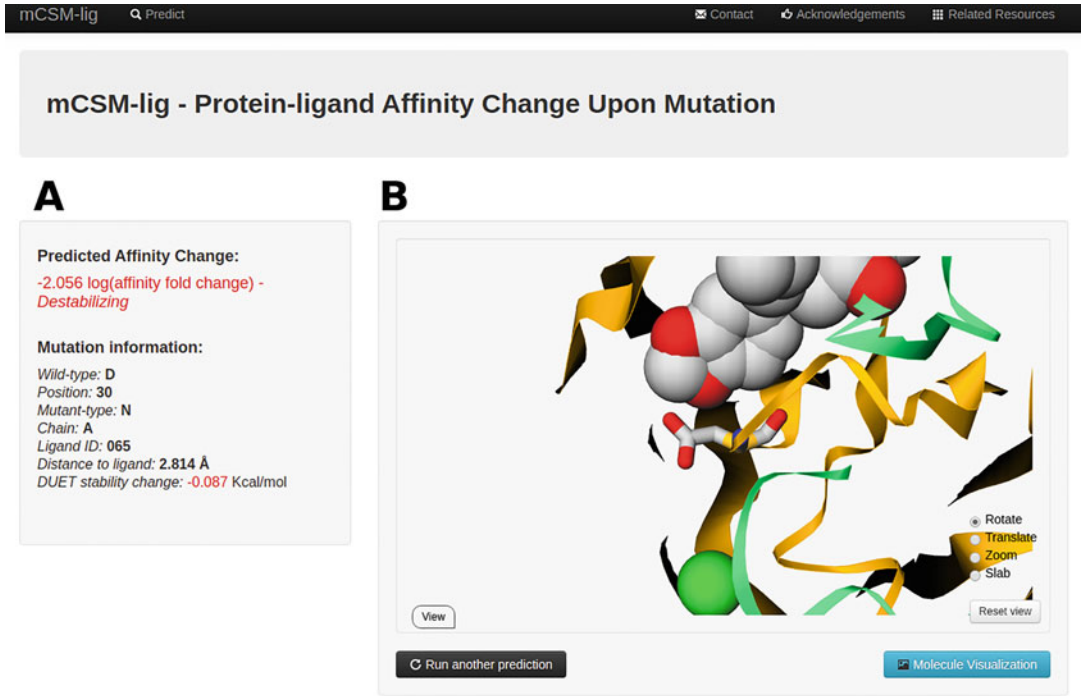


Fig. 12 mCSM-lig result page. (**a**) The predicted affinity change between protein and ligand upon mutation is shown in logarithm scale. (**b**) The protein and ligand are displayed in 3D viewer with a ribbon (for protein), a stick (for wild-type amino acid), and a sphere (for ligand) representation

combination of multiple molecular changes. For coding mutations, we initially ask ourselves three questions:

1. Is the mutation within 5 Å of an interface? If so, is the mutation more likely to disrupt the interaction ($\Delta\Delta G < \pm 0.5$ kcal/mol) based on the corresponding mCSM output (e.g., mCSM-PPI, mCSM-DNA, mCSM-NA, mCSM-Lig)? If the mutation is further than 12 Å away, it is less likely to disrupt the interaction directly, so the mCSM predictions are less reliable.
2. Is the mutation likely to disrupt protein folding and stability? mCSM-Stability, SDM, DUET, and DynaMut provide insight into this, with mutations leading to $\Delta\Delta G < \pm 0.5$ kcal/mol more likely to have a significant biological effect. Mutations at buried residues are more likely to have a larger effect on protein stability.
3. Is the mutation a special case that is more likely to lead to disruption of the protein due to unique geometry restraints of the residues (*see* **Notes 3** and **4**)?

To more exhaustively explore how mutations in a protein lead to a phenotype, and to identify those molecular features that best

capture the driving of the molecular mechanisms, an investigation into the performance of each inputted feature should be conducted in order to construct the highest performing predictive model.

A more robust method for selecting which features are most informative can be performed using feature selection in R, a statistical programming language. While R is powerful enough itself to create classification models, we can also use it to measure which features from our predictive tools' output are most effective in stratifying mutations. Two effective approaches are:

1. A random forest classification algorithm to measure feature importance using a set of mutations with known class labels (e.g., pathogenic/nonpathogenic, deleterious/nondeleterious).
2. The Boruta Algorithm performs permutations of the data to statistically compare each feature's importance with that attainable at random, and uses this to eliminate uninformative features. The package in R provides a graphical output using boxplots.

Features that score highly provide evidence that the molecular consequence that they measure is relevant to how mutations lead to the phenotype of interest. The algorithm can also highlight correlation between features. When two or more features are highly correlated and are likely measuring the same information, only one should be used in subsequent predictive model development to remove redundancy, minimize noise and avoid bias from weighting a model in favor of a particular attribute. The model should also have the fewest possible features that perform best. Using too many features may generate a model that performs accurately on training data but cannot be generalized to real-world data.

3.4 Machine Learning Phenotypes: Building a Predictive Classifier

An initial understanding of molecular mechanisms imparted by disease-causing mutations is a crucial step toward establishing a genotype–phenotype correlation. However, manual analysis of different results can often miss underlying, statistically significant relationships among different mutational measurements, which can help relate them to the phenotype. Machine learning, and in particular supervised learning, addresses this issue by providing a set of tools for the efficient analysis of labeled data (e.g., experimentally characterized mutations) in order to derive a model that describes a phenomenon, aiming for generalization (applying it to unseen data). The identification of patterns and associations within the data will further help the predictive model establish a distinction between mutations within the same gene leading to different phenotypes, and hence the development of an effective predictive tool that can be used to interpret novel clinical variants.

Here, our goal is to build a machine learning classifier to distinguish between pathogenic vs. nonpathogenic mutations in a

given gene. Multiple steps are required to obtain a nonbiased, accurate predictor:

1. Dataset curation: Machine learning algorithms require a well-curated dataset. In a supervised machine learning approach, all data labels (here, pathogenic or nonpathogenic for each mutation) must be known in order to enable correlations to be assessed between labels (e.g., phenotypes) and features/properties used as evidence to represent each data point (e.g., mutations). The quality of a classifier directly depends on the quality of the data used to build it, so accurate clinical sources are required to justify labeling mutations as pathogenic or nonpathogenic. In this case, generally, nonpathogenic variants can be curated from population variant databases such as GnomAD, usually taking into account frequent mutations. Even common variants, however, may still be linked to a disease, especially if it is a weakly penetrative mutation or recessive condition, which would add noise to the data set and thus complicate the task of building a general predictive model. In situations where other biologically relevant information is present, such as cellular fitness cost, it is essential that this type of information is present for every mutation in a dataset, as a supervised algorithm cannot handle missing data labels. The initial dataset should contain a representative set of mutations within all phenotype classes (pathogenic and nonpathogenic), and ideally, present a balanced number of instances between classes, to minimize biases toward overrepresented classes in the resultant model. More details on metrics used to evaluate the performance of predictive models on an imbalanced dataset are discussed below.
2. Feature generation: The feature generation stage is crucial as it provides descriptive information about each mutation, to be used by the learning algorithm to finally classify the phenotype of a mutation. As described above, features can encompass a diverse range of mutational information:
   (a) Protein stability and dynamics (mCSM-Stability, DUET, SDM, Dynamut).
   (b) Protein functional changes such as changes in affinity for other proteins (mCSM-PPI2), nucleic acids (mCSM-NA), and ligands (mCSM-lig).
   (c) At the residue level, changes in protein pharmacophore and local residue environment such as changes in interatomic interactions (Arpeggio) are also important, as some mutations at the same locus can have different phenotypes.
   (d) Sequence-level predictors (SIFT, Polyphen, SNAP2).

(e) Evolutionary-based predictors (ConSurf), population based mutational tolerance (MTR-Viewer), as well as amino acid substitution matrices (e.g., PAM30, BLOSUM62, PSSM) offer added information on the likelihood of one mutation to change into another.

Feature generation is directly dependent on the wild-type biological functions of the protein, which is why an understanding of the biological relevance is important at the very beginning of this process.

3. Training and Testing sets: The data collected must be divided into training and testing sets to assess the generalization power of a classifier, that is, its ability to correctly predict on new data, and to ensure that it has not been over- or undertrained. Data used to train the model should be different, nonredundant, from the data used to test the model. It is common practice to divide the original dataset into Training and Test sets at the start of learning. For small datasets, a large proportion of the data may need to be segregated into the Test set to provide sufficient data to accurately measure performance of the trained model. This can be done in a bootstrapping procedure or through cross-validation, when the original data set is divided into *k*-folds and each is taken iteratively as the test set while remaining data are used in training (*k*-fold cross-validation).
4. Feature selection: The features selected for training can strongly influence accuracy, so it is important to select only informative features, and eliminate irrelevant or nondiscriminative ones, which are a common source of noise. Feature selection can also help reduce overfitting and reduce training time, as it aims to generate simpler, more concise models. Feature selection methods provided in the Python machine learning library, Scikit-Learn [58], include univariate selection, feature importance, correlation matrix, and recursive feature elimination or addition. Alternatively, forward stepwise selection can be performed as a greedy heuristic in which features are included iteratively, one at a time, based on their individual performance contributions.
5. Machine learning platforms: Different tools have been developed for implementing machine learning. Some offer a graphical user interface (GUI), such as Weka [59], while some run as python packages through the command line, such as Scikit-Learn. Different packages for different programming languages offer similar algorithms and options to adjust the algorithm parameters according to specific tasks. The major classification algorithms we test are Naive Bayes, Decision Trees, K-Nearest Neighbor, Support Vector Machines, and Ensemble Classifiers. It is good practice to compare

representative algorithms of each class, provided that the algorithm is compatible with the dataset type. Within weka, this can be done automatically using the auto-weka function. In cases where the training set is unbalanced, oversampling or undersampling of the training data can be used to achieve a better representation of classes within the classification model-building stage, preventing model bias in always detecting the predominant class and achieving a false high performance.

6. Model validation: The primary tool in the validation of a model is the use of a nonredundant independent test set, also called blind test.

   Validation can be furthered using internal data testing such as k-fold cross validation, in which the dataset is divided into k subsets. One subset is used as a test set, while the remaining $(k - 1)$ subsets are used to train a model. The process is repeated k times, until all the data have been used in both training and test sets. The final model performance is calculated as the average of the performances of all k iterations. We will often vary k based on the size of the dataset. When the training set is small (e.g., ~200 data points), we may use leave-one-out validation, where k is equal to the size of the dataset. An important aspect when selecting predictive models is consistency in performance between the training and test sets. This usually indicates a robust model, within which discrepancies (e.g., a much higher performance on training than with the test set) might indicate overfitting.

7. Model evaluation: Several different evaluation metrics may be used for classification tasks. These are generally calculated on values obtained from a confusion matrix, which is a summary of the data points, and their actual and predicted phenotypes (Table 6).

8. From the distributions of data points within the matrix, descriptive metrics can be calculated:

   (a) accuracy (number of correct predictions: [(TP + TN)/TOTAL]),

   (b) precision (rate of correctly predicted positive instances from all assigned as positives: [TP/(TP + FP)],

Table 6
Description of a confusion matrix

Predicted value	Actual value Positive	Negative
Positive	True positive	False positive
Negative	False negative	True negative

(c) recall (rate of correctly predicted positive instances from all real positive instances: [TP/(TP + FN)],

(d) *f*-score (a weighted average of recall and precision), and,

(e) Matthews correlation coefficient (MCC) a balanced measure between true positives and true negatives

$$\left[(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN}) / \sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}\right]$$

where TP = True positive; TN = True negative; FP = False positive; and FN = False negative.

Classifier performance can also be described graphically using a Receiver Operating Characteristic curve, which compares the TP Rate and TN Rate. The closer the area under the curve is to 1, the better the classifier performance.

These metrics should be used in a combinatorial fashion across all elements of training, test, and cross-validation stages to compare model performance during different stages of classifier optimization. When the dataset is imbalanced, balanced measures such as MCC should be prioritized, as other measures might bias for an overtrained model on the dominant dataset.

4 Notes

1. Often following curation, the distribution of number of pathogenic and benign mutations is unbalanced, which can affect efforts to build predictive tools using machine learning. Two approaches that can help include oversampling of the underrepresented class, or undersampling of the overrepresented class. Evaluation metrics that are less biased toward unbalanced classes, such as the Matthew's correlation coefficient, precision-recall curves, and Kendall correlations, should also be preferentially used.
2. The chain ID for the provided PDB file is a mandatory field for all the structure-based methods; blank characters are not allowed. It is possible that homology modeling tools might not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. Several tools exist to perform this task (e.g., http://www.canoz.com/sdh/renamepdbchain.pl).
3. Special cases: Mutations to and from prolines. Prolines are the only amino acid whose amino group is connected to the side-chain, which in the context of the peptide bond greatly limits torsional angles. The nature of this residue therefore needs to be taken into account while analyzing mutation effects. For instance, (1) mutations to prolines in the middle of alpha-

helices can introduce kinks, affecting local structure and (2) since prolines are commonly found in turns and loops, their substitution might interfere with the formation of secondary structures such as hairpins.

4. Special cases: mutations of positive-phi glycines. Similarly to prolines, positive phi glycines, while rare in experimental structures, deserve special consideration due to their torsional angles. Glycines are the only residues capable of adopting positive-phi angles. These glycines are usually conserved across evolution, meaning that mutations on positive-phi glycines, especially on loops and hairpins, tend to be destabilizing.

Acknowledgments

This work was supported by Australian Government Research Training Program Scholarships [to S.P., M.K., Y.M., C.H.M.R.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; and the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.].

References

1. Jatana N, Ascher DB, Pires DEV et al (2019) Human LC3 and GABARAP subfamily members achieve functional specificity via specific structural modulations. Autophagy:1–17. https://doi.org/10.1080/15548627.2019.1606636
2. Abayakoon P, Jin Y, Lingford JP et al (2018) Structural and biochemical insights into the function and evolution of sulfoquinovosidases. ACS Cent Sci 4(9):1266–1273. https://doi.org/10.1021/acscentsci.8b00453
3. Ascher DB, Cromer BA, Morton CJ et al (2011) Regulation of insulin-regulated membrane aminopeptidase activity by its C-terminal domain. Biochemistry 50(13):2611–2622. https://doi.org/10.1021/bi101893w
4. Portelli S, Phelan JE, Ascher DB et al (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Sci Rep 8(1):15356. https://doi.org/10.1038/s41598-018-33370-6
5. Silk M, Petrovski S, Ascher DB (2019) MTR-Viewer: identifying regions within genes under purifying selection. Nucleic Acids Res 47(W1):W121–W126. https://doi.org/10.1093/nar/gkz457
6. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43(Database issue):D387–D391. https://doi.org/10.1093/nar/gku966
7. Lucy G, Douglas EVP, Álvaro O-N et al (2014) An integrated computational approach can classify VHL missense mutations according to risk of clear cell renal carcinoma. Human Molecular Genetics, 23(22):5976–5988. https://doi.org/10.1093/hmg/ddu321
8. Blaszczyk M, Harmer NJ, Chirgadze DY et al (2015) Achieving high signal-to-noise in cell regulatory systems: spatial organization of multiprotein transmembrane assemblies of FGFR and MET receptors. Prog Biophys Mol Biol 118(3):103–111. https://doi.org/10.1016/j.pbiomolbio.2015.04.007
9. Jafri M, Wake NC, Ascher DB et al (2015) Germline mutations in the CDKN2B tumor suppressor gene predispose to renal cell carcinoma. Cancer Discov 5(7):723–729. https://doi.org/10.1158/2159-8290.CD-14-1096

10. Pacitto A, Ascher DB, Wong LH et al (2015) Lst4, the yeast Fnip1/2 orthologue, is a DENN-family protein. Open Biol 5 (12):150174. https://doi.org/10.1098/rsob.150174
11. Pires DE, Chen J, Blundell TL et al (2016) In silico functional dissection of saturation mutagenesis: interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. https://doi.org/10.1038/srep19848
12. Albanaz ATS, Rodrigues CHM, Pires DEV et al (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12(6):553–563. https://doi.org/10.1080/17460441.2017.1322579
13. Casey RT, Ascher DB, Rattenberry E et al (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5(3):237–250. https://doi.org/10.1002/mgg3.279
14. Jubb HC, Pandurangan AP, Turner MA et al (2017) Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002
15. Pandurangan AP, Ascher DB, Thomas SE et al (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45(2):303–311. https://doi.org/10.1042/BST20160422
16. Sibanda BL, Chirgadze DY, Ascher DB et al (2017) DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science 355 (6324):520–524. https://doi.org/10.1126/science.aak9654
17. Rodrigues CH, Ascher DB, Pires DE (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res 46(W1): W127–W132. https://doi.org/10.1093/nar/gky375
18. Hnizda A, Fabry M, Moriyama T et al (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia 32 (6):1393–1403. https://doi.org/10.1038/s41375-018-0073-5
19. Andrews KA, Ascher DB, Pires DEV et al (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55 (6):384–394. https://doi.org/10.1136/jmedgenet-2017-105127
20. Usher JL, Ascher DB, Pires DE et al (2015) Analysis of HGD gene mutations in patients with alkaptonuria from the United Kingdom: identification of novel mutations. JIMD Rep 24:3–11. https://doi.org/10.1007/8904_2014_380
21. Nemethova M, Radvanszky J, Kadasi L et al (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24(1):66–72. https://doi.org/10.1038/ejhg.2015.60
22. Ramdzan YM, Trubetskov MM, Ormsby AR et al (2017) Huntingtin inclusions trigger cellular quiescence, deactivate apoptosis, and lead to delayed necrosis. Cell Rep 19(5):919–927. https://doi.org/10.1016/j.celrep.2017.04.029
23. Traynelis J, Silk M, Wang Q et al (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27 (10):1715–1729. https://doi.org/10.1101/gr.226589.117
24. Trezza A, Bernini A, Langella A et al (2017) A computational approach from gene to structure analysis of the human ABCA4 transporter involved in genetic retinal diseases. Invest Ophthalmol Vis Sci 58(12):5320–5328. https://doi.org/10.1167/iovs.17-22158
25. Ascher DB, Spiga O, Sekelska M et al (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. Eur J Hum Genet 27(6):888–902. https://doi.org/10.1038/s41431-019-0354-0
26. Soardi FC, Machado-Silva A, Linhares ND et al (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med 2:7. https://doi.org/10.1038/s41525-017-0009-4
27. Phelan J, Coll F, McNerney R et al (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14:31. https://doi.org/10.1186/s12916-016-0575-9
28. Silvino AC, Costa GL, Araujo FC et al (2016) Variation in human cytochrome P-450 drug-metabolism genes: a gateway to the understanding of Plasmodium vivax relapses. PLoS One 11(7):e0160172. https://doi.org/10.1371/journal.pone.0160172

29. White RR, Ponsford AH, Weekes MP et al (2016) Ubiquitin-dependent modification of skeletal muscle by the parasitic nematode, Trichinella spiralis. PLoS Pathog 12(11): e1005977. https://doi.org/10.1371/journal.ppat.1005977
30. Hawkey J, Ascher DB, Judd LM et al (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom 4(3). https://doi.org/10.1099/mgen.0.000165
31. Holt KE, McAdam P, Thai PVK et al (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet 50(6):849–856. https://doi.org/10.1038/s41588-018-0117-9
32. Karmakar M, Globan M, Fyfe JAM et al (2018) Analysis of a Novel pncA mutation for susceptibility to pyrazinamide therapy. Am J Respir Crit Care Med 198(4):541–544. https://doi.org/10.1164/rccm.201712-2572LE
33. Vedithi SC, Malhotra S, Das M et al (2018) Structural implications of mutations conferring rifampin resistance in Mycobacterium leprae. Sci Rep 8(1):5016. https://doi.org/10.1038/s41598-018-23423-1
34. Karmakar M, Rodrigues CHM, Holt KE et al (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS One 14(5):e0217169. https://doi.org/10.1371/journal.pone.0217169
35. Ascher DB, Wielens J, Nero TL et al (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. https://doi.org/10.1038/srep04765
36. Jubb HC, Higueruelo AP, Ochoa-Montano B et al (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. J Mol Biol 429(3):365–371. https://doi.org/10.1016/j.jmb.2016.12.004
37. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. https://doi.org/10.1093/bioinformatics/btt691
38. Pandurangan AP, Ochoa-Montano B, Ascher DB et al (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45(W1):W229–W235. https://doi.org/10.1093/nar/gkx439
39. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42(Web Server issue):W314–W319. https://doi.org/10.1093/nar/gku411
40. Douglas EVP, Carlos HMR, David BA et al (2020) mCSM-membrane: predicting the effects of mutations on transmembrane proteins, Nucleic Acids Research, gkaa416. https://doi.org/10.1093/nar/gkaa416
41. Rodrigues CH, Pires DE, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46(W1):W350–W355. https://doi.org/10.1093/nar/gky300
42. Rodrigues CHM, Myung Y, Pires DEV et al (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. Nucleic Acids Res 47(W1):W338–W344. https://doi.org/10.1093/nar/gkz383
43. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res 44(W1):W469–W473. https://doi.org/10.1093/nar/gkw458
44. Yoochan M, Carlos HMR, David BA, Douglas EVP et al (2020) mCSM-AB2: guiding rational antibody design using graphbased signatures, Bioinformatics. 36(5):1453–1459. https://doi.org/10.1093/bioinformatics/btz779
45. Yoochan M, Douglas EVP, David BA et al. mmCSM-AB: guiding rational antibody engineering through multiple point mutations, Nucleic Acids Research, gkaa389. https://doi.org/10.1093/nar/gkaa389
46. Pires DEV, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45(W1):W241–W246. https://doi.org/10.1093/nar/gkx236
47. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. https://doi.org/10.1038/srep29575
48. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44(W1):W557–W561. https://doi.org/10.1093/nar/gkw390
49. Douglas EVP et al (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. BMC genomics (12) No. S4. BioMed Central
50. Douglas EVP, Raquel CM-M, Carlos HS, Frederico FC, Wagner M Jr (2013) aCSM: noise-free graphbased signatures to large-scale receptor-based ligand prediction, Bioinformatics 29(7):855–861. https://doi.org/10.1093/bioinformatics/btt058

51. Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311. https://doi.org/10.1093/nar/29.1.308
52. Stenson PD, Mort M, Ball EV et al (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136(6):665–677. https://doi.org/10.1007/s00439-017-1779-6
53. Landrum MJ, Lee JM, Benson M et al (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46(D1):D1062–D1067. https://doi.org/10.1093/nar/gkx1153
54. Karczewski KJ, Francioli LC, Tiao G et al (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv:531210. https://doi.org/10.1101/531210
55. Sudlow C, Gallacher J, Allen N et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3): e1001779. https://doi.org/10.1371/journal.pmed.1001779
56. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res 46(5):2699. https://doi.org/10.1093/nar/gky092
57. Rose PW, Prlic A, Altunkaya A et al (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res 45(D1):D271–D281. https://doi.org/10.1093/nar/gkw1000
58. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
59. Witten IH, Frank E, Hall MA et al (2016) Data mining, fourth edition: practical machine learning tools and techniques. Morgan Kaufmann, Burlington

APPENDIX 6: QUANTIFYING THE RATES OF LATE REACTIVATION TUBERCULOSIS: A SYSTEMATIC REVIEW

Quantifying the rates of late reactivation tuberculosis: a systematic review

Katie D Dale, Malancha Karmakar, Kathryn J Snow, Dick Menzies, James M Trauer, Justin T Denholm**

The risk of tuberculosis is greatest soon after infection, but *Mycobacterium tuberculosis* can remain in the body latently, and individuals can develop disease in the future, sometimes years later. However, there is uncertainty about how often reactivation of latent tuberculosis infection (LTBI) occurs. We searched eight databases (inception to June 25, 2019) to identify studies that quantified tuberculosis reactivation rates occurring more than 2 years after infection (late reactivation), with a focus on identifying untreated study cohorts with defined timing of LTBI acquisition (PROSPERO registered: CRD42017070594). We included 110 studies, divided into four methodological groups. Group 1 included studies that documented late reactivation rates from conversion (n=14) and group 2 documented late reactivation rates in LTBI cohorts from exposure (n=11). Group 3 included 86 studies in LTBI cohorts with an unknown exposure history, and group 4 included seven ecological studies. Since antibiotics have been used to treat tuberculosis, only 11 studies have documented late reactivation rates in infected, untreated cohorts from either conversion (group 1) or exposure (group 2); six of these studies lasted at least 4 years and none lasted longer than 10 years. These studies found that tuberculosis rates declined over time, reaching approximately 200 cases per 100 000 person-years or less by the fifth year, and possibly declining further after 5 years but interpretation was limited by decreasing or unspecified cohort sizes. In cohorts with latent tuberculosis and an unknown exposure history (group 3), tuberculosis rates were generally lower than those seen in groups 1 and 2, and beyond 10 years after screening, rates had declined to less than 100 per 100 000 person-years. Reinfection risks limit interpretation in all studies and the effect of age is unclear. Late reactivation rates are commonly estimated or modelled to prioritise tuberculosis control strategies towards tubuculosis elimination, but significant gaps remain in our understanding that must be acknowledged; the relative importance of late reactivation versus early progression to the global burden of tuberculosis remains unknown.

Lancet Infect Dis 2021

Published Online
April 20, 2021
https://doi.org/10.1016/S1473-3099(20)30728-3

*Co-senior authors

Victorian Tuberculosis Program, Royal Melbourne Hospital (K D Dale MPH, J M Trauer PhD, J T Denhom PhD) **and Department of Microbiology and Immunology** (K D Dale, M Karmakar MSc, J T Denhom), **Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, VIC , Australia; Baker Heart and Diabetes Institute, Melbourne, VIC, Australia** (M Karmakar); **Centre for International Child Health, Department of Paediatrics, Royal Children's Hospital** (K J Snow PhD) **and Australia Department of Paediatrics** (K J Snow), **University of Melbourne, Parkville, VIC, Australia; Respiratory Epidemiology and Clinical Research Unit, McGill International TB Centre, Montreal, QC, Canada** (D Menzies MD); **School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC , Australia** (J M Trauer)

Correspondence to:
Katie Dale, Victorian Tuberculosis Program, The Royal Melbourne Hospital, The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC 3000, Australia
katie.dale@mh.org.au

Introduction

Tuberculosis is a major global health problem,[1] and yet the natural history of tuberculosis remains poorly understood.[2] The risk of developing tuberculosis is known to be greatest in the first few months following infection with *Mycobacterium tuberculosis*,[3] and many studies have presented empirical data to quantify the high rates of disease in this period.[4–6] However, *M tuberculosis* can persist in a latent state, referred to as latent tuberculosis infection (LTBI), with individuals remaining at risk of developing tuberculosis more than 2 years after infection (late reactivation).[7] Quantifying late reactivation risk is difficult for a variety of reasons, including the need for extended follow-up periods, imperfect diagnostic tools, uncertainty in attributing tuberculosis disease episodes to a specific exposure, and deliberate modification of natural reactivation rates through preventive therapy. Thus, although the phenomenon of late reactivation is well established,[7,8] and groups at increased risk have been identified,[9] uncertainty regarding the absolute magnitude of late reactivation risk persists.

In the past decade, increased emphasis has been placed on the detection and treatment of LTBI to prevent future active tuberculosis disease, particularly given the global efforts being made towards tuberculosis elimination.[10] Uncertainty regarding the magnitude of late reactivation risk is problematic because accurate estimates are crucial for predicting the benefits of preventive therapy at both the individual and population level.[11] The evaluation of programmatic strategies frequently adopt modelling approaches to estimate public health impact and cost-effectiveness, and the parameter values that are used to simulate late reactivation can have a significant effect on outputs.[11–14]

A narrative review argued that tuberculosis reactivation more than 2 years after infection was rare,[15] but no systematic review of late reactivation has yet been done. We did a systematic review of the evidence quantifying the rate of late reactivation from LTBI to tuberculosis disease in the general population.

Methods

Definitions

We defined late reactivation conceptually as any form of tuberculosis disease occurring at least 2 years from new infection, in the absence of reinfection. We considered a positive LTBI test to indicate infection, although we recognise that LTBI diagnostic tools are imperfect and provide a correlate of infection rather than definitive proof.[2] We defined a LTBI diagnosis as an "immune response to prior acquired *M tuberculosis* antigens without evidence of clinically manifest tuberculosis",[16] as determined by any tuberculin skin test (TST; Mantoux, von Pirquet, or Heaf) or interferon-γ release assay (IGRA). Conversion was defined as when an individual is found to have a positive LTBI test following a negative one. We use the term reactivation to refer to progression from LTBI to active tuberculosis, rather than the occasional historical use of the term for what is now termed recurrent tuberculosis.[17] The terms primary and endogenous refer to disease occurring before or after 5 years from infection,

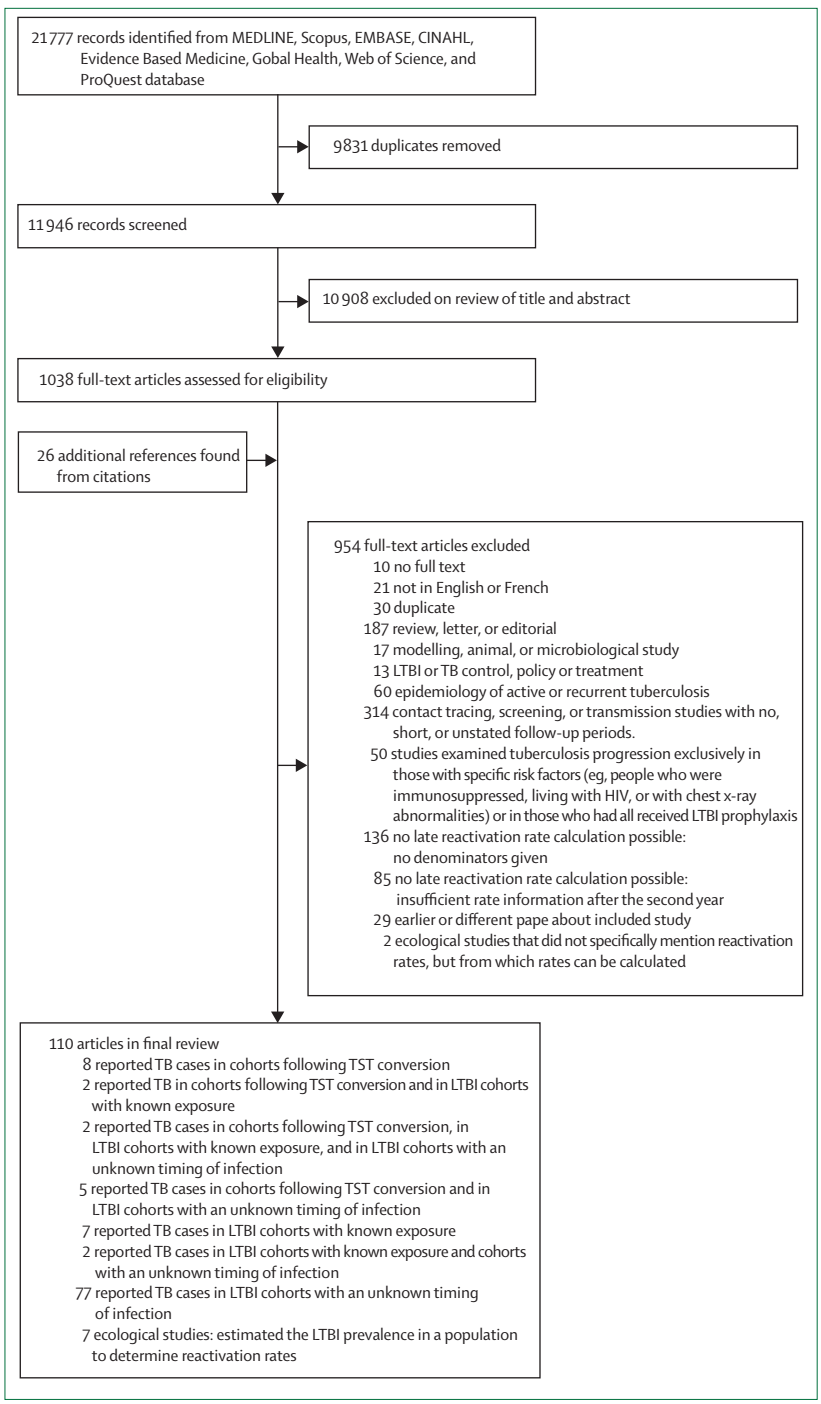


Figure 1: **Systematic review article selection**
LTBI=latent tuberculosis infection. TB=tuberculosis. TST=tuberculin skin test

See **Online** for appendix

For more on **WebPlotDigitizer** see https://automeris.io/WebPlotDigitizer/

respectively (Holm in 1969[18] appears to have first used this cutoff, although the terms were in use much earlier[17,19]). We use tuberculosis to refer to all manifestations of active tuberculosis, unless otherwise specified (eg, pulmonary tuberculosis).

Data extraction and analysis

We grouped studies by methodology and arranged these hierarchically into four groups according to the clarity of evidence provided on infection and its timing, and hence late reactivation rates.

Cohort studies documenting late reactivation in populations with TST conversion were discussed first (group 1), because they provide the clearest evidence of both infection and its timing. Group 2 included studies in LTBI cohorts with known *M tuberculosis* exposure, for which an uncertain number of participants might have been infected from a previous exposure. Group 3 were studies that documented tuberculosis progression in LTBI cohorts with an unknown timing of infection. Finally, group 4 included ecological studies that estimated population LTBI prevalence to determine reactivation rates. In group 4, the prevalence of LTBI is uncertain and the timing of infection is unknown, although some studies attempted to exclude cases attributable to recent infection.

Two reviewers (KDD and MK) independently did data extraction and quality assessment using a selection of questions from the Cochrane Risk Of Bias In Non-randomized Studies—of Interventions (ROBINS-I) assessment tool in the first two methodological groups (appendix p 31). Disagreements were resolved through discussion and consensus. KDD extracted data and assessed quality in the other methodological groups. Extracted information included study setting, participant selection, characteristics and exclusions, screening method(s), length and nature of follow-up (active with or without repeated assessments, or passive), follow-up and disease definitions, missing data, outcomes, and estimated tuberculosis incidence in study settings, extracted into a standardised form.

Numerical data were extracted by KDD into Microsoft Excel 2010, and analyses were done in R (version 3.5.2). WebPlotDigitizer was used to extract data from published figures in which numerical data was not presented.

Annual reactivation rates were calculated as the annual number of tuberculosis cases in each cohort divided by the number of LTBI cases. Where studies provided only the initial sample size and follow-up period without numbers observed over time, each annual denominator was estimated by subtracting the number of tuberculosis cases in the previous year. Where possible, 95% CIs were calculated using the exact Poisson test.

Results

We identified 11 946 unique studies and excluded 10 908 following title and abstract review. Following full-text review, we identified 26 additional studies through citation references and excluded a further 954, leaving 110 studies in the final review (figure 1). Studies are discussed below under their methodological grouping, with several appearing more than once because they met criteria for more than one group. A description of

	Strengths	Limitations
Strongest design: cohorts followed for incidence of tuberculosis after possible new infection (groups 1 and 2)		
Cohort studies of populations following TST conversion	TST conversion provides the clearest evidence of infection and its timing, which is strengthened by the requirement for known tuberculosis exposure. Genotyping information can be used to provide evidence of transmission.[5]	Documenting conversion in study populations is laborious, typically requiring repeated assessments of at-risk populations and so limiting cohort size. Varying definitions of conversion and inversion might be used with varying accepted times between negative and positive LTBI tests. Varying definitions of the timing of infection might be used (eg, time of first positive test,[20,21] midpoint between first positive and last negative test,[22,23] 2 months before erythema nodosum,[24] or 3 months before hilus adenitis[24]). Varying definitions of disease reactivation can be used (eg, date of first abnormal chest x-ray,[25] diagnosis,[20,26] treatment initiation,[27] or midpoint between last normal chest x-ray and first abnormal chest x-ray[22,23]). Varying follow-up intervals and methods might be used (eg, passive versus active, with or without TST, with or without chest x-ray monitoring). Radiological monitoring might identify more tuberculosis cases than passive case detection methods.[28,29] Loss to follow-up (through migration or death) is seldom quantified.
Cohort studies in populations with LTBI identified following Mtb exposure	Mtb infection and its timing are more clearly defined in individuals with a known exposure and LTBI diagnosis than in populations with LTBI without known exposure. However, an uncertain number might have been infected at a previous exposure. Genotyping information can be used to provide evidence of transmission.[5]	Because participants must have a known exposure, a long period of recruitment is required to obtain a sufficiently large cohort size. Varying definitions of Mtb exposure (eg, household[20,30] or close circle contact status,[31] and index patient might have any active[32,33] or only infectious[26] tuberculosis), timing of infection (LTBI diagnosis,[26] index diagnosis or notification,[5] index treatment initiation[27]), and timing of reactivation. Varying follow-up intervals and methods.
Weaker design: cohorts followed for incidence of tuberculosis after cross-sectional study to determine prevalent tuberculosis infection (group 3)		
Cohort studies in other populations with LTBI and unknown timing of Mtb exposure	Without the need for a known exposure, large cohorts can be opportunistically screened. Efforts to exclude recent infection can be made; eg, asking about recent contact[34] or excluding genotypically clustered cases.[35–37] It can be assumed that recent infection is unlikely in certain cohorts in low-incidence settings; eg, for migrants from high-incidence settings, time of migration provides a point beyond which infection is much less likely.	The timing of infection is unknown, such that the proportion of recently infected individuals is unknown (except possibly for some cohorts in low-incidence settings). Varying definitions of disease reactivation. Varying follow-up intervals and methods. Because TST or IGRA reversion can occur, limiting cohorts to individuals with positive test results might overestimate reactivation rates in all who have been infected.
Weakest design: ecological studies (group 4)		
Using LTBI survey data to infer LTBI prevalence in a population and tuberculosis notifications to quantify reactivation episodes arising from the inferred pool	Without the need for known exposure or follow-up, the LTBI and case cohorts can be larger, allowing subgroup comparison. Efforts to exclude recent infection can be made.	Timing of infection is unknown, such that the proportion recently infected is also unknown, and how reactivation rates change in cohorts over time cannot be observed. Accuracy depends on how well the LTBI cohort represents the case cohort (eg, uncertainty will be increased if the base cohort is small,[37] or only includes those that attended mass screening,[38] or only includes those who were tested at a different time to the case cohort[36]). Because TST and IGRA reversion can occur, limiting reference cohorts to individuals with positive test results might overestimate late reactivation rates in all who have been infected.

TST=tuberculin skin test. LTBI=latent tuberculosis infection. Mtb=Mycobacterium tuberculosis. IGRA=interferon-γ release assay

Table 1: **Classification and rating of the major strengths and limitations of the included studies**

Panel: Factors that influence the interpretation of reactivation rates in all studies

- Varying LTBI diagnosis methods, which have differing validity (sensitivity and specificity), and NTM exposure in the study population. For example, a Danish study found reactivation rates in participants aged 15–24 years to be 3 times higher in those who lived in an area with a low prevalence of bovine tuberculosis versus a high-prevalence area.[39]
- Whether chest x-ray was done in conjunction with screening for LTBI or not, to exclude active disease, and whether those with chest x-ray abnormalities were included in the study cohort.
- Varying definitions of disease; eg, pulmonary, respiratory, or all forms. The definition of disease is particularly variable and difficult to interpret in pre-antibiotic era (early 20th century) studies.[40]
- Ongoing infection or reinfection risks, which might inflate reactivation rates. Ongoing cases in some studies, particularly pre-antibiotic era studies, might be relapse episodes rather than reactivations.[21,41]
- In study settings where preventive treatment is routinely used, sample sizes might be limited, which could introduce bias if groups at highest risk are most likely to receive preventive treatment.
- There are recognised limitations with using genotyping to classify transmission.[42,43]
- Whether and how reactivation rates are disaggregated by age, time since study entry, sex, and screening method, preventive treatment, and risk factors.
- The prevalence of individual and population risk factors; eg, BCG vaccination status, chest x-ray status, HIV, diabetes, or smoking.

LTBI=latent tuberculosis infection. NTM= non-tuberculous mycobacteria.

key strengths and limitations of each methodological approach is provided in table 1, and the factors that influence the interpretation of reactivation rates in all studies are in the panel. Individual studies and their risk of bias score, are described in the appendix (pp 5–10) and summarised in table 2.

	Years of study (recruited; end)	Setting	Population	Age group on recruitment (years)	Follow-up (years)	Active or passive follow up; numbers observed over time given (yes/no)	LTBI therapy (yes/no)	LTBI screening method (method: TST cutoff)	Sample size (number of tuberculosis cases/number with conversion or TST positive at study entry)*
Group 1: prospective cohort studies documenting tuberculosis progression following TST conversion (some populations had a known exposure)									
Myers et al, 1964[44]	1921–41; 1960	USA	Child contacts	6–12	NS	Active; no	No	Pirquet, Mantoux 1928 on: ≥5–10 mm	11/154
Myers et al, 1965[45]	1921–41; 1960	USA	Schoolchildren	13–17	NS	Active; no	No	Pirquet, Mantoux 1928 on: ≥5–10 mm	11/129
Meyer 1949[24]	1929–44; 1947	Norway	General population	≥4	<17	Active; yes	No	Pirquet: ≥3 mm or ≥4 mm	98/889
Madsen et al 1942[25]	1934–36; 1940	Denmark	Medical and high school students	"About 18"	<5	Active†; yes	No	Mantoux: ≥10 mm 48 h or 8 mm 72 h	11 with x-ray changes/167
Badger and Ayvazian 1948[46]	1932–43; 1948	USA	Student nurses	NS	5–15	Active†; no	No	Saranac Old Tuberculin: NS	40/285
Daniels et al 1948[28]	1935; 1944	England	Nurses	18–25	5	Active†; yes	No	Mantoux: ≥5 mm	44/347
Hertzberg 1948[20]	1936–46; NS	Norway	General population	All	<10	Active; yes	No	Pirquet or Mantoux: ≥10 mm	727/1829
Gedde-Dahl 1952[21]	1937–44; 1945	Norway	General population	All	NS	Active; yes	No	Pirquet: ≥3 mm or 4 mm	32 progressive pulmonary tuberculosis/214
Hyge et al 1956[41]	1942; 1954	Denmark	Exposed female schoolchildren	12–18	12	Active; no	No	Mantoux: NS	55 progressive pulmonary tuberculosis/70
Sutherland 1967 and 1968;[23,47] Styblo 1991[48] (British MRC trial)	1951–52; NS	UK	Schoolchildren	14·0–15·5	10	Active†; no	No	TST: ≥6 mm‡	243/2085‡
Ferebee and Mount 1962;[30] Ferebee 1970[3] (USPHS trials)	1957–60; NS	USA	Household contacts	All	<10	Active; no	No	TST: ≥5 mm at 12 months	32/867
Debre et al 1973[49]	1959–66; 1969	France	General population, mostly children	5–24	3–10	Active†; yes	No	TST: NS	24/1451
Veening 1968[50]	1960; 1967	Netherlands	Exposed male Navy recruits	18–20	7	Active; yes	No	Mantoux: NS	12/128
Stead 1987;[51] Stead and Dutt 1989[52]	1979–87; NS	USA	Nursing home residents	≥50	NS	>70% retested several times	No	TST: ≥12 mm change	89/965
Group 2: cohorts with LTBI followed from recent Mtb exposure									
Myers 1963[53]	1921–41; 1960	USA	Child contacts, no pulmonary infiltration	0–5	NS	Active; no	No	Pirquet, Mantoux 1928 on: ≥5–10 mm	41/599
Hertzberg 1948[20]	1936–46; 1946	Norway	Household contacts	All	>10	Active; yes	No	Pirquet or Mantoux: ≥10 mm	552/1043
Hyge et al 1956[41]	1942; NS	Denmark	Exposed female schoolchildren	12–18	12	Active; no	No	Mantoux: NS	9/105
Ferebee and Mount 1962;[30] Ferebee 1970[3] (USPHS trials)	1957–60; NS	USA	Household contacts	All	<10	Active†; no	No	TST: ≥5 mm	472 (to 8 years, 479 to 10 years)/7744
Dobler and Marks 2013[54]	2000–09; 2009	Australia	Contacts	All	Mean 4·6	Passive	No	TST: ≥10 mm	38/3942
Reichler et al 2018[27]	2002–06; 2011	USA and Canada	Contacts of culture-positive pulmonary cases§	NS	<8	Passive	No	TST: ≥5 mm	89/499

(Table 2 continues on next page)

www.thelancet.com/infection Published online April 20, 2021 https://doi.org/10.1016/S1473-3099(20)30728-3

Cohort studies of populations following TST conversion

Group 1 included 14 studies that reported tuberculosis progression in cohorts more than 2 years after TST conversion,[3,21,24,25,28,41,44–47,49–52] with four additionally indicating that all participants were known to have been recently exposed (table 2).[3,41,44,50] Most of these studies were done in moderate-incidence to high-incidence settings from the early to mid-20th century (ie, during the pre-antibiotic era).

In the nine studies that began during the pre-antibiotic era (pre-1945), the studies by Myers and colleagues in 1964 and 1965 had the longest follow-up duration.[44,45] Numbers observed over time were not reported in these studies, but the follow-up probably ranged from 19 years to 39 years given the study recruitment and end dates.[44,45] In most pre-antibiotic era studies, rates of tuberculosis disease decreased substantially over the first 2 years from conversion, with annual rates after 2 years varying from 0 cases per 100 000 person-years to more than 1000 cases per 100 000 person-years (appendix p 25). In studies that considered reactivation by age at conversion, Meyer 1949 and Hertzberg 1948 found late reactivation to be more common in those infected after the age of 12 years than before,[20,24] although this difference was not seen in the studies by Myers and colleagues.[44,45] Meyer noted that this difference was chiefly due to the high rates in those infected after the age of 12 in the first 3–5 observation years (appendix p 25).[24] Meyer also reported that "all cases of pulmonary tuberculosis with a latent period of more than 3 years (as well as several with a shorter latent period) broke out after the age of 14".[25] Reactivation occurred more frequently in females than in male patients in the studies by Myers and colleagues,[44,45] but not in those by Meyer or Hertzberg.[20,24]

Since the advent of antibiotics for tuberculosis from the mid-1940s, four studies have documented tuberculosis progression in the control arms of trials of preventive interventions (table 2, figure 2). In a British Medical Research Council (MRC) tuberculosis vaccine trial, beginning in 1951, 14·0–15·5 year-olds were followed-up for 15 years, receiving TST and chest radiographs approximately every 14 months until 1960 (approximately 8–9 years after study commencement) and postal enquiries thereafter.[56] Sutherland (1968) reported reactivation rates among participants who were found to have converted during the trial for up to 10 years post-conversion.[23,47] In isoniazid chemoprophylaxis trials run by the US Public Health Service (USPHS), from 1957, household contacts of active cases received TST and chest radiographs at study entry and 12 months later, and tuberculosis cases among those individuals found to have converted at 12 months were documented by Ferebee (1970) for 10 years after exposure.[3,30] Neither of these two studies reported the size of the cohorts that remained under observation beyond the first year, and the ages of the USPHS converters were not given.[3,30] In a smaller study by Debre and colleagues (1973), yearly chest radiographs were recorded from French residents with

	Years of study (recruited; end)	Setting	Population	Age group on recruitment (years)	Follow-up (years)	Active or passive follow up; numbers observed over time given (Yes/No)	LTBI therapy (Yes/No)	LTBI screening method (method: TST cutoff)	Sample size (number of tuberculosis cases/number with conversion or TST positive at study entry)*
(Contiued from previous page)									
Sloot et al 2014[5]	2002–11; 2012	Netherlands	Contacts of pulmonary cases	All	<11	Passive	Yes, 45% treated	IGRA or TST: ≥10 mm	14¶/739
Erkens et al 2016[26]	2005–13; 2014	Netherlands	Contacts of infectious cases	All	≥5	Passive	No	TST and, since 2010, IGRA: NS	41¶/2251
Altet et al 2015[31]	2007–09; NS	Spain	First circle contacts of smear positive cases	All	4	Active; yes	No	IGRA or TST: ≥5 mm	QFT: 14¶/81; TST: 14¶/340
Abubakar et al 2018[2,33]**	2010–15; 2016	UK	Contacts of active tuberculosis	>16	Median 2·9‖	Passive >24 months	No	IGRA or Mantoux: multiple cutoffs	TSpot.TB: 31/648; QFT-GIT: 30/793; TST ≥5 mm: 43/1704; TST ≥10 mm: 38/1323; TST ≥15 mm: 34/899
Heiden et al 2017[55]	2013; 2017	Flight (Turkey to Germany)	Crew and co-passengers of tuberculosis case with haemoptysis	All	3·7	Passive	No	TST or IGRA: >10mm	0/14

LTBI=latent tuberculosis infection. TST=tuberculin skin test. Mtb=Mycobacterium tuberculosis. NS=not stated. USPHS=US Public Health Service. MRC=Medical Research Council. IGRA=interferon-γ release assay. QFT=QuantiFERON. QFT-GIT=QuantiFERON-TB Gold In-Tube test. TSpot. TB. *Includes tuberculosis classification if some sites of disease were excluded. †Active follow-up in these studies included roentgenography or chest x-ray for some or all participants for some or the whole of follow-up. ‡We did not have information on the number of tuberculosis cases occurring over time, so to calculate reactivation rates over time we relied on figure 14 in Styblo 1991,[48] which shows the proportion of all cases (n=243) that reactivated over time, with an unspecified TST cutoff. We assumed a TST induration cutoff of 6 mm or greater to be most probable given case numbers from table 1 in Sutherland 1968[47] and a statement in the provisional analysis[23] that 125 participants developed tuberculosis before conversion was observed. Therefore, conversion was assumed for 125 participants. [23] §Who had "shared air space…in the household or other indoor setting for more than 15 hours per week or more than 180 hours total during an infectious period, defined as the interval from 3 months before collection of the first culture-positive sputum specimen or the date of onset of cough (whichever was longer) through 2 weeks after the initiation of appropriate antituberculosis treatment". [27] The index case also needed to be aged 15 years or older. [27] ¶Tuberculosis cases identified within 180 days after index diagnosis; [5] less than 100 days after screen,[26] or during contact tracing[31] were not included in the numerator. ‖Additional data were obtained from personal communication with the study's corresponding author, so data included in the table and figure 2 might not appear in the referenced manuscript. **This follow-up period was taken from the published manuscript and refers to the follow-up of both contacts and recent migrants.[32,33]

Table 2: **Description of cohort studies included in the review that documented late reactivation following TST conversion or Mtb exposure**

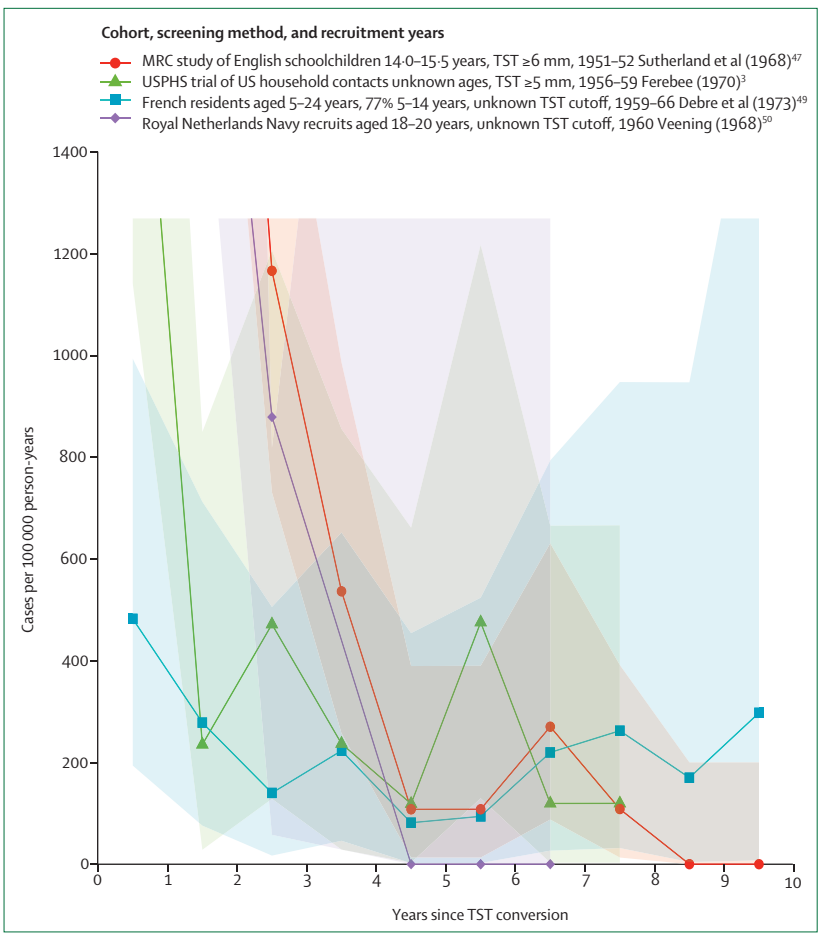


Figure 2: **Tuberculosis reactivation rates over time, in cohorts with TST conversion**
MRC=British Medical Research Council. USPHS=US Public Health Service. TST=tuberculin-skin test. Shaded areas represent 95% CIs. In the USPHS trial, observation was reported as incomplete beyond the seventh year, so these results have been excluded.[3] All studies reported active follow-up, but only the study in the Royal Netherlands Navy[50] and the French study[49] provided numbers observed over time. Different definitions of conversion were used, and results in the USPHS trial are shown from exposure, rather than conversion. The reactivation rate plotted at 2·5 years for Veening (1968) is the average rate 1–4 years after conversion.

recent conversion for up to 10 years. The numbers of cases and reactors observed over time was also documented.[49] Finally, a study by Veening (1968) documented tuberculosis cases among 128 recruits in the Royal Netherlands Navy at 1 year, 4 years, and 7 years after exposure to a marine with open tuberculosis.[50]

Late reactivation rates varied between studies. In all studies except Debre and colleagues (1973; which had a high proportion of young children,[49] unlike the others), late reactivation rates generally decreased as time went on following conversion (figure 2). In the studies by Sutherland (1968) and Veening (1968), both of which exclusively included youths, reactivation rates remained relatively high in the second to fourth year after conversion, compared with other studies (figure 2).[49–50] During the fifth follow-up year, late reactivation rates in all studies dropped to below 200 case per 100 000 person-years. Sutherland reported no cases among converters in the MRC trial after more than 7 years from conversion, despite observation extending to at least 10 years.[23] Similarly, Ferebee (1970) documented no progression in the USPHS trial after the eighth year, although observation was noted to have been incomplete beyond the seventh year.[3]

In no study since the advent of antibiotics for tuberculosis has it been possible to calculate late reactivation rates over time from conversion by either age group or TST status. However, Sutherland reported that the 10-year reactivation rates were higher in individuals who were younger (14 years old) at the time of conversion than those who were older (20 years old; appendix p 26).[47] This study also noted that reactivation rates increased slightly in those individuals showing a greater TST response, noting that this difference was only observable in the first 5 years.[23,47] Debre and colleagues also reported that an increased proportion of 15–24-year olds (3·9%) progressed over the 10 years from conversion, as compared with 10–14-year olds (1·9%) and 5–9-year olds (2·6%).[49]

The most recent study in this category followed a cohort of 965 converters in nursing home residents in Arkansas (USA) aged more than 50 years for an unspecified period (possibly up to 8 years).[51,52] This study found an average reactivation rate of 339 cases per 100 000 person-years in years 2 to 4, and no episodes thereafter.[51,52]

LTBI cohorts followed from known exposure

Late reactivation rates in populations with LTBI following a known exposure (group 2) can be observed in 11 studies: four from the mid-20th century that have already been introduced,[3,20,30,41,53] and seven others published after 2011 (table 2, appendix pp 8–10).[5,26,27,31–33,55,57]

In three of the studies already introduced, no clear difference in late reactivation rates can be seen when comparing cohorts with demonstrated conversion to those with LTBI following a known exposure (appendix p 27).[3,20,41] In a pre-antibiotic era study of female high-school students by Hyge in 1956, late reactivation rates can be compared between recent converters and those known to be TST positive before their so-called massive exposure, with no clear differences observed (appendix p 27).[41]

Considering late reactivation rates by age, a pre-antibiotic era study by Myers (1963) reported on 599 child contacts in Minnesota, USA, presenting before the age of 6 years.[53] In these children, all but three of 41 TB reactivations occurred either before 8 years of age or at ages 15–21 years, despite follow-up to an average age of 32 years (appendix p 28).[53] Hyge (1956) also noted that no case of post-primary pulmonary tuberculosis occurred before the age of 15.[41] Ferebee (1970) presented the number of tuberculosis cases by age group over time in all participants in the USPHS trial who were TST-positive on entry or had converted by 12 months.[3] The highest reactivation rates between 2 and 5 years post-exposure were seen in 15–54-year olds compared with younger or

older groups, but rates were similar thereafter (appendix p 28).[3] In addition to the USPHS trial, five more recent studies of untreated contacts were identified with results summarised in figure 3.[26,27,31–33,57] Across studies, screening methods and cutoffs varied, but rates ranged from 0 cases per 100 000 person-years to 500 cases per 100 000 person-years from 2 to 5 years after exposure, reaching approximately 200 cases per 100 000 person-years or lower by the fifth year, and appearing to decline further beyond this point. However, the decreasing or unspecified number of participants under observation in several studies makes interpretation difficult. Three included studies reported on genotypic concordance of reactivation episodes, although none reported concordance of the cases of late reactivation specifically.[5,31,58]

LTBI cohorts with unknown timing of infection

86 studies (group 3) have documented tuberculosis progression in populations with LTBI for whom the timing of infection was not established (appendix pp 12–23). The largest of these studies were done in the mid-20th century as part of vaccine trials,[56,59–66] military screening,[67–70] or mass screening or vaccination campaigns.[39,71,72]

These studies typically presented the average reactivation rate across all observation years. High average annual reactivation rates (>400 cases per 100 000 person-years) were reported in numerous studies done in settings known to have a high incidence of tuberculosis,[34,63–65,73–80] but several studies in low-incidence settings among recent migrants from high-incidence settings also found high reactivation rates.[81–83] The lowest reactivation rates were seen in studies that followed untreated residents of low-incidence countries on LTBI registries.[84,85]

Several studies presented average reactivation rates across all observation years stratified by TST response, predominantly finding increasing rates with greater TST diameter (appendix p 28).[39,61,64,66,67,86,87] Studies that presented rates by TST status over time typically showed declining rates, with the most rapid declines occurring in those with the greatest TST responses, such that rates gradually converged (figure 4A).[39,56,59,88]

Horwtiz and colleagues (1969) also observed this pattern across multiple age groups in a study of adult residents of Denmark, with the highest rates occurring among the youngest, 15–24-year-old, group (figure 4B).[39]

In most studies that observed reactivation rates in cohorts for more than 10 years (up to 20 years), rates beyond 10 years were lower than 50–100 cases per 100 000 person-years.[39,56,59,81] Higher rates more than ten years after screening were observed in one study by Gernez-Rierux and Gervois (1973), with distinguishing features of this study including that high rates were observed in initial non-reactors, follow-up was active and included annual chest radiographs, and participants were the youngest compared with others in this group.[66]

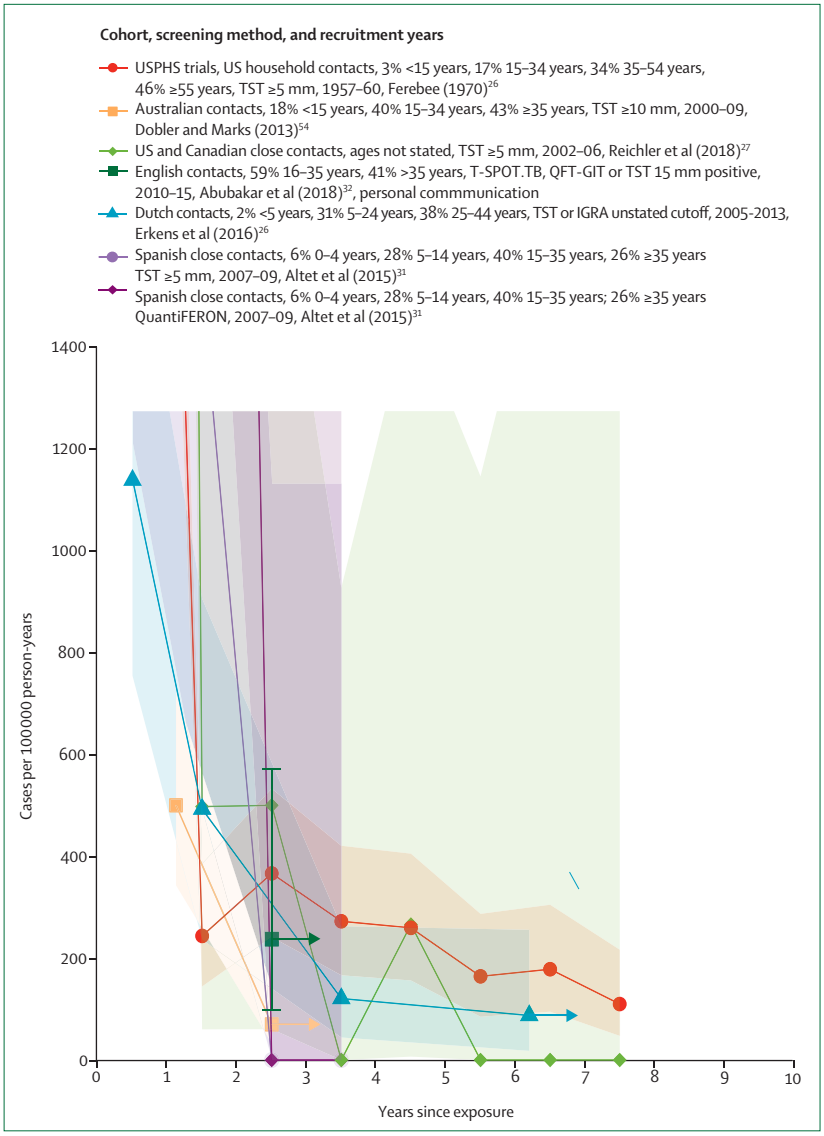
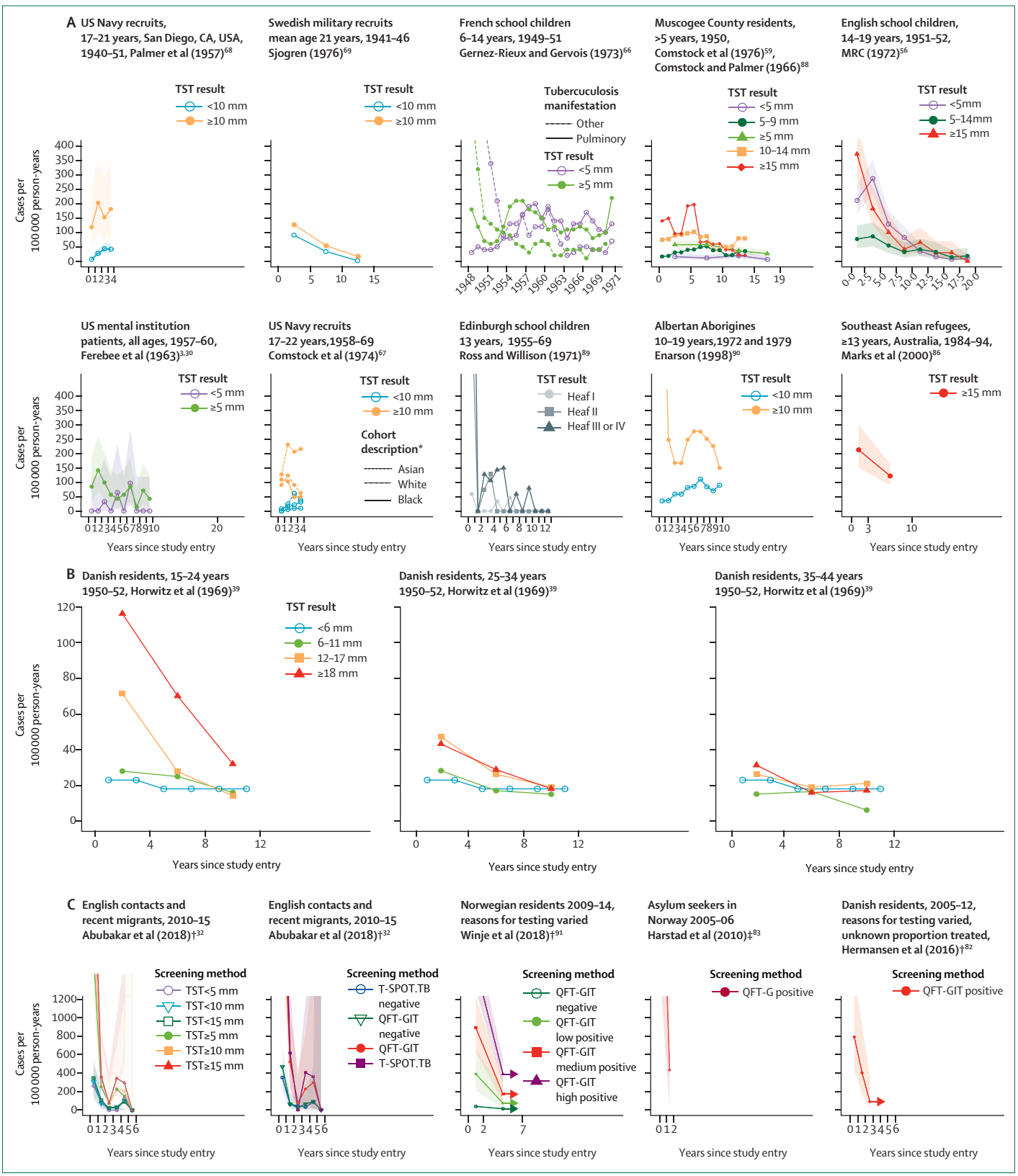


Figure 3: **Tuberculosis reactivation rates over time in untreated cohorts with known *Mycobacterium tuberculosis* exposure and LTBI diagnosis**

LTBI=latent tuberculosis infection. USPHS=US Public Health Service. TST=tuberculin skin test. QFT-GIT=QuantiFERON-TB Gold In-Tube. IGRA=interferon-γ release assay. QFT-G=QuantiFERON-TB Gold. The arrows indicate that the reactivation rates given in these studies beyond the previous year label were not right-censored. Shaded areas represent 95% CIs. Reactivation rates were provided by Erkens and colleagues[26] in the first, second, third to fifth, and 5 or more years. All but two studies had passive follow-up beyond 2 years: the study in Spanish close contacts[31] followed up participants for 4 years, and the USPHS trials reported active follow up, but did not provide observed cohort sizes over time.[330]

Two studies of child reactors have observed differing reactivation rates by age. In a large study of Puerto Rican children from 1949 with 19 years of follow-up, results were presented by age at tuberculosis diagnosis rather than time from screening, with the explanation that the "changes that occurred with the passage of time…[were] small and inconstant compared with the effect associated with age."[60] The greatest disease risk was observed in the very young, with a second high-risk period in youth


A
US Navy recruits, 17–21 years, San Diego, CA, USA, 1940–51, Palmer et al (1957)[68]
Swedish military recruits mean age 21 years, 1941–46 Sjogren (1976)[69]
French school children 6–14 years, 1949–51 Gernez-Rieux and Gervois (1973)[66]
Muscogee County residents, >5 years, 1950, Comstock et al (1976)[59], Comstock and Palmer (1966)[88]
English school children, 14–19 years, 1951–52, MRC (1972)[56]
US mental institution patients, all ages, 1957–60, Ferebee et al (1963)[3,30]
US Navy recruits 17–22 years,1958–69 Comstock et al (1974)[67]
Edinburgh school children 13 years, 1955–69 Ross and Willison (1971)[89]
Albertan Aborigines 10–19 years,1972 and 1979 Enarson (1998)[90]
Southeast Asian refugees, ≥13 years, Australia, 1984–94, Marks et al (2000)[86]
TST result
<10 mm
≥10 mm
<5 mm
≥5 mm
5–9 mm
10–14 mm
≥15 mm
<5mm
5–14mm
Tuberculosis manifestation
Other
Pulmonary
Cohort description*
Asian
White
Black
Heaf I
Heaf II
Heaf III or IV
Cases per 100 000 person-years
Years since study entry
B
Danish residents, 15–24 years 1950–52, Horwitz et al (1969)[39]
Danish residents, 25–34 years 1950–52, Horwitz et al (1969)[39]
Danish residents, 35–44 years 1950–52, Horwitz et al (1969)[39]
<6 mm
6–11 mm
12–17 mm
≥18 mm
C
English contacts and recent migrants, 2010–15 Abubakar et al (2018)†[32]
English contacts and recent migrants, 2010–15 Abubakar et al (2018)†[32]
Norwegian residents 2009–14, reasons for testing varied Winje et al (2018)†[91]
Asylum seekers in Norway 2005–06 Harstad et al (2010)‡[83]
Danish residents, 2005–12, reasons for testing varied, unknown proportion treated, Hermansen et al (2016)†[82]
Screening method
TST<5 mm
TST<10 mm
TST<15 mm
TST≥5 mm
TST≥10 mm
TST≥15 mm
T-SPOT.TB negative
QFT-GIT negative
QFT-GIT
T-SPOT.TB
QFT-GIT negative
QFT-GIT low positive
QFT-GIT medium positive
QFT-GIT high positive
QFT-G positive
QFT-GIT positive

(appendix p 29).[60] In a more recent study among those with a positive TST aged 6–10 years in Hong Kong, reactivation rates were also observed to increase after the age of 15 years (from 38 cases per 100 000 person-years to 608 cases per 100 000 person-years).[92] Neither study reported whether participants received chest x-ray screening at study commencement.[60,92]

Studies that reported tuberculosis progression in cohorts with LTBI who were exclusively diagnosed by IGRA are listed in the appendix (pp 21–22) and the four largest studies (n>39) that presented rates over time are illustrated in figure 4C.[32,82,83] In a UK cohort, reactivation rates were found to be similarly high when participants were screened with T-SPOT.TB or TST of 15 mm or more (>300 cases per 100 000 person years in the fourth and fifth year after screening), but lower when QuantiFERON-TB Gold In-Tube (QFT-GIT) or lower TST cutoffs were used.[93,94]

Ecological studies

In group 4, seven studies have estimated reactivation rates in the years after infection indirectly by using TST/IGRA survey data to infer the prevalence of infection, and tuberculosis notifications to quantify reactivation episodes. Three studies performed in low-incidence settings additionally used various approaches to exclude cases suspected of being due to recent infection (eg, by excluding genotypically clustered cases).[35–37] It was possible to calculate reactivation rates from the available data in a further two ecological studies,[95,96] but they did not discuss rates and so were excluded from the main analysis, and are described in the appendix (p 30).

Grzybowski and Allen (1964) used industrial and community TST-survey data in Ontario, Canada, to estimate the population LTBI prevalence, and categorised TB cases as either reactivation (meaning recurrent TB), recent or remote on the basis of chest radiograph result, age and clinical manifestation (citing the work of Wallgren [1948][97] and stating that certain manifestations generally occur within 2 years of infection).[35] They estimated a remote reactivation rate of 57 cases per 100 000 person-years in those with a TST response of more than 5 mm without radiographical evidence of inactive disease.[35] They also estimated reactivation rates of all pulmonary tuberculosis in Ontario in 1960, finding different rates by age and sex (figure 5A).[35] Stead (1983) found similar age-specific reactivation rates in Arkansas, USA, in 1961, 1971, and 1981[98] when using the age-specific TST prevalence found in Ontario in 1958–60,[35] together with the assumption of a 5% reversion rate over time (figure 5A).[35]

Barnett and colleagues (1971) used rates of TST positivity (≥6 mm) found during mass tuberculin screening, and tuberculosis notifications from 1960–69 in Saskatchewan, Canada, to estimate reactivation rates.[38] Rates were lower than in the Ontario study, ranging from 14 cases per 100 000 person-years in 50–59-year-olds to 46 cases per 100 000 person-years in 0–14-year olds (figure 5A).[38]

More recently, Horsburgh and colleagues (2010)[37] and Shea and colleagues (2014)[36] presented reactivation rate estimates in the USA from Palm Beach (Florida) and nationwide, respectively, by using TST surveys to provide denominator estimates and non-genotypically clustered tuberculosis notifications to determine numerators.[36,37] Rates of 70 cases per 100 000 person-years (95% CI 48–100) were found for Palm Beach and 84 cases per 100 000 person-years (83–85) for national estimates, with results varying by age (figure 5B).

Finally, in two studies by Mulder and colleagues, screening results, using the IGRA, QFT-GIT,[99] and TST[100] from a sample of recent immigrants to the Netherlands, were projected onto the entire immigrant cohort, who were monitored for 2 years after arrival. A Bayesian analysis of published data was used to provide sensitivity estimates for their methods.[99,100] Reactivation rates were found to be 193–247 cases per 100 000 person-years in those screened with QFT-GIT, 114–212 in those with TST of 15 mm or greater, and 97–173 in those with TST of 10 mm or greater, depending on the age, sex, and tuberculosis incidence in the country of birth, with no marked differences by these strata.[99,100]

Discussion

This systematic review presents the evidence for rates of LTBI reactivation beyond 2 years from infection (late reactivation). The evidence is diverse and must be interpreted in the context of each study's methodological approach rather than meta-analysed, but there are some consistent trends, with time from infection and age appearing to be key influences. Decreasing reactivation rates were seen in almost all studies that observed cohorts over several years, and in antibiotic-era studies in untreated populations rates reached approximately 200 cases per 100 000 person-years or below by the fifth year from exposure or conversion, and appeared to decline further beyond this point, although the decreasing or unspecified number of participants under

Figure 4: **Tuberculosis reactivation rates over time in populations screened for latent tuberculosis by TST result (A), by age and TST result (respiratory and pleural tuberculosis; B),[39] and in studies that included IGRA (C)**
TST=tuberculin skin test. IGRA=interferon-γ release assay. MRC=British Medical Research Council. QFT-GIT=QuantiFERON-TB Gold In-Tube. QFT-G=QuantiFERON-TB Gold. X-axis labels indicate the years over which the results were averaged. Shaded areas represent 95% CIs. The arrows indicate that the reactivation rates given in these studies beyond the previous year label are not right-censored. In the study of French school children, participants were recruited from 1949–51 and rates were calculated for each year in relation to the number of participants followed over 3 consecutive years.[66] Research on Muscogee County, GA, USA, residents reported adjustment for losses using a modified life-table method.[88] *Terms have been updated. †Populations were known to include close contacts. ‡Populations known to include those with chest x-ray abnormalities.

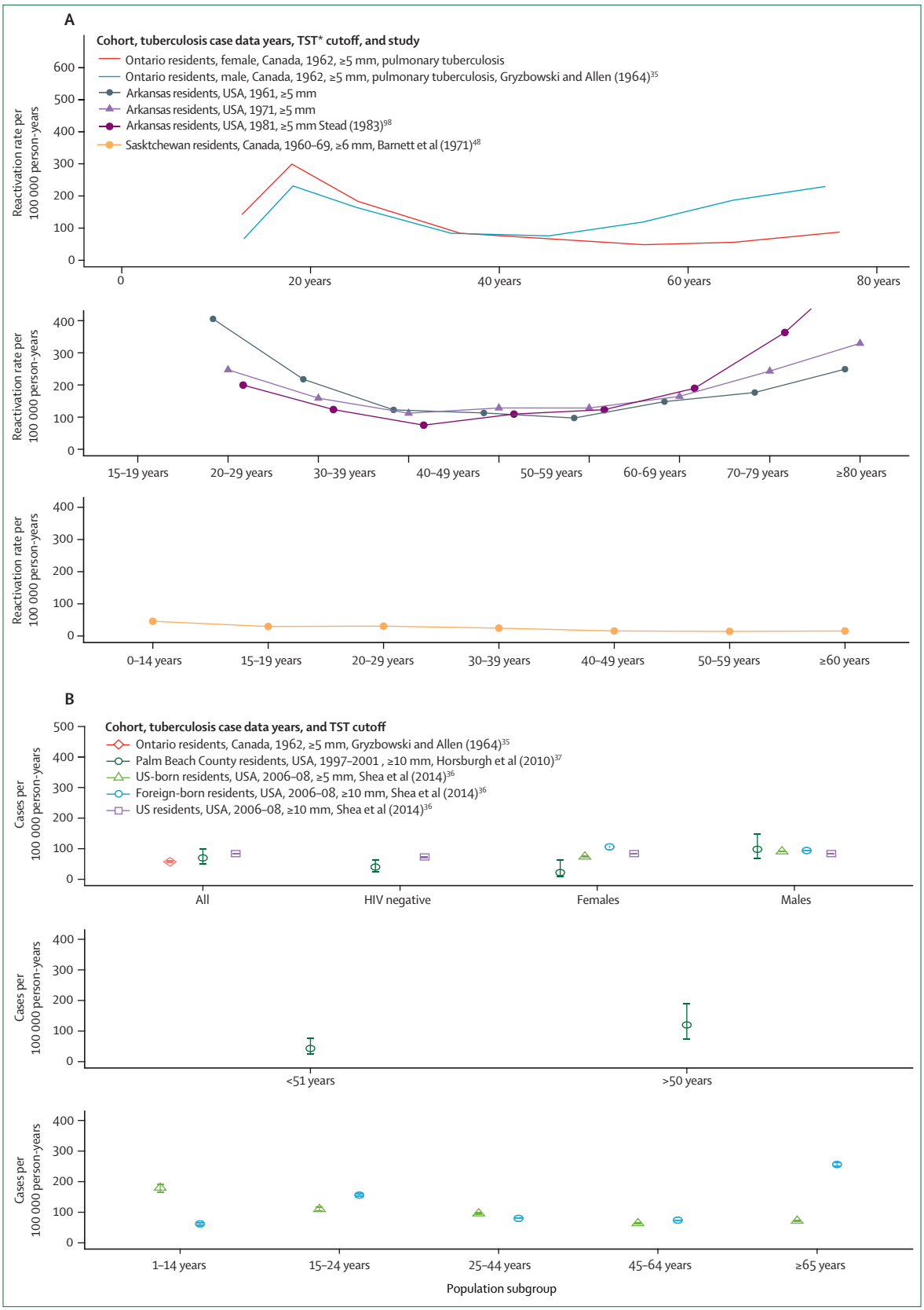


Figure 5: **Annual tuberculosis reactivation rates from ecological studies that estimated LTBI prevalence in a population and then used tuberculosis notifications to quantify reactivation episodes**
Studies that included all tuberculosis cases in their estimates–ie, they made no adjustments to exclude cases of recent infection (A) and studies that excluded cases of recent infection (B). LTBI=latent tuberculosis infection. TST=tuberculin skin test. The first panel of figure 5A is adapted from Grzybowski and Allen,[101] by permission of the American Thoracic Society.

observation in most studies limits interpretation. Consistently, rates were lower in studies of those participants with unknown timing of exposure, and studies that followed these cohorts beyond 10 years typically reported rates below 50–100 cases per 100 000 person-years.[56,59]

In addition to time from infection, there is evidence of a possible association between age and late reactivation rates, although the contribution of each is difficult to discern. Most studies did not publish age-disaggregated data and of those that did, quantification of reactivation rates could have been influenced by differences in the background prevalence of reactivity (which commonly increases with age[30,35,52,102]) and age-related differences in reinfection risk, which can vary with time and social mixing patterns.[21,35] For example, in the few studies that estimated late reactivation rates over time in cohorts aged under approximately 13 years, late reactivation rates remained relatively low from 6 to 14 years of age,[24,45,53,92] and several studies observed a subsequent increase in risk during youth,[45,53,60,92,103] perhaps indicating that late reactivation might not continuously decline with time from infection. Consistent with this theory, the study of TST converters by Debre (1973) had the youngest cohort (77% 5–14 years of age), and was also the only study for which reactivation rates did not continuously decrease over time. However, these studies were all done in high-incidence settings and so the degree to which the increase during youth (or the absence of a decrease) was due to increasing reinfection as social contacts changed,[21,35] rather than late reactivation, is impossible to determine.

Distinct patterns of reactivation by age can also be observed in several of the ecological studies in low-incidence settings, with high rates in youth, relatively low rates for most of adult life, and then an increase again into old age (figure 5).[35–37,98] However, the relative contribution of time from infection and age cannot be discerned in these studies. Although the pattern was observable in the foreign-born group of one study in a low-incidence setting that excluded genotypically clustered tuberculosis cases,[36] tuberculosis due to recent infection acquired overseas (before migration or during travel) would not have been identified as part of a cluster and hence recent infection, rather than late reactivation, might still have contributed to the observable peaks.[36]

Although reactivation can occur decades after infection,[7] we found no cohort studies with sufficient follow-up time and large enough samples to quantify this risk adequately beyond 10 years. Styblo suggested that the endogenous reactivation rate in Dutch residents aged 65–74 years and 75 years or older in 1973–76 would simply be their incidence of bacillary pulmonary tuberculosis (10 cases per 100 000 person-years for individuals aged years 65–74 years and 20 cases per 100 000 person-years for those aged 75 years or older), on the assumption that they had all been infected earlier in life when tuberculosis incidence was very high.[48] It is also commonly quoted that the lifetime risk of tuberculosis infection is 10%, with half that risk occurring in the first 5 years after infection, which would imply, for example, an annual endogenous rate of 93 cases per 100 000 person-years if infected at 20 years of age with a life expectancy of 80 years. However, although this assertion is commonly referenced to one of several sources,[3,48,60,104,105] we found no data in these or other sources to support it. Although Comstock and colleagues suggested that "The lifetime risk for a young child who is a strongly positive reactor may run as high as 10 per cent," and might have reached this conclusion by extrapolating the annual rate of tuberculosis found among Puerto-Rican 7–12 year olds (123 cases per 100 000 person-years) or 13–18 year-olds (149 cases per 100 000 person-years) with TST indurations of 16 mm or more by 71–86 years, the authors stated that this was an upper estimate among strong reactors.[60] Furthermore, the mean follow-up in their cohort was only 19 years and the reinfection risk in Puerto Rico was likely to be high during the study period (tuberculosis mortality was 179 per 100 000 person-years in 1948).[106] By contrast, Vynnycky and Fine's age-structured deterministic tuberculosis transmission models used to explore the dynamics of infection and pulmonary disease in White male patients in England and Wales in the second half of the 20th century[107–111] suggested that most risk occurs within 5 years from infection, with a constant low rate of endogenous pulmonary tuberculosis thereafter.[107] They estimated the endogenous rate to be negligible in males infected as children and 30 cases per 100 000 person-years in males infected as adults, acknowledging that, although there might have been an increase into old age, the "magnitude and pattern of the increases is unknown".[107]

Our review also highlighted the strengths and limitations of LTBI diagnosis. Included studies demonstrated that the relative size of the TST response reliably stratified reactivation risk in the first few years after diagnosis,[39,56,59,88] and one study also showed this stratification was possible with differing interferon-γ concentrations.[112] However, in those studies that followed up populations from the point of LTBI diagnosis with a TST, the relative size of the TST response became less meaningful with the passing of time.[39,56,59,88] Furthermore, many included studies demonstrated TST reversion (a change in an individual's LTBI test result from positive to negative),[3,21,25,28,35] and some found it to be more common among younger age groups,[3,112,113] in those without radiographical abnormalities,[3] in those with smaller indurations, and where infection rates are low.[25,28] For example, in the USPHS trial, it was reported that 6·5% of the initial reactors in the placebo group and 7·9% in the isoniazid group had reverted at 12 months,[3] and their 10-year reactivation rates were lower than in those who remained positive (64·0% lower in the placebo group and 45·2% lower in the isoniazid group).[3] Therefore, the likelihood of LTBI resolution in a population over time and the factors that

Search strategy and selection criteria

We did a systematic review, following PRISMA guidelines,[114] searching Medline, Scopus, CINAHL, EMBASE, Evidence Based Medicine Reviews, Global Health, Web of Science and PROQUEST Dissertation and Thesis for primary research studies that estimated tuberculosis incidence rates in cohorts of people identified as having LTBI. We searched databases from inception to June 25, 2019. We included grey literature, and only included articles in English or French. We combined the following three groups of terms with the Boolean operator AND: (1) tuberculosis, (2) incidence (or inciden* or rate or rates) adjacent (within three words) to tuberculosis or tb, (3) a range of terms that might be expected in studies meeting our search strategy combined with the Boolean operator OR. The complete search strategy is in the appendix.

We included primary research studies that reported rates of progression to clinically significant tuberculosis more than 2 years after *M tuberculosis* exposure and LTBI diagnosis, or following a tuberculin-skin test (TST) test or interferon-gamma release assay (IGRA) conversion (test result changes from negative to positive). We also included studies for which these rates could be calculated; that is, those which included tuberculosis case numbers and follow-up period or person-years of observation beyond 2 years. We included screening studies that documented tuberculosis progression in populations with LTBI whose exposure history was not explicitly known, and cohort studies that used alternative methodologies to estimate reactivation rates. Additional studies were identified from the reference lists of initially identified studies.

We excluded modelling studies and studies that examined tuberculosis progression exclusively in individuals who had been recently BCG vaccinated because such vaccination can affect LTBI diagnosis, and also studies where participants had received LTBI prophylaxis because of the effect on progression risk. Studies exclusively in individuals with specific risk factors (eg, people who were immunosuppressed, living with HIV, or with chest x-ray abnormalities) were also excluded because our aim was to quantify rates in the general population. Where studies included the aforementioned populations and stratified results by these attributes, we only reported those results from unvaccinated and untreated cohorts without known risk-factors.

KDD reviewed the search results (by title, abstract and full text) for inclusion, and MK duplicated screening by reviewing a random sample of 250 results to ensure agreement. The review protocol was registered with PROSPERO (CRD42017070594).

influence it are likely to affect ongoing reactivation rates, and although the late reactivation rate in an entire cohort might be low many years after infection, the risk in those that retain reactivity could be higher. Correspondingly, ecological and other studies that inevitably limit their reference cohorts to contemporary reactors might overestimate the risk of reactivation many years from exposure, since they effectively exclude those who have reverted from their denominators.

Quantifying late reactivation rates is methodologically challenging and the available evidence reflects this difficulty. Our review provides an opportunity to summarise the limitations of the existing evidence and to highlight key aspects requiring further study. All included studies were limited by the possibility that reinfection might have contributed to observed reactivation rates, a contribution that could have varied with time given the declining tuberculosis incidence in many study settings. Future studies, such as reanalyses of existing datasets or studies of contacts who do not receive preventive treatment, have the opportunity to use both genomic and epidemiological data to better distinguish late reactivation from disease due to reinfection. Other key limitations included the frequent omission of details regarding the size of cohorts remaining under observation over time, the ages of participants, screening methods, chest x-ray status, and the disaggregation of results by these factors and over time. Given the wide use of IGRAs it should also be highlighted that few studies have specifically documented late reactivation rates in populations following an IGRA assessment. Future studies should ideally be disaggregated by screening method, immune status, receipt of preventive therapy, chest x-ray status, age, and time since infection. Disaggregation of results by sex would also be valuable, particularly as there is evidence to suggest that rates might differ by sex[35] and yet two of the studies that are most commonly used to estimate late reactivation exclusively considered males.[67,109]

We endeavoured to make this review systematic, transparent, and thorough, although the historical nature of many potentially relevant studies made comprehensiveness challenging. The review is also limited by the language restriction, which might have meant that important studies in languages other than English or French were missed. Furthermore, the paucity of the literature made it necessary to compare studies with heterogeneous demographics, screening and follow-up methods, and disease definitions, despite recognising that these all could affect reactivation rates. The varying use of chest x-ray screening is one example. Most studies excluded active disease using such screening at study commencement, but several additionally included screening during follow-up.[23,30,47,49] Chest x-ray screening can identify disease in asymptomatic patients,[28,115] and if radiological abnormalities cyclically wax and wane,[2] it is impossible to know whether and when such cases would have otherwise presented to health care. Therefore, the inclusion and frequency of active chest x-ray follow-up, versus passive follow-up, might affect observed reactivation rates.

The relative and potentially interacting contribution of age and time remain unclear and results must be viewed with caution, but existing evidence shows late reactivation rates to decline to approximately 200 cases per 100 000 person-years or below by the fifth year from exposure or conversion in untreated cohorts, they may be lower between 5 years and 10 years, and are unknown beyond this. In light of the debate surrounding the timeline of tuberculosis,[14] and with global efforts towards tuberculosis elimination, acknowledging the significant gaps that remain in our understanding of late reactivation is important, and attempts should be made to redress them. The importance of late reactivation, relative to early progression, remains unknown.

Contributors

JMT conceived the study. KDD designed the study, did the systematic review and analysed the results, and MK duplicated several steps of the

systematic review and analysis. KDD drafted the Article with input from JTD, JMT, KJS and DM. All authors contributed to Article revisions.

Declaration of interests

KDD is a recipient of the Miller Foundation Scholarship for doctoral studies in Infection and Immunity. This work was also supported by an Early Career Fellowship from the National Health and Medical Research Council (grant number: APP1142638) granted to JMT. MK is supported by a Melbourne Research Scholarship. KS is supported by the Centre for Research Excellence in Tuberculosis (TB-CRE).

Acknowledgments

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication. We also acknowledge and thank Ibrahim Abubakar, Rishi Gupta, the librarians of the University of Melbourne (VIC, Australia) and all those associated with the research that is included in this review, most particularly the scrupulous researchers from decades past.

References

1 WHO. Global tuberculosis report 2020. Geneva, Switzerland: World Health Organization, 2020.
2 Esmail H, Barry CE 3rd, Young DB, Wilkinson RJ. The ongoing challenge of latent tuberculosis. *Philos Trans R Soc Lond B Biol Sci* 2014; **369:** 20130437.
3 Ferebee SH. Controlled chemoprophylaxis trials in tuberculosis. A general review. *Bibl Tuberc* 1970; **26:** 28–106.
4 Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur Respir J* 2013; **41:** 140–56.
5 Sloot R, Schim van der Loeff MF, Kouw PM, Borgdorff MW. Risk of tuberculosis after recent exposure. A 10-year follow-up study of contacts in Amsterdam. *Am J Respir Crit Care Med* 2014; **190:** 1044–52.
6 Trauer JM, Moyo N, Tay EL, et al. Risk of active tuberculosis in the five years following infection 15%? *Chest* 2016; **149:** 516–25.
7 Lillebaek T, Dirksen A, Baess I, Strunge B, Thomsen VO, Andersen AB. Molecular evidence of endogenous reactivation of Mycobacterium tuberculosis after 33 years of latent infection. *J Infect Dis* 2002; **185:** 401–04.
8 Borgdorff MW, Sebek M, Geskus RB, Kremer K, Kalisvaart N, van Soolingen D. The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. *Int J Epidemiol* 2011; **40:** 964–70.
9 Campbell JR, Winters N, Menzies D. Absolute risk of tuberculosis among untreated populations with a positive tuberculin skin test or interferon-gamma release assay result: systematic review and meta-analysis. *BMJ* 2020; **368:** m549.
10 Lönnroth K, Migliori GB, Abubakar I, et al. Towards tuberculosis elimination: an action framework for low-incidence countries. *Eur Respir J* 2015; **45:** 928–52.
11 Dobler CC, Martin A, Marks GB. Benefit of treatment of latent tuberculosis infection in individual patients. *Eur Respir J* 2015; **46:** 1397–406.
12 Menzies NA, Wolf E, Connors D, et al. Progression from latent infection to active disease in dynamic tuberculosis transmission models: a systematic review of the validity of modelling assumptions. *Lancet Infect Dis* 2018; **18:** e228–38.
13 Ragonnet R, Trauer JM, Scott N, Meehan MT, Denholm JT, McBryde ES. Optimally capturing latency dynamics in models of tuberculosis transmission. *Epidemics* 2017; **21:** 39–47.
14 Oxlade O, Pinto M, Trajman A, Menzies D. How methodologic differences affect results of economic analyses: a systematic review of interferon gamma release assays for the diagnosis of LTBI. *PLoS One* 2013; **8:** e56044.
15 Behr MA, Edelstein PH, Ramakrishnan L. Revisiting the timetable of tuberculosis. *BMJ* 2018; **362:** k2738.
16 Mack U, Migliori GB, Sester M, et al. LTBI: latent tuberculosis infection or lasting immune responses to *M. tuberculosis*? A TBNET consensus statement. *Eur Respir J* 2009; **33:** 956–73.
17 Frost WH. The age selection of mortality from tuberculosis in successive decades. *Milbank Q* 1940; **18:** 61–66.
18 Holm J. Development from tuberculous infection to tuberculous disease. The Hague, Holland: KNVC, 1969.
19 Myers JA. The establishment and use of fundamental procedures in tuberculosis control. *Public Health Rep* 1946; **61:** 1563–83.
20 Hertzberg G. The achievements of BCG vaccination illustrated by material at the tuberculosis department of the Oslo public health service. Oslo, Norway: Johan Grundt Tanum Forlag, 1948.
21 Gedde-Dahl T. Tuberculous infection in the light of tuberculin matriculation. *Am J Hyg* 1952; **56:** 139–214.
22 Medical Research Council. B.C.G. and vole bacillus vaccines in the prevention of tuberculosis in adolescents; first (progress) report to the Medical Research Council by their Tuberculosis Vaccines Clinical Trials Committee. *BMJ* 1956; **1:** 413–27.
23 Sutherland I. Progress Report 1967, Part 1. The Hague, Holland: KNVC, 1967.
24 Meyer SN. Statistical Investigations of the relationship of tuberculosis morbidity and mortality to infection. *Acta Tuberculosea Scandinavica* 1949; **18:** 222.
25 Madsen T. Studies on the epidemiology of tuberculosis in Denmark. Copenhagen: Ejnar Munksgaard, 1942.
26 Erkens CG, Slump E, Verhagen M, Schimmel H, Cobelens F, van den Hof S. Risk of developing tuberculosis disease among persons diagnosed with latent tuberculosis infection in the Netherlands. *Eur Respir J* 2016; **48:** 1420–28.
27 Reichler MR, Khan A, Sterling TR, et al. Risk and timing of tuberculosis among close contacts of persons with infectious tuberculosis. *J Infect Dis* 2018; **218:** 1000–08.
28 Daniels M, Ridehalgh F, Springett VH, Hall IM. Tuberculosis in young adults: report on the Prophit Tuberculosis Survey, 1935–1944. London, UK: H.K. Lewis, 1948.
29 Kushigemachi M, Schneiderman LJ, Barrett-Connor E. Racial differences in susceptibility to tuberculosis: risk of disease after infection. *J Chronic Dis* 1984; **37:** 853–62.
30 Ferebee SH, Mount FW. Tuberculosis morbidity in a controlled trial of the prophylactic use of isoniazid among household contacts. *Am Rev Respir Dis* 1962; **85:** 490–510.
31 Altet N, Dominguez J, Souza-Galvão ML, et al. Predicting the development of tuberculosis with the tuberculin skin test and quantiFERON testing. *Ann Am Thorac Soc* 2015; **12:** 680–88.
32 Abubakar I, Drobniewski F, Southern J, et al. Prognostic value of interferon-γ release assays and tuberculin skin test in predicting the development of active tuberculosis (UK PREDICT TB): a prospective cohort study. *Lancet Infect Dis* 2018; **18:** 1077–87.
33 Abubakar I, Lalvani A, Southern J, et al. Two interferon gamma release assays for predicting active tuberculosis: the UK PREDICT TB prognostic test study. *Health Technol Assess* 2018; **22:** 1–96.
34 Mahomed H, Hawkridge T, Verver S, et al. The tuberculin skin test versus QuantiFERON TB Gold in predicting tuberculosis disease in an adolescent cohort study in South Africa. *PLoS One* 2011; **6:** e17984.
35 Grzybowski S, Allen EA. The challenge of tuberculosis in decline. a study based on the epidemiology of tuberculosis in Ontario, Canada. *Am Rev Respir Dis* 1964; **90:** 707–20.
36 Shea KM, Kammerer JS, Winston CA, Navin TR, Horsburgh CR Jr. Estimated rate of reactivation of latent tuberculosis infection in the United States, overall and by population subgroup. *Am J Epidemiol* 2014; **179:** 216–25.
37 Horsburgh CR Jr, O'Donnell M, Chamblee S, et al. Revisiting rates of reactivation tuberculosis: a population-based approach. *Am J Respir Crit Care Med* 2010; **182:** 420–25.
38 Barnett GD, Grzybowski S, Stýblo K. The current risk of contracting evolutive tuberculosis, in Saskatchewan, according to the state of previous tuberculin tests and x-ray image. *Bull Int Union Tuberc* 1971; **45:** 55–79.
39 Horwitz O, Wilbek E, Erickson PA. Epidemiological basis of tuberculosis eradication. 10. Longitudinal studies on the risk of tuberculosis in the general population of a low-prevalence area. *Bull World Health Organ* 1969; **41:** 95–113.
40 Marais BJ, Gie RP, Schaaf HS, et al. The clinical epidemiology of childhood pulmonary tuberculosis: a critical review of literature from the pre-chemotherapy era. *Int J Tuberc Lung Dis* 2004; **8:** 278–85.
41 Hyge TV. The efficacy of BCG-vaccination; epidemic of tuberculosis in a state school, with an observation period of 12 years. *Acta Tuberc Scand* 1956; **32:** 89–107.

42 Houben RM, Yates TA, Moore DA, McHugh TD, Lipman M, Vynnycky E. Re: "Estimated rate of reactivation of latent tuberculosis infection in the United States, overall and by population subgroup". *Am J Epidemiol* 2014; **180:** 450–51.

43 Borgdorff MW, van der Werf MJ, de Haas PE, Kremer K, van Soolingen D. Tuberculosis elimination in the Netherlands. *Emerg Infect Dis* 2005; **11:** 597–602.

44 Myers JA, Bearman JE, Dixon HG. Natural history of tuberculosis in the human body VI. Prognosis among tuberculin reactor children of six to twelve years. *Am Rev Respir Dis* 1964; **90:** 359–69.

45 Myers JA, Bearman JE, Dixon HG. Natural history of tuberculosis in the human body VIII. Prognosis among tuberculin reactor girls and boys of thirteen to seventeen years. *Am Rev Respir Dis* 1965; **91:** 896–908.

46 Badger TL, Ayvazian LF. Tuberculosis in nurses; clinical observations on its pathogenesis as seen in a 15 year follow-up of 745 nurses. *Am Rev Tuberc* 1949; **60:** 305–31.

47 Sutherland I. The ten-year incidence of clinical tuberculosis following "conversion" in 2550 individuals aged 14 to 19 years. The Hague, Holland: KNVC, 1968.

48 Styblo K. Epidemiology of tuberculosis. The Hague: Royal Netherlands Tuberculosis Association, 1991.

49 Debre R, Perdrizet S, Lotte A, Naveau M, Lert F. Isoniazid chemoprophylaxis of latent primary tuberculosis: in five trial centres in France from 1959 to 1969. *Int J Epidemiol* 1973; **2:** 153–60.

50 Veening GJ. Long term isoniazid prophylaxis. Controlled trial on INH prophylaxis after recent tuberculin conversion in young adults. *Bull Int Union Tuberc* 1968; **41:** 169–71.

51 Stead WW, To T. The significance of the tuberculin skin test in elderly persons. *Ann Intern Med* 1987; **107:** 837–42.

52 Stead WW, Dutt AK. Tuberculosis in the elderly. *Semin Respir Infect* 1989; **4:** 189–97.

53 Myers JA, Bearman JE, Dixon HG. The natural history of the tuberculosis in the human body. V. Prognosis among tuberculin-reactor children from birth to five years of age. *Am Rev Respir Dis* 1963; **87:** 354–69.

54 Dobler CC, Marks GB. Risk of tuberculosis among contacts in a low-incidence setting. *Eur Respir J* 2013; **41:** 1459–61.

55 An der Heiden M, Hauer B, Fiebig L, et al. Contact investigation after a fatal case of extensively drug-resistant tuberculosis (XDR-TB) in an aircraft, Germany, July 2013. *Euro Surveill* 2017; **22:** 30493.

56 Medical Research Council. BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. *Bull World Health Organ* 1972; **46:** 371–85.

57 Dobler CC, Marks GB. Risk of tuberculosis among contacts in a low-incidence setting. *Eur Respir J* 2013; **41:** 1459–61.

58 Denholm JT, LeslDE, Jenkin GA, et al. Long-term follow-up of contacts exposed to multidrug-resistant tuberculosis in Victoria, Australia, 1995–2010. *Int J Tuberc Lung Dis* 2012; **16:** 1320–25.

59 Comstock GW, Woolpert SF, Livesay VT. Tuberculosis studies in Muscogee County, Georgia. Twenty-year evaluation of a community trial of BCG vaccination. *Public Health Rep* 1976; **91:** 276–80.

60 Comstock GW, Livesay VT, Woolpert SF. The prognosis of a positive tuberculin reaction in childhood and adolescence. *Am J Epidemiol* 1974; **99:** 131–38.

61 Comstock GW, Livesay VT, Woolpert SF. Evaluation of BCG vaccination among Puerto Rican children. *Am J Public Health* 1974; **64:** 283–91.

62 National Tuberculosis Institute, Bangalore. Tuberculosis in a rural population of South India: a five-year epidemiological study. *Bull World Health Organ* 1974; **51:** 473–88.

63 Gothi GD, Nair SS, Chakraborty AK, Ganapathy KT. Five year incidence of tuberculosis and crude mortality in relation to non specific tuberculin sensitivity. *Indian J Tuberc* 1976; **23:** 58–63.

64 Radhakrishna S, Frieden TR, Subramani R. Association of initial tuberculin sensitivity, age and sex with the incidence of tuberculosis in south India: a 15-year follow-up. *Int J Tuberc Lung Dis* 2003; **7:** 1083–91.

65 Bunyasi EW, Luabeya AKK, Tameris M, et al. Impact of isoniazid preventive therapy on the evaluation of long-term effectiveness of infant MVA85A vaccination. *Int J Tuberc Lung Dis* 2017; **21:** 778–83.

66 Gernez-Rieux C, Gervois M. Protection conferred by BCG during the twenty years following vaccination. *Bull World Health Organ* 1973; **48:** 139–54.

67 Comstock GW, Edwards LB, Livesay VT. Tuberculosis morbidity in the U.S. Navy: its distribution and decline. *Am Rev Respir Dis* 1974; **110:** 572–80.

68 Palmer CE, Jablon S, Edwards PQ. Tuberculosis morbidity of young men in relation to tuberculin sensitivity and body build. *Am Rev Tuberc* 1957; **76:** 517–39.

69 Sjögren I. Tuberculosis in BCG-vaccinated and unvaccinated young Swedish men. A comparative study. *Scand J Respir Dis* 1976; **57:** 208–22.

70 Large SE. BCG vaccination in the brigade of Gurkhas. *J R Army Med Corps* 1965; **111:** 246–58.

71 Olsen HC. The use of the tuberculin test in a tuberculosis scheme; experience in Bornholm. *Tubercle* 1956; **37:** 47–57.

72 Härö AS. Twenty years later–evaluation of the results of a national mass BCG-vaccination in Finland. *Scand J Respir Dis Suppl* 1972; **80:** 153–69.

73 Martin V, Guerra JM, Cayla JA, Rodriguez JC, Blanco MD, Alcoba M. Incidence of tuberculosis and the importance of treatment of latent tuberculosis infection in a Spanish prison population. *Int J Tuberc Lung Dis* 2001; **5:** 926–32.

74 Large SE. Tuberculosis in the Gurkhas of Nepal. *Tubercle* 1964; **45:** 320–35.

75 Chigbu LN, Iroegbu CU. Incidence and spread of *Mycobacterium tuberculosis*-associated infection among Aba Federal prison inmates in Nigeria. *J Health Popul Nutr* 2010; **28:** 327–32.

76 Nduba V, Van't Hoog AH, Mitchell EMH, Borgdorff M, Laserson KF. incidence of active tuberculosis and cohort retention among adolescents in western Kenya. *Pediatr Infect Dis J* 2018; **37:** 10–15.

77 Andrews JR, Hatherill M, Mahomed H, et al. The dynamics of QuantiFERON-TB gold in-tube conversion and reversion in a cohort of South African adolescents. *Am J Respir Crit Care Med* 2015; **191:** 584–91.

78 Scolari C, El-Hamad I, Matteelli A, et al. Incidence of tuberculosis in a community of Senegalese immigrants in Northern Italy. *Int J Tuberc Lung Dis*. 1999; **3:** 18–22.

79 Grzybowski S, Galbraith JD, Styblo K, Chan-Yeung M, Dorken E, Brown A. Tuberculosis in Canadian Eskimos. *Arch Environ Health* 1972; **25:** 329–32.

80 Grzybowski S, Styblo K, Dorken E. Tuberculosis in Eskimos. *Tubercle* 1976; **57** (suppl): 1–58

81 Choudhury IW, West CR, Ormerod LP. The outcome of a cohort of tuberculin-positive predominantly South Asian new entrants aged 16–34 to the UK: Blackburn 1989-2001. *J Public Health (Oxf)* 2014; **36:** 390–95.

82 Hermansen TS, Lillebaek T, Langholz Kristensen K, Andersen PH, Ravn P. Prognostic value of interferon-γ release assays, a population-based study from a TB low-incidence country. *Thorax* 2016; **71:** 652–58.

83 Harstad I, Winje BA, Heldal E, Oftung F, Jacobsen GW. Predictive values of QuantiFERON-TB Gold testing in screening for tuberculosis disease in asylum seekers. *Int J Tuberc Lung Dis* 2010; **14:** 1209–11.

84 Mojazi-Amiri H, Larppanichpoonphol P, Nugent K. Tuberculosis reactivation in referrals to public health clinics in Texas. *Am J Med Sci* 2013; **346:** 442–46.

85 Roth DZ, Ronald LA, Ling D, et al. Impact of interferon-γ release assay on the latent tuberculosis cascade of care: a population-based study. *Eur Respir J* 2017; **49:** 1601546.

86 Marks GB, Bai J, Simpson SE, Sullivan EA, Stewart GJ. Incidence of tuberculosis among a cohort of tuberculin-positive refugees in Australia: reappraising the estimates of risk. *Am J Respir Crit Care Med* 2000; **162:** 1851–54.

87 Marks GB, Bai J, Stewart GJ, Simpson SE, Sullivan EA. Effectiveness of postmigration screening in controlling tuberculosis among refugees: a historical cohort study, 1984–1998. *Am J Public Health* 2001; **91:** 1797–99.

88 Comstock GW, Palmer CE. Long-term results of BCG vaccination in the southern United States. *Am Rev Respir Dis* 1966; **93:** 171–83.

89 Ross JD, Willison JC. Tuberculosis in Edinburgh: B.C.G. vaccination and Heaf tuberculin grades at school. *Scott Med J* 1971; **16:** 443–49.

90 Enarson DA. Tuberculosis in Aboriginals in Canada. *Int J Tuberc Lung Dis* 1998; **2**(9 suppl 1)**:** S16-S22.

91 Winje BA, White R, Syre H, et al. Stratification by interferon-γ release assay level predicts risk of incident TB. *Thorax* 2018; **73:** 652-61.

92 Leung CC, Yew WW, Au KF, et al. A strong tuberculin reaction in primary school children predicts tuberculosis in adolescence. *Pediatr Infect Dis J* 2012; **31:** 150–53.

93 Grinsdale JA, Islam S, Tran OC, Ho CS, Kawamura LM, Higashi JM. Interferon-gamma release assays and pediatric public health tuberculosis screening: the San Francisco program experience 2005 to 2008. *J Pediatric Infect Dis Soc* 2016; **5:** 122–30.

94 Tsou P-H, Huang W-C, Huang C-C, et al. Quantiferon TB-Gold conversion can predict active tuberculosis development in elderly nursing home residents. *Geriatr Gerontol Int* 2015; **15:** 1179–84.

95 Borgdorff MW, van den Hof S, Kremer K, et al. Progress towards tuberculosis elimination: secular trend, immigration and transmission. *Eur Respir J* 2010; **36:** 339–47.

96 Winje BA, Grøneng GM, White RA, Akre P, Aavitsland P, Heldal E. Immigrant screening for latent tuberculosis infection: numbers needed to test and treat, a Norwegian population-based cohort study. *BMJ Open* 2019; **9:** e023412.

97 Wallgren A. The time-table of tuberculosis. *Tubercle* 1948; **29:** 245–51.

98 Stead WW, Lofgren JP. Does the risk of tuberculosis increase in old age? *J Infect Dis* 1983; **147:** 951–55.

99 Mulder C, van Deutekom H, Huisman EM, et al. Role of the QuantiFERON(R)-TB Gold in-tube assay in screening new immigrants for tuberculosis infection. *Eur Respir J* 2012; **40:** 1443–49.

100 Mulder C, Mulleners B, Borgdorff MW, van Leth F. Predictive value of the tuberculin skin test among newly arriving immigrants. *PLoS One* 2013; **8:** e60130.

101 Grzybowski S, Allen EA. The challenge of tuberculosis in decline: a study based on the epidemiology of tuberculosis in Ontario, Canada. *Am Rev Respir Dis* 1963; **90:** 707–720.

102 Zeidberg LD, Gass RS, Dillon A, Hutcheson RH. The Williamson County tuberculosis study. A twenty-four-year epidemiologic study. *Am Rev Respir Dis* 1963; **87**: 1–88.

103 Pope AS, Sartwell PE, Zacks D. Development of tuberculosis in infected children. *Am J Public Health Nations Health* 1939; **29:** 1318–25.

104 Comstock GW. Epidemiology of tuberculosis. *Am Rev Respir Dis* 1982; **125:** 8–15.

105 Styblo K. The relationship between the risk of tuberculosis infection and the risk of developing infectious tuberculosis. *Bull Int Union Tuberc Lung Dis* 1985; **60:** 117–19.

106 Palmer CE, Shaw LW, Comstock GW. Community trials of BCG vaccination. *Am Rev Tuberc* 1958; **77:** 877–907.

107 Vynnycky E, Fine PE. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. *Epidemiol Infect* 1997; **119:** 183–201.

108 Vynnycky E, Fine PE. The annual risk of infection with Mycobacterium tuberculosis in England and Wales since 1901. *Int J Tuberc Lung Dis* 1997; **1:** 389–96.

109 Vynnycky E, Fine PE. Lifetime risks, incubation period, and serial interval of tuberculosis. *Am J Epidemiol* 2000; **152:** 247–63.

110 Vynnycky E, Fine PE. Interpreting the decline in tuberculosis: the role of secular trends in effective contact. *Int J Epidemiol* 1999; **28:** 327–34.

111 Vynnycky E, Fine PEM. The long-term dynamics of tuberculosis and other diseases with long serial intervals: implications of and for changing reproduction numbers. *Epidemiol Infect* 1998; **121:** 309–24.

112 Wiker HG, Mustafa T, Bjune GA, Harboe M. Evidence for waning of latency in a cohort study of tuberculosis. *BMC Infect Dis* 2010; **10:** 37.

113 Zacks D, Sartwell PE. Development of tuberculosis and changes in sensitivity to tuberculin in an institution for the feebleminded-a ten years' study. *Am J Public Health Nations Health* 1942; **32:** 732–38.

114 Liberati A, Altman D, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies thatevaluate health care interventions: explanation and elaboration. *Plos Med* 2009; **6:** e1000100.

115 Myers JA, Bearman JE, Botkins AC. The natural history of tuberculosis in the human body. X. Prognosis among students with tuberculin reaction conversion before, during and after school of nursing. *Dis Chest* 1968; **53:** 687–98.

© 2021 Elsevier Ltd. All rights reserved.

THE LANCET
Infectious Diseases

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Dale KD, Karmakar M, Snow KJ, Menzies D, Trauer JM, Denholm JT. Quantifying the rates of late reactivation tuberculosis: a systematic review. *Lancet Infect Dis* 2021; published online April 20. http://dx.doi.org/10.1016/S1473-3099(20)30728-3.

Supplementary material

Quantifying the rates of late reactivation tuberculosis: A systematic review

Dale, K.D., MPH[1,2]; Karmakar, M., MSc,[2,3] Snow, K.J., PhD[4]; Menzies, D. MD[5]; Trauer J.M., PhD[1,6*]; Denholm, J. T., PhD[1,2*].

[1] Victorian Tuberculosis Program, Royal Melbourne Hospital, at the Peter Doherty Institute for Infection and Immunity, Level 5, 792 Elizabeth St, Melbourne, Victoria, Australia, 3000

[2] Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, 792 Elizabeth Street, Melbourne, Victoria, Australia, 3000

[3] Baker Heart and Diabetes Institute, PO Box 6492, Melbourne, Victoria, Australia, 3004

[4] Department of Paediatrics, University of Melbourne, 50 Flemington Rd, Parkville, Victoria, 3052

[5] Respiratory Epidemiology and Clinical Research Unit, McGill International TB Centre, 5252 de Maisonneuve West, Room 3D.58, Montreal, Quebec, Canada, H4A 3S5

[6] School of Public Health and Preventive Medicine, Monash University, 553 St Kilda Road, Melbourne, Victoria, Australia, 3004

* Co-senior authors

Search Strategy

On the 25th of June 2019 we searched Medline, Scopus, CINAHL, EMBASE, Evidence Based Medicine Reviews, Global Health, Web of Science and PROQUEST Dissertation and Thesis for studies that estimated TB case rates in human cohorts identified as having LTBI. We applied no date restrictions and only included articles in English or French. We combined the following three groups of terms with the Boolean operator "AND": 1) tuberculosis; 2) incidence or (inciden* or rate or rates) adjacent (within three words) to (tuberculosis or tb)); 3) the following terms combined with "OR": endogenous, reactivat*, latency, time lag, late progression, late disease, lifetime risk, incubation period, post primary, postmigration, post migration, post immigration, postimmigration, post arrival, postarrival, unclustered, non clustered, latent period, cohort effect, ongoing risk, persistent risk, infectious disease incubation period, latent tuberculosis, tuberculin test, Interferon-gamma Release Tests, tuberculin skin test, t spot, igra, mantoux, quantiferon, qft ((years or time) adjacent (within three words) (arrival or migration)), (progress*) adjacent (within three words) (tuberculosis or tb or disease or active). Additional studies were identified from the reference lists of identified studies.

Specific search strategies for each database:

Medline, EMBASE and Evidence-based medicine (OVID)

tuberculosis.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

AND

incidence/ or ((inciden* or rate or rates) adj3 (tuberculosis or tb)).ab,ti.

AND

endogenous.ab.ti or (reactivat* or latency or time lag or late progression or late disease or lifetime risk or incubation period or post primary or postmigration or post migration or post immigration or postimmigration or post arrival or postarrival or unclustered or non clustered or latent period or cohort effect or ongoing risk or persistent risk).mp. or Infectious Disease Incubation Period/ or ((years or time) adj3 (arrival or migration)).mp. or (progress* adj3 (tuberculosis or tb or disease or active)).mp. OR (develop* adj3 (tuberculosis or tb or disease or active)).mp. or latent tuberculosis.mp. or tuberculin test.mp. or Interferon-gamma Release Tests.mp. or tuberculin skin test.mp. or t spot.mp or igra.mp or mantoux.mp or quantiferon.mp or qft.mp

limited to English and French

SCOPUS

TITLE-ABS-KEY (tuberculosis) AND (KEY (incidence) OR TITLE-ABS-KEY ((inciden* OR rate OR rates) W/3 (tuberculosis OR tb))) AND TITLE-ABS-KEY (endogenous OR reactivat* OR latency OR "time lag" OR "late progression" OR "late disease" OR "lifetime risk" OR "incubation period" OR "post primary" OR postmigration OR "post migration" OR "post immigration" OR postimmigration OR "post arrival" OR postarrival OR unclustered OR "non clustered" OR "latent period" OR "cohort effect" OR "ongoing risk" OR "persistent risk" OR "infectious disease incubation period" OR "latent tuberculosis" OR "tuberculin test" OR "Interferon-gamma Release Tests" or "tuberculin skin test" or "t spot" or "t-spot" OR "igra" OR "mantoux" OR "quantiferon" OR "qft" OR

((years OR time) W/3 (arrival OR migration)) OR

((progress*) W/3 (tuberculosis OR tb OR disease OR active)) OR

((develop*) W/3 (tuberculosis OR tb OR disease OR active))) AND (LIMIT-TO (LANGUAGE , "English")) OR LIMIT-TO (LANGUAGE , "French"))

CINAHL

tuberculosis

AND

(MH "incidence" or ((inciden* or rate or rates) N3 (tuberculosis or tb)))

AND

endogenous OR reactivat* OR latency OR "time lag" OR "late progression" OR "late disease" OR "lifetime risk" OR "incubation period" OR "post primary" OR postmigration OR "post migration" OR "post immigration" OR postimmigration OR "post arrival" OR postarrival OR unclustered OR "non clustered" OR "latent period" OR "cohort effect" OR "ongoing risk" OR "persistent risk" OR "infectious disease incubation period" OR "latent tuberculosis" OR "tuberculin test" or "Interferon-gamma Release Tests" or "tuberculin skin test" or "t spot" or "igra" or "mantoux" or "quantiferon" or "qft" OR ((years OR time) N3 (arrival OR migration)) ((progress*) N3 (tuberculosis OR tb OR disease OR active)) OR ((develop*) N3 (tuberculosis OR tb OR disease OR active))

limited to English and French

Global Health

(((tuberculosis) AND (inciden* or rate or rates) AND (((title:(endogenous) OR ab:(endogenous)) OR reactivat* OR latency OR "time lag" OR "late progression" OR "late disease" OR "lifetime risk" OR "incubation period" OR "post primary" OR postmigration OR "post migration" OR "post immigration" OR postimmigration OR "post arrival" OR postarrival OR unclustered OR "non clustered" OR "latent period" OR "cohort effect" OR "ongoing risk" OR "persistent risk" OR "years after migration" OR "years after arrival" OR "time after migration" or "time after arrival" OR "progression to tuberculosis" OR "progression to active" OR "progression to disease" OR "progressing to tuberculosis" OR "progressing to active" OR "progressing to disease" OR "progressed to tuberculosis" OR "progressed to active" OR "progressed to disease" OR "developed tuberculosis" OR "developed active" OR "developed disease" OR "developing tuberculosis" OR "developing active" OR "developing disease" OR "developed tuberculosis" OR "developed active" OR "developed disease" OR "development of tuberculosis" OR "development of active" OR "development of disease" OR "latent tuberculosis" OR "tuberculin test" OR "Interferon-gamma Release Tests" OR "tuberculin skin test" OR "t spot" OR "t-spot" OR "igra" OR "mantoux" OR "quantiferon" OR "qft")))

I forgot to limit the Global Health search to English and French language articles and so did this during title, abstract and full-text review.

Web of Science

TS=(tuberculosis)

AND

TS= (inciden* OR rate OR rates)

AND

(TS= (endogenous OR reactivat* OR latency OR "time lag" OR "late progression" OR "late disease" OR "lifetime risk" OR "incubation period" OR "post primary" OR postmigration OR "post migration" OR "post immigration" OR postimmigration OR "post arrival" OR postarrival OR unclustered OR "non clustered" OR "latent period" OR "cohort effect" OR "ongoing risk" OR "persistent risk" OR "latent tuberculosis" OR "tuberculin test" OR "Interferon-gamma Release Tests" OR "tuberculin skin test" OR "t spot" OR "t-spot" OR "igra" OR "mantoux" OR "quantiferon" OR "qft") OR

TS= ((years OR time) NEAR/3 (arrival OR migration)) OR

TS=((progress*) NEAR/3 (tuberculosis OR tb OR disease OR active)) OR

TS= ((develop*) NEAR/3 (tuberculosis OR tb OR disease OR active)))

limited to English and French

PROQUEST

all(tuberculosis) AND (su(incidence) OR ((inciden* OR rate OR rates) NEAR/3 (tuberculosis OR tb))) AND ((ab(endogenous) OR ti(endogenous)) OR all(reactivat* OR latency OR "time lag" OR "late progression" OR "late disease" OR "lifetime risk" OR "incubation period" OR "post primary" OR postmigration OR "post migration" OR "post immigration" OR postimmigration OR "post arrival" OR postarrival OR unclustered OR "non clustered" OR "latent period" OR "cohort effect" OR "ongoing risk" OR "persistent risk" OR "latent tuberculosis" OR "tuberculin test" OR "Interferon-gamma Release Tests" OR "tuberculin skin test" OR "t spot" OR "igra" OR "mantoux" OR "quantiferon" OR "qft" OR ((years OR time) NEAR/3 (arrival OR migration)) OR ((progress*) NEAR/3 (tuberculosis OR tb OR disease OR active)) OR ((develop*) NEAR/3 (tuberculosis OR tb OR disease OR active))))

limited to English and French

Table S1 Description of cohort studies included in the review that documented late reactivation following TST conversion or *Mycobacterium tuberculosis* exposure (This is a more detailed version of Table 2 in the manuscript).

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years)	LTBI screening method (method: cut-off)	Active or passive follow up; numbers observed over time given	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with conversion at study entry	Number of cases beyond two years	Approximate TB incidence per 100,000 per annum (setting and years)	Risk of bias assessment (low, moderate, serious, critical)
Prospective cohort studies documenting TB progression following TST conversion (some populations had a known exposure)														
[P]Myers *et al.* 1964[1]	1921-1941; 1960	Minneapolis, Minnesota, USA	Child contacts	6-12	NS, ~16-39	Pirquet initially and then Mantoux in 1928, 0·1mg then 1·0mg; edema or induration ≥5-10mm at least one year after negative test (up to 12 yrs)	Active, some appeared to be radiologically monitored; N, only total numbers lost given	Y``	N	N	11/154, four demonstrated primary infiltrates, but they resolved initially without symptoms	6	~250-62 (New York, 1921-1960)[2]	Serious/Critical: screening; reinfection;
[P]Myers *et al.* 1965[3]	1921-1941; 1960	Minneapolis, Minnesota, USA	Lymanhurst School and Health Center, exposure not stated	13-17	NS, ~19-39	OT scarification method from 1921, puncture method in 1927, intracutaneous in 1928, 0·1mg then 1·0mg; edema or induration ≥5-10mm at least one year after negative test (up to 8yrs)	Active, some appeared to be radiologically monitored; N, only total numbers lost given	Y`	N	N	11/129	6	~250-62 (New York, 1921-1960)[2]	Critical: screening; reinfection;
[P]Meyer 1949[4]	1929-1944;1947	Oslo, Norway	Residents, some with a known exposure	4-20+	<17	Most Pirquet, Norwegian tuberculin· Where quantified: ≥3mm infiltration or ≥4mm rubor· < 5 yrs between negative and positive but <25% >2 yrs·	Active; Y	Y``	N	N	Progressive pulmonary, extrathoracic and exudative pleuritic TB - 98/889	32	Pulmonary TB: ~280-125 (Norway, 1929-1944)[5]	Serious/Critical: screening; reinfection;
Madsen *et al.* 1942[6]	1934 and 1936; 1940	Copenhagen, Denmark	Medical students, Denmark University. High school students at technical college	NS	<5	Mantoux 1TU and 100TU PPD from 1935: 10mm at 48hrs or 8mm at 72hrs	Active, some had repeated roentgenogram; Y	Y`	N	N	11 with X-ray changes/167	0	New cases: 163-70 (1921-1940)[7] cited by[8]	Critical: screening; reinfection; outcome
[P]Badger & Ayvazian 1948[9]	1932-1943; 1948	Boston, USA	Nurses, presumably exposed	NS	5-15	Saranac OT: NS	Active, TST and roentgenogram every six months; N	Y`	NS	N	40/285	7	~130-170 (New York, 1932-1948)[2]	Critical: reinfection; missing outcome
[P]Daniels *et al.* 1948[10]	1934-1943; 1944	England	Nurses, exposed	18-25	5	Mantoux reaction <1·0mg: 5mm	Active, annual CXR; Y	Y`	N	N	44/347	3	~130-145 (London, 1935-1944)[2]	Critical: reinfection

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; TST=tuberculin skin test; USA=United States of America; NS=not stated; N=No; Y=Yes; OT=Old Tuberculin; TU=tuberculin units; PPD=purified protein derivative.

[P] Results are plotted and appear either in the figures below or in the main manuscript.

`` Cohort known to include those with certain CXR abnormalities.

` Unclear whether cohort included those with CXR abnormalities, in the case of several of the pre-chemotherapy studies the development of primary foci/calcifications was discussed in detail.

[n] Only those with normal/negative/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or the study presented progression rates separately in those with and without abnormalities.

Table S1 Continued. Description of cohort studies included in the review that documented late reactivation following TST conversion or *Mtb* exposure (This is a more detailed version of Table 2 in the manuscript).

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years, unless stated)	Follow-up (years)	LTBI screening method (method: cut-off)	Active or passive follow up; numbers observed over time given	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with conversion at study entry	Number of cases beyond two years	Approximate TB incidence per 100,000 per annum (setting and years)	Risk of bias assessment (low, moderate, serious, critical)
Prospective cohort studies documenting TB progression following TST conversion (some populations had a known exposure)														
[P]Hertzberg 1948[11]	1936-1946; NS	Oslo, Norway	Residents with varying exposure	all	<10, small number "over 10"	Pirquet to 1944, then Mantoux 1mg OT: conversion to ≥10mm	"We are in touch with practically all persons"	Y`	N	N	727/1,829	Males:14 Females: 16	Pulmonary TB: ~120 (Norway, 1948)[5]	Critical: screening; reinfection; missing
[P]Gedde-Dahl 1952[12]	1937-1944; 1945	Kinn, Norway	Residents, some with household exposure	all	mean: 3·8, from TST negative test	Pirquet: Danish tuberculin 3mm; Norwegian plus adrenalin 4mm; Danish plus adrenalin 4mm	Active; Y	Y`	NS	N	32 "progressive pulmonary"/214	3	TB morbidity from pulmonary TB: 300 & 201 (Kinn, 1937-40 &-1941-44)	Critical: screening; reinfection
[P]Hyge *et al.* 1956[13]	1942; 1954	Denmark	Female school children with "massive exposure to infection"	12-18	12	Mantoux <100TU: NS	Active; N	Y``	N	N	Evidence of primary TB: 41/70 Post primary progressive pulmonary TB: 14/70	5	Respiratory TB: 146-66 (Copenhagen, 1947-1954)[14]	Critical: reinfection; missing
[P]Sutherland 1967;[15] Sutherland 1968;[16]Styblo 1991[8]	1951-52; NS	Great Britain	School children	14-15·5	10	TST: 0-4mm to 100TU to ≥5mm to 3 TU	Active with CXR and TST for 8-10 years following start of trial; N	Y[n]	N	N	243/2085[*] (≥6mm)	NS, ~43[*]	~145-~30 (London, 1951-1972)[2]	Serious: screening; reinfection; missing; outcome

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; NS=Not stated; OT=Old Tuberculin; Y=Yes; N=No; TST=tuberculin skin test; TU=tuberculin units.

[P] Results are plotted and appear either in the figures below or in the main manuscript.

[*] We did not have information on the number of TB cases occurring over time. Figure 14 in Styblo 1991[8] shows the proportion of all cases (n=243) that reactivated over time, with an unspecified TST cut-off. We used this to estimate numbers of TB cases over time. To calculate reactivation rates we assumed a 6+mm cut-off and used denominators from Table 1 in Sutherland 1968,[16] with 125 cases added to account for those participants that developed TB before conversion was observed (this figure was stated in the provisional analysis[15]), i.e. conversion was assumed for 125 participants.

` Unclear whether cohort included those with CXR abnormalities, in the case of several of the pre-chemotherapy studies the development of primary foci/calcifications was discussed in detail.

`` Cohort known to include those with certain CXR abnormalities.

[n] Only those with normal/negative/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or the study presented progression rates separately in those with and without abnormalities.

Table S1 Continued. Description of cohort studies included in the review that documented late reactivation following TST conversion or Mtb exposure (This is a more detailed version of Table 2 in the manuscript).

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years, unless stated)	Follow-up (years)	LTBI screening method (method: cut-off)	Active or passive follow up; numbers observed over time given	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with conversion at study entry	Number of cases beyond two years	Approximate TB incidence per 100,000 per annum (setting and years)	Risk of bias assessment (low, moderate, serious, critical)
Prospective cohort studies documenting TB progression following TST conversion (some populations had a known exposure)														
[P]Ferebee & Mount 1962;[17]Ferebee 1970[18]	1957-1960; NS	USA	Household contacts of active cases	>2months	10 (incomplete after 7)	TST 5TU: conversion to ≥5mm at 12 months	Active (incomplete after seven years), and repeated CXR at 12 months; N	Y``	N	N	32/867	13	41-32 (USA, 1956-1959)[19]	Moderate: selection screening; reinfection; missing
[P]Debre *et al*. 1973[20]	1959-1966; 1969	France	Residents of France, mostly young children	77% 5-14; 19% 15-19; 4% 20-24	3-10	"recent conversion of their tuberculin test"	Active with annual CXR and bacteriological examination; Y	Y[n]	N	N	24/1451	13	60 (France,1972)[21]	Moderate/ Serious: screening; reinfection
[P]Veening 1968[22]	1960; 1967	Netherlands	Male Netherlands Navy recruits exposed to open TB	18-20	7	Mantoux Danish PPD RT 23+ tween, 1 TU in 0·1 mI; NS	Active; Y	Y`	NS	N	12/128	3 from 0-4 years, 0 beyond	New cases: 42 (Netherlands, male 15-19 year olds, 1961)[8]	Moderate: reinfection; outcome
Stead 1987;[23] Stead & Dutt 1989[24]	1979-1987; NS	Arkansas, USA	Nursing home residents	≥50	NS, presumably <8	TST 5 units in 0·1ml: ≥12mm from last negative to first positive	"More than 70% of all residents have been retested several times"	Y``	NS	N	89/965	9 in second to fourth years	234 (study setting)[24] 12-9 (USA, 1956-1959)[19]	Critical: selection; reinfection; missing

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; USA=United States of America; TST=tuberculin skin test; TU=tuberculin units; Y=Yes; N=No; PPD=purified protein derivative; NS=not stated.

[P] Results are plotted and appear either in the figures below or in the main manuscript.

`` Cohort known to include those with certain CXR abnormalities.

[n] Only those with normal/negative/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or the study presented progression rates separately in those with and without abnormalities.

` Unclear whether cohort included those with CXR abnormalities.

Table S1 Continued. Description of cohort studies included in the review that documented late reactivation following TST conversion or *Mycobacterium tuberculosis* exposure (This is a more detailed version of Table 2 in the manuscript).

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years)	LTBI screening method (method: cut-off)	Active or passive follow up; numbers observed over time given	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Number of cases beyond two years	Approximate TB incidence per 100,000 per annum (setting and years)	Risk of bias assessment (low, moderate, serious, critical)
Cohorts with LTBI followed from *Mycobacterium tuberculosis* exposure														
[P]Myers *et al.* 1963[25]	1921-1941; 1960	Minneapolis, Minnesota, USA	Child contacts	<6, mean=3	NS, mean age of 3 to mean age of 32, ~16-39	Pirquet initially and then Mantoux in 1928, 0·1mg then 1·0mg; edema or induration of 5 to 10mm or equivalent diameter	Active; only total numbers lost given	Y`	N	N	41/599 with no pulmonary infiltration	19	~250-62 (New York, 1921-1960)[2]	Critical: screening; reinfection; outcome
[P]Hertzberg 1948[11]	1936-1946;1946	Oslo, Norway	Residents exposed to destructive or non-destructive TB in family	19·9% 0-2; 50·1% 3-12; 9·5% 13-16; 12·2% 17-25; 8·4% 25+·	>10, small number "over 10"	Pirquet to 1944, then Mantoux 1mg old tuberculin: ≥10mm	"We are in touch with practically all persons"	Y`	N	N	272/498 males; 280/545 females	Males: 7 Females: 15	Pulmonary TB: ~120 (Norway, 1948)[5]	Critical: screening; reinfection; missing
[P]Hyge *et al.* 1956[13]	1942; NS	Denmark	Female school children	12-18	12	Mantoux <100TU: NS	Active; N	Y``	N	N	Progressive pulmonary TB: 9/105	5	Respiratory TB: 146-66 (Copenhagen, 1947-1954)[14]	Critical: reinfection; missing
[P]Ferebee & Mount 1962;[17]Ferebee 1970[18]	1956-1959; NS	USA	Household contacts of active cases	3·2% <15; 17·2% 15-34; 33·6% 35-54; 46·0% 55+·	10 (incomplete after 7)	TST 5TU: ≥5mm	Active (incomplete after seven years), and repeated CXR at 12 months; N	Y``	N	N	472 (to eight years, 479 to ten)/7,744	99	41-32 (USA, 1956-1959)[19]	Low-Moderate: reinfection; missing
[P]Dobler & Marks 2013[26]	2000-2009;2009	Sydney West and South West, Australia	Contacts of "TB patients"	Mean: 33 17·6% 0-14; 39·7% 15-34; 42·7% ≥35	mean: 4·6 std dev: 2·9	TST: ≥10mm	Passive	NS	NS	N	Time after screening: 3 months to 2 years, 30/3,942; ≥2 years, 8/3,912	8	5-6 (Australia, 2000-2009)[27,28]	Moderate/Serious: selection; reinfection; missing; outcome

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; USA=United States of America; NS=not stated; Y=Yes; N=No; TU=tuberculin units; NS=not stated; TST=tuberculin skin test; std dev=standard deviation.
[P] Results are plotted and appear either in the figures below or in the main manuscript.
`` Cohort known to include those with certain CXR abnormalities.
` Unclear whether cohort included those with CXR abnormalities.

Table S1 Continued. Description of cohort studies included in the review that documented late reactivation following TST conversion or *Mycobacterium tuberculosis* exposure (This is a more detailed version of Table 2 in the manuscript).

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years, unless stated)	LTBI screening method (method: cut-off)	Active or passive follow up; numbers observed over time given	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Number of cases beyond two years	Approximate TB incidence per 100,000 per annum (setting and years)	Risk of bias assessment (low, moderate, serious, critical)
Cohorts with LTBI followed from *Mycobacterium tuberculosis* exposure														
[P]Reichler *et al.* 2018[29]	2002-2006; 2011	Nine sites, USA and Canada	Close contacts of culture-positive pulmonary TB ≥15 years old[#]	NS	some <2, some <8	TST: ≥5mm	Passive	Y`	NS	N	89/499	3	~5 (USA, 2002-2006)[19] ~5 (Canada, 2002-2006)[30]	Moderate/Serious: selection; reinfection; missing
[P]Sloot *et al.* 2014[31]	2002-2011; 2012	Amsterdam, Netherlands	Contacts of pulmonary cases	16·3% 0-14; 53·9% 15-44; 25·9% 45-64; 3·8% ≥65·	<11 (incomplete 5-11)	IGRA TST 2TU PPD RT23: ≥10mm	Passive	Y`	VC	45%	14/739 (+57 coprevalent[*])	2	~9-6 (Netherlands, 2002-2012)[32]	Moderate/Serious: selection; screening; reinfection; missing
[P]Erkens *et al.* 2016[33]	2005-2013; 2014	Netherlands	Contacts of infectious cases	2% <5; 31% 5-24; 38% 25-44	≥5 (IQR 3·0-7·4)	TST and confirmatory IGRA (since 2010): NS	Passive	Y`	VC	N	45/2251 (+4 cases developed TB <100 days after LTBI diagnosis)	9	5 (Netherlands, 2015)[33]	Serious: confounding; screening; reinfection; missing

Abbreviations: LTBI=latent tuberculosis infection; Mtb=Mycobacterium tuberculosis; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; USA=United States of America; NS=not stated; TST=tuberculin skin test; Y=Yes; N=No; IGRA=interferon-gamma release assay; TU=tuberculin units; PPD=purified protein derivative; VC= variable, and considered in analysis; IQR=interquartile range.

[P] Results are plotted and appear either in the figures below or in the main manuscript.

[#] "shared air space with an individual with pulmonary tuberculosis in the household or other indoor setting for >15 hours per week or >180 hours total during an infectious period, defined as the interval from 3 months before collection of the first culture-positive sputum specimen or the date of onset of cough (whichever was longer) through 2 weeks after the initiation of appropriate antituberculosis treatment"[29]

[*] ≤ 180 days after index diagnosis.[31]

` Unclear whether cohort included those with CXR abnormalities.

Table S1 Continued. Description of cohort studies included in the review that documented late reactivation following TST conversion or *Mycobacterium tuberculosis* exposure (This is a more detailed version of Table 2 in the manuscript).

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years, unless stated)	LTBI screening method (method: cut-off)	Active or passive follow up; numbers observed over time given	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Number of cases beyond two years	Approximate TB incidence per 100,000 per annum (setting and years)	Risk of bias assessment (low, moderate, serious, critical)
Cohorts with LTBI followed from *Mycobacterium tuberculosis* exposure														
[P]Altet *et al.* 2015[34]	2007-2009; NS	Barcelona, Spain	Contacts (from first circle) of smear positive cases	6·1% 0-4; 28·2% 5-14; 40·1% 15-35; 25·8% ≥35	4	QuantiFERON TST 2TU PPD RT23: previously positive excluded, ≥5mm	Active	Y[n]	VC	N	QuantiFERON positive: 14/81 TST ≥5mm: 14/340 (+~70-80 identified during contact study)	0	~17-19 (Spain, 2007-2009)[32]	Low/Moderate selection; reinfection
[P]Abubakar *et al.* 2018[35,36]^ UK Predict study	2010-2015; 2016	London, Birmingham and Leicester, UK	Recent contacts of active TB	58·6% 16-35; 41·3% >35	median for both contacts and migrants in study: 2·9 (range 21 mths to 5·9 yrs)	QFT-GIT, T-SPOT·TB, Mantoux TST; multiple cut-offs	Telephone contact at 12 and 24 months and passive beyond	Y`	VC	N	TSpot·TB: pos 31/648, neg 20/2,916 QFT-GIT: pos, 30/793, neg 21/2,771 TST: ≥5mm 43/1,704; <5mm 8/1860; ≥10mm 38/1323; <10mm 13/2,241; ≥15mm 34/899; <15mm 17/2,665	5 (positive to T-SPOT·TB QFT-GIT or TST 15mm)	14-10 (United Kingdom, 2010-2016)[32]	Moderate: selection; screening; reinfection; missing
Heiden *et al.* 2017[37]	2013;2017	3-hour flight Turkey to Germany	Crew and passengers	8·7% 0-14; 69·1% 15-49; 22·2% ≥50	3 yrs, 8 mths	TST or IGRA; >10mm TST or conversion > 5 mm	Passive	"supposed to"	VC	N	0/14	0	20-17 (Turkey, 2013-2017)[38] ~6-8 (Germany, 2013-2017)[38]	Serious: screening; missing

Abbreviations: LTBI=latent tuberculosis infection; Mtb=Mycobacterium tuberculosis; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; NS=Not stated; TST=tuberculin skin test; TU=tuberculin units; PPD=purified protein derivative; VC= variable, and considered in analysis; QFT-GIT=QuantiFERON-TB Gold In-Tube test; pos=positive; neg=negative; mths=months; yrs=years; IGRA=interferon-gamma release assay;

[P] Results are plotted and appear either in the figures below or in the main manuscript.

^ Additional data was obtained from the study's corresponding author, so data included in the table and Figure 2 may not appear in the study manuscript.

` Unclear whether cohort included those with CXR abnormalities.

[n] Only those with normal/negative/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or the study presented progression rates separately in those with and without abnormalities.

Table S2 Probability of reactivation in cohorts of studies that documented late reactivation following conversion or *Mycobacterium tuberculosis* exposure, over specific time periods.

Publication (first author and year of publication for clarity)	Age (years)	Sample size: Number of TB cases/number with conversion or reactivity at study entry	LTBI screening method and cut-off	Approximate annual TB incidence in study setting per 100,000 persons.	Number and percentage of all TB cases occurring beyond two years	Probability of reactivation, as a percentage and number of TB cases, in brackets, over specific time periods from conversion/exposure. The number of participants under observation is used as the denominator where possible, but many studies did not actively observe participants for the whole follow-up period or, if they did, didn't give the numbers followed up overtime, so the term "probability" is potentially, technically inappropriate: see those marked *.				
						0-2 years	2-5 years	5-10 years	10-20 years	20-30 years
						0-1 years / 1-4 years		5-8 years	10-12 years	
						3mths to 2 years		5-7 years	10-17 years	
									10-15 years	
Prospective cohort studies documenting TB progression following TST conversion (some populations had a known exposure)										
Myers 1964[1]	6-12	10/195†	TST ≥5-10mm	250-60[2]	6 (54·5)	2·06 (4)	1·05 (2)	1·59 (3)	0.00 (0)	0.51 (1)‡
Myers 1965[3]	13-17	11/129§	TST ≥5-10mm	250-60[2]	6 (54·5)	3·12 (4)	1·61 (2)	1·63 (2)	0.83 (1)	0.83 (1)‡
Meyer[4]	0-3	7/~57	TST ≥3-4mm	280-125 (ptb)[5]	0 (0)	21·09 (7)	0·00 (0) ??			
Meyer[4]	4-12	15/~339	TST ≥3-4mm	280-125 (ptb)[5]	3 (20·0)	3·55 (12)	0·36 (1)	1·72 (2)	0.00 (0)	
Meyer[4]	13-19	42/~236	TST ≥3-4mm	280-125 (ptb)[5]	16 (38·1)	11·22 (26)	6·58 (13)	2·69 (3)	0.00 (0)	
Meyer[4]	20+	34/~240	TST ≥3-4mm	280-125 (ptb)[5]	32 (32·7)	8·92 (21)	5·98 (12)	1·54 (1)	0.00 (0)	
Madsen[6]	~18	11 pulm/167	TST ≥8-10mm	163-70[7,8]	0 (0)	6·62 (11)	0·00 (0)			
Badger[9]	NS	40/285	TST: NS	130-170[2]	7 (17·5)	11·84 (33)	1·18 (3)	1·24 (3)	1.44 (1)	
Daniels[10]	18-25	44/347	TST ≥5mm	130-145[2]	3 (6·8)	12·47 (41)	6·15 (3)			
Hertzberg[11]	0-2	102/172	TST ≥10mm	120 (ptb)[5]	2 (2·0)	61·53 (100)	8·16 (2)	0·00 (0)	0.00 (0) ??	
Hertzberg[11]	3-12	288/726	TST ≥10mm	120 (ptb)[5]	10 (3·5)	40·15 (277)	4·68 (7)	11·54 (3)	50.00 (1) ??	
Hertzberg[11]	13-16	114/325	TST ≥10mm	120 (ptb)[5]	7 (6·1)	34·96 (107)	12·32 (5)	21·98 (2)	0.00 (0) ??	
Hertzberg[11]	17-24	154/350	TST ≥10mm	120 (ptb)[5]	8 (5·2)	48·36 (146)	20·34 (7)	11·11 (1)	0.00 (0) ??	
Hertzberg[11]	25+	69/256	TST ≥10mm	120 (ptb)[5]	2 (2·9)	28·55 (67)	6·90 (2)	0·00 (0)		
Gedde-Dahl[12]	0-14	4 pulm/62	TST ≥3 or 4mm	300-200 (ptb)	2 (50·0)	3·92 (2)	6·52 (2)			
Gedde-Dahl[12]	15-29	28 pulm/152	TST ≥3 or 4mm	300-200 (ptb)	3 (9·1)	20·53 (27)	3·00 (1)			
Hyge[13]	12-18	55 pulm TB/70	TST NS	145-65 (respTB)[14]	5 (9·1)	13·04 (9)	3·31 (2)	0·00 (0)	5.08 (3)	
Sutherland[15,16]	14-15·5	243/2,085¶	TST ≥6mm¶	145-30[2]	~44 (~18·1)¶	9·78 (~200)¶	1·79 (~34)¶	0·52 (~10)¶		
Ferebee 1962[17,18]*	All	32/867	TST ≥5mm	40-30[19]	13 (40·6)	2·20 (19)	0·83 (7)	0·71 (6)		
Debre[20]	5-24	24/1,451	TST NS	60[21]	13 (54·2)	0·76 (11)	0·44 (6)	1·04 (7)		
Veening[22]	18-20	12/128	TST NS	40[8]	yrs 1-4: 3 (25·0)	7·03 (9)	2·52 (3)	0·00 (0)		
Stead[23,24]	≥50	89/965	TST ≥12mm change	10[19]	yrs 1-4: 9 (?)	8·29 (80)	1·02 (9)	0 ??		
Cohorts with LTBI followed from recent *Mtb* exposure										
Myers 1963[25]	0-5	41/599‖	TST ≥5-10mm	250-60[2]	19 (46·3)	3·7 (22)	4·9 (7)	0·2 (1)	1.9 (11)	0.00 (0)‡
Hertzberg[11]	0-2	143/207	TST ≥10mm	120 (ptb)[5]	3 (2·1)	72·68 (140)	7·33 (3)	0·00 (0)	0.00 (0) ??	
Hertzberg[11]	3-12	295/ 522	TST ≥10mm	120 (ptb)[5]	10 (3·4)	58·19 (285)	4·20 (6)	9·55 (4)	0.00 (0) ??	
Hertzberg[11]	13-16	49/99	TST ≥10mm	120 (ptb)[5]	1 (2·0)	49·49 (44)	14·97 (4)	8·33 (1)	0.00 (0) ??	
Hertzberg[11]	17-24	48/127	TST ≥10mm	120 (ptb)[5]	3 (6·3)	40·71 (45)	15·23 (3)	0·00 (0)	0.00 (0) ??	
Hertzberg[11]	25+	17/88	TST ≥10mm	120 (ptb)[5]	1 (5·9)	21·16 (16)	0·00 (0)	16·67 (1)	0.00 (0) ??	
Hyge[13]	12-18	9 /105	TST NS	145-65 (respTB)[14]	5 (55·5)	3·85 (4)	1·98 (2)	3·06 (3)	0.00 (0)	
Ferebee 1962[17,18]*	all	472/7744	TST ≥5mm	40-30[19]	99 (46·0)	4·83 (373)	0·90 (66)	0·45 (33)		
Dobler[26]*	all	38/3942	TST ≥10mm	5-6[27,28]	8 (21·1)	0·87 (30)	0·20 (8) ??			
Reichler[29]*	NS	89/499	TST ≥5mm	5[19,30]	3 (3·6)	17·78 (86)††	0·77 (4)	0·00 (0)		
Sloot[31]*	all	14**/739	IGRA, TST ≥10mm	9-6[32]	2 (14·3)	9·44 (69)††	0·16 (1)	0·26 (1)		
Erkens[33]*	all	41**/2251	IGRA, TST NS	5[33]	9 (20·0)	1·65 (36)††	0·36 (6)	0·44 (3) ??		
Altet[34]	all	14**/81	IGRA	15-20[32]	0 (0)	18·06 (14)**	0·00 (0)			
Altet[34]	all	14**/340	TST ≥5mm	15-20[32]	0 (0)	4·16 (14)**	0·00 (0)			

Abbreviations: LTBI=latent tuberculosis infection; TST=tuberculin skin test; TB=tuberculosis; *Mtb=Mycobacterium tuberculosis*; USPHS=United States Public Health Service; MRC=Medical Research Council; mths=months; yrs=years; IGRA=interferon-gamma release assay; pulm=pulmonary TB; ptb=pulmonary TB; respTB=respiratory TB.

Colour of cells: Purple=Results from studies performed in the pre-antibiotic era; Blue=Results from studies in first two years after conversion or exposure; Pink=Results from studies for which a certain proportion of participants received preventive therapy.

† We excluded one TB case from this paper because Myers *et al.* 1964 reported that they had converted at the age of 19yrs. It was also reported that 6 died of causes other than TB, and 41 were lost to follow up before the study end, but it is unknown when these cases were lost, so these losses could not be incorporated into rate calculations.

‡ The number of participants who were followed for different periods of time was not reported, but presumably the minimum length of follow-up was 19 years (given study recruitment ended in 1941 and the study end was 1960), with the exception of cases that died or were lost to follow-up.

§ Myers *et al.* 1965 reported that 6 died of causes other than TB and 14 were lost to follow up before study end, but it is unknown when these cases were lost, so this loss could not be incorporated into rate calculations.

¶ We did not have information on the number of TB cases occurring over time. Figure 14 in Styblo 1991[8] shows the proportion of all cases (n=243) that reactivated over time, with an unspecified TST cut-off. We used this to estimate numbers of TB cases over time. To calculate reactivation rates we assumed a 6+mm cut-off and used denominators from Table 1 in Sutherland 1968,[16] with 125 cases added to account for those participants that developed TB before conversion was observed (this figure was stated in the provisional analysis[15]), i.e. conversion was assumed for 125 participants.

‖ Myers *et al.* 1963 reported that 22 died of causes other than TB and 102 were lost to follow up before study end, but it is unknown when these cases were lost, so this loss could not be incorporated into rate calculations.

** TB cases identified ≤ 180 days after index diagnosis;[31] <100 days after screen;[33] or during contact tracing[34] were not included in the numerator.

†† Includes cases identified during contact tracing,[33] ≤180 days after index diagnosis;[31] <100 days after screen;[33] or during contact tracing.[34]

Note: we were unable to present the results for Abubakar *et al.* 2018[35,36]^ as we didn't have sufficient data, and the study by Heiden *et al.* 2017[37] was not presented because no TB progression was documented at all.

?? Upper cut-offs were not given. For example, results were only shown for 2+ years, 5+ years, 10+ years, etc.

Table S3 Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Myers *et al.* 1964[1]	1921-1944; NS	Minneapolis, Minnesota, USA	Children	6-12	~16-39, active	Pirquet, then Mantoux in 1928, 0·1mg then 1·0mg; edema or induration ≥5-10mm or equivalent	Y``	N	N	4/108 with pulmonary infiltration (3,196 pyrs); 62/1,583 with no pulmonary infiltration (42,233pyrs)	With pulmonary infiltration: 125 With no pulmonary infiltration: 147	~250-65 (New York, 1921-1961)[2]
[P]Myers *et al.* 1965[3]	1921-1941; 1960	Minneapolis, Minnesota, USA	Lymanhurst School and Health Center	13-17	~19<39, active	OT scarification, puncture method in 1927 intracutaneous in 1928, 0·1mg then 1·0mg; edema or induration ≥5-10mm	Y`	N	N	55/715	~263	~250-62 (New York, 1921-1960)[2]
Pope 1939[39]	1924-1934; 1936	Massachusetts, USA	School children	6-19 mean : 11·43	<12 mean : 3·4, active	Pirquet	Y[n]	N	N	241/99,769	242	~200-138 (New York, 1924-1936)[2]
Heimbeck 1938[40]	1924; NS	Oslo, Norway	Residents	13-24	mean : ~4, "follow them up methodically"	Pirquet, tuberculin slit into epidermis, not drilled	NS	NS	N	13-24 yrs old: 14/467 (2,111pyrs) females; 10/447 (1,831pyrs) males; 20-30 yrs old: 6/403 (1,364pyrs) females.	13-24 yr old females: 660; 13-24 yr old males: 550; 20-30 yr old females: 440	Pulmonary TB: ~280 (Norway, 1924)[5]
Scheel 1935[41]	1926; 1935	Oslo, Norway	Medical students	16-25	3	von Pirquet tuberculin test, and a few Mantoux	Y`	N	N	361	1,350	Pulmonary TB: ~280-190 (Norway, 1926-1935)[5]
Myers *et al.* 1941[42]	1929-1936	University of Minnesota, USA	Medical school students	NS	4, retrospective, questionnaires sent to past students	Tuberculin test; NS	Y`	N	N	2/160	313	~173-138 (New York, 1929-1936)[2]
Myers *et al.* 1940[43]	1929; 1938	University of Minnesota (?) USA	Nursing students	NS	3, retrospective, questionnaires sent to past students	Tuberculin test; NS	Y`	N	N	7/281	yr 1: 356 yr 2: 714; yr 3: 1439; total 834	~173-128 (New York, 1929-1938)[2]
Hastings & Behn 1941[44]	1929-1935;	Minneapolis, USA	Student nurses	NS	3, unclear, but likely to be active	TST; NS	Y`	N	N	1/198	168	~175-144 (New York, 1928-1935)[2]
Geer 1934[45]	1928; 1930	Ancker Hospital, Mississippi, USA	Student nurses	NS	2, unclear, but likely to be active	0·1mgm and then 1mgm OT intracutaneous	Y`	N	N	1/33·6 (30% of 112)	1488	~175-170 (New York, 1928-1930)[2]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; USA=United States of America; Y=Yes; N=N; OT=Old tuberculin; NS=Not stated; yr=year; pyrs=person years; yrs=years; TST=tuberculin skin test; USA=United States of America.

[P] Results are plotted and appear either in the figures below or in the main manuscript.

`` Cohort known to include those with certain CXR abnormalities.

` Unclear whether cohort included those with certain CXR abnormalities, in the case of pre-chemotherapy studies the development of primary foci/calcifications was sometimes discussed in detail.

[n] Only those with normal/negative/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or progression rates in those with and without abnormalities were presented separately.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Myers *et al.* 1968[46]	1930-1953; NS	Minnesota, USA	Nursing and medicine graduates	NS	NS, presumably <~30, active with periodic roentgenograms during training	NS, tuberculin test	Y`	NS	N	Fairview Hospital: 4 recrudescent pulmonary lesions /82 (2,175 pyrs); St Mary's Nursing: 4 pleural or pulmonary lesions /150 (4,307 pyrs); The Swedish School: 4 pleural or pulmonary lesions 1 pulmonary lesion/151 (4,015 pyrs).	Fairview Hospital School 184; St Mary's School of Nursing 93; The Swedish School of Nursing 25;	~170-65 (New York, 1930-1961)[2]
Heimbeck 1949;[47] Heimbeck 1951[48]	1924-1936; 1948	Oslo, Norway	Residents and nursing students exposed to "massive" infection during training	20	NS, presumably <24, unclear whether active or passive	Tuberculin Pirquet	NS	N	N	Nurses: during training 0-3 yrs 22/668 (1772 pyrs); after graduation 3-NS yrs 18/504 (5,677 pyrs). Residents by sex and age: females 20-22 yrs 4/398 (1126 pyrs); males 20-22 yrs 5/436 (1,244 pyrs); females 22-NS yrs 5/318 (2,698 pyrs); males 22-NS yrs 10/372 (3,346 pyrs)	Nurses: during training 0-3 yrs 1241; after graduation 3-NS yrs 317. Residents by sex and age: females 20-22 yrs 1,005; males 20-22 yrs 402; females 22-NS yrs 185; males 22-NS yrs 299;	Pulmonary TB: ~280-120 (Norway, 1924-1948)[5]
Badger & Ayvazian 1948[9]	1932; 1948	Boston, USA	Nurses	NS, ~18-19	5-15, active, TST and roentgenogram every six months	Saranac Old Tuberculin; NS	Y`	NS	N	31/374	0-2378 (only annual rates provided)	~130-~170 (New York, 1932-1948)[2]
Madsen *et al.* 1942[6]	1934; 1935-1936	Nakshov and Ronne, Denmark	Residents	all	<2, active, some had repeated roentgenography	TST; NS	Y`	N	N	Ronne by age: 7-14 yrs, 4 X-ray change at initial examination/165; 15-35 yrs: 26 X-ray change at initial examination/1112 Nakshov by age: 7-14 yrs, 12 X-ray change at initial examination/441	0 "X-ray changes" in children after initial examination Ronne 15-35 years of age: 92	163-70 (1921-1940)[7] cited by[8]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; NS=not stated; N=No; pyrs=person years; yrs=years; Y=Yes; USA=United States of America.
` Unclear whether cohort included those with certain CXR abnormalities, in the case of pre-chemotherapy studies the development of primary foci/calcifications was sometimes discussed in detail.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Madsen *et al.* 1942[6]	1934, 1936; 1940	Copenhagen, Denmark	University and High School students	"about 18"	<5, active, some had repeated roentgenography	TST; NS	Y`	N	N	17 "X-ray changes" and 10 with TB on gastric lavage or expectorate/2,071	"X-ray changes": 301	163-70 (1921-1940)[7] cited by[8]
Ferguson 1946[49]	1934; 1943	Saskatchewan and Winnipeg, Canada	Female nursing students	~20	mean 2.43, unclear	TST	NS	N	N	5/478 (1,165·2 pyrs)	429	~80-88 (Canada, 1934-1943)[30]
Ferguson 1946[49]	1934; 1943	Saskatchewan, Canada	Saskatchewan Sanatoria employees	mean: 23.3	mean: 1.44, unclear	TST	NS	N	N	13/462 (665·69 pyrs)	1953	~80-88 (Canada, 1934-1943)[30]
Israel and Hethrington 1941[50]	1935; 1939	Philadelphia, USA	Nursing students in Philadelphia General Hospital	17-21	<3, active, fluoroscopy at four-month intervals	0·00002mg. and 0·005mg. PPD tuberculin	Y``	NS	N	Pos to 0·006mg: 23/183 Pos to 0·00002mg: 11/177 Total pos: 34/360	Pos to 0·006mg: 6,050; Pos to 0·00002mg: 2,700; Total pos: 4,320	~150-125 (New York, 1935-1939)[2]
Daniels *et al.* 1948[10]	1935; 1944	England	Nurses; Medical students; controls	18-25	5, active, annual CXR	Mantoux reaction <1·0mg: ≥5mm	Y`	N	N	64/6946. Years of observation: 0-1 yrs: 21/2934; 1-2 yrs 14/1955; 3-4 yrs 15/1216; 4-5 yrs 11/643; 5-6 yrs 3/198	730	~130-145 (London, 1935-1944)[2]
Wright 1941[51]	1936-1937; 1939-1940	Montreal, Canada	Undergraduate nurses in Royal Victoria Hospital	NS	3, CXR and TST six monthly	1/10mg and 1mg of old tuberculin	Y`	N	N	0/36	0	~80-86(Canada, 1936-1940)[30]
Olsen 1956[52]	1936; 1945	Bornholm, Denmark	Residents	all	<9, unclear, isolated island community	Mantoux 1,10 and 100 units, then latterly 3 and 100 units: ≥6mm	NS	N	N	15 pulmonary TB and 2 TB deaths/3,994 (28,264pyrs)	Pulmonary TB: 60	~57 (Bornholm, 1936-1940)[52]
Hertzberg 1948[11]	1936-1946; 1946	Oslo, Norway	Residents without known TB in their families	5·2% 0-2; 38·2% 3-12; 22·4% 13-16; 20·1% 17-25; 14·1% 25+.	>10, "We are in touch with practically all persons"	Pirquet to 1944, then Mantoux 1mg old tuberculin: ≥10mm	Y`	N	N	240/684 males; 245/666 females	Males: 16,173 Females: 17,488	Pulmonary TB: ~120 (Norway, 1948)[5]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; TST=tuberculin skin test; NS=not stated; N=No; pyrs;person years; USA=United States of America; PPD=purified protein derivative; Pos=positive; yrs=years.
`` Cohort known to include those with certain CXR abnormalities.
` Unclear whether cohort included those with certain CXR abnormalities, in the case of pre-chemotherapy studies the development of primary foci/calcifications was sometimes discussed in detail.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
[P]Palmer *et al.* 1957[53]	1940-1951; 1955	San Diego, USA	Male Navy recruits	17-21	3·43-5.16, active & annual roentgenograms whilst in service	TST 5TU: ≥10mm	Y[n]	NS	N	37/5,910	157	~120- 80 (New York, 1940-1955)[2] 52.6-46.6 (USA, 1953-1955)[19]
[P]Sjögren 1976[54]	1941-1946; 1960	Sweden	Male military recruits	mean: 20.9	15, appears to be active	Mantoux, 0·1mg or 1mg of OT; ≥10mm	NR	N	N	10+ mm: 223/ 25,525 <10mm: 39/7,013	10+ mm: 64 <10mm: 42	Males: 140-56; Females: 140-51 (Sweden, 1951-1960)[55]
Härö 1972[56]	1945-1949; follow up 1957-1969	Finland	Males born 1926-1941 found to be natural reactors during BCG mass vaccination.	~4-23	10-20 (no follow-up <10 yrs), passive	Modified Trambusti tuberculin test	NS	N	N	1938 pulmonary TB/82,012 (1,066,156 pyrs)	182	Respiratory TB: Males: 247-120, Females: 130-79 (Finland, 1954-1969)[55]
Stephanopoulos & Costeletos 1957[57]	1946; 1954	Sotira, Greece	Student nurses	NS	<3, active, regular radiology	Mantoux	Y`	NS	N	17/474	1,554	unknown
Gernez-Rieux & Gervois 1973[58]	1949-1951; 1971	Lille, France	School children	6-14	4-9, active, annual CXR & TST	Mantoux 10 TU; >5mm	Y`	N	N	6-12 mm: 1,447 >12mm: 262	306	60 (France,1972)[21]
Comstock 1974[59]	1949-1951; 1969	Puerto Rico	Children	1≤18	mean: 18·87, passive	TST 10TU: ≥6mm	NS	N	N	1,400/82,269	90	Mortality: 179-33 (Puerto Rico-1948-1955)[60]
Frimodt-Miller *et al.* 1964[61]	1950; 1958	Villages surrounding Madanapalle, India	Residents	all	<7, active, periodic CXR surveys	1-10-100 TU and then 5 and 100 TU	Y`	N	N	Pulmonary TB Round I to IV 52/22,403 pyrs; Round I to VI 132/45,846 pyrs; Males round I to IV 37/10,719 pyrs; round I to VI 91/22,007 pyrs; Females round I to IV 15/11,684 pyrs; round I to VI 41/23,839 pyrs.	Round I to IV 232; Round I to VI 288; Males, round I to IV 345; Males, round I to VI 413; Females, round I to IV 129; Females, round I to VI 172.	Pulmonary TB Round I-VI: 171 (India, 1950-1958)[61]
[P]Horwitz *et al.* 1969[62]	1950-52; 1964	Denmark	Residents	15-44	12, passive	TST 5 or 10 TU: ≥6mm	Y[n]	N	N	987 respiratory or pleural TB cases/286,250	29	Respiratory TB ~50-10 (in study areas, 1950-1964)[62]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; USA=United States of America; TST=tuberculin skin test; TU=tuberculin units; Y=Yes; NS=not stated; N=No; OT=old tuberculin; pyrs=person years; AFT=awaiting further results..

[P] Results are plotted and appear either in the figures below or in the main manuscript.

`` Cohort known to include those with certain CXR abnormalities.

` Unclear whether cohort included those with certain CXR abnormalities.

[n] Only those with normal/neg/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or progression rates in those with and without abnormalities were presented separately.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
[P]Comstock 1976[63]; Comstock & Palmer 1966[64]	1950; 1970	Muscogee county, USA	Residents	>5	20, passive	TST 5 TU: ≥5mm	Y[n]	N	N	207/22,027	47	53-18 (USA, 1953-1970)[19]
Comstock 1964[65]	1947; 1959	Muscogee county, USA	School children	NS	12, passive	TST 5 TU and then 100 TU: ≥5mm	NS	N	N	5TU: 24/1,492 100TU: 5/3,768	5TU: 134 100TU: 11	~100 (New York, 1947)[2] 32.4 (USA, 1959)[19]
Large 1965[66]	1951-1961	British Malaya	Male recruits to British Army from Nepal (Gurkhas)	15-17	< 9, possibly active	Heaf Grade1+	Y``	N	N	89/6,280 (29,186pyrs)	304	unknown
[P]MRC 1972 trial,[67] Hart & Sutherland 1977[68]	1951-52; 1971	UK	School children	14-15·5	20, active	TST 3 TU and 100 TU; ≥5mm	Y[n]	N	N	3 TU, 5-14mm 178/8,838; 3 TU, ≥15mm 140/6,866; pos to 100 TU only 56/6,253; neg, unvaccinated 248/12,867.	3 TU, 5-14mm 44; 3 TU, ≥15mm 102; pos to 100 TU only 45; neg, unvaccinated 96.	~145-30 (London, 1951-1972)[2]
[P]Ross & Willison 1971 [69]	1955-1969; 1968	Edinburgh, Scotland	School children	13	<13, passive	Heaf test	N	N	N	Neg 27/328,250 pyrs; Heaf I 2/24,639 pyrs; Heaf II 5/13,990 pyrs; Heaf III & IV 11/17,589 pyrs.	Neg 8; Heaf I 8; Heaf II 36; Heaf III & IV 63.	~120-36 (London, 1955-1968)[2]
Mount & Ferebee 1962^[70]	1956; ~1962	USA	Known close contacts of previous, predominantly, pulmonary TB cases	All	4, active	TST 5TU; ≥5mm	Y	N	N	8/609 in the first year and a possible 3 more cases in the next three years	1,313 in first year after screening	39-31 (USA, 1957-1960)[19]
[P]Ferebee *et al.* 1963;[71] Ferebee 1970[18]	1957-1960; NS	USA	Mental institution patients	All	10, active, incomplete >7	TST 5TU; ≥5mm	Y[n]	NS	N	≥5mm: 49/7074	≥5mm: 69	39-31 (USA, 1957-1960)[19]
[P]Comstock 1974[72]	1958-1969; 1972	USA	Male Navy recruits	95% 17-22	4, active, annual examination	TST 5 TU; ≥10mm	NS	Only in some "Asians"	N	"Blacks": 32/8,810 "Asians": 97/12,467 "Whites": 132/42,547	"Blacks": 93 "Asians": 196 "Whites": 79	36-16 (USA, 1958-1972)[19] Vietnam war?

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; USA=United States of America; TST=tuberculin skin test; TU=tuberculin units; Y=Yes;N=No; NS=not stated; pyrs=person years; UK=United Kingdom; pos=positive; neg=negative.

[P] Results are plotted and appear either in the figures below or in the main manuscript.

`` Cohort known to include those with certain CXR abnormalities.

[n] Only those with normal/neg/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or progression rates in those with and without abnormalities were presented separately.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years, unless stated) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Ross & Willison 1971[73]	1959; 1970	Edinburgh, Scotland	School children	13	<11, passive	Heaf test	N	N	N	Heaf I 5/25,288 pyrs; Heaf II 9/8,019 pyrs; Heaf III 19/6,356 pyrs; Heaf IV 12/2,911 pyrs; All 45/42,574 pyrs.	Heaf I 20; Heaf II 112; Heaf III 299; Heaf IV 412; All 106.	~90-30 (London, 1959-1970)[2]
Lotte *et al.* 1971[74]	1961; 1966	France, Poland, Switzerland (Geneva) and Yugoslavia	School children	6-14	<4 in Geneva and France, 2 in Yugoslavia and 1 in Poland; active, regular radiology	2 TU of PPD RT23, with addition of Tween 80, an initial patch test in Geneva for <10 years; ≥10mm	Y`	N	N	Yugoslavia: 117/58,123 Poland: 55/34,532 France: 67/127,307 Geneva: 48/23,047 (Denominators are the total populations tested, not the reactors)	≥14mm: 320 Recent increase of reactivation ≥18mm: 1190	60 (France,1972).[21] Quoted estimates in text of bacillary TB : 70 (Poland, 1967)[i]; 55 (Yugoslavia, 1967), but noted to be "certainly underestimated"[74]
National Tuberculosis Institute, Bangalore 1974[75]	1961; 1968	Bangalore, India	Villlagers	all	~18 months, active, between TST and radiological surveys I and II.	1 TU RT 23 with 0·05% Tween 80 in 0·1ml; ≥10mm	Y[n]	N	N	Culture positive TB with no CXR abnormality: 0-9mm 15/19,419; 10mm-19mm 2/4,767; 20+mm 12/4,201; 10+mm 14/8,968;	0-9mm 50; 10mm-19mm 27; 20+mm 185; 10+mm 101.	Average annual incidence: 103 (study cohort)[75]
Gothi *et al.* 1976[76]	1961-1963; 1966-1968	Bangalore, India	Villlagers	all	5, active, with four TST and radiological surveys	1 TU RT 23 with Tween 80; ≥10mm	Y`	N	N	197 culture pos and suspect disease/9,786	By age ≥10mm to 1TU: 5-14 yrs 205; 15-54 yrs 332; 55+ yrs 908; By age ≥8mm to 23TU: 5-14 yrs 46; 15-54 yrs 115; 55+ yrs 472; By age 7mm to 20TU: 5-14 yrs 118; 15-54 yrs 165; 55+ yrs 588.	Culture pos TB: 145 (study cohort, 1961-1968)[77]
Gothi *et al.* 1978[77]	1961-1963; 1966-1968	Bangalore, India	Villlagers	all	5, active, with four TST and radiological surveys	1 TU RT 23 with Tween 80; ≥10mm	Y[n]	N, without scar	N	Culture pos TB: 29 males/2,505; 12 females/2,148.	5-14 yrs: 52; 15-34 yrs: 158 35-54 yrs: 227; 55+ yrs: 242 male: 227; female: 110	Culture pos TB: 145 (study cohort, 1961-1968)[77]
Grzybowski *et al.* 1972[78]; Grzybowski *et al.* 1976[79]	1964; 1969	Northwest Territories, Canada	Five Eskimo community settlements	all	5, passive	TST; NS	Y`	N	NS	Age: <20 yrs 21/152; ≥20 yrs 20/267; all 39/NS; 0-14 yrs 12/NS; 15-24 yrs 13/NS; 25-34 yrs 6/NS; 35+ yrs 8/NS.	Age: <20 yrs 2,780; ≥20 yrs 1,490; all 1,814; 0-14 yrs 2,637; 15-24 yrs 2,301; 25-34 yrs 1,348; 35+ yrs 1,168.	1,310 (Eskimos, North West Territories, 1967-1969)[78]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; N=No; pyrs=person years; Tu=tuberculin units; yrs=years; pos=positive; PPD=purified protein derivative.
[p] Results are plotted and appear either in the figures below or in the main manuscript.
` Unclear whether cohort included those with certain CXR abnormalities.
[n] Only those with normal/neg/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or progression rates in those with and without abnormalities were presented separately.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Narmada *et al.* 1977[80]	1968: 1972	Madras City, India	Children	1month<1 2 yrs	4, active, 12 household visits	5 IU 0·0001mg /0·1ml PPD-S	N	N	N	Age: 0-4 yrs 11/126; 5-9 yrs 33/408; 10-12 yrs 15/409; 0-5mm 24/3055; 6-11mm 4/610; 12-17mm 3/190; 18-23mm 34/541; 24+mm 22/212.	Age: 0-4 yrs 2183; 5-9 yrs 2022; 10-12 yrs 917; 0-5mm 196; 6-11mm 164; 12-17mm 395; 18-23mm 1571; 24+mm 2594.	Culture pos TB: 352 (rural Tamil Nadu survey, 1971-1973)[81]
Radhakrishna *et al.* 2003 (Tuberculosis Research Centre (ICMR)[82]	1968; NS	Chennai, India	No BCR scar, without smear/culture positivity	>1 mth	<15, active, surveys every 2.5 years with radiography	TST 3 IU of PPD-S and 10 units of PPD-B: ≥12mm	Y[n], ≥10 yrs of age, & 5-9 yrs at 2.5 yrs in.	VC	N	Culture pos TB: NS/111,224	All: 332 Males: 469 Females: 170	Culture pos TB: 191 (study cohort, 1968-1983)[82]
Radhakrishna *et al.* 2007[83]	1968; NS	Chennai, India	Residents	all	<15, active, surveys every 2.5 years with radiography	Dual testing with PPD-S and PPD-B (Batty strain); PPD-S = 8–11 mm; PPD-S minus PPD-B ≥2 mm or PPD-S ≥12mm.	Y[n], ≥10 yrs of age, & 5-9 yrs at 2.5 yrs in.	VC	N	Culture pos TB, no TB case at home: NS/114,445	Culture pos TB, no TB case at home: 370	Culture pos TB: 191 (study cohort, 1968-1983)[82]
Grzybowski *et al.* 1976[79]	1969; 1975	Northwest Territories, Canada	Eskimo communities	all	5, passive	TST; NS	Y[n]	N	800/2300 from 1967-1973	All ages 83/2229; 0-14 yrs 3/462; 15-24 yrs 18/494; 25-34 yrs 24/448; 35+ yrs 38/825.	All ages 621; 0-14 yrs 108; 15-24 yrs 607; 25-34 yrs 893; 35+ yrs 768.	960-240 (North West Territories, 1969-1974)[79]
Capewell *et al.* 1986[84]	1970-1983; 1971-1983	Edinburgh, Scotland	School children	13	<13, passive	Heaf test	N	N	N	Neg 1/882; Heaf I 1/4363; Heaf II 1/555 Heaf III 6/235; Heaf IV 7/155	Neg 19; Heaf pos 41; Heaf I 3; Heaf II 27 Heaf III 305; Heaf IV 628	~30 (London, 1970-1983)[85]
[P]Enarson 1998[86]	1972 & 1979; 1989	Canada	Alberta Aboriginals	10-19	<10, results come from "regular recording of routine testing"	TST; ≥10mm	NS	N	NS	≥10mm: NS/25,972 pyrs <10mm: NS/37,369 pyrs	See Figure 3 in manuscript for rates over time. Annual averages not stated.	NS[86]
Enarson 1998[86]	1980; 1982	Northwest Territories, Canada	Eskimo communities	all	2, unclear, possibly passive	TST; NS	Y`	N	N	9/1,718	263	68 (North West Territories, 1980-1982)[86]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; TST=tuberculin skin test; IU=International units; PPD=purified protein derivative; yrs=years; pos=positive; VC= variable, and considered in analysis; N=No; NS=not stated; Y=Yes, pyrs=person years.

[P] Results are plotted and appear either in the figures below or in the main manuscript.

[n] Only those with normal/neg/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or progression rates in those with and without abnormalities were presented separately.

` Unclear whether cohort included those with certain CXR abnormalities.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years);	Follow-up (years, unless stated) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Nolan *et al.* 1988[87]	1980-1981; 1986	Seattle-King County, USA	Recently arrived southeast Asian refugees	all	5, passive	TST 5 units PPD-T; >10mm	Y[n]	NS	15/25 cases prescribed, 10 completed, 5 did not	22/3,300 PPD pos; 78 cases found at initial screening are not included in the results	<10mm 10; >10mm 133; CXR normal 103; CXR abnormal 649;	39 (Seattle-King County, 1981-1986)[87]
Fine *et al.* 1994[88]	1980-1984; 1989	Karonga District, Malawi	Residents	NS	5, active, two TST surveys	TST 2IU PPD RT23	NS	N	N	0mm 19/71,055 pyrs; 1–5mm 2/5,442 pyrs; 6–10mm 3/15,059 pyrs; 11–15mm 14/21,561 pyrs; 16–20mm 9/13,488 pyrs; >20mm 6/2,801 pyrs.	0mm 27; 1–5mm 37; 6–10mm 20; 11–15mm 65; 16–20mm 67; >20mm 412.	41 in BCG neg study cohort[88]
MacIntyre & Plant 1999[89]	1989; 1994	Victoria, Australia	South-East Asian Refugees	87·9% ≤35 mean: 33	5, passive	TST: 15 mm post-BCG or 10 mm without past BCG,	Y`	VNC	Y, 22%	0-4mm 0/191; 5–9mm 0/183; 10–14mm 2/283; 15–19mm 1/181; >19mm 2/97; >20mm	0-4mm 0; 5–9mm 0; 10–14mm 141; 15–19mm 110; >19mm 2/97; >20mm	~6 (Australia, 1989-1994)[90]
[P]Marks *et al.* 2000[91] and Marks *et al.* 2001[92]	1984-1994; 1998	Sydney, Australia	Recently arrived southeast Asian refugees	>12	mean: 10·3, passive	TST: multiple cut-offs	Y[n]	VC	N	≥10mm: 98/NS[91]	≥10mm: 122; ≥15mm: 160 ≥20mm: 192	6-5 (Australia, 1984-1998)[93]
Choudhury *et al.* 2014[94]	1989-2001; 2008	England & Wales	Recent migrants	16-34	mean: 10·16, GP registration records used	No BCG: Heaf grade 2–4 (equiv. ≥6 mm); BCG history: Heaf grade 3–4 (equiv. ≥15 mm)	Y``	VC	N	53/402	0-5 yrs post-migration: 1,800 10-15 yrs post-migration: 1,000	7-9 (England and Wales excluding London, 1989-2001)[95]
Daley *et al.* 1998[96]	1990-1994;	San Francisco, USA	Injecting drug users in methadone maintenance clinic	NS	median: 22 months, active	TST; ≥10mm	Y`	N	N	1/259 HIV neg	390	~11-10 (USA, 1990-1994)[97]
Moss *et al.* 2000[98]	1990-1994; 1996	San Francisco, USA	Homeless population	median: 38	median: 3·2, passive	TST; ≥10mm	NS	NS	N	12/695 (2,524pyrs)	475	270 (in study population 1990-1994)[98]
Cook *et al.* 2008[99]	1990-2002; 2006	British Columbia, Canada	Residents without risk factors for TB	All, mean: 32	<17, passive	TST 5 TU: ≥10mm	Y``	35%	DNC – did not complete	7/25,035	10-14mm: 0·7^^ 15-19mm: 2·8^^ ≥mm: 3·0^^	~7-5 (Canada, 1990-2006)[100]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; USA=United States of America; TST=tuberculin skin test; PPD=purified protein derivative; Y=Yes; NS=Not stated; GP=general practitioner; pos=positive; N=No; IU=International units; VNC=variable, not considered in analysis; VC= variable, and considered in analysis; neg=negative.

[P] Results are plotted and appear either in the figures below or in the main manuscript.

`` Cohort known to include those with certain CXR abnormalities.

` Unclear whether cohort included those with certain CXR abnormalities.

^^These rates were calculated by dividing the number of TB cases by the cohort initially tested, and then dividing this by 17 (monitoring years from 1990-2006). Participant years of observation weren't given.

[n] Only those with normal/neg/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or progression rates in those with and without abnormalities were presented separately.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years, unless stated) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Scolari *et al.* 1999[101]	1991; 1995	Brescia, Northern Italy	Sengalese immigrants in residential compound	mean: 32	mean: 34 months, passive	Test-Tin 5IR tuberculin S; ≥2 points showing infiltration	Y	NS	N	4/115	1,230	15 (study region, 1993-1995) cited in [101]
Sanchez /Martin *et al.* 2001[102]	1991- 1999	Spain	Prison inmates	≥16	mean: 3·4, active, twice-yearly radiograph	TST 2 TU PPD TST RT-23 with Tween 80: ≥15mm	Y`	VNC	N	18/632 (number who were HIV negative and untreated NS)	HIV-neg: 488	639 (study setting)[102]
Klein *et al.* 2001[103]	1995	Bronx, New York City, USA	Current and former Injecting drug users in methadone maintenance program	median: 40	mean: 2·5, active	Mantoux 5 TU of PPD; Tubersol	NS	NS	"some subjects"	HIV neg: 0/203	0	~9 (USA, 1995)[97]
Mojazi-Amiri *et al.* 2013[104]	1995-2002; 2008	Texas, USA	Patients entered in LTBI registry, "most referred for contact investigation"	all	7-14, passive	TST; NS	Y`	NS	DNC	NS/20,353 DNC	58	3·5 (USA, 2011)[104]
Leung *et al.* 2006[105]	1999; 2003	Hong Kong	School children	mean: 10	mean: 4·5, passive	1 U PPD RT-23: ≥20mm	NS	Y	N	10/662	341	104-100 (Hong Kong, 2000-2003)[38]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; Y=yes; NS=Not stated; N=No; PPD=purified protein derivative; TU=tuberculin units; N=No; USA=United States of America; HIV=human immunodeficiency virus; neg=negative; VNC=variable, not considered in analysis; DNC=did not complete.
` Unclear whether cohort included those with certain CXR abnormalities.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years, unless stated) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Leung *et al.* 2012[106]	1999-2000; 2010	Hong Kong	School children	6-10	10-11, passive	TST 1-unit PPD RT-23: multiple cut-offs	NS	Y	N	≥15mm: 13/4637 pyrs	≥15mm: 280.4 <15 yrs of age: 1,000 ≥15 yrs of age: 608·1	91 (Hong Kong, 2004)[107]
Chan-Yeung *et al.* 2007[108]	2000: 2004	Hong Kong	Old age home residents	91·9% ≥70	mean: 2·5, homes contacted every six months to determine deaths, discharges or TB cases	PPD-RT23; ≥10mm	Y	NS	N	<5mm: 8/1.387; ≥5mm: 19/2,218 <10mm: 8/1,936; ≥10mm: 19/1,669 <15mm: 12/2,526; ≥15mm: 15/1,079	<5mm: 229; ≥5mm: 323 <10mm: 160; ≥10mm: 432 <15mm: 181; ≥15mm: 542	104-91 (Hong Kong, 2000-2004)[38,107]
Hemmati *et al.* 2011[109]	2002; 2007	Kermanshah, Iran	Primary school children	7-11	5, active with periodic radiography "if necessary"	TST 0·1ml 5TU PPD: 10-14 and ≥15mm	Y'	99.2 %	N	10-14mm 0/301; ≥15mm 0/529	10-14mm 0; ≥15mm 0	13-24 (Iran, 2004) www.cdc.hb.ir cited by authors
Joshi *et al.* 2011[110]	2004;	Sevagram, India	Health care workers, Mahatma Gandhi Institute of Medical Sciences	≥18	~6, active, face to face, telephone or email at ~6 years	QFT-GIT IFN-Ɣ ≥0·35IU/ml and TST (1 TU PPD RT23); ≥10mm	NS	NS	Y, "a small proportion"	TST or QFT-GIT pos, 6/360	TST pos 363; QFT-GIT pos 369; TST neg 342; QFT-GIT neg 338.	282 (India, 2004)[38]
Roth *et al.* 2017[111]	2004-14 (excluding 2009); 2014	British Columbia, Canada	LTBI pos people assessed at British Columbia Centre for Disease Control clinics, including contacts	All	<53 months, passive	TST, QFT-GIT and T-SPOT.TB; ≥10mm	NS	V	N	TST pos 2004-2008 18/25,244 pyrs; TST pos 2010-2014 16/19,518 pyrs; TST and IGRA pos 2010-2014 2/1,445 pyrs; TST pos and IGRA neg 2010-2014: 0/6,015 pyrs.	TST pos 2004-2008 71; TST pos 2010-2014 82; TST and IGRA pos 2010-2014 138; TST pos and IGRA neg 2010-2014 0	~5 (Canada, 2004-2012)[100]
Tsou *et al.* 2015[112]	2004; NS	Central Taiwan	Veteran nursing home residents	≥65	5, active, interviewed at six-month intervals	TST 2 TU PPD RT-23: conversion, <10 then ≥10 mm increase at 2-yrs. QFT-G conversion: IFN-γ of <0·35 IU/mL & ≥0·35 IU/mL≥ at 2yrs	Y'	N	N	TST pos: 3/100 QFT-G pos: 1/39	TST pos: 600 QFT-G pos: 513	659 (study setting)[112]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; mth=month; TST=tuberculin skin test; PPD=purified protein derivative; NS=Not stated; Y=Yes; N=No; pyrs=person years; yrs=years; TU=tuberculin units; QFT-GIT=QuantiFERON-TB Gold In-Tube test; pos=positive; neg=negative.
` Unclear whether cohort included those with certain CXR abnormalities.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years, unless stated) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Grinsdale *et al.* 2016[113]	2005-2008	San Francisco, USA	Foreign-born children, and child contacts	<15	4-7·8, passive	QFT-G and QFT-GIT; TST: ≥10mm or ≥5mm in contacts	Y’	VC	N	QFT pos: 0/11 TST pos: 0/153	0	16 (study setting, 2005-2012)[113]
Harstad *et al.* 2010[114]	2005-2006; 2008	Norway	Asylum seekers	≥18	23-32 months, passive	QFT-GIT and TST RT 23, 2 TU: ≥6mm	Y``	NS	N	QFT pos: 8/238 TST pos: 8/415	QFT pos: 1680 TST pos: 964 (assuming 24 mths follow up)	~7-8 (Norway, 2005-2008)[32]
Andrews *et al.* 2015[115]	2005-2007	Worcester, South Africa	School children	12-18	<5, active until two years	QFT & TST RT 23, 2 TU at baseline and after two yrs: ≥5mm	N	Not available	N	TST: pos 58/6,519·1 pyrs; neg 3/3,502·5 pyrs; conversion 7/654·3pyrs; reversion 0/235·7 pyrs; QFT: pos 46/4,371·1 pyrs; neg 7/3,994·6 pyrs; conversion 17/1,223·1 pyrs; reversion 3/203·9 pyrs.	TST: pos 890; TST neg 90; conversion 1,070; reversion 0; QFT: pos 970; neg 180; conversion 1,390; reversion 1,470.	~900-1000 (South Africa, 2005-2009)[32]
Mahomed *et al.* 2011[116]	2005-2009	Worcester, South Africa	School children	12-18	median: 2.4, active follow-up at two years and have had three monthly visits	QFT-GIT; TST RT 23, 2 TU: ≥5mm	CXR for smear pos	Y (93.8%)	N	QFT-GIT pos 39/2,669; TST pos 40/2,894.	QFT-GIT pos 640; TST pos 600.	~900-1000 (South Africa, 2005-2009)[32]
Mahomed *et al.* 2013[117]	2005-2009	Worcester, South Africa	School children	12-18	22 months-3·8yrs, active follow-up at two years and have had three monthly visits	QFT-GIT; TST RT 23, 2 TU: ≥5mm	“screened for TB at baseline”	40.3% with scar	N	QFT-GIT pos 48/3,233; QFT-GIT neg 17/2,804; TST pos 44/3,115; TST neg 12/2,456.	QFT-GIT pos 650; QFT-GIT neg 270; TST pos 610; TST neg 200.	~900-1000 (South Africa, 2005-2009)[32]
[P]Hermansen *et al.* 2016[118]	2005-2012	Denmark	Residents with QFT-GIT test (for various reasons)*	All	median 3·36, passive	QFT-GIT	NS	NS	Y, unknown proportion, estimated ~35%	20/1520 (>90 days from screening, 183 developed TB within 90 days)	Age: <15 yrs 565; 15-24 yrs 746; 25-34 yrs 654; 35+ yrs 284; All 383	~6-8 (Denmark, 2005-2012)[32]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; QFT-G=QuantiFERON-TB Gold; QFT-GIT=QuantiFERON-TB Gold In-Tube test; Y=Yes; VC=variable, and considered in analysis; N=No; pos=positive; neg=negative; TST=tuberculin skin test; TU=tuberculin units.
[P] Results are plotted and appear either in the figures below or in the main manuscript.
* These cohorts contained participants that had been identified as recent contacts of TB cases.
`` Cohort known to include those with certain CXR abnormalities.
` Unclear whether cohort included those with certain CXR abnormalities.

Table S3 Continued. Description of cohort studies included in the review that documented reactivation following latent tuberculosis screening and unknown timing of infection.

Publication	Years of study (recruited; end)	Setting	Population	Age-group when recruited (years)	Follow-up (years, unless stated) and method	LTBI screening method and cut-off	TB disease excluded with CXR	BCG	LTBI therapy	Sample size: Number of TB cases/number with reactivity at study entry	Average annual reactivation rates per 100,000	Approximate TB incidence per 100,000 per annum (setting and years)
Cohorts defined by initial cross-sectional latent tuberculosis screening and unknown timing of infection												
Chigbu & Iroegbu 2010[119]	2006:2007	Nigeria	Prison inmates	mean: 33·8	1, active	Mantoux - 0·1ml 5TU PPD: ≥10mm for HIV neg	NS	NS	N	8 sputum pos TB/58 HIV neg	13,793	219 (Nigeria, 2006-2007)[38]
Nduba *et al.* 2018[120]	2008-2009; 2010	Siaya County, Kenya	School children	12-18	<2, mean: 1·2, active, repeat TSTs	TST; ≥10mm	Y`	85·5% had scar	N	8/1777·8 pyrs	450	400 (Nyanza province, Kenya, 2013)[121]
Azoulay *et al.* 2015[122]	2008; 2011	Paris, France	Health care workers	mean:39	2, active, repeat QFT and CXR	QFT-GIT	Y`	87%	N	0/99 initially QFT positive followed to two years	0	11-9 (France, 2008-2011)[38]
Bunyasi *et al.* 2017[123]	2009-2012: 2014	South Africa	HIV-neg children	≤4	median: 5, active, repeat LTBI assessments	QFT-GIT or TST; ≥10mm	NS	Y	N	Age: 0<1 yrs 10/111 pyrs; 1<2 yrs 69/378 pyrs; 2<3 yrs 30/183 pyrs; 3<4 yrs 1/37 pyrs; All 110/708 pyrs.	Age: 0<1 yrs 27,600; 1<2 yrs 33,200;2<3 yrs 36,900; 3<4 yrs 8,500; All: 32,700	967-820 (South Africa, 2009-2014)[38]
[P]Winje *et al.* 2018[124]	2009-2014; 2016	Norway	Residents who had received a QFT-GIT (for various reasons)*	All	NS; presumably <7, passive	QFT-GIT	NS	NS	N	<2 yrs after QFT: low pos 14/2166; medium pos 38/2670; high pos 124/5042; neg 24/34,128 ≥2 yrs after QFT: low pos 3/1,679; medium pos 8/1,910; high pos 32/3,543; neg: 9/32,124[#]	<2 yrs after QFT: low pos 390; med pos 890; high pos 1.560; neg 40; ≥2 yrs after QFT: low pos 70; medium pos 170; high pos 390; neg 10[#]	8-6 (Norway, 2009-2016)[38]
Du *et al.* 2016[125]	2010; 2013	Changping District, Beijing, China	College students	15-28	3, active annual CXR	ELISPOT assay & TST 0·1ml of 5IU PPD: ≥10mm	Y'	78·7% had scar	N	ELISPOT pos and TST pos: 0/171	0	77 (China, 2010)[38]
[P]Abubakar *et al.* 2018[35,36]^ UK Predict study	2010-2015; 2016	London, Birmingham & Leicester, UK	UK contacts and recent migrants*	≥16	median: 2.9, telephone contact at 12 and 24 months and passive beyond	T-SPOT.TB QFT-GIT Mantoux TST: multiple cut-offs	Y`	VC	N	Migrants (see Table 2 for contacts): TSpot.TB pos 21/587; QFT-GIT pos 17/651; TST5mm 21/1253 TST10mm 20/828; TST15mm 18/586	Migrants: TSpot.TB pos 1,150; QFT-GIT pos 830; TST5mm 540; TST10mm 790; TST15mm 990.	~10-14 (UK, 2010-2016)[32]
Gao *et al.* 2018[126]	2015; NS, 2017	Zhongmu County, China	Rural residents	50-69	2, active	QFT-GIT	Y`	NS	N	10 pulmonary TB/1,127 (2009pyrs)	498	66-63 (China, 2015-2017)[38]

Abbreviations: LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; PPD=purified protein derivative; HIV=human immunodeficiency virus; neg=negative; NS=not stated; N=No; pos=positive; Y=Yes; TST=tuberculin skin test; pyrs=person years; QFT-GIT=QuantiFERON-TB Gold In-Tube test; yrs=years; IU=international units; UK= United Kingdom; VC=variable, and considered in analysis.
^ Additional data was obtained from the study's corresponding author, so data included in the table may not appear in the available study manuscript.
[P] Results are plotted and appear either in the figures below or in the main manuscript.
* These cohorts contained participants that had been identified as recent contacts of TB cases.
` Unclear whether cohort included those with certain CXR abnormalities.
[#] Low pos, IFN-γ 0.35 to <1.0; medium pos, IFN-γ 1.0 to <4.0; and high pos, IFN-γ>4.0 IU/mL.

Table S4 Ecological studies that estimate TB reactivation by using all TB cases in a population, and an estimation of the latent tuberculosis prevalence.

Publication	Setting	Population (TB case cohort; base cohort sample)	Age-groups (years)	LTBI screening method and cut-off	Active disease excluded with CXR in base cohort	BCG	LTBI therapy	Sample size: TB cases/case source cohort; positive in base cohort/base cohort	Approximate TB incidence per 100,000 per annum in setting (setting and years)
[P]Gryzbowski & Allen 1964[127]	Ontario, Canada	All notified TB cases in residents 1962; mass-screening 1958-1960	all	TST 1:2,000 dilution, with 0·05mg of tuberculin in 0·1ml (equiv. 5 TU): ≥5mm	Y`	NS	[#]	1,766/6,342,000; 32,441/177,259	28 (study setting, 1962)[127]
[P]Barnett *et al.* 1971[128]	Sasktchewan, Canada	All notified TB cases in residents 1960-1969 (divided by 10) and population 1964; mass-screening 1960-1969	all	No.2 dilution Old Tuberculin (1:1000 = 1/10mgm) to 1964, 5 TU in first half of 1964, then 1/20 mgm: ≥6mm	Y[n]	N	N	1,468 (1960-1969)/947,000 (1964); 250,704/1,420,056 (all, CXR status not stated)	16 (study setting, 1960-1969)[128] ~55-25 (Canada, 1960-1969)[30]
[P]Stead 1983[129]	Arkansas, USA	TB cases in Arkansas 1961, 1971 and 1981; TST survey in Ontario, Canada, 1958-1960 with annual reversion of 5% assumed.[127]	all	As used in Grzybowski & Allen 1964[127]	NS	NS	N	NS/NS. See Figure 5 for results.	234 (study setting)[24] 12.3-9.3 (USA, 1956-1959)[19]
[P]Horsburgh *et al.* 2010[130]	Palm Beach County, Florida, USA	All non-genotypically clustered cases in residents 1997-2001; TST survey in rural, western Palm Beach County, Florida 1998-2000	≥1	TST 5 TU: ≥10mm	NS	NS	NS	80 (16-23 unclustered)/34,759; 135/447 (15 had previously completed 6 months treatment)	46 (study setting,1997-2001)[130]
Mulder *et al.* 2012[131]	Netherlands	Immigrants Apr 2009-Mar 2011; screening results from a sample of the same cohort. They used "a Bayesian model to obtain a posterior distribution of the sensitivity for the QFT-GIT"	≥18	QFT-GIT	Y``	NS	N	30/26,317; 296/1,468	46 in immigrants and ~2 in native population in 2010[132]
Mulder *et al.* 2013[132]	Netherlands	Immigrants Apr 2009-Mar 2011; screening results from a sample of the same cohort. "The numerator…was modelled using the Poisson distribution …The denominator…took into account differences in sensitivity of the TST reported in the published data"	≥18	TST 0·1ml RT23: ≥10mm & ≥15mm	NS	85% of base cohort	N	30/26,317; ≥10mm: 273/643 ≥15mm: 145/643	46 in immigrants and ~2 in native population in 2010[132]
[P]Shea *et al.* 2014[133]	USA	CDC Non-genotypically clustered cases 2006-2008; TST survey from 1999-2000 NHANES survey	all	TST 0·1 mL PPD S-1: ≥10mm	NS	NS	NS	39,920/907,272,727; NS/7,386	~4 (USA, 2006-2008)[133]

Abbreviations: TB=tuberculosis; LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; TB=tuberculosis; equiv.=equivalent; TU=tuberculin units; Y=Yes; NS=not stated; N=No; USA=United States of America; NA=not applicable; QFT-GIT=QuantiFERON-TB Gold In-Tube test; TST=tuberculin skin test; TU=tuberculin units; NHANES=National Health and Nutrition Examination Survey; PPD=purified protein derivative; pop=population; IGRA=Interferon Gamma Release Assay.

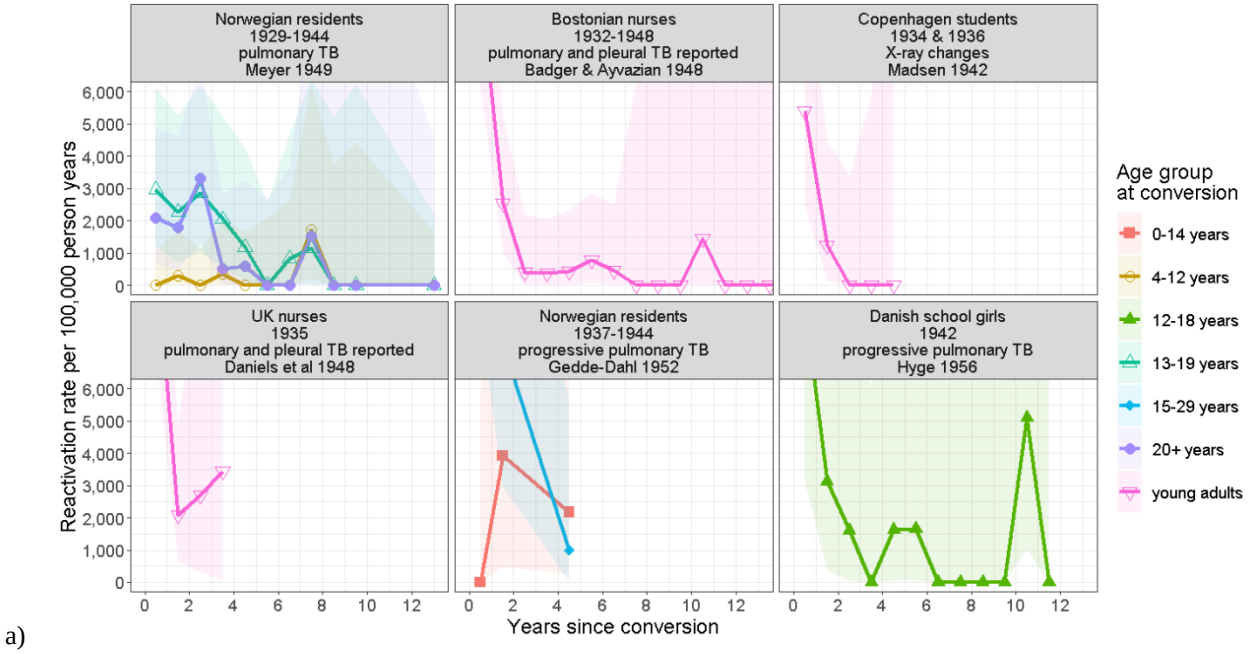
[P] Results are plotted and appear either in the figures below or in the main manuscript.

[#] "…all persons known to have had tuberculosis in the past were recalled; and prophylatic chemotherapy was liberally used, in both recent tuberculin converters and persons with inactive tuberculosis"[127]

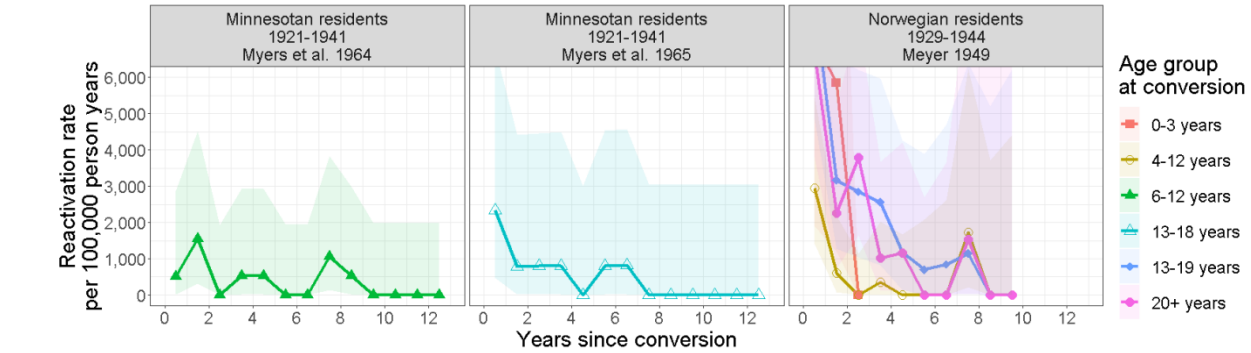
`` Cohort known to include those with certain CXR abnormalities.

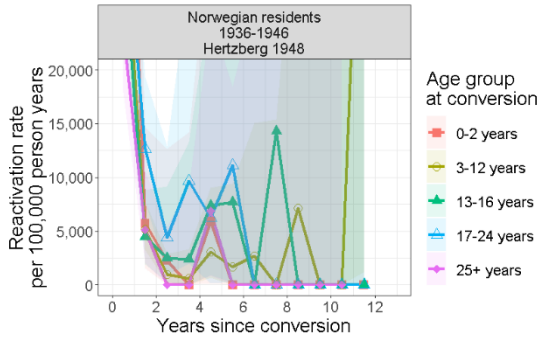
` Unclear whether cohort included those with certain CXR abnormalities, in the case of pre-chemotherapy studies the development of primary foci/calcifications was often discussed in detail.

[n] Only those with normal/negative/satisfactory radiograph, or those without radiologic manifestations/changes/lesions were included in the study, or progression rates in those with and without abnormalities were presented separately.



a)





b)

Figure S1 Pre-chemotherapy era studies of a) TB reactivation rates of pulmonary or pleural TB among converters (Gedde-Dahl 1952[12] and Hyge 1956[13] reported on progressive pulmonary TB cases and Madsen 1948[6] reported on "X-ray changes"), and **b) TB reactivation rates among converters.** Participant details, recruitment years, study lead author and publication year are given in the figure headings. Shaded areas represent 95% confidence intervals. The length of follow-up in Myers *et al.* 1964[1] and 1965[3] may have ranged from 19-39 years given the study recruitment and end dates (1921-1941 and 1960), however numbers followed over time were not given. Not shown in the figures are one case in Myers *et al.* 1964[1] that reactivated at 24 years from conversion, and one case from Myers 1965 that reactivated at 26 years from conversion. In the study by Meyer 1949, observation of the 0-3 year group continued into school-age for an unspecified period with no cases observed, and a small number from the other cohorts were observed up until the 17th year, again, with no cases being observed.[4] A very small number of study participants were followed in the study by Hertzberg for 11+ years, with an unspecified upper limit.[11]

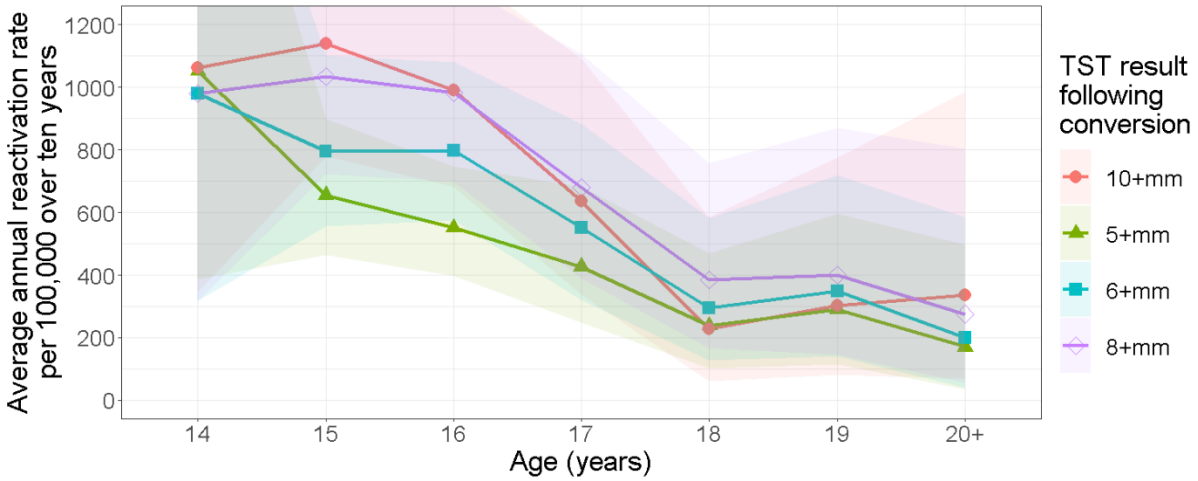
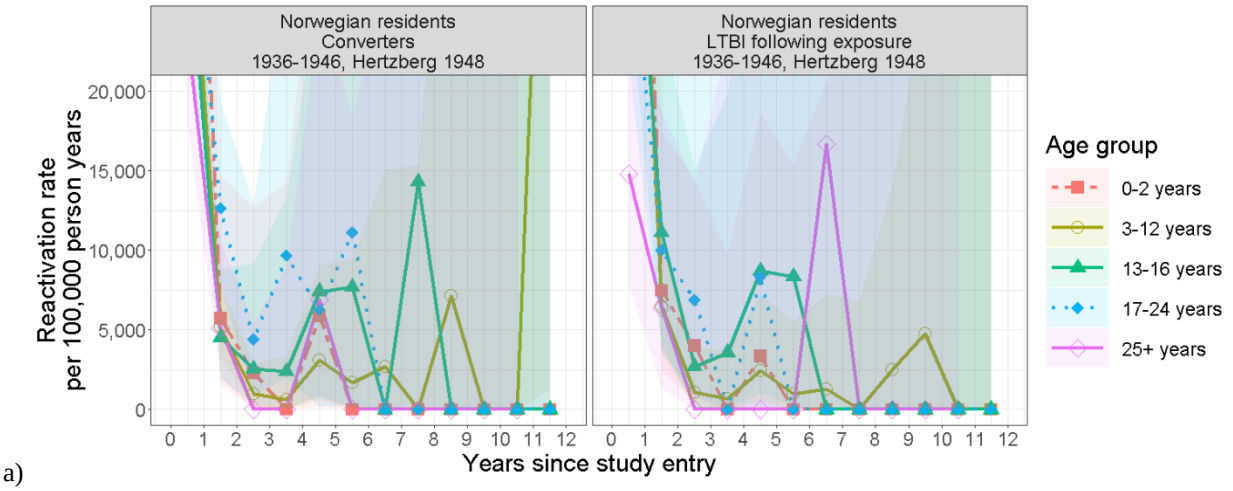
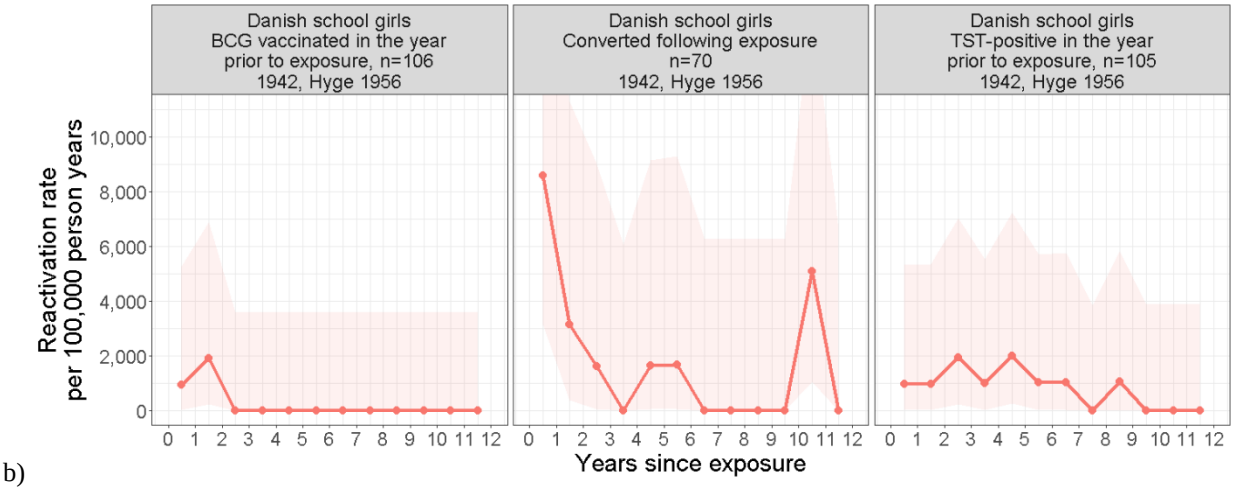


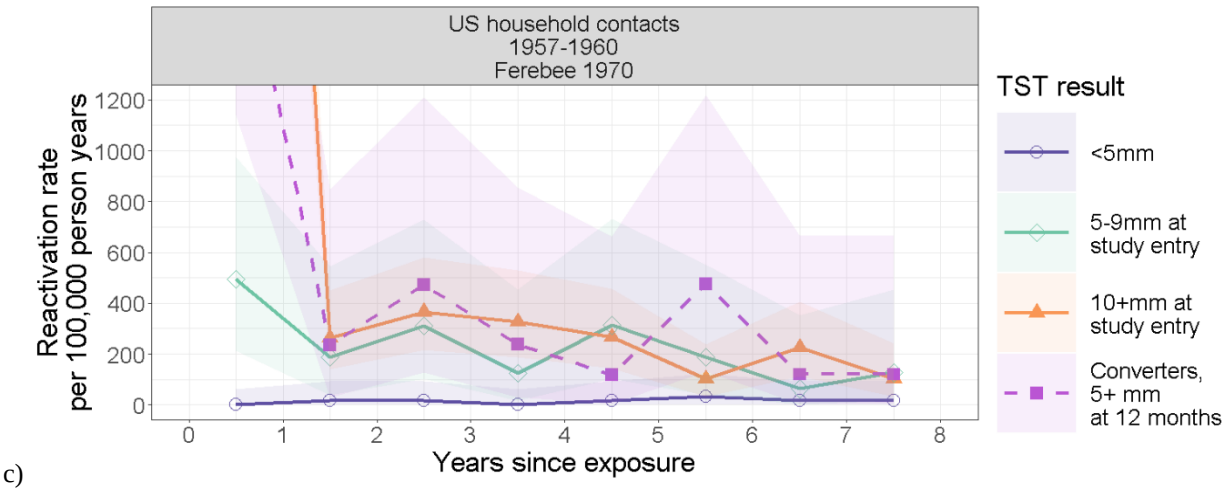
Figure S2 Average annual reactivation rates in first ten years following TST conversion, by age and TST result in English school children in the from the MRC (Medical Research Council) study[67], as described by Sutherland.[16] Shaded areas represent 95% confidence intervals.



a)



b)



c)

Figure S3 Comparing TB reactivation rates in cohorts that have shown conversion to those with LTBI following a known TB exposure in a) Norwegian residents (1936-1946),[64] b) groups of female high school students, 12-18 years of age, following a "massive" exposure in a blacked out air-raid shelter in 1943 (progressive pulmonary TB only),[13] and c) in the placebo cohort of the USPHS (United States Public Health Service) trial among household contacts (observation was incomplete beyond the seventh year and so these results have been excluded).[17] Shaded areas represent 95% confidence intervals. A very small number of study participants were followed in the study by Hertzberg for 11+ years, with an unspecified upper limit. Participant characteristics, recruitment years and study lead author and publication year are given in the figure headings.

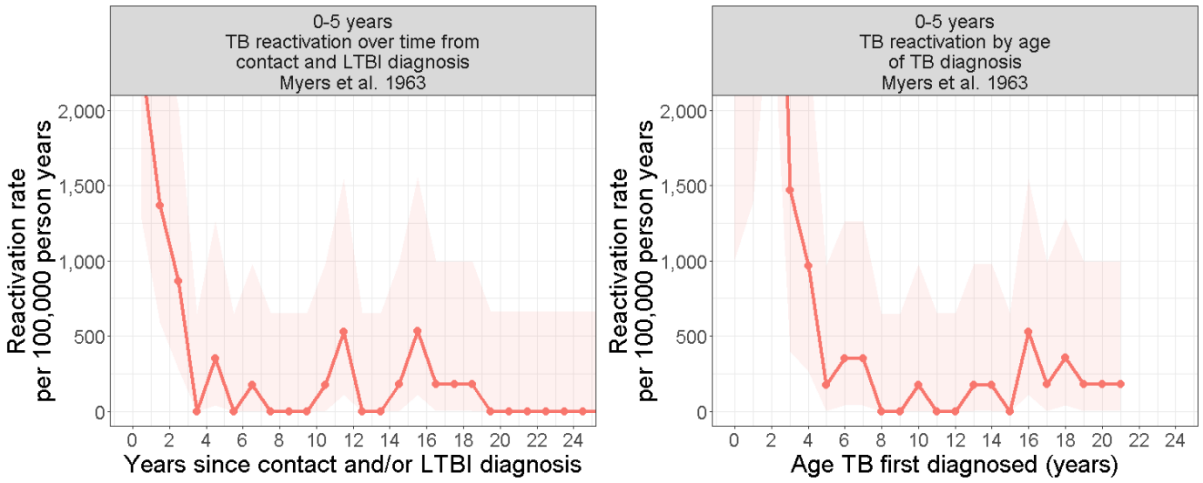


Figure S4 TB reactivation rates in child contacts identified with LTBI before the age of six in Minnesota, USA (recruited 1921-1941).[25] The left panel shows the TB reactivation rates over time from contact and LTBI diagnosis and the right panel shows TB reactivation rates by age of TB diagnosis. Myers *et al.* 1963 actively followed up the 0-5 year-olds to a mean age of 32 years, but did not give numbers observed overtime (given the years of study recruitment and study end, the follow-up period possibly varied from ~19-39 years[25]), and 124 of the cohort of 599 reportedly either died or were lost to follow-up over time, so the rate estimates in the above panels, particularly in the later years, will be inaccurate.[25] Shaded areas represent 95% confidence intervals.

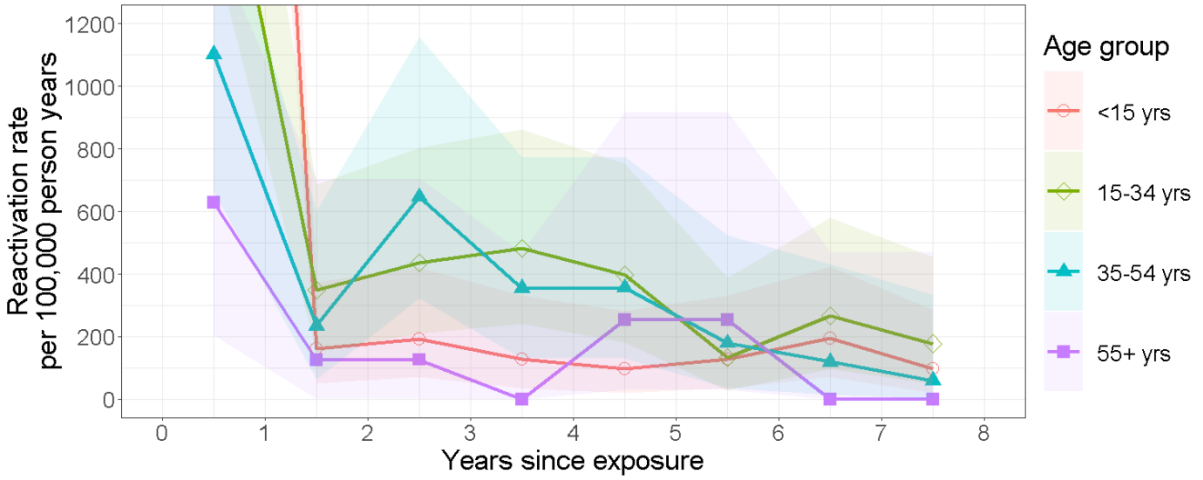


Figure S5 TB reactivation rates in the placebo cohort of the USPHS (United States Public Health Service) trial among household contacts by age-group for all "reactors" (TST 5+ mm at study entry or after 12 months follow-up). Observation was incomplete beyond the seventh year so these results have been excluded.[17] Shaded areas represent 95% confidence intervals.



Figure S6 Average annual TB reactivation rates per 100,000 by TST induration diameter. Note that the study among Danish residents only considers respiratory and pleural TB,[62] and the study by Radhakrishna et al 2003 considers only culture-positive TB.[82] Participant details, years of follow-up, lead author and publication year are given in the figure headings Shaded areas represent 95% confidence intervals.

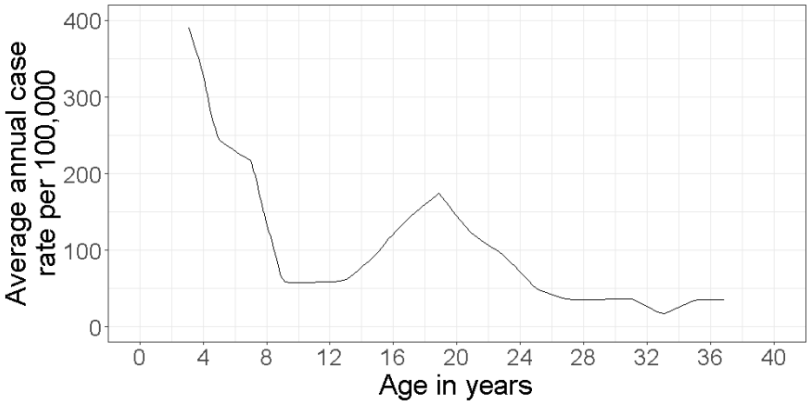


Figure S7 "Incidence of tuberculosis among initial reactors to tuberculin, by age when tuberculosis was first diagnosed"(adapted from figure 1 in Comstock *et al.* 1974).[63]

Table S5 Ecological studies excluded from the main analysis because they did not strictly meet our study criteria. Studies were not designed to estimate reactivation rates and nor were they discussed, but data provided in the studies can be used to calculate reactivation rates.

Publication	Setting	Population (TB case cohort; base cohort sample)	Age-groups (years)	Active disease excluded with CXR in base cohort	BCG	LTBI therapy	Sample size: TB cases/case source cohort; positive in base cohort/base cohort	Estimated reactivation rates	Approximate TB incidence per 100,000 per annum in setting (setting and years)
Borgdorff *et al.* 2010[134]	Netherlands	RFLP used to identify pulmonary TB cases among Netherlands-born residents in 1995 and 2005 with unique strains (not seen in the previous two years); age-specific proportion with latent TB based on annual risks of infection, see Styblo 1990.[135]	All	NA	NS	NS	1995: 170 pulmonary TB/1,702,000 2005: 91 pulmonary TB/982,000	See Figure S9 (calculated from data provided in Table 3 of the manuscript)	9 (Netherlands, 2000)[38] 8 (Netherlands, 2005)[38]
Winje *et al.* 2019[136]	Norway	TB case data from Norwegian immigrants 2008-2016, who arrived 2008-2011; estimated IGRA prevalence using age and country specific LTBI prevalence data from the published literature, and Norwegian data on asylum seekers[137]*	All	NS	NS	Unclear^	948/14,852	Calculated from data provided in Table 2[#]: 1,527/100,000py in the first five years after arrival, varying widely (509-2,771/100,000py) depending on country of origin. A high proportion of cases occurred in the first year after migration (33.3-83.3% depending on country of origin).	7.2-6.1 (Norway, 2008-2016)[38]

Abbreviations: TB=tuberculosis; LTBI=latent tuberculosis infection; CXR=Chest X-ray; BCG=Bacillus Calmette–Guérin vaccinated; RFLP=Restriction-fragment-length-polymorphism; TB=tuberculosis; NA=not applicable; NS=not stated; Y=yes; py=person years.
* IGRA positivity estimates ranged from 19.6%-28.4% depending on country of origin.
^ "The monitoring and evaluation system of the long-standing TB and LTBI screening programme is weak"[136]
[#] Assuming an IGRA sensitivity of 84%, as the study authors did, and assuming that the IGRA positivity estimated by the authors did not differ between immigrants remaining in Norway and those that emigrate.

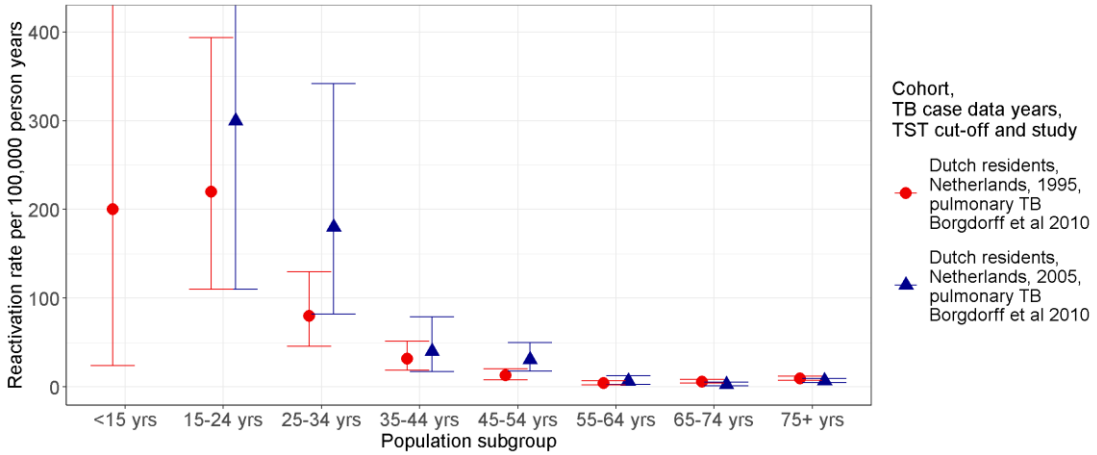


Figure S8 Annual pulmonary TB reactivation rates per 100,000 by age group. These values were calculated using data in Table 3 of Borgdorff *et al.* 2010.[134] This study used restriction-fragment-length-polymorphism to identify pulmonary TB cases among Netherlands-born residents in 1995 and 2005 with unique strains (not seen in the previous two years) by age group, and estimated the age-specific proportion with latent TB based on methods used by Styblo 1990;[135] these were used as the numerators and denominators for reactivation rate calculations, respectively. The missing value for 2005 in the <15 yrs age group 2005 was infinity: two cases in an estimated population of none with latent TB.

Table S6 The template used to assess data quality in our systematic review, based on the Cochrane Risk Of Bias In Non-randomized Studies - of Interventions (ROBINS-I) assessment tool

Study			
Bias due to confounding (confounding)	1.1 Is there potential for confounding of the effect of intervention in this study? Did the study include participants on LTBI treatment? No: Low: PN / N Yes: Serious / Critical: Y/PY	Confounding: Direction of bias:	Low: PN / N Serious / Critical: Y/PY
Selection bias (selection)	2.1. Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention? Were cohort characteristics representative of the general "infected" population, or if the cohort did include those with abnormal CXR, BCG vaccinated, or different age groups, results were disaggregated by these characteristics? Yes: Y / PY No: PN / N… - The cohort characteristics were either unstated/unclear; or - The study population included those on latent TB treatment, latent TB treatment was not randomly assigned, and treated participants were excluded, meaning the considered cohort (untreated) may have differed from the excluded (treated) in some way; of - The cohort included varying characteristics likely to influence reactivation rates (those with abnormal CXR, BCG vaccinated, or different age groups) but results were not disaggregated by these characteristics.	Bias in intervention classification: Direction of bias:	Low: Y / PY Moderate: PY/PN Serious / Critical: PN / N
Bias in classification of interventions (screening)	3.1 Were intervention (infection) groups clearly defined? LTBI screening method: - One clearly described method: Y / PY - Several methods, results disaggregated: Y / PY - Several methods, results not disaggregated: PN / N - Time of infection/conversion/exposure may have preceded the beginning of follow-up by more than 6 months (depending on the time between negative and positive screening results or the time of exposure and study commencement): PN / N	Bias in intervention classification: Direction of bias:	Low: Y / PY Moderate: PY/PN Serious / Critical: PN / N
Bias due to deviations from intended interventions (reinfection)	4.1. Were there deviations from the intended intervention beyond what would be expected in usual practice? TB incidence in the study setting: - low (<40/100,000 persons per year): PY - moderate (40-100/100,000 persons per year): PN - high (>100/100,000 persons per year): N	Bias due to deviations in intended interventions: Direction of bias:	Low: Y / PY Moderate: PY/PN Serious / Critical: PN / N
Bias due to missing data (missing)	5.1 Were outcome data available for all, or nearly all, participants? Follow-up and percentage loss to follow-up over time: - Active follow-up, numbers over time given, <25% loss to follow-up: Y; >25% loss to follow-up: PY/ PN - Active, numbers over time not given: PN - Passive follow-up: N - Not clearly described: NI	Bias due to missing data: Direction of bias:	Low: Y / PY Moderate: PY/PN Serious / Critical: PN / N
Bias in measurement of outcomes (outcomes)	- The method of TB disease diagnosis and considered manifestations was described and comparable across time in the follow-up period: Y / PY - The method of TB disease diagnosis and considered manifestations was unclear or the follow-up was active and included CXR screening, which may have affected the reactivation rate. PN / N	Bias in measurement of outcomes: Direction of bias:	Low: Y / PY Moderate: PY/PN Serious / Critical: PN / N
RISK OF OVERALL BIAS	Confounding: Low Moderate Serious Critical Selection bias: Low Moderate Serious Critical Classification of interventions: Low Moderate Serious Critical Deviations from intended interventions: Low Moderate Serious Critical Missing data: Low Moderate Serious Critical Measurement of outcome: Low Moderate Serious Critical Overall: Low Moderate Serious Critical		Low / Moderate / Serious / Critical
DIRECTION OF OVERALL BIAS			

References

1. Myers A, Bearman JE, Dizon HG. Natural History of Tuberculosis in the Human Body VI. Prognosis among Tuberculin Reactor Children of Six to Twelve Years. *Am Rev Respir Dis* 1964; **90**(3): 359.
2. Hermans S, Horsburgh CR, Jr., Wood R. A Century of Tuberculosis Epidemiology in the Northern and Southern Hemisphere: The Differential Impact of Control Interventions. *PLoS One* 2015; **10**(8): e0135179.
3. Myers JA, Bearman JE, Dixon HG. Natural History of Tuberculosis in the Human Body. VIII. Prognosis among Tuberculin Reactor Girls and Boys of Thirteen to Seventeen Years. *Am Rev Respir Dis* 1965; **91**(6): 896-908.
4. Meyer SN. Statistical Investigations of the relationship of Tuberculosis Morbidity and Mortality to Infection. *Acta Tuberculosea Scandinavica* 1949; (18): 222 pp.
5. Kinander W, Bruvik T, Dahle UR. Dominant Mycobacterium tuberculosis lineages in elderly patients born in Norway. *PLoS One* 2009; **4**(12): e8373.
6. Madsen T. Studies on the epidemiology of tuberculosis in Denmark; 1942.
7. Iversen E. The incidence of pulmonary tuberculosis in Denmark by sex and age 1921-1955. *Dan Med Bull* 1957; **4**(6): 191-6.
8. Styblo K. Epidemiology of Tuberculosis. The Hague: Royal Netherlands Tuberculosis Association, 1991.
9. Badger TL, Ayvazian LF. Tuberculosis in nurses; clinical observations on its pathogenesis as seen in a 15 year follow-up of 745 nurses. *American Review Of Tuberculosis* 1949; **60**(3): 305-31.
10. Daniels M, Ridehalgh F, Springett VH, Hall IM. Tuberculosis in young adults: report on the Prophit Tuberculosis Survey, 1935-1944. London: H.K. Lewis; 1948.
11. Hertzberg G. The Achievements of BCG Vaccination illustrated by Material at the Tuberculosis Department of the Oslo Public Health Service: Oslo : I Kommisjon Hos Johan Grundt Tanum Forlag; 1948.
12. Gedde-Dahl T. Tuberculous infection in the light of tuberculin matriculation. *Am J Hyg* 1952; **56**(2): 139-214.
13. Hyge TV. The efficacy of BCG - Vaccination. Epidemic of tuberculosis in a state school, with an observation period of 12 years. *Acta Tuberculosea Scandivanica* 1956; **32**(2): 89-107.
14. Franks H. Tuberculosis control in Denmark. *Br Med J* 1959; **2**(5142): 88-92.
15. Sutherland I. Progress Report 1967, Part 1. The Hague, Holland: Tuberculosis Surveillance Research Unit, International Union against Tuberculosis, 1967.
16. Sutherland I. The ten-year incidence of clinical tuberculosis following "conversion" in 2550 individuals aged 14 to 19 years: KNCV, The Hague, Netherlands, 1968.
17. Ferebee SH, Mount FW. Tuberculosis morbidity in a controlled trial of the prophylactic use of isoniazid among household contacts. *Am Rev Respir Dis* 1962; **85**: 490-510.
18. Ferebee S. Controlled chemoprophylaxis trials in tuberculosis. A general review. *Bibl Tuberc* 1970; **26**: 28-106.
19. Centers for Disease Control and Prevention. TB Incidence in the United States, 1953-2017. October 19, 2018 2018. https://www.cdc.gov/tb/statistics/tbcases.htm (accessed 12 November 2018).
20. Debre R, Perdrizet S, Lotte A, Naveau M, Lert F. Isoniazid chemoprophylaxis of latent primary tuberculosis: in five trial centres in France from 1959 to 1969. *Int J Epidemiol* 1973; **2**(2): 153-60.
21. Santé publique France. Tuberculose, Données épidémiologiques, Données sur les déclarations de tuberculose. 23/03/2018 2018. http://invs.santepubliquefrance.fr/Dossiers-thematiques/Maladies-infectieuses/Maladies-a-declaration-obligatoire/Tuberculose/Donnees-epidemiologiques (accessed 12 December 2018).
22. Veening GJ. Long term isoniazid prophylaxis. Controlled trial on INH prophylaxis after recent tuberculin conversion in young adults. *Bull Int Union Tuberc* 1968; **41**: 169-71.
23. Stead WW, To T. The significance of the tuberculin skin test in elderly persons. *Ann Intern Med* 1987; **107**(6): 837-42.
24. Stead WW, Dutt AK. Tuberculosis in the elderly. *Semin Respir Infect* 1989; **4**(3): 189-97.
25. Myers JA, Bearman JE, Dixon HG. The Natural History of Tuberculosis in the Human Body. V. Prognosis among Tuberculin-Reactor Children from Birth to Five Years of Age. *Am Rev Respir Dis* 1963; **87**(3; Pt 1): 354-69.
26. Dobler CC, Marks GB. Risk of tuberculosis among contacts in a low-incidence setting. *Eur Respir J* 2013; **41**(6): 1459-61.
27. Lin M, Spencer J, Roche P, McKinnon M, National TBACftCDNA. Tuberculosis notifications in Australia, 2000. *Commun Dis Intell Q Rep* 2002; **26**(2): 214-25.
28. Barry C, Waring J, Stapledon R, Konstantinos A. Tuberculosis notifications in Australia, 2008 and 2009. *Commun Dis Intell* 2012; **36**(1): 82-94.

29. Reichler MR, Khan A, Sterling TR, et al. Risk and Timing of Tuberculosis Among Close Contacts of Persons with Infectious Tuberculosis. *J Infect Dis* 2018.
30. Gallant V, Ogunnaike-Cooke S, McGuire M. Tuberculosis in Canada: 1924-2012. *Can Commun Dis Rep* 2014; **40**(6): 99-107.
31. Sloot R, Schim van der Loeff MF, Kouw PM, Borgdorff MW. Risk of tuberculosis after recent exposure. A 10-year follow-up study of contacts in Amsterdam. *Am J Respir Crit Care Med* 2014; **190**(9): 1044-52.
32. World Health Organization. Tuberculosis country profiles. 2018. http://www.who.int/tb/country/data/profiles/en/ (accessed 12 November 2018).
33. Erkens CG, Slump E, Verhagen M, Schimmel H, Cobelens F, van den Hof S. Risk of developing tuberculosis disease among persons diagnosed with latent tuberculosis infection in the Netherlands. *Eur Respir J* 2016; **48**(5): 1420-8.
34. Altet N, Dominguez J, Souza-Galvao ML, et al. Predicting the Development of Tuberculosis with the Tuberculin Skin Test and QuantiFERON Testing. *Ann Am Thorac Soc* 2015; **12**(5): 680-8.
35. Abubakar I, Drobniewski F, Southern J, et al. Prognostic value of interferon-gamma release assays and tuberculin skin test in predicting the development of active tuberculosis (UK PREDICT TB): a prospective cohort study. *Lancet Infect Dis* 2018; **18**(10): 1077-87.
36. Abubakar I, Lalvani A, Southern J, et al. Two interferon gamma release assays for predicting active tuberculosis: the UK PREDICT TB prognostic test study. *Health Technol Assess* 2018; **22**(56): 1-96.
37. Heiden Mad, Hauer B, Fiebig L, et al. Contact investigation after a fatal case of extensively drug-resistant tuberculosis (XDR-TB) in an aircraft, Germany, July 2013. *Eurosurveillance* 2017; **22**(12): 30493.
38. The World Bank. Incidence of tuberculosis (per 100,000 people): World Health Organization, Global Tuberculosis Report. 2019. https://data.worldbank.org/indicator/SH.TBS.INCD?view=chart (accessed 20th March 2019).
39. Pope AS, Sartwell PE, Zacks D. Development of Tuberculosis in Infected Children. *Am J Public Health Nations Health* 1939; **29**(12): 1318-25.
40. Heimbeck J. Incidence of tuberculosis in young adult women, with special reference to employment. *British Journal of Tuberculosis* 1938; **32**(3): 154-66.
41. Scheel. Tuberculosis among Medical Students in Oslo; the Prophylactic Use of BCG. *Bull Acad Med* 1935; **114**: 149-51.
42. Myers J, Diehl H, Boynton R, Ch'iu P, Streukens T, Trach B. Tuberculosis among students and graduates of medicine. *Ann Intern Med* 1941; **14**: 1575-94.
43. Myers J, Boynton R, Diehl H, Streukens T, Ch'iu P. Tuberculosis among students and graduates in nursing. *Ann Intern Med* 1940; **14**: 873-97.
44. Hastings D, Behn B. Tuberculosis among nurses: a study of the effect of tuberculosis service on the incidence of tuberculosis infection and disease among student nurses. *American Review of Tuberculosis* 1941; **44**: 681-94.
45. Geer E. Primary tuberculosis among nurses. *American Review of Tuberculosis* 1934; **29**: 88-97.
46. Myers JA, Bearman JE, Botkins AC. The natural history of tuberculosis in the human body. X. Prognosis among students with tuberculin reaction conversion before, during and after school of nursing. *Dis Chest* 1968; **53**(6): 687-98.
47. Heimbeck J. Tuberculous Superinfection. *Acta Tuberculosea Scandinavica* 1949; (21): 36-41.
48. Heimbeck J. The Relation between Tuberculous Infection and Tuberculous Disease. *Acta Med Scand* 1951; **140**(Suppl. 259): 144-8.
49. Ferguson RG. BCG vaccination in hospitals and sanatoria of Saskatchewan. *Canadian Journal of Public Health* 1946; **37**(11): 435-51.
50. Israel HL, Hetherington HW, Ord JG. A Study of Tuberculosis among Students of Nursing. *J Am Med Assoc* 1941; **117**(10): 839-43.
51. Wright HP. The Comparative Value of Various Tuberculin Tests in Children, Medical Students, and Nurses-in-Training. *Can Med Assoc J* 1941; **44**(1): 44-6.
52. Olsen HC. The Use of the Tuberculin Test in a Tuberculosis Scheme. Experience in Bornholm. *Tubercle* 1956; **37**(1): 47-57.
53. Palmer CE, Jablon S, Edwards PQ. Tuberculosis morbidity of young men in relation to tuberculin sensitivity and body build. *Amer Rev Tuberc* 1957; **76**(4): 517-39.
54. SjÖGren I. Tuberculosis in BCG-vaccinated and unvaccinated young Swedish men. A comparative study. *Scand J Respir Dis* 1976; **57**(5): 208-22.

55. Haro AS. Cohort approach in tuberculosis surveillance: comparison of the situation in Sweden and Finland. *Tuber Lung Dis* 1994; **75**(4): 271-82.
56. HÄRÖ AS. Twenty years later-evaluation of the results of a national mass BCG-vaccination in Finland. *Scand J Respir Dis* 1972; (80): 153-69.
57. Stephanopoulos C, Costeuetos E. Tuberculosis Morbidity in Student Nurses trained at the " Sotiria " Sanatorium and " Evangelismos " General Hospital Schools. *Acta Tuberculosea Scandinavica* 1957; **33**(1/2): 211-18.
58. Gernez-Rieux C, Gervois M. Protection conferred by BCG during the 20 years following vaccination. *Bull World Health Organ* 1973; **48**(2): 139-54.
59. Comstock GW, Livesay VT, Woolpert SF. Evaluation of BCG vaccination among Puerto Rican children. *Am J Public Health* 1974; **64**(3): 283-91.
60. Palmer CE, Shaw LW, Comstock GW. Community trials of BCG vaccination. *Am Rev Tuberc* 1958; **77**(6): 877-907.
61. Frimodt-Moller J, Thomas J, Parthasarathy R. Observations on the protective effect of BCG vaccination in a South Indian rural population. *Bull Wld Hlth Org* 1964; **30**(4): 545-74.
62. Horwitz O, Wilbek E, Erickson PA. Epidemiological basis of tuberculosis eradication. 10. Longitudinal studies on the risk of tuberculosis in the general population of a low-prevalence area. *Bull World Health Organ* 1969; **41**(1): 95-113.
63. Comstock GW, Woolpert SF, Livesay VT. Tuberculosis studies in Muscogee County, Georgia. Twenty-year evaluation of a community trial of BCG vaccination. *Public Health Rep* 1976; **91**(3): 276-80.
64. Comstock GW, Palmer CE. Long-term results of BCG vaccination in the southern United States. *Am Rev Respir Dis* 1966; **93**(2): 171-83.
65. Comstock GW. Community Research in Tuberculosis Muscogee County, Georgia. *Public Health Rep* 1964; **79**(12): 1045-56.
66. Large SE. B. C. G. vaccination in the brigade of gurkhas. *J R Army Med Corps* 1965; **111**(4): 246-58.
67. BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. *Bull World Health Organ* 1972; **46**(3): 371-85.
68. Hart PD, Sutherland I. BCG and vole bacillus vaccines in the prevention of tuberculosis in adolescence and early adult life. *Br Med J* 1977; **2**(6082): 293-5.
69. Ross JD, Willison JC. Tuberculosis in Edinburgh: B.C.G. vaccination and Heaf tuberculin grades at school. *Scott Med J* 1971; **16**(10): 443-9.
70. Mount FW, Ferebee SH. The effect of isoniazid prophylaxis on tuberculosis morbidity among household contacts of previously known cases of tuberculosis. *Am Rev Respir Dis* 1962; **85**: 821-7.
71. Ferebee SH, Mount FW, Murray FJ, Livesay VT. A Controlled Trial of Isoniazid Prophylaxis in Mental Institutions. *Am Rev Respir Dis* 1963; **88**: 161-75.
72. Comstock GW, Edwards LB, Livesay VT. Tuberculosis morbidity in the U.S. Navy: its distribution and decline. *Am Rev Respir Dis* 1974; **110**(5): 572-80.
73. Ross JD, Willison JC. The relationship between tuberculin reactions and the later development of tuberculosis: an investigation among Edinburgh school children in 1960-70. *Tubercle* 1971; **52**(4): 258-65.
74. Lotte A, Perdrizet S, Hatton F. Epidemiology of tuberculosis and failures of tuberculosis control in children. *Bull World Health Organ* 1971; **44**(Suppl): 229 pp.
75. Anonymous. Tuberculosis in a rural population of South India: a five-year epidemiological study. *Bull World Health Organ* 1974; **51**(5): 473-88.
76. Gothi GD, Nair SS, Chakraborty AK, Ganapathy KT. Five year incidence of tuberculosis and crude mortality in relation to non specific tuberculin sensitivity. *Indian J Tuberc* 1976; **23**(2): 58-63.
77. Gothi GD, Chakraborty AK, Jayalakshmi MJ. Incidence of sputum positive tuberculosis in different epidemiological groups during five year follow up of a rural population in South India. *Indian J Tuberc* 1978; **25**(2): 83-91.
78. Grzybowski S, Galbraith JD, Styblo K, Chan-Yeung M, Dorken E, Brown A. Tuberculosis in canadian eskimos. *Arch Environ Health* 1972; **25**(5): 329-32.
79. Grzybowski S, Styblo K, Dorken E. Tuberculosis in Eskimos. *Tubercle* 1976; **57**(4): 58 pp.
80. Narmada R, Narain R, Raju VB, Naganna K, Sundaram RS. Incidence of tuberculosis among infected and non-infected children. *The Indian journal of medical research* 1977; **65**(2): 171-83.
81. Radhakrishna S, Frieden TR, Subramani R, Kumaran PP. Trends in the prevalence and incidence of tuberculosis in south India. *Int J Tuberc Lung Dis* 2001; **5**(2): 142-57.

82. Radhakrishna S, Frieden T, Subramani R. Association of initial tuberculin sensitivity, age and sex with the incidence of tuberculosis in south India: a 15-year follow-up. *Int J Tuberc Lung Dis* 2003; **7**(11): 1083.
83. Radhakrishna S, Frieden TR, Subramani R, Santha T, Narayanan PR, Indian Council of Medical R. Additional risk of developing TB for household members with a TB case at home at intake: a 15-year study. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2007; **11**(3): 282-8.
84. Capewell S, France A, Uzel N, Leitch AG. The current value of tuberculin testing and BCG vaccination in school children. *Br J Dis Chest* 1986; **80**(3): 254-64.
85. The Scottish Office Department of Health. The control of tuberculosis in Scotland, 1998.
86. Enarson DA. Tuberculosis in Aboriginals in Canada. *Int J Tuberc Lung Dis* 1998; **2**(9 SUPPL. 1): S16-S22.
87. Nolan CM, Elarth AM. Tuberculosis in a cohort of Southeast Asian Refugees. A five-year surveillance study. *Am Rev Respir Dis* 1988; **137**(4): 805-9.
88. Fine PE, Sterne JA, Ponnighaus JM, Rees RJ. Delayed-type hypersensitivity, mycobacterial vaccines and protective immunity. *Lancet* 1994; **344**(8932): 1245-9.
89. MacIntyre CR, Plant AJ. Longitudinal incidence of tuberculosis in South-East Asian refugees after re-settlement. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 1999; **3**(4): 287-93.
90. Oliver G. Tuberculosis notifications in Australia, 1994. *Commun Dis Intell* 1996; **20**(5): 108-15.
91. Marks GB, Bai J, Simpson SE, Sullivan EA, Stewart GJ. Incidence of tuberculosis among a cohort of tuberculin-positive refugees in Australia: reappraising the estimates of risk. *Am J Respir Crit Care Med* 2000; **162**(5): 1851-4.
92. Marks GB, Bai J, Stewart GJ, Simpson SE, Sullivan EA. Effectiveness of postmigration screening in controlling tuberculosis among refugees: a historical cohort study, 1984-1998. *Am J Public Health* 2001; **91**(11): 1797-9.
93. Roche P, Merianos A, Antic R, et al. Tuberculosis notifications in Australia, 1999. *Commun Dis Intell Q Rep* 2001; **25**(4): 254-60.
94. Choudhury IW, West CR, Ormerod LP. The outcome of a cohort of tuberculin-positive predominantly South Asian new entrants aged 16-34 to the UK: Blackburn 1989-2001. *J Public Health (Oxf)* 2014; **36**(3): 390-5.
95. Anderson SR, Maguire H, Carless J. Tuberculosis in London: a decade and a half of no decline [corrected]. *Thorax* 2007; **62**(2): 162-7.
96. Daley CL, Hahn JA, Moss AR, Hopewell PC, Schecter GF. Incidence of tuberculosis in injection drug users in San Francisco: impact of anergy. *Am J Respir Crit Care Med* 1998; **157**(1): 19-22.
97. Murray JF. A century of tuberculosis. *Am J Respir Crit Care Med* 2004; **169**(11): 1181-6.
98. Moss AR, Hahn JA, Tulsky JP, Daley CL, Small PM, Hopewell PC. Tuberculosis in the homeless - A prospective study. *Am J Respir Crit Care Med* 2000; **162**(2): 460-4.
99. Cook VJ, Hernández-Garduño E, Elwood RK. Risk of tuberculosis in screened subjects without known risk factors for active disease. *Int J Tuberc Lung Dis* 2008; **12**(8): 903-8.
100. Public Health Agency of Canada. Tuberculosis in Canada, 2012. Ottawa (Canada): Minister of Public Works and Government Services Canada;, 2015.
101. Scolari C, El-Hamad I, Matteelli A, et al. Incidence of tuberculosis in a community of Senegalese immigrants in Northern Italy. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 1999; **3**(1): 18-22.
102. Martin V, Guerra JM, Cayla JA, Rodriguez JC, Blanco MD, Alcoba M. Incidence of tuberculosis and the importance of treatment of latent tuberculosis infection in a Spanish prison population. *Int J Tuberc Lung Dis* 2001; **5**(10): 926-32.
103. Klein RS, Gourevitch MN, Teeter R, Schoenbaum EE. The incidence of tuberculosis in drug users with small tuberculin reaction sizes. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2001; **5**(8): 707-11.
104. Mojazi-Amiri H, Larppanichpoonphol P, Nugent K. Tuberculosis reactivation in referrals to public health clinics in Texas. *The American journal of the medical sciences* 2013; **346**(6): 442-6.
105. Leung CC, Yew WW, Chang KC, et al. Risk of active tuberculosis among schoolchildren in Hong Kong. *Arch Pediatr Adolesc Med* 2006; **160**(3): 247-51.
106. Leung CC, Yew WW, Au KF, et al. A strong tuberculin reaction in primary school children predicts tuberculosis in adolescence. *Pediatr Infect Dis J* 2012; **31**(2): 150-3.
107. Government of the HKSAR CfHP. Tuberculosis Manual. Hong Kong SAR: Tuberculosis and Chest Service,Public Health Services Branch, 2006.

108. Chan-Yeung M, Dai DL, Cheung AH, et al. Tuberculin skin test reaction and body mass index in old age home residents in Hong Kong. *J Am Geriatr Soc* 2007; **55**(10): 1592-7.
109. Hemmati M, Ghadiri K, Rezaei M. Tuberculin reactivity in school Age children; five-year follow-up in Iran. *Iranian Journal of Pediatrics* 2011; **21**(1): 39-44.
110. Joshi R, Narang U, Zwerling A, et al. Predictive value of latent tuberculosis tests in Indian healthcare workers: a cohort study. *Eur Respir J* 2011; **38**(6): 1475-7.
111. Roth DZ, Ronald LA, Ling D, et al. Impact of interferon-gamma release assay on the latent tuberculosis cascade of care: A population-based study. *Eur Respir J* 2017; **49**(3): 1601546.
112. Tsou P-H, Huang W-C, Huang C-C, et al. Quantiferon TB-Gold conversion can predict active tuberculosis development in elderly nursing home residents. *Geriatrics & gerontology international* 2015; **15**(10): 1179-84.
113. Grinsdale JA, Islam S, Tran OC, Ho CS, Kawamura LM, Higashi JM. Interferon-Gamma Release Assays and Pediatric Public Health Tuberculosis Screening: The San Francisco Program Experience 2005 to 2008. *Journal of the Pediatric Infectious Diseases Society* 2016; **5**(2): 122-30.
114. Harstad I, Winje BA, Heldal E, Oftung F, Jacobsen GW. Predictive values of QuantiFERON-TB Gold testing in screening for tuberculosis disease in asylum seekers. *Int J Tuberc Lung Dis* 2010; **14**(9): 1209-11.
115. Andrews JR, Hatherill M, Mahomed H, et al. The dynamics of QuantiFERON-TB gold in-tube conversion and reversion in a cohort of South African adolescents. *Am J Respir Crit Care Med* 2015; **191**(5): 584-91.
116. Mahomed H, Hawkridge T, Verver S, et al. The tuberculin skin test versus QuantiFERON TB Gold® in predicting tuberculosis disease in an adolescent cohort study in South Africa. *PLoS One* 2011; **6**(3).
117. Mahomed H, Ehrlich R, Hawkridge T, et al. TB incidence in an adolescent cohort in South Africa. *PLoS One* 2013; **8**(3): e59652.
118. Hermansen TS, Lillebaek T, Langholz Kristensen K, Andersen PH, Ravn P. Prognostic value of interferon-gamma release assays, a population-based study from a TB low-incidence country. *Thorax* 2016; **71**(7): 652-8.
119. Chigbu LN, Iroegbu CU. Incidence and spread of Mycobacterium tuberculosis-associated infection among Aba Federal prison inmates in Nigeria. *Journal of Health, Population and Nutrition* 2010; **28**(4): 327-32.
120. Nduba V, Van't Hoog AH, Mitchell EMH, Borgdorff M, Laserson KF. Incidence of Active Tuberculosis and Cohort Retention Among Adolescents in Western Kenya. *The Pediatric infectious disease journal* 2018; **37**(1): 10-5.
121. World Health Organization. Global tuberculosis report 2014: World Health Organization, 2014.
122. Azoulay D, Abiteboul D, Gangloff C, et al. Two-year follow-up study of a cohort of hospital health care workers with a positive QuantiFERON test. *Archives des Maladies Professionnelles et de l'Environnement* 2015; **76**(6): 559-67.
123. Bunyasi EW, Luabeya AKK, Tameris M, et al. Impact of isoniazid preventive therapyonthe evaluation of longterm effectiveness of infant MVA85A vaccination. *Int J Tuberc Lung Dis* 2017; **21**(7): 778-83.
124. Winje BA, White R, Syre H, et al. Stratification by interferon-gamma release assay level predicts risk of incident TB. *Thorax* 2018.
125. Du F, Zhang Z, Gao T, et al. Diagnosis of latent tuberculosis by ELISPOT assay and tuberculin skin test. *Med Mal Infect* 2016; **46**(3): 150-3.
126. Gao L, Zhang HR, Xin HN, et al. Short-course regimens of rifapentine plus isoniazid to treat latent tuberculosis infection in older Chinese patients: a randomised controlled study. *Eur Respir J* 2018; **52**(6).
127. Grzybowski S, Allen EA. The Challenge of Tuberculosis in Decline. A Study Based on the Epidemiology of Tuberculosis in Ontario, Canada. *Am Rev Respir Dis* 1964; **90**: 707-20.
128. Barnett GD, Grzybowski S, Styblo K. [The current risk of contracting evolutive tuberculosis, in Saskatchewan, according to the state of previous tuberculin tests and x-ray image]. *Bull Int Union Tuberc* 1971; **45**: 55-79.
129. Stead WW, Lofgren JP. Does the risk of tuberculosis increase in old age? *J Infect Dis* 1983; **147**(5): 951-5.
130. Horsburgh CR, O'Donnell M, Chamblee S, et al. Revisiting Rates of Reactivation Tuberculosis. *Am J Respir Crit Care Med* 2010; **182**(3): 420-5.
131. Mulder C, van Deutekom H, Huisman EM, et al. Role of the QuantiFERON(R)-TB Gold In-Tube assay in screening new immigrants for tuberculosis infection. *Eur Respir J* 2012; **40**(6): 1443-9.
132. Mulder C, Mulleners B, Borgdorff MW, van Leth F. Predictive value of the tuberculin skin test among newly arriving immigrants. *PLoS ONE [Electronic Resource]* 2013; **8**(3): e60130.
133. Shea KM, Kammerer JS, Winston CA, Navin TR, Horsburgh CR, Jr. Estimated rate of reactivation of latent tuberculosis infection in the United States, overall and by population subgroup. *Am J Epidemiol* 2014; **179**(2): 216-25.
134. Borgdorff MW, van den Hof S, Kremer K, et al. Progress towards tuberculosis elimination: secular trend, immigration and transmission. *Eur Respir J* 2010; **36**(2): 339-47.

135. Styblo K. The elimination of tuberculosis in The Netherlands. *Bull Int Union Tuberc Lung Dis* 1990; **65**(2-3): 49-55.
136. Winje BA, Groneng GM, White RA, Akre P, Aavitsland P, Heldal E. Immigrant screening for latent tuberculosis infection: numbers needed to test and treat, a Norwegian population-based cohort study. *BMJ open* 2019; **9**(1): e023412.
137. Winje BA, Oftung F, Korsvold GE, et al. Screening for tuberculosis infection among newly arrived asylum seekers: comparison of QuantiFERONTB Gold with tuberculin skin test. *BMC Infect Dis* 2008; **8**: 65.

APPENDIX 7: HGDISCOVERY: AN ONLINE TOOL PROVIDING FUNCTIONAL AND PHENOTYPIC INFORMATION ON NOVEL VARIANTS OF HOMOGENTISATE 1,2-DIOXIGENASE

HGDiscovery: an online tool providing functional and phenotypic information on novel variants of homogentisate 1,2- dioxigenase

Malancha Karmakar[1,2,3,#], Vittoria Cicaloni[1,2,4,#], Carlos H.M. Rodrigues[1,2,3], Ottavia Spiga[4], Annalisa Santucci[4], David B. Ascher[1,2,4,5,*]

[1] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[2] Structural Biology and Bioinformatics, Department of Biochemistry, University of Melbourne, Melbourne, Victoria, Australia

[3] Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia

[4] Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Siena, Italy

[5] Department of Biochemistry, Bio21 Institute, University of Cambridge, Cambridge, UK

# These authors contributed equally.

*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au.

Abstract

Alkaptonuria (AKU), a rare genetic disorder, is characterized by the accumulation of homogentisic acid (HGA) in the body. Affected individuals lack enough functional levels of an enzyme required to breakdown HGA. Mutations in the *HGD* gene cause AKU and they are responsible for deficient levels of functional homogentisate 1,2-dioxygenase (HGD), which, in turn, leads to excess levels of HGA. Although HGA is rapidly cleared from the body by the kidneys, in the long term it starts accumulating in various tissues, especially cartilage. Over time (rarely before adulthood), it eventually changes the color of affected tissue to slate blue or black. Here we report a comprehensive mutation analysis of 111 pathogenic and 190 non-pathogenic HGD missense mutations using protein structural information. Using our comprehensive suite of graph-based signature methods, mCSM complemented with sequence-based tools, we studied the functional and molecular consequences of each mutation on protein stability, interaction and evolutionary conservation. The scores generated from the structure and sequence-based tools were used to train a supervised machine learning algorithm with 84% accuracy. The empirical classifier was used to generate the variant phenotype for novel HGD missense mutations. All this information is deployed as a user friendly freely available web server called HGDiscovery (http://biosig.unimelb.edu.au/hgdiscovery/).

Introduction

Alkaptonuria (AKU) is a rare recessive metabolic disorder which was used by Sir Archibald Garrod in his Croonian lectures to describe inborn errors of metabolism [1]. It is a hereditary disorder, resulting from mutations in the enzyme homogentisate 1,2 dioxygenase (HGD) (EC 1.13.11.5), responsible for the breakdown of homogentisic acid (HGA) which is an intermediate metabolite in the tyrosine degradation pathway [2]. With blockage in tyrosine metabolism, elevated levels of HGA leads to deposition of its own polymers as an ochronotic pigment in the connective tissue including cartilage, heart valves, and sclera [3]. Manifestation of disease during early childhood is seen as "homogentisic aciduria", which is darkening of the urine upon standing. Delayed symptoms can be seen after 30 years of age which involves "ochronosis" – pigmentation of collagenous tissues like cardiac valves, eyes, ears and skin [4]. Current estimates of the disease occurrence in the Unites States obtained from the National Organisation of Rare Disorders is 1 in 250,000 – 1,00,000 live births [5].

HGD gene located on chromosome 3q21-q23 [6], is a single copy gene composed of 14 exons [7]. Due to compound heterozygosity or homozygosity of HGD gene variants, the enzymatic defect in HGD is autosomal recessive [6, 8]. Information on all variants identified till date globally have been documented in the HGD mutation database (http://hgddatabase.cvtisr.sk/).
The experimental crystal structure of the HGD protein has been solved (PDB code 1EY2 and 1EYB) in 2000. The HGD protein protomer (NP_000178.2), is composed of 445 amino acids, which includes a 280 residue N-terminal domain, a central β-sandwich and a 140 residue C-terminal domain [8]. It is a complex hexameric protein arranged as a dimer of trimers [9]. It is principally expressed in osteoarticular compartment cells (i.e. chondrocytes, synoviocytes and osteoblasts) [10] and in prostate, small intestine, colon, kidney and liver [7]. The spatial structure of the protomer, two-disc like trimers and the hexamer are maintained by an intricate network of non-covalent inter and intra-molecular interaction. This makes the protein structure extremely vulnerable to mutations [11].

The major obstacle in studying an ultra-rare and complex disease like AKU is the lack of a standardized methodology to assess disease severity and response to treatment [12], which is complicated by the fact that AKU symptoms differ from one individual to another. Detailed evaluation and comparison of clinical and genomic data of AKU patient can play a key role to understand AKU variability. An in-depth molecular characterization of the disease is needed in pharmacogenomics prediction for suitable medical treatment. To address the issue we developed ApreciseKUre platform, which includes data on potential biomarkers, patients' quality of life, biochemical outcomes and clinical information facilitating their integration and

analysis in order to shed light on pathological characterization of every AKU patient in a typical Precision Medicine perspective [13-16] .

We wanted to further elaborate and build a new database which would complement the existing ApreciseKUre database. The new database would provide the necessary underlying molecular information for novel and known clinical HGD variants. We have tried to exploit structural and sequence based information to build a predictive tool using supervised machine learning algorithm. The model has been implemented through the webserver HGDiscovery, providing functional and phenotypic consequences of HGD non-synonymous variations to better guide clinical decisions.

Methods

Data curation

After removal of duplicate mutations, we curated a dataset composed of 301 non-synonymous substitutions. It included 190 non-pathogenic non-synonymous variations retrieved from gnomAD v.3 (Genome build GRCh38/hg38, Ensembl gene ID: ENSG00000113924.11, Region 3:120628173-120682571) [17] and 111 AKU-causing clinical mutations. The 111 variants were first described in the study of Ascher et al. 2019 [18] and included in HGD Mutation Database (http://hgddatabase.cvtisr.sk) [19], which summarizes results of mutation analysis from approximately 530 AKU patients reported so far.

HGD protein structure

The X-ray crystallographic 3D structure of *Homo sapiens* holo-HGD (holo-HGDHs, PDB ID: 1EY2) is incomplete; thus, it needed structural reconstruction of the missing residues of the monomer and then of the whole hexamer in order to be able to perform a complete evaluation of variants effect on protein stability and flexibility. The missing loop in the human protein structure (residues 348–355) was reconstructed by homology modeling using the *Pseudomonas putida* HGD (HGDPp) structure. By using protein BLAST [20] software we found three structures belonging to *Pseudomonas putida* with a sequence identity (the amount of characters which match exactly between two different sequences) larger than 49% and with root-mean-square deviation (RMSD) amounting to 1.8 Å for Cα [21]. We opted for HGDPp, with PDB ID 4AQ2 since, similarly to 1EY2, as it had no substrate. The structures of holo-HGDHs (PDB ID: 1EY2) and its homologous HGDPp (PDB ID: 4AQ2) were retrieved from the Protein Data Bank (PDB) [22]. Thereafter at the 1EY2 and 4AQ2 sequences alignment on BLAST web server [20], we modelled the missing residues. The modelling of the loop 348-355 was carried out using a homology model approach in which an elucidated structure of HGDPp loop was employed as template to

model the structure of the protein of interest. The completed monomer structure served as a starting point for the reconstruction of the whole HGDHs oligomeric protein on the template of the asymmetric units of PDB entry 1EY2. The structure reliability was validated using PROCHECK [23]. Additionally, the energy minimization of the hexameric protein was performed using GROMACS 5.0.2 [24] in order to obtain an optimized 3D structure, a relaxation of the highly energetic conformations and a correct geometry for the following simulations (for additional information see Supplementary Methods in [18]).

Biophysical and evolutionary score generation

A thorough structural and sequence based assessment was performed for all the HGD variants to account for the potential effects of AKU-causing mutations. Variations in protein-protein interactions between the different monomers of the hexamer HGD upon mutation was determined using mCSM-PPI2 [25]. Changes in protein stability and folding were determined using our in-house tools like mCSM-Stability [26], SDM [27] and DUET [28] and conformational flexibility changes using the normal mode analysis tool called DynaMut [29]. Effects of mutations on binding affinity of HGD to its substrate homogentisic acid were analyzed using mCSM-Lig [30]. All these are novel machine learning approaches that use graph-based signatures to represent the structural and biochemical environment of the wild-type 3D structure of a protein to quantitatively predict the effects of point mutation. To complement the above methods we used sequence based feature like SNAP2 (Screening for Non-Acceptable Polymorphisms) [31], ConSurf [32] and Provean (Protein Variation Effect Analyzer) [33] which provides valuable evolutionary information. To enrich the analysis we included protein's wild type structural information such as residue depth, dihedral angles of the HGD chain ϕ (phi) and ψ (psi), relative solvent accessibility and secondary structure information. We calculated changes in molecular interactions such as hydrophobic, ionic, van der Waals', halogen and hydrogen bonds and π interactions (cation–π, donor–π, halogen–π, carbon–π, π–π) between the wild type and mutant structures using Arpeggio [34]. We also included population-based variability using the missense tolerance ratio (MTR) [35] scoring system.

Supervised Machine learning for empirical model building

We evaluated different supervised machine learning algorithms for classification which is available within the scikit-learn Python library. These include – K-Nearest Neighbors (KNN), Random Forest, Decision Trees, Extra Trees, AdaBoost, Gradient Boosting, SVM, Gaussian Naïve Bayes, and Stochastic Gradient Descent. The best performing model was chosen by assessing metrics like Matthews correlation co-efficient (MCC), Receiver Operating Characteristic (AUROC) curve, accuracy, F1-score and precision. The model was trained using stratified 10-fold cross validation. We carefully split the train and blind test dataset non-redundantly with respect to the amino acid residue position.

To address the issue of imbalance between the pathogenic and non-pathogenic mutations in the data, we evaluated the model performance by both under-sampling the non-pathogenic mutations and oversampling pathogenic mutations in the train dataset [36]. The performance was compared for above mentioned scenario and the normal dataset and best results were obtained when the pathogenic mutations were oversampled using the Extra Tree algorithm. **Ex**tremely **ra**ndomized **tree** classifier (or Extra Tree) is an ensemble machine learning algorithm and a variation of the random forest algorithm. The empirical binary classifier built using this algorithm highlights a set of structural and evolutionary features which can be used to discriminate between AKU-causing and non-pathogenic variations.

Webserver development

We have implemented HGDiscovery as a user-friendly and freely available webserver (http://biosig.unimelb.edu.au/hgdiscovery/). The front-end of the server was developed using Materializecss framework version 1.0.0, while the back-end was built in Python using the Flask framework version 1.0.2. The server is hosted on a Linux server running Apache 2.

Results

In this work we have used the 3D protein structure to understand the functional and molecular consequences of mutations in HGD leading to AKU disease and using the information generated from these analyses we have trained a supervised machine learning algorithm to develop a predictive tool to determine novel variants which could lead to AKU manifestation. Figure 1 depicts the novel methodological pipeline we have developed.

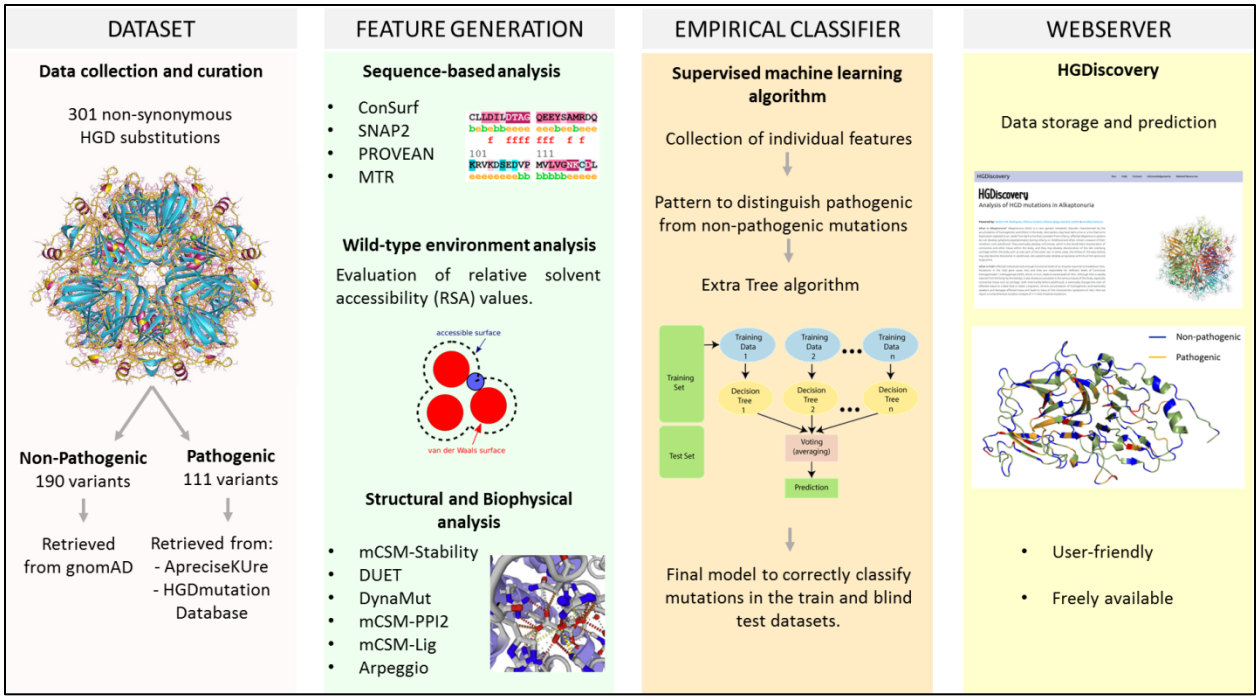


Figure 1: HGDiscovery workflow. The first step involves scoping published literature and clinical databases to prepare a curated list of non-synonymous HGD mutations. The second step involves generating various structure and sequence based features for the curated missense mutations. In the third step, we use these features in a supervised machine learning algorithm to build a binary classifier, which can distinguish between pathogenic and non-pathogenic missense mutations. Finally, we develop a free available user friendly webserver which contains phenotypic information on all HGD variants.

Sequence-based analysis of HGD variants

ConSurf, SNAP2 and PROVEAN are sequence-based predictors and consider evolutionary information to predict functionally important non-synonymous mutation. The prediction helps us understand the biological impact of a mutation on the protein structure. A consistent pattern was observed from all of the sequence based features. The pathogenic mutations were associated with deleterious scores and the non-pathogenic mutations scored neutral. All the features were statistically significant to be used to train the predictive algorithm to build the empirical tool (p-values SNAP2: 4.6 e^{-14}, PROVEAN: 1.1 e^{-9}, ConSurf: 2.4 e^{-10}). Population-based variability was considered using the missense tolerance ratio (MTR) scoring system. Majority of the pathogenic mutations were in the bottom 25th percentile, reflecting intolerance and hence associated with altering protein function.

Wild-type environment analysis

The wild-type environment analysis includes data on relative solvent accessibility (RSA), residue depth, dihedral angles and secondary structure information for both pathogenic and non-pathogenic variants. Looking into the relative solvent accessibility values for the pathogenic and non-pathogenic mutations (p-value: 2.2 e^{-8}), we see pathogenic mutations tend to be more exposed than non-pathogenic variants. It has been previously described that the HGD protomer structure constitutes of a pore in which the side chains of large number of residues are exposed [21]. These residues are thought to play an important part in the complex HGD catalytic function and we see subtle changes in the side chains as non-synonymous substitution can affect the active site functionality [18]. The residue depth values reveal pathogenic mutations are more buried than non-pathogenic mutations. This observation is congruous with earlier observation where point mutations on the surface were better tolerated in the globular hexameric HGD protein structure.

Structural and Biophysical analysis

Our in-house biophysical tools mCSM-Stability [26], DUET [28] and DynaMut [29] were used to study and understand the impact of missense mutations on protein stability, folding and conformational flexibility. These tools are novel machine-learning algorithms which rely on graph-based signatures to calculate changes in Gibb's free energy upon non-synonymous mutations. We observed pathogenic mutations to be associated with highly destabilizing scores affecting protein stability and dynamics. The effects of mutation on the substrate binding affinity to active site were determined using mCSM-Lig [30]. Pathogenic mutations altered the active / substrate binding pocket. mCSM-PPI2 [25] was used to assess changes in protein-protein interaction and we observed pathogenic mutations hindered the formation of the symmetrical homohexamer. Therefore, pathogenic mutations either reduced or disrupted the HGD protein activity.

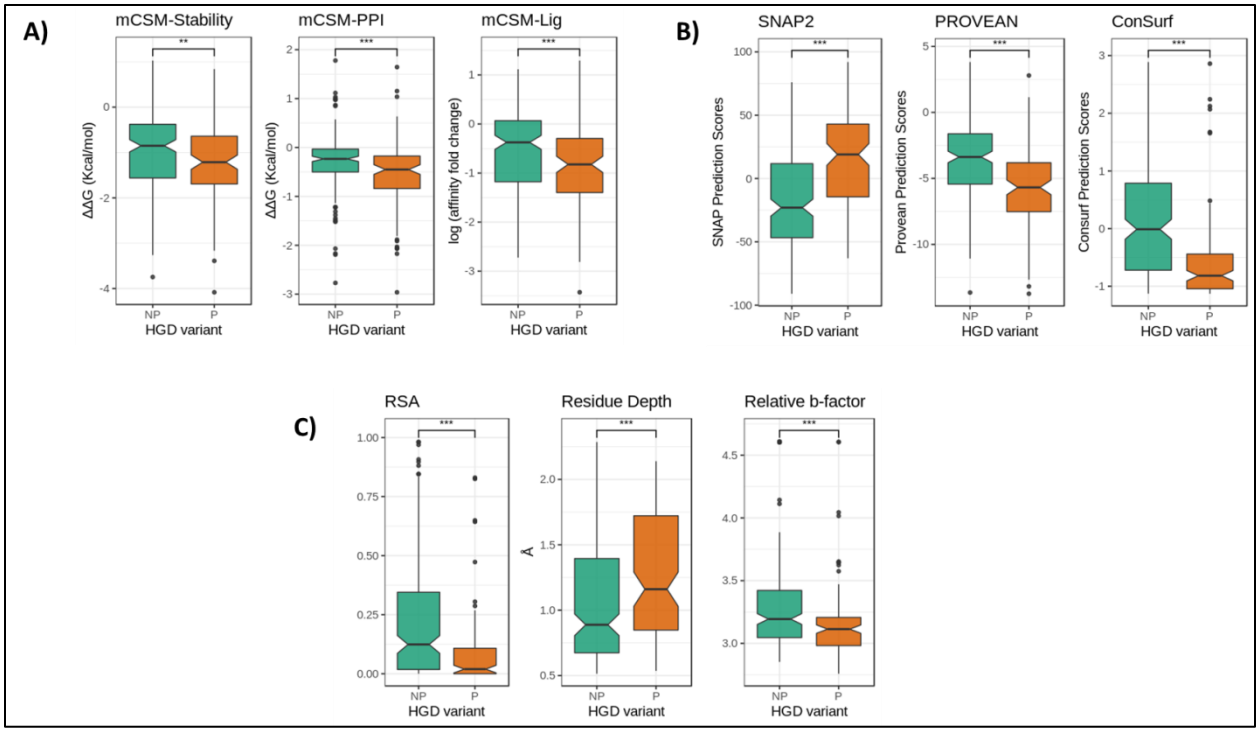


Figure 2: Boxplot representation of features. A) Structural features. B) Sequence based features. C) Wild-type environment features. The non-pathogenic mutations (NP) are represented as sea green and pathogenic mutations (P) as dark orange. (*** $p < 0.0001$, ** $p < 0.001$, Welch two sample t-test).

Supervised machine learning algorithm: Extra Tree

Our features could be grouped into eight distinct categories – protein stability, protein-protein interactions, ligand affinity, evolutionary conservation scores, distance parameters, MTR scores, molecular interaction and backbone geometry. Each category of features was initially used to build and evaluate the performance of the predictive model. After a thorough analysis of the individual features, we combined them together to see if there is a pattern which could be used to distinguish pathogenic from non-pathogenic HGD mutations. We observed that when different categories of features were combined together, in addition to using stratified 10-fold cross validation with Extra Tree algorithm, yielded a more robust and balanced performance. The Extra Tree algorithm implements a meta estimator that fits randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and reduces over-fitting [37].

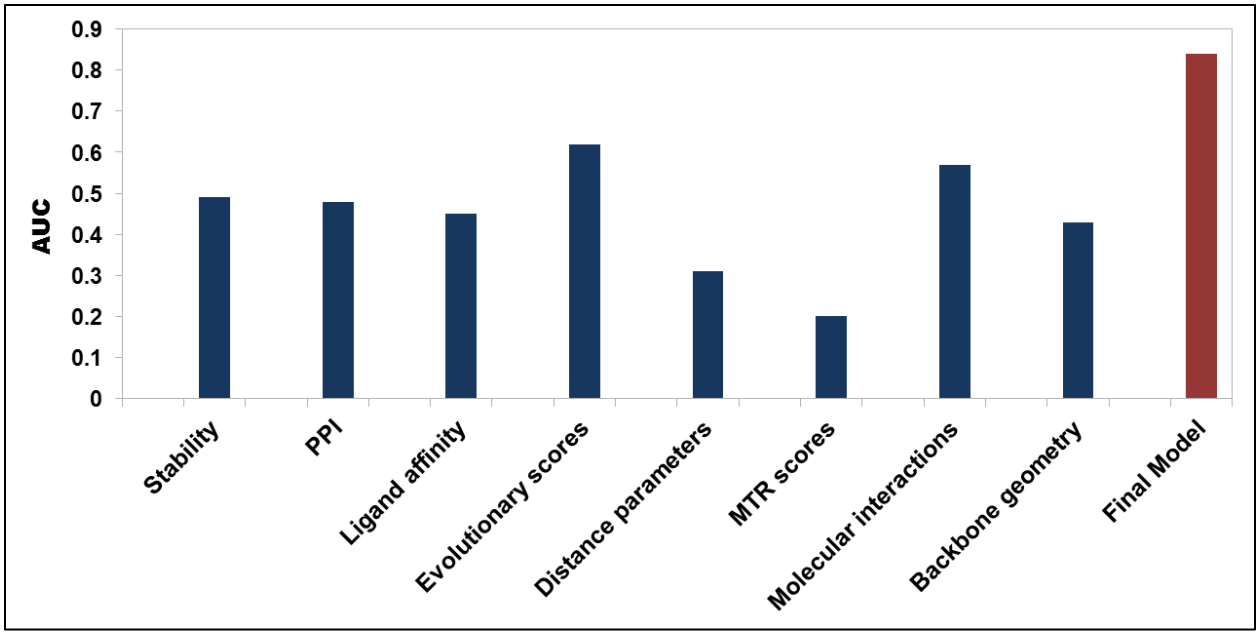


Figure 3: Empirical model performance trained on individual class of features. The Extra Tree algorithm was trained using stratified 10-fold cross validation using eight distinct class of features (first eight bars from left to right; dark blue bars) and with a combination of all features (red bar). The AUC scores is low when a single class of feature is used for training the binary classifier, however, a significant improvement is noticed when all the eight different features are combined to build the model.

190 non-pathogenic and 111 pathogenic mutations were split into non-redundant train and blind test datasets with respect to their amino acid position. Initially we observed poor performance on the model's ability to predict pathogenic mutation. We concluded that the train data set was imbalanced as there were more non-pathogenic mutations than pathogenic mutations. We improved the metric scores by oversampling (duplicating) [36] the pathogenic mutations in the train dataset. The final model correctly classified 84% and 73% of mutations in the train and blind test datasets respectively.

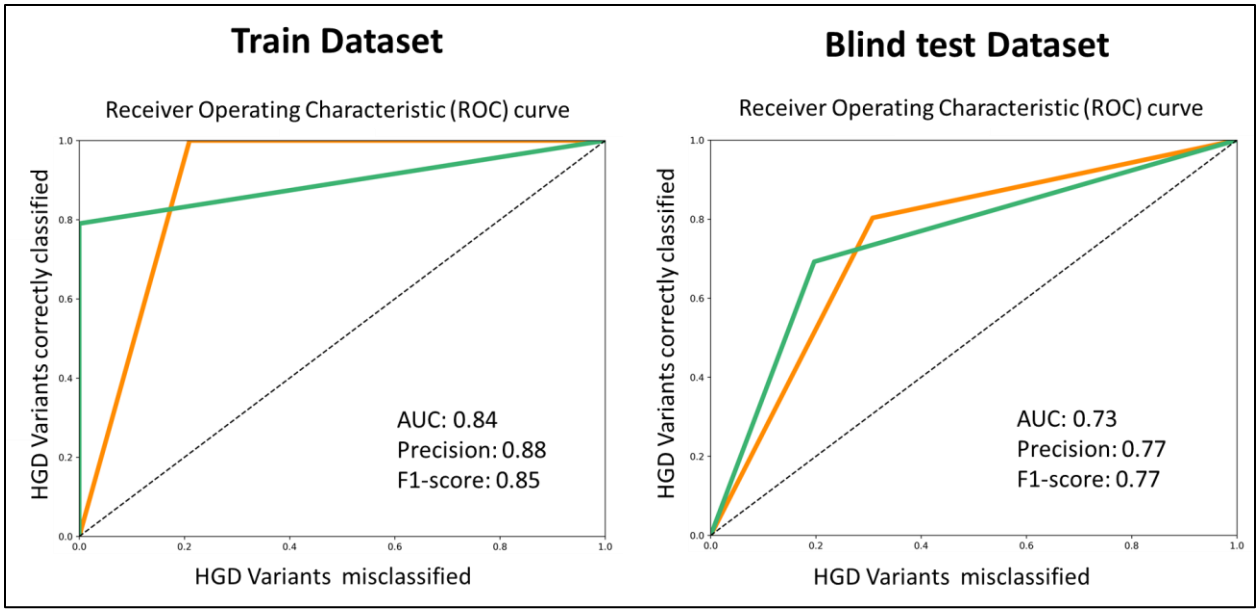


Figure 4: Receiver Operating Characteristic (ROC) curves of HGD classifier. The evaluation metrics shown for train and test dataset where pathogenic mutations are represented in dark orange and non-pathogenic mutations in sea green. (AUC = area under the curve).

HGDiscovery Webserver

HGDiscovery allows for users to query for a single point mutation or submit a list of mutations to be analysed in batch. For the "Single Mutation" option users are asked to provide the point mutation as a string containing the wild-type reside one-letter code, its corresponding residue number and the mutant residue one-letter code. The "Mutation List" option requires that a text file is submitted with the list of mutations (one per line).

The results page for the "Single Mutation" option displays the predicted outcome on the top alongside with details of the input mutation, wild-type residue environment, the variables and scores used by our predictive model and external links to experimental evidence (when available). An interactive 3D viewer using the NGL-viewer [38] shows the molecular contacts generated by Arpeggio [34] for wild-type and mutant structures.

On the "Mutation List" option, the results are displayed as a downloadable table. Individual analysis for each variant on the table can be analysed similarly to "Single Mutation" option by clicking the "Details" button. An interactive viewer is also shown at the bottom of the page highlighting Pathogenic and Non-pathogenic mutations on the 3D structure.

Discussion

Here we present an empirical classifier HGDiscovery, which has phenotypic information on all variants of homogentisate 1,2 dioxygenase, (EC 1.13.11.5), an enzyme involved in the metabolism of tyrosine, whose deficiency leads to Alkaptonuria [OMIM 203500]. We combine structural, evolutionary and molecular information from known HGD variations and look to investigate a pattern to distinguish non-pathogenic from AKU-causing non-synonymous variants. So along with physiological information from ApreciseKUre platform, we have an additional AKU-dedicated database which provides new insight into functional and phenotypic consequences of novel HGD non-synonymous variations, crucial for a genetic disease like AKU to support clinical decisions.

The 3D crystal structure of the HGD active form reveals a highly complex and dynamic hexameric organization comprising two disk-like trimers [9]. An intricate network of noncovalent interactions is needed to maintain the spatial structure firstly of the protomer, the trimer and then the hexamer. This delicate structure presents a very low tolerance to mutations and can be easily disrupted mainly by missense variants compromising enzyme function. In case of HGD, missense variants represent approximately 65% of all known AKU substitutions [4, 11, 39] and 93 distinct amino acid residue positions within the structure are affected by the 111 AKU-causing missense changes. Recent studies on evolutionary conservation revealed that AKU variants were mainly located at more conserved residue positions [18] and, consequently, HGD missense changes can influence protein folding and stability or interactions with other protomers or substrate. Specifically, they can decrease stability of individual protomers, disrupt protomer–protomer interactions, or modify residues in the active-site region. Thus, when a novel HGD missense mutation is identified, it is important to distinguish causal AKU variants from non-pathogenic ones.

With sequence-based tools such as ConSurf, SNAP2 and PROVEAN we have evaluated evolutionary information in order to predict functionally important non-synonymous mutations and the biological impact of a mutation on HGD protein structure. The obtained results supported our hypothesis: the pathogenic mutations were associated with deleterious scores whereas the non-pathogenic mutations with neutral scores. Additionally, using MTR score system we have analyzed population-based variability and most of the pathogenic mutations resulted to be in the bottom 25^{th} percentile, reflecting intolerance and alteration of protein function. With the help of biophysical tools (i.e. mCSM-Stability, DUET and DynaMut) we investigated the impact of missense mutations on protein stability, folding and conformational flexibility. AKU-causing mutations appear to reduce or disrupt the HGD protein activity by destabilizing its structure and altering the active site/substrate binding pocket.

It is not uncommon that AKU patients carry compound heterozygotes for two HGD gene variants. In such cases, the estimation of the role of each missense variant is not trivial, since the hexamer could be assembled with monomers all affected by the same variant (homo-oligomer) or by two different ones (heterooligomer) [40]. Variants affecting two different regions could have additive destructive effect, on the contrary, the effects could partially compensate for those that belong to the same region. However, we do not have any tools able to evaluate such events so far [12]. Compound heterozygosity could have even interfered with our analysis, where a variant labelled as non-pathogenic could actually be pathogenic. This was the limitation of our study. But with increasing availability of genomic and clinical data after patient analysis in future, we can always update our tool and re-label the mislabeled non-synonymous variants.

The information available from the above study can be used to develop new treatment strategies, for example, use of small molecules. We know that a pathogenic mutation with destabilizing scores for stability and flexibility leading to reduced enzyme activity can be rescued partially or totally with the help of a small molecule and hence might decrease the severity of the disease [18]. Moreover, understanding the protein structure and function would also help in designing tailored drugs and therapies.

Therefore, this framework may represent an online tool that can be turned into a best practice model for Rare Diseases. We believe this is not limited to the study of AKU, but it represents a proof of principle study that could be applied to other rare diseases, allowing data management, analysis and interpretation. We applied this novel methodological pipeline to understand and determine novel drug resistant mutations in tuberculosis [41, 42] and even performed a real-time analysis [43] on tuberculosis patient. Hence, HGDiscovery is a user friendly freely available tool which could help with faster and more accurate diagnosis of AKU.

Acknowledgements

M.K and C.H.M.R were funded by Melbourne Research Scholarships. D.B.A. was funded by a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council and Fundacao de Amparo a Pesquisa do Estado de Minas Gerais(FAPEMIG) [MR/M026302/1]; the Jack Brockhoff Founda-tion [JBF 4186, 2016]; and an Investigator Grant from the National Health and Medical Research Council of Australia[GNT1174405]. Supported in part by the Victorian Government's OIS Program.

References

1. Garrod, A.E., *The incidence of alkaptonuria: a study in chemical individuality. 1902 [classical article].* The Yale journal of biology and medicine, 2002. **75**(4): p. 221-231.
2. Phornphutkul, C., et al., *Natural history of alkaptonuria.* N Engl J Med, 2002. **347**(26): p. 2111-21.
3. Damarla, N., et al., *Alkaptonuria: A case report.* Indian journal of ophthalmology, 2017. **65**(6): p. 518-521.
4. Zatkova, A., L. Ranganath, and L. Kadasi, *Alkaptonuria: Current Perspectives.* Appl Clin Genet, 2020. **13**: p. 37-47.
5. Disorders, N.O.f.R., NORD, 2019. **https://rarediseases.org/rare-diseases/alkaptonuria/**.
6. Pollak, M.R., et al., *Homozygosity mapping of the gene for alkaptonuria to chromosome 3q2.* Nat Genet, 1993. **5**(2): p. 201-4.
7. Fernandez-Canon, J.M., et al., *The molecular basis of alkaptonuria.* Nat Genet, 1996. **14**(1): p. 19-24.
8. Janocha, S., et al., *The human gene for alkaptonuria (AKU) maps to chromosome 3q.* Genomics, 1994. **19**(1): p. 5-8.
9. Titus, G.P., et al., *Crystal structure of human homogentisate dioxygenase.* Nat Struct Biol, 2000. **7**(7): p. 542-6.
10. Laschi, M., et al., *Homogentisate 1,2 dioxygenase is expressed in human osteoarticular cells: implications in alkaptonuria.* J Cell Physiol, 2012. **227**(9): p. 3254-7.
11. Nemethova, M., et al., *Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy.* Eur J Hum Genet, 2016. **24**(1): p. 66-72.
12. Ranganath, L.R. and T.F. Cox, *Natural history of alkaptonuria revisited: analyses based on scoring systems.* J Inherit Metab Dis, 2011. **34**(6): p. 1141-51.
13. Cicaloni, V., et al., *Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease.* Faseb j, 2019. **33**(11): p. 12696-12703.
14. Spiga, O., et al., *Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease.* Orphanet J Rare Dis, 2020. **15**(1): p. 46.
15. Spiga, O., et al., *A new integrated and interactive tool applicable to inborn errors of metabolism: Application to alkaptonuria.* Comput Biol Med, 2018. **103**: p. 1-7.
16. Spiga, O., et al., *ApreciseKUre: an approach of Precision Medicine in a Rare Disease.* BMC Med Inform Decis Mak, 2017. **17**(1): p. 42.
17. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-91.
18. Ascher, D.B., et al., *Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU.* Eur J Hum Genet, 2019. **27**(6): p. 888-902.
19. Zatkova, A., et al., *Identification of 11 Novel Homogentisate 1,2 Dioxygenase Variants in Alkaptonuria Patients and Establishment of a Novel LOVD-Based HGD Mutation Database.* JIMD Rep, 2012. **4**: p. 55-65.
20. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
21. Jeoung, J.-H., et al., *Visualizing the substrate-, superoxo-, alkylperoxo-, and product-bound states at the nonheme Fe(II) site of homogentisate dioxygenase.* Proceedings of the National Academy of Sciences, 2013. **110**(31): p. 12625.
22. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

23. Laskowski, R., et al., *PROCHECK: A program to check the stereochemical quality of protein structures.* Journal of Applied Crystallography, 1993. **26**: p. 283-291.
24. Berendsen, H.J.C., D. van der Spoel, and R. van Drunen, *GROMACS: A message-passing parallel molecular dynamics implementation.* Computer Physics Communications, 1995. **91**(1): p. 43-56.
25. Rodrigues, C.H.M., et al., *mCSM-PPI2: predicting the effects of mutations on protein–protein interactions.* Nucleic Acids Research, 2019. **47**(W1): p. W338-W344.
26. Pires, D.E.V., D.B. Ascher, and T.L. Blundell, *mCSM: predicting the effects of mutations in proteins using graph-based signatures.* Bioinformatics (Oxford, England), 2014. **30**(3): p. 335-342.
27. Pandurangan, A.P., et al., *SDM: a server for predicting effects of mutations on protein stability.* Nucleic Acids Res, 2017. **45**(W1): p. W229-w235.
28. Pires, D.E.V., D.B. Ascher, and T.L. Blundell, *DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach.* Nucleic acids research, 2014. **42**(Web Server issue): p. W314-W319.
29. Rodrigues, C.H.M., D.E.V. Pires, and D.B. Ascher, *DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability.* Nucleic Acids Research, 2018. **46**(W1): p. W350-W355.
30. Pires, D.E., T.L. Blundell, and D.B. Ascher, *mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance.* Sci Rep, 2016. **6**: p. 29575.
31. Hecht, M., Y. Bromberg, and B. Rost, *Better prediction of functional effects for sequence variants.* BMC Genomics, 2015. **16 Suppl 8**: p. S1.
32. Ashkenazy, H., et al., *ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.* Nucleic Acids Res, 2016. **44**(W1): p. W344-50.
33. Choi, Y. and A.P. Chan, *PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels.* Bioinformatics, 2015. **31**(16): p. 2745-2747.
34. Jubb, H.C., et al., *Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures.* Journal of molecular biology, 2017. **429**(3): p. 365-371.
35. Traynelis, J., et al., *Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation.* Genome Res, 2017. **27**(10): p. 1715-1729.
36. Krawczyk, B., *Learning from imbalanced data: open challenges and future directions.* Progress in Artificial Intelligence, 2016. **5**(4): p. 221-232.
37. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine Learning, 2006. **63**(1): p. 3-42.
38. Rose, A.S. and P.W. Hildebrand, *NGL Viewer: a web application for molecular visualization.* Nucleic Acids Research, 2015. **43**(W1): p. W576-W579.
39. Zatkova, A., *An update on molecular genetics of Alkaptonuria (AKU).* J Inherit Metab Dis, 2011. **34**(6): p. 1127-36.
40. Gallagher, J.A., et al., *Alkaptonuria: An example of a "fundamental disease"--A rare disease with important lessons for more common disorders.* Semin Cell Dev Biol, 2016. **52**: p. 53-7.
41. Karmakar, M., et al., *Structure guided prediction of Pyrazinamide resistance mutations in pncA.* Scientific Reports, 2020. **10**(1): p. 1875.
42. Karmakar, M., et al., *Empirical ways to identify novel Bedaquiline resistance mutations in AtpE.* PLoS One, 2019. **14**(5): p. e0217169.
43. Karmakar, M., et al., *Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy.* Am J Respir Crit Care Med, 2018. **198**(4): p. 541-544.

APPENDIX 8: BIOINFORMATIC APPROACHES TO PREDICT MUTATION EFFECTS IN THE BINDING SITE OF THE PROANGIOGENIC MOLECULE CD93

Bioinformatic approaches to predict mutation effects in the binding site of the proangiogenic molecule CD93

[3]Cicaloni V.*, [4,5]Karmakar M.*, [1]Frusciante L., [2]Pettini F., [1]Trezza A., [1]Orlandini M., [1]Galvagni F., [1]Nardi F, [7]Mongiat M., [4,5,6]Ascher DB., [1]Santucci A. and [1]Spiga O#.

[1]*Department of Biotechnology, Chemistry and Pharmacy, University of Siena, ITALY*

[2]*Department of Medical Biotechnologies, University of Siena, ITALY*

[3]*Toscana Life Sciences Foundation, Siena, ITALY*

[4]*Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia*

[5]*Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia*

[6]*Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, UK*

[7]*Department of Translational Research, Division of Molecular Oncology, CRO Aviano National Cancer Institute, Aviano, ITALY*

**equally contribution*

#*corresponding author*

Abstract

The transmembrane glycoprotein CD93 has been identified as a potential new target to inhibit tumour angiogenesis. Recently, Multimerin-2, a pan-endothelial extracellular matrix protein, has been identified as a specific ligand for CD93, but the interaction mechanism between these two proteins has still to be clarified. In this paper, we aim to computationally investigate the structural and functional effects of induced mutations on the binding domain of CD93. Starting from experimental data, we provide a workflow to analyse the "non-canonical" disulphide bridge disruption in the C-type lectin-like domain (CTLD). In addition, we investigate how mutation of Phe238 in the CD93 sushi-like domain affects the overall mobility of the CTLD domain inducing changes in the residue interaction network. The pivotal role of these aminoacid residues was also confirmed by Protein-Protein Interaction (PPI) docking analyses, which was used to predict effects of variations on the inter-residue interaction network at the binding site. The comprehensive molecular insight obtained from this study might provide an useful tool to drug design in cancer therapy.

Introduction

CD93 (also known as C1qRP) is a single-pass transmembrane glycoprotein belonging to group XIV family of the C-type lectin-like domain (CTLD) superfamily (Zelensky and Gready 2005). This group also includes Thrombomodulin (TM), Endosialin (TEM1/CD248), and CLEC14A, which share similar molecular structures from N- to C-terminus, consisting of a CTLD (designated as D1), one to six Epidermal growth factor (EGF)-like repeats (designated as D2), a sushi-like domain (designated as DX), a highly glycosylated serine/threonine-rich mucin-like domain (designated as D3), a transmembrane domain (designated as D4) and a short cytoplasmic domain (designated as D5) (Orlandini et al. 2014).

The CTLD canonical structure features a characteristic double-loop, which is stabilized by highly conserved disulphide bridges along with hydrophobic and polar interactions. CTLDs can bind different ligands simultaneously due to the flexible loop also referred to as the "long loop region" which is considered a key structure for carbohydrate binding. Thus, the CTLD is highly adaptable, conferring multiple functions to the protein and binding not only to sugars, but also to other structures, including proteins, lipids and inorganic molecules (Zelensky and Gready, 2005).

CD93 is predominantly expressed in endothelial cells (ECs) with expression also observed in monocytes, natural killer cells, platelets, myeloid cells, hematopoietic stem cells, and several lymphocyte subtypes (Greenlee et al. 2008; Khan et al. 2019). Notably, CD93 is highly expressed in blood vessels within tumours and has been identified as a key regulator of glioma angiogenesis (Galvagni et al. 2017; Tosi et al. 2017; Langenkamp et al. 2015), making it suitable as a potential target for anti-angiogenic treatment. In addition, we have identified a new signalling pathway involved in regulating EC adhesion and migration (Galvagni et al. 2016) but much remains to be clarified about the role of CD93 in the control of EC physiology. Recently, the pan-endothelial extracellular matrix (ECM) protein Multimerin 2 was identified as the interacting partner of CD93 (Khan et al. 2017; Galvagni et al. 2017). EMILINs/Multimerins form a small protein family, which is part of the superfamily of collagenous and non-collagenous proteins containing the gC1q signature (Colombatti et al. 2012). They are characterized by an N-terminal EMI domain, a central part of the molecule formed by a long region with a high probability of a coiled-coil structure, and a region homologous to the gC1q domain (Braghetta et al., 2004). We observed the CD93/Multimerin-2 interaction to be highly specific, since no interaction was seen with other ECM molecules including EMILIN2, which shares similar molecular domains with Multimerin-2 (Galvagni et al. 2017).

CD93 and Multimerin-2 are both up-regulated in tumour vasculature during tumour progression suggesting that the CD93/Multimerin-2 interaction regulates tumour angiogenesis. Indeed, disruption of this interaction strongly impaired EC migration and *in vitro* angiogenesis (Galvagni et al. 2017). Recent work has suggested that inhibition of CD93/Multimerin-2 interaction may lead to disruption of vascular integrity in tumours, showing

that the CD93/Multimerin-2 complex is required for activation of β1 integrin, phosphorylation of focal adhesion kinase, and fibronectin fibrillogenesis in ECs (Lugano et al. 2018). These observations strengthen the hypothesis that CD93 plays a key role in vascular maturation and organization of the ECM in tumours.

Binding of CD93 to Multimerin-2 is dependent on a long-loop region in the CTLD of CD93 and this interaction is abrogated by point mutations in the CTLD and sushi-like domains (Galvagni et al. 2017; Khan et al. 2017). Here, the application of computational approaches, combined with experimental data, allowed us to gain more in-depth molecular insights into the CD93/Multimerin-2 interaction, offering a platform for developing innovative therapeutics able to target these molecules and block their interaction.

Materials and Methods

Experimental Data

The chimeric constructs containing the extracellular domains of CD93 fused to Myc and the Multimerin-2 wild type fused to a His tag were generated as previously described (Orlandini et al 2014; Colladel et al 2016). The CD93 point mutants were obtained using the QuikChange II XL Site-Directed Mutagenesis kit (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's instructions. All constructs were confirmed by sequencing.

To obtain conditioned media (CM) containing recombinant proteins, human Lenti-X 293T cells (Clontech Laboratories Inc., Mountain View, CA, USA) were transiently transfected using Lipofectamine 2000 Transfection Reagent (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions. After 48 h, cells were rinsed with PBS and grown in DMEM without supplements for 10 h. The media were collected and centrifuged to remove cellular debris and then aliquots were stored at -80° C. The protein concentration was determined by immunoblotting experiments, performed with mouse anti-Myc antibodies (9E10, Santa Cruz Biotechnology, Dallas, TX, USA) (Orlandini et al 2008). ELISA-based solid-phase binding assays were performed to detect the interaction between recombinant proteins (Galvagni et al. 2017).

Structural Biology procedure

Homology Modeling

The primary sequences of human CD93 and Multimerin-2 were retrieved from the UniProt database (www.uniprot.org), with accession number Q9NPY3 and Q9H8L6 respectively. Suitable template structures for homology modelling were initially acquired by online submission of the FASTA sequence to I-TASSER and PHYRE2 web servers (Yang et al. 2015; Kelley et al, 2015). Based on the percentage of identity and alignment coverage, we identified the crystal structure of the C-type mannose receptor 2 (PDB ID: 5AO6), (Paracuellos et al. 2015) as templates for the CTLD; EMILIN2 (accession number: Q9BXX0) was chosen as a template for Multimerin-2. Three-dimensional atomic coordinates were generated with MODELLER, implemented in PyMod2.0 (Webb and Sali, 2014; Janson et al. 2017), following the target-template alignment suggested by

PHYRE2. For each domain, the best out of 5 models based on lowest value of DOPE (Discrete Optimized Protein Energy) was chosen as final model. Ultimately, individual constructed domains for each protein were connected by means of the Swiss PDB Viewer software (www.expasy.org/spdbv). The 3D molecular conformation of the two proteins was minimized using the Amber99SB (Lindorff-larsen et al. 2010) force field, until a final convergence of 0.01 kcal mol^{-1} $Å^{-1}$ was achieved. Validation of the models was carried out using Ramachandran plot calculations computed with the PROCHECK program (Laskowski et al. 1993). The three-dimensional structures of CD93 mutants were obtained submitting wild type CD93 structure to the DUET web server (biosig.unimelb.edu.au/duet/). The multiple sequence alignments were performed using ClustalO (Sievers et al. 2011) and colour figures were generated using PyMOL (Schrodinger 2015). The complex CD93/Multimerin-2 was constructed by assembling modelled structures of the interacting components through an exhaustive search using the GRAMM-X server (Tovchigrechko and Vakser 2006). The resulting complex structure was analysed using the PDBePISA tool for the exploration of macromolecular interfaces (Krissinel and Henrick 2007).

MD simulations

Molecular dynamics simulations of CD93 wild type and mutants were carried out in GROMACS 2016 (Abraham et al. 2015). The protein structures were solvated in a triclinic box filled with TIP3P water molecules and Na^+/Cl^- ions were added to neutralize the system. The whole systems were then minimized with a maximal force tolerance of 1000 kJ mol^{-1} nm^{-1} using the steepest descendent algorithm.

The optimized systems were gradually heated to 310 K in 1 ns in the NVT ensemble, followed by 10 ns equilibration in the NPT ensemble at 1 Atm and 310 K, using the V-Rescale thermostat and Berendsen barostat (Bussi et al. 2007; Berendsen et al. 1984). Subsequently, further 100 ns MD simulations were performed for data analysis. Newton's equations of atomic motion were integrated by the Verlet algorithm with 2 fs time step. LINCS algorithm (Hess et al. 1997) to constrain all the covalent bonds involving hydrogen atoms and the Particle Mesh Ewald (PME) algorithm was employed for long-range interactions computation (Darden et al. 1993).

The analysis tools implemented in GROMACS were applied in order to calculate Root-mean-square deviation (RMSD) and Root-mean-square fluctuation (RMSF). The graphs were plotted by the XMGrace software (Turner 2005).

The effects of mutations on the flexibility and global correlated motion of the CD93 CTLD domain were assessed by Principal component analysis (PCA) and Dynamic cross-correlation map (DCCM), using the Bio3D package in R (Grant et al. 2006). At first, we extracted 5000 conformations of each (native and mutants) MD production run trajectory and removed translational and rotational motions, in order to be able to construct a structure of 3x3 N positional covariance matrix of the Cα atoms of the CTLD domain. Diagonalization of the

atomic displacement correlation matrix returned a set of eigenvectors, which provides the direction of motion of each component, and the correspondent eigenvalue, representing the magnitude of such motions. Eigenvectors were ranked by the highest eigenvalue. Dynamic cross-correlation analysis allowed the construction of a matrix, that is a graphical representation of all atom-wise cross-correlations, in which the correlation values varies from −1 to +1, with positive numbers showing correlated motions (i.e. residues that move in the same direction) and negative numbers showing anti-correlated 22 motion (i.e. residues that move in opposite direction) (Grant et al. 2006).

The Residue interaction network (RIN) for all residues of the CTLD domain was constructed by submitting the average structure extracted from the equilibrium phase of each MD simulation trajectory to the RING 2.0 web server (http://protein.bio.unipd.it/ring/). RING enables to analyse mutation effects, protein folding, domain–domain communication and catalytic activity through the identification of covalent and non-covalent bonds in protein structures, including π–π stacking and π–cation interactions (Piovesan et al. 2016).

Protein-protein Interface Docking

We used the EVolutionary Couplings server (https://evcouplings.org/) to provide functional and structural information about proteins derived from the evolutionary sequence record, using methods from statistical physics (Hopf et al. 2014). By using FASTA sequence (UniProt code: Q9NPY3 and Q9H8L6 for CD93 and Multimerin-2 respectively), EVcomplex (https://evcouplings.org/complex) was used to determine co-evolved residues in our selected PPI complex poses and to provide the information if a protein interaction is conserved across enough sequenced genomes using a single pair per genome (Hopf et al. 2014).

We performed an *in silico* validation of the effect of CD93 mutants (Galvagni et al. 2017) by mCSM-PPI2 (Rodrigues et al. 2019). Using the transcripts ENST00000246006.5 and ENST00000372027 for CD93 (Chr20) Multimerin-2 (Chr10) respectively, we then mapped the gnomAD missense variants to the structure (Karczewski et al. 2019). PPI interface of the complex was analysed using the PDBePISA tool (Krissinel and Henrick 2007). Finally, we calculated the measure of regional intolerance to missense variation for CD93 in both docked- poses by using MTR. (Traynelis et al. 2017; Silk et al. 2019)

Results and Discussion

Validation of Homology Model

In order to identify the amino acid residues critical to the CD93/Multimerin-2 binding, homology modeling was performed to predict their three-dimensional (3D) structure. To model the CTLD domain of CD93 glycoprotein we started from structural information and FASTA sequence retrieved from UniProtKB Q9NPY3. The closest homologous structure suggested by PHYRE2 webserver was human C-type mannose receptor 2 (PDB ID:

5AO6). Visual inspection of 5AO6 PDB structure in PyMOL exhibited the canonical fold of “long form” CTLDs consisting of the long loop region and six conserved cysteines (Figure 1).

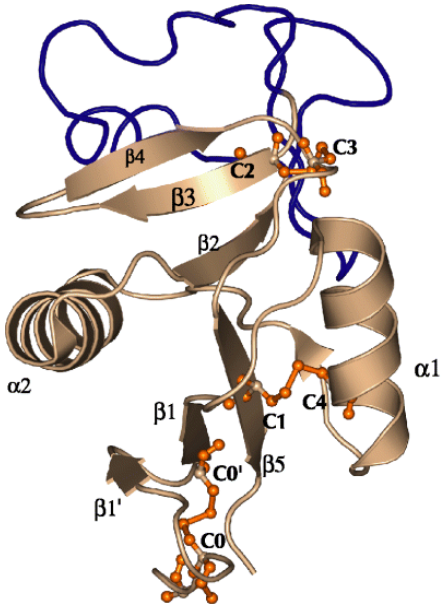


Figure 1. *Cartoon representation of a typical CTLD structure (PDB 1K9I). In blue is shown the LLR, orange sticks represent cysteine bridges (C0-C0' being specific for long form CTLDs). (Source: Zelensky and Gready 2005)*

On the other hand, the two unique cysteine residues distinctive of the group XIV family members were missing. Conflicting hypothesis have been devised whether these “non-canonical” cysteine residues are involved in disulphide bond formation. Previous studies on the members of group XIV family showed how the CTLD of these four proteins contains 8 conserved cysteine resisdues, which are likely involved in four disulphide bonds (Figure 2). Nativel et al. performed specific tests, which demonstrated that the CTLD domain of CD93 did not comprise any free thiol groups. In 2016, they evaluated the presence of reduced cysteines in the CD93 CTLD, by means of Ellman's assay for the quantification of thiol groups, and speculated that the CTLD of CD93 did not contain any free cysteines, all being engaged in disulphide bonds (Nativel et al. 2016).

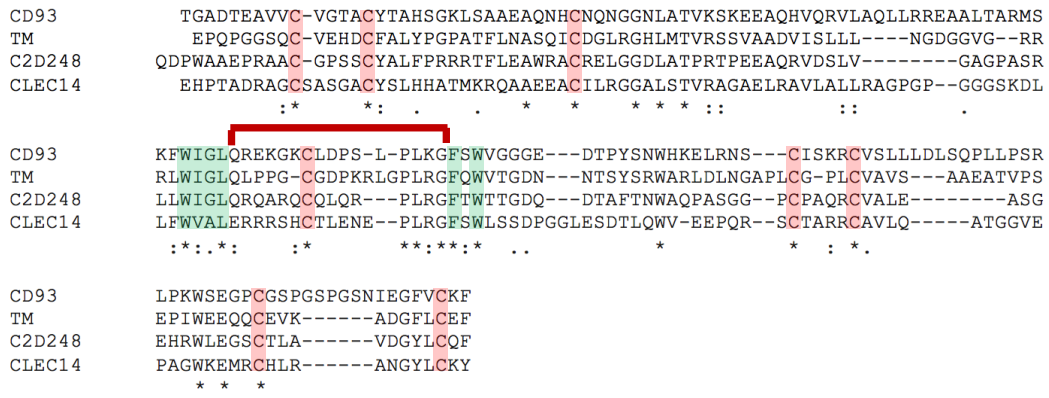


Figure 2. *Amino acid sequence alignment of the CD93, TM, CD248, CLEC14A CTLD domain using ClustalO. Red-boxed amino acid represents the conserved cysteine residues, green-boxed amino acid represents highly conserved amino acid residues at the base of the long loop region, the latter being highlighted with a red bracket.*

To predict the structure of the interacting complex we integrated homology modeling and docking simulations with molecular binding information (Galvagni et al. 2017). At first, we performed a blind protein-protein docking simulation, obtaining a cluster of potential CD93/Multimerin-2 complexes, which showed for CD93 a potential interaction region between CTLD and sushi-like domains. Next, to narrow down the molecular surface of Multimerin-2 responsible for the interaction with CD93 we combined previous data (Galvagni et al. 2017) with Immunoprecipitation analysis using Multimerin-2 deletion mutants. By using the complete extracellular region of CD93 as a ligand, we were able to restrict the surface of Multimerin-2 interacting with CD93 to a region spanning from the amino acids 563 to 618 (Figure 3).

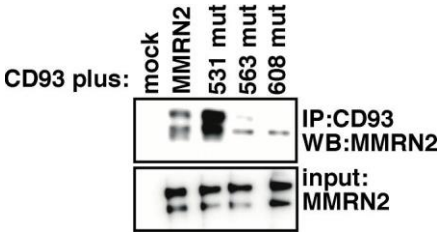


Figure 3. *A portion of the Multimerin-2 coiled-coil region is required for binding to CD93. Western blot analysis of the Immunoprecipitation experiment performed using Multimerin-2 deletion mutants: wild type Multimerin-2 (MMRN2); Multimerin-2 deleted from amino acids 531 to 539 (531 mut); Multimerin-2 deleted from amino acids 563 to 574 (563 mut); Multimerin-2 deleted from amino acids 608 to 618 (608 mut). CM containing Myc-tagged CD93 and Multimerin-2 mutants were incubated and immunoprecipitaed using anti-CD93 antibodies. Bands were revealed with anti-Multimerin-2 and anti-Myc antibodies. CM from cells transfected with the empty vector (mock) was used as a control. The experiment was repeated three times.*

To verify whether the correct folding of CD93 was necessary for binding to Multimerin-2, amino acid residues, which were far from the putative binding site of CD93 and predicted to be pivotal to correct folding, were deleted. Even the deletion of few amino acids in the N-terminal of the CTLD-sushi-like deletion mutant highly reduced the binding strength of the mutant to Multimerin-2, indicating that the correct folding of CTLD is instrumental to proper CD93/Multimerin-2 binding (Galvagni et al. 2017). To test whether the non-canonical cysteines (C104 and C136) within the long loop region were important for MMRN2 binding, the single mutants $CD93^{C104S}$ and $CD93^{C136S}$ were generated. The relative migration of the two CD93 mutants, assessed by Western Blotting, showed the presence of two bands in SDS-PAGE, suggesting that they were correctly folded (Figure 4). However, in solid phase analysis these mutants failed to bind to Multimerin-2, highlighting the importance of these residues for a proper CD93/Multimerin-2 interaction.

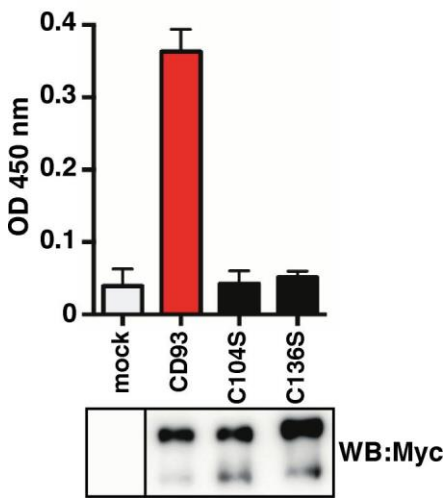
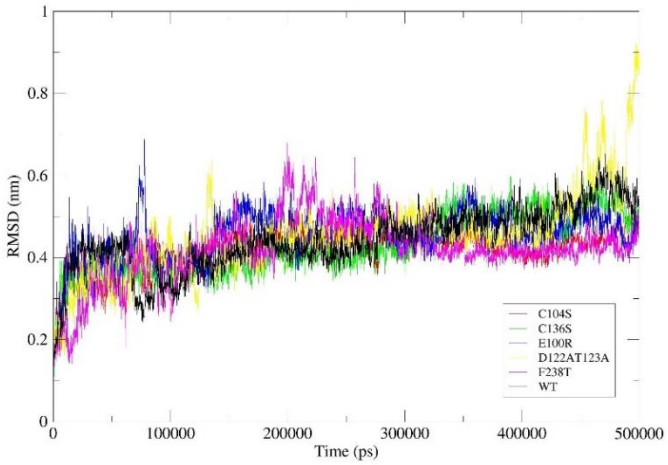


Figure 4. *Mutation analysis of the CD93/Multimerin-2 interaction. Solid phase analysis of the interaction of the C104S and C136S CD93 mutants with Multimerin-2. The expression of the CD93 mutants in the CM from 293T transfected cells was comparable, as assessed by immunoblotting using an anti-Myc antibody (bottom panel). CM were applied to ELISA plates coated with purified Multimerin-2 and CM from 293T cells transfected with the empty vector (mock) were used as a control. Data represent the means ±SD of three independent experiments.*

Recent studies carried out by Khan et al., gauged how Multimerin-2 is the ligand for CLEC14A, CD93 and CD248 of group XIV family C-type lectins, but not for TM (Khan et al. 2017). Like us, they assessed correct

folding of CLEC14A and CD93 upon mutation of the considered cysteines (C103 and C138 for CLEC14A) and evaluated their binding to Multimerin-2 obtaining comparable outcomes, which show their important role for protein-protein interaction. These “non-canonical” cysteines are unique in CTLDs but highly conserved amongst the group XIV family, suggesting disulphide bond formation (Nativel et al. 2016). Furthermore, the statement that CD93 along with CLEC14A cysteine mutants were correctly folded but failed to bind Multimerin-2 allowed us to speculate that these residues are likely to be important in the local conformation of the long-loop region and disulphide bond formation may be essential for binding to Multimerin-2.

Site-directed mutagenesis studies verified the involvement of key amino acid residues in the proper formation of the CD93/Multimerin-2 complex. Experiments have shown that the binding strength of the extracellular domain

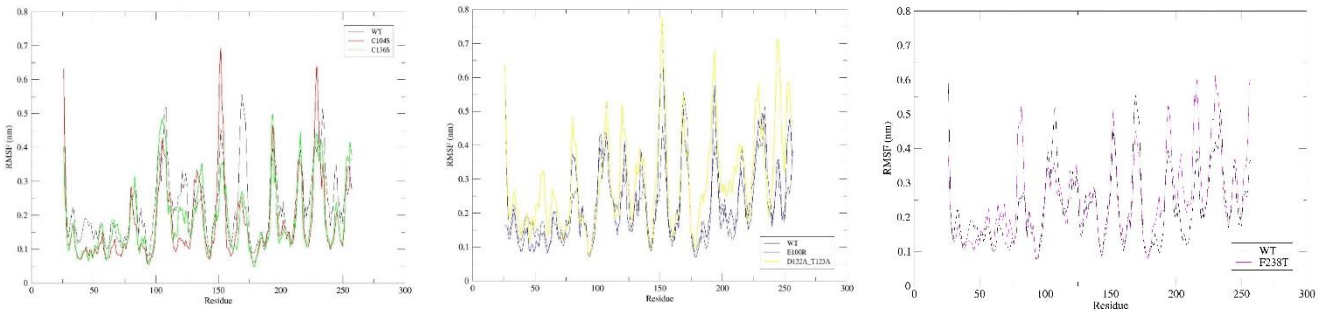


of CD93 to Multimerin-2 is sensitive to amino acid substitution of the two cysteines in the long loop region of CD93 (C104 and C136) and of F238 in the sushi-like domain. Based on these results, MD simulations of the CTLD-sushi-like region of $CD93^{WT}$ and $CD93^{C104S}$, $CD93^{C136S}$ and $CD93^{F238T}$ mutants where carried out to assess the structural impact of these mutations on human CD93. Furthermore, MD simulations of CD93 mutants showing increase (E100R) and preservation (D122A/T123A) of interaction with Multimerin-2 (Galvagni et al. 2017) were also performed in order to carry out a comparative study to better understand the possible mechanism behind the loss of interaction upon C104S, C136S and F238T mutations in CD93. Simulations of native and mutant CD93 were performed for 500 ns.

Structural Analysis

The RMSD was calculated on the backbone. At around 150 ns, the RMSDs for the wild type and mutants trajectories are at equilibrium, with a value of 4.3 nm. CD93 C104S mutant did not show significant deviation from the initial structure, showing a RMSD profile very stable until the end; CD93 wild type and C136S had a progressive increase of 0.1 nm after 250 ns. E100R and D122A/T123A showed a trend comparable to the wild type with an evident fluctuation of the double mutant suggesting some influence of the mutation on the structural

stability of the domains. Mutation F238T, after reaching the point of equilibrium, followed a trend comparable to



the C104S mutant with a stable RMSD profile until the end.

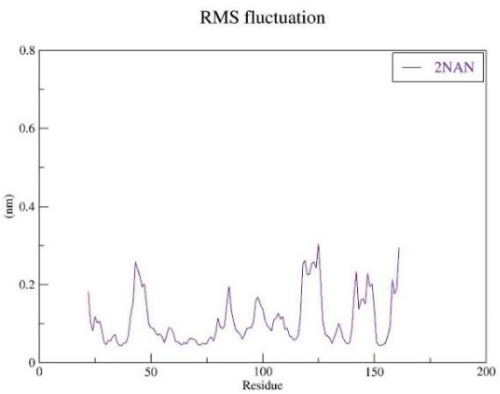
Figure 5. *Backbone RMSD values during MD simulation.*

The RMSF profile for $CD93^{WT}$ and for the $CD93^{C104S}$, $CD93^{C136S}$, $CD93^{E100R}$, $CD93^{F238T}$ $CD93^{D122A/T123A}$ mutants was computed to reveal the structural fluctuations that occur following induced mutations.

Overall, the average RMSF of the native conformation and the $CD93^{E100R}$ and $CD93^{D122A/T123A}$ mutants were quite comparable (Figure 6); it can be observed that mutation of the Glu100 in Arg resulted in a global increase of the CTLD domain stability. Double mutation of Asp122 and Thr123 both in Ala did not impact the overall dynamics of the domain, resulting in a residue wise RMSF profile comparable to the wild type. On the contrary, significant variations in fluctuations compared to the wild type CTLD were observed in both $CD93^{C104S}$ and $CD93^{C136S}$ following mutations of Cys104 and Cys136 in Ser, and thus the cleavage of the S-S bond. More specifically, in Figure 6 is shown that a decrease in fluctuations involved the regions between residues 86-90 and 117-125 while an increase was detected in the regions between residues 127-137 in both mutants.

Figure 6. *Per residue RMSF profile of each mutant MD simulations against wild type CD93 CTLD domain.*

These results suggest that the disruption of the disulphide bond between these highly conserved cysteine residues does not impact the CD93/Multimerin-2 binding by affecting the folding of the CD93 CTLD. Rather, these cysteines are likely to be important in the local conformation of regions closed to the long-loop region, which is believed to be fundamental not only for carbohydrate binding but also to many other structures (Zelensky et al.



2005). In order to corroborate these assumptions, an additional MD simulation under the same conditions was carried out using as representative structure a resolved CTLD (PDB ID: 2NAN; Popsilova et al. 2017) lacking the fourth "non-canonical" S-S bond. As shown in Figure 7, the fluctuation profile of the CD302 antigen CTLD follows a trend quite comparable to that one showed by $CD93^{C104S}$ and $CD93^{C136S}$ mutants, with an average RMSF value of 1 Å. The results obtained using a resolved CTLD, in addition to validating the correctness of the homology model, confirmed the supposition of a correct folding of the CTLD domain despite the disruption of the disulphide bond.

Figure 7. *Per residue RMSF profile of PDB:2NAN CTLD crystal structure.*

Interestingly, mutation F238T in the sushi-like domain of CD93 also alters the overall flexibility of the CTLD, showing a similar trend in fluctuation of specific regions of the domain as compared to $CD93^{C104S}$ and $CD93^{C136S}$ (Figure 6). Though we observed a unique increase in the flexibility of the region from residue 78 to residue 84 of the CTLD, which does not occur in the $CD93^{WT}$ or in the $CD93^{C104S}$ and $CD93^{C136S}$ mutants, a trend comparable to the two cited mutants was instead noticed in the region between residues 40-50. Importantly, a more evident increase in fluctuation of the sushi-like domain, when compared to the wild type, was noticed. It is interesting to observe that the regions, which increase the flexibility of the CTLD domain, are about the same as the $CD93^{C104S}$ and $CD93^{C136S}$ mutants, considering that the disulfide bridge has been kept intact. These observations suggest a long-range effect on the motion of the CTLD following mutation of the $CD93^{F238T}$ on the sushi-like domain, confirming the pivotal role of this residue, as indicated by next docking studies. Remarkably, despite the presence of the disulphide bridge (between cysteine residues 96-133), the long loop region of TM, spanning from residue 91 to 107, shows more flexibility in comparison with the long loop region of the wild type CD93 (data not shown). TM is the only member of the group XIV family unable to bind Multimerin-2 in in vitro assays (Khan et al. 2017). As mentioned before, sequence alignment shows the absence of two cysteines in the sushi-like domain likely involved in disulfide bond formation. Based on these assumptions, it is possible to hypothesize a long-range effect on the overall mobility of the TM CTLD due to differences in the motion of the sushi-like domain, as supposed for $CD93^{F238T}$.

We performed Principal Component Analysis, a multivariate statistical technique used to reduce number of data produced by MDs, in order to study in deep how conformational flexibility could influence biological functions of CD93 (Ichiye et al. 1991, Amadei et al. 1993). To determine the number of dimensions to which the data is reduced we located the dimension prior to the point where the variance rapidly falls to a relatively stable value. This can be accomplished by creating a scree plot in which the eigenvalues, determined in the diagonalization of the covariance matrix, are ordered from the strongest to weakest.

As shown in Figure 8, we have reported each system with a cross-plot, representing the projection of each trajectory coordinates onto the first three PCs, and the corresponding scree plots. The first three PCs accounted for 56.3%, 50%, 52.4%, 55.7%, 48.3% and 52% of the variance in the motion observed for $CD93^{WT}$, $CD93^{C104S}$, $CD93^{C136S}$, $CD93^{F238T}$ $CD93^{E100R}$, and $CD93^{D122A/T123A}$, respectively. It can be seen that mutations that impair the CD93/Multimerin-2 interaction have an impact on the conformational space that the CTLD of CD93 occupies during the simulation, when compared to the two mutants able to retain such interaction which show a behavior comparable to the wilde type. C104S mutation had the strongest effect on the CTLD motion. When compared to the wild type structure, $CD93^{C104S}$ occupies a smaller phase space, with a contraction of the conformational space and a decrease in the percentage of correlated motion along the first three PCs (30.56 %, 11.31% and 8.15 %). This is visible in the two-dimensional cross-plot and suggests that the mutant CTLD domain is less flexible. PCA analysis of C136S mutation shows a slighter decrease in the conformational space compared to the wild type, with the first three PCs capturing 26%, 11.65% and 8.09 % of the motions, and for both C104S and C136S mutants a change in direction of motion of the domain was observed. These results are in accordance with the RMSF analysis and suggest that the disruption of the disulfide bridge determines a decrease in the mobility of a loop of the CTLD that could thus influence the overall behavior of the domain. If we consider that cysteine 136 is found at the top of one of the loops peculiar of the group XIV family CTLD domains, we can infer that the same loop, following the disruption of the bridge could no longer be held in direction of the C104, allowing for new interactions with residues in the inner part of the domain.

F238T mutation shows a proportion of variance for the first three PC (41.12%, 8.13% and 6.45%) comparable to the wilde type with a change in the direction of motion that suggests a long-distance effect on the CTLD (Figure 8). On the other hand, the increase in the overall mobility, when compared to C104S and C136S mutants, advises a different impact on the interaction network of the CTLD. As anticipated above, both E100R single mutation and D122A/T123A double mutation PCA analysis show a proportion of variance in the motion of the CTLD domain comparable to the wilde type. Once again, these behaviors reflect the RMSF profile.

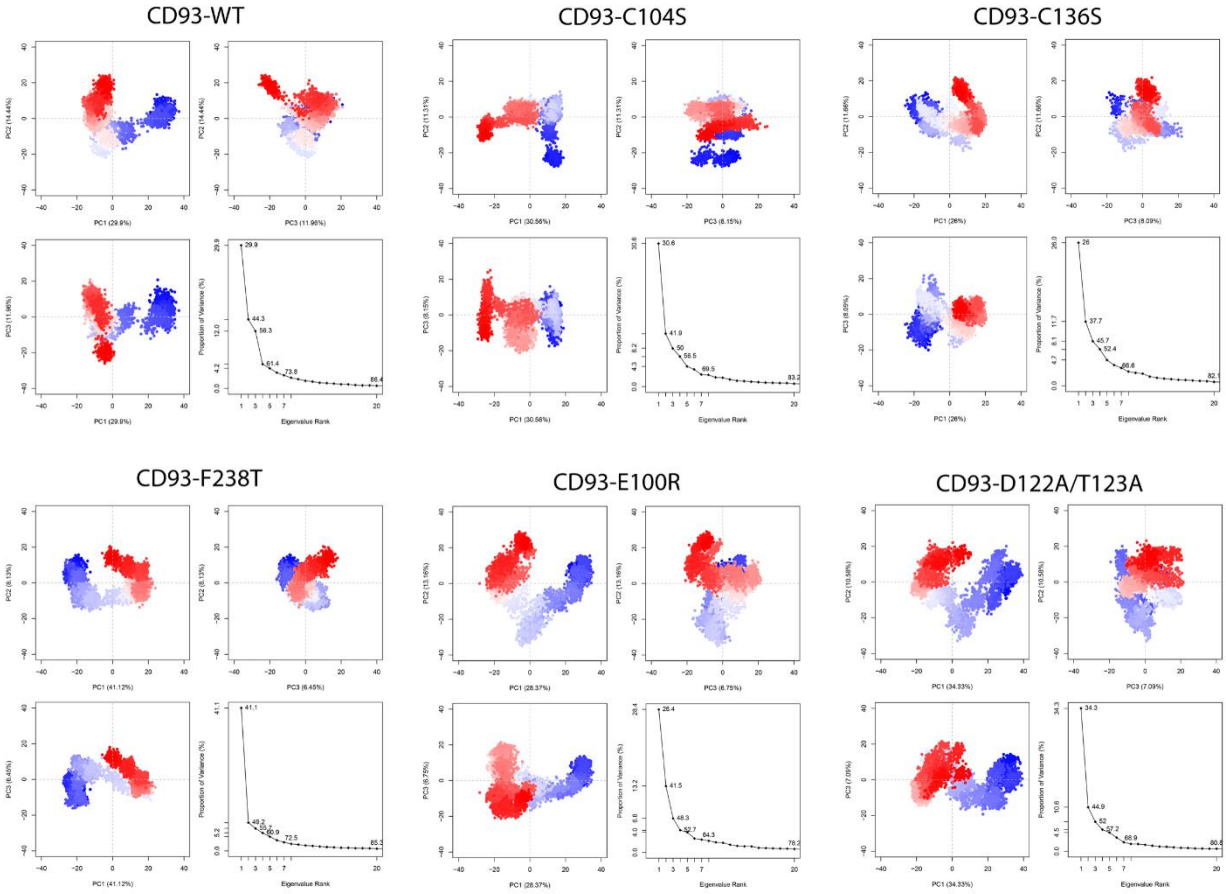


Figure 8. *Projections of trajectories onto the subspace by the first three eigenvectors. Projection of trajectories into PC1, PC2, and PC3 for the CTLD of each system. The converged stable conformation and unstable scattered state are shown with red and blue dots, respectively. The white dots indicate the intermediate states observed in both complexes*

Also, we performed Dynamic cross-correlation analysis (DCCM), in order to investigate fluctuations and domain motions. Comparing the $CD93^{WT}$ with C104S, C136S, in Figure 9 we observed a decrease in the CTLD fluctuations, which translated in the reduction of both correlated and anti-correlated motions. This is particularly true for the regions spanning amino acid residues 65-80, 98-150 (including the mutated cysteine residues) and 145-175, which all seem to move in an anti-correlated manner with respect to the N-terminal region of the domain. This loss in correlated movements, in particular in the region of the loop containing C136, is consistent with the hypothesis inferred above. As for F238T mutant, the behavior showed by the PCA analysis corresponds to an escalation of correlated and anti-correlated motions spread all over the CTLD when compared to the wild type but also to the C104S and C136S mutants, despite the mutation is not in the CTLD domain itself.

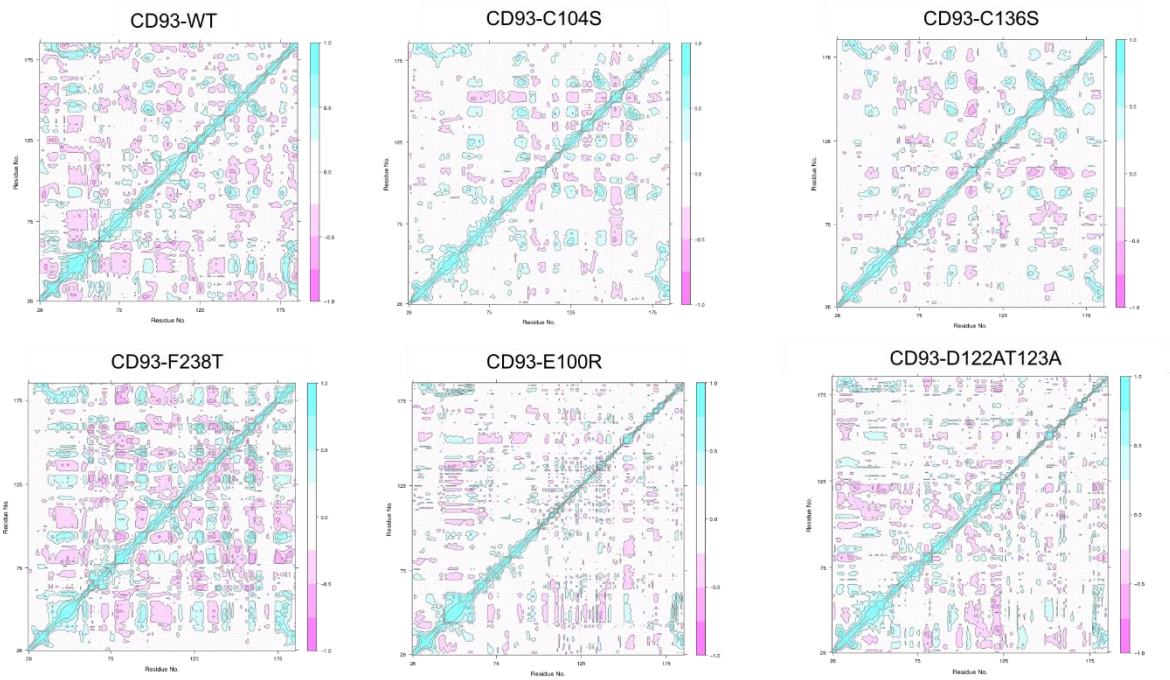


Figure 9. *Dynamical cross-correlation map (DCCM) calculated using the MD simulation of native and mutant – CD93 CTLD as input. The color scheme follows the range of correlation: cyan corresponds to positive correlation values (from 0.25 to 1); pink corresponds to negative correlation values (from −0.25 to −1); white corresponds to weak or no-correlation values spanning from −0.25 to +0.25. The variation in the intensity of cyan or pink color tracks the magnitude of correlation or anti-correlation.*

We observed that the cleavage of the disulfide bridge in $CD93^{C104S}$, $CD93^{C136S}$, namely mutants which are not able to interact with Multimerin-2, did not seem to remarkably disrupt the overall residue interaction network when compared to the wild type. On the other hand, the histograms in Figure 10 highlighted how the network of these systems is characterized by a PIPISTACK interaction involving Trp128 and Phe114 that seems to be lacking in the wild type. This behavior reflects the increase in stability showed by the RMSF profile in the region near residue 125 of these mutants compared to the native protein. Once again, this could be due to the disruption of the disulfide bridge, as Trp128 is on the same loop containing Cys136. Moreover, mutation of residue 136 from Cysteine to Serine determines the loss of Van der Waals interactions between Phe114 and residues 136 itself and Val142 in both C104S and C136S mutants, with respect to the wild type. This is quite interesting since Phe114 is a highly conserved residue in the group XIV C-type lectins. From sequence alignment among the family (Figure 2) it has been seen that the long loop of the CTLD of these proteins lies between two highly conserved hydrophobic regions (94WIGL97 and F114 for CD93), thus we speculate that Phe114 might be pivotal for the right folding of the domain.



Figure 10. *Column chart displaying the total amount of hydrogen bonds (blue) and disulphide bonds (red), Van der Waals forces (green), π–π stacking and π–cation (orange), ionic interactions (yellow) only considering the loop region containing Cys136.*

Furthermore, Surface Accessible Solvent Area (SASA) calculated for all the residues during the simulation (Figure 11) shows that the SASA value for Pro124 and Tyr125 is higher in the MD simulation of WT respect to C104S, C136S mutants, which consistently shows a higher number of interaction established for these residues, confirming that this region of the loop is more densely held in the core of the domain.

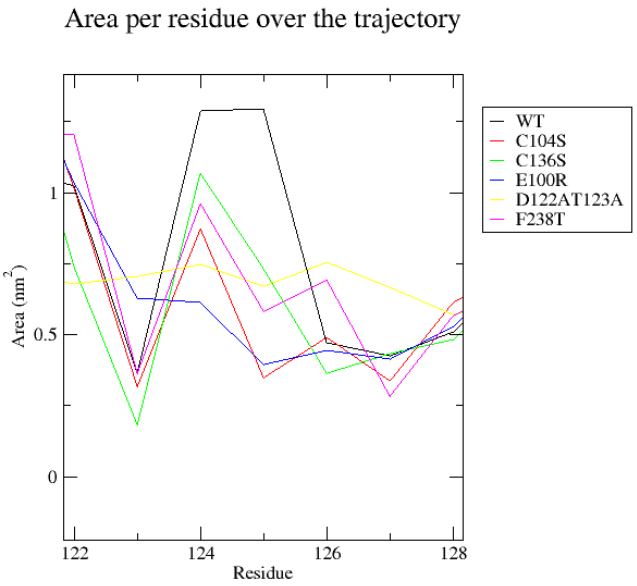


Figure 11. *Surface Accessible Solvent Area (SASA) calculated for all the residues during the simulation*

As for F238T, the residue interaction network calculated for the sushi-like domain (Figure 12) shows an increase in the interaction established in the region spanning amino acids 234-244, which included the mutated residue. This can be explained by the loss of a PIPISTACK following mutation of the Phenylalanine in Threonine but interestingly a decompensation, which translates in an increase of established interaction in that same region, can be observed for CD93^{C104S} and CD93^{C136S} mutants, despite the presence of the π–π stacking interaction.



Figure 12. *Column chart displaying the total amount of hydrogen bonds (blue) and disulphide bonds (red), Van der Waals forces (green), π–π stacking and π–cation (orange), ionic interactions (yellow) only considering the DX domain.*

Protein-Protein Interaction (PPI) Docking Analyses

The aim of PPI docking procedure is to predict correct poses and to score them according to the strength of interaction in a reasonable time frame. In this study we presented an extended approach to evaluate the reliability of protein–protein complex structures, confirming by new experimental data. Starting from best two plausible docking poses, we have applied a four-steps workflow for screening the most trustworthy one. Based on an evolutionary statistical approach, our aim was to find co-evolved residues between CD93 and Multimerin-2: if a protein-protein interaction is conserved across enough sequenced genomes, using a single pair per genome can give accurate predictions of the interacting residues.

Initially, the paired sequences were concatenated and statistical co-evolution analysis were performed using EVcouplings (Marks et al., 2011; Morcos et al., 2011; Aurell and Ekeberg, 2012), that applies a pseudolikelihood maximization (PLM) approximation to determine the interaction parameters in the underlying maximum entropy probability model (Balakrishnan et al., 2011; Ekeberg et al., 2013; Kamisetty et al., 2013), simultaneously generating both intra- and inter-Evolutionary Coupling scores for all pairs of residues within and across the protein pairs. Thus, this program predicts interacting residues in protein complexes from sequence covariation for the complex of interest. The analysis of correlated evolutionary sequence changes across proteins can identify residues that are close in space with enough accuracy to determine the three-dimensional structure of the protein complexes; as a consequence, it can be used to screen the reliability of the selected poses (Hopf et al. 2014). All results are summarized in Table 1 and represented in Figure 11. For Pose 1, the most interesting co-evolved residues couples are closer or equal than 8 Å: Leu-610 (in Multimerin-2) and Glu-131 (in CD93) with a distance of 5.5 Å and a high probability score of 0.82; Ala-585 (in Multimerin-2) and Pro-245 (in CD93) with a distance of 8 Å and a probability score of 0.70. Although the other two pairs (Ala-597 in Multimerin-2 and Tyr-125 in CD93; Ala-585 in Multimerin-2 and Pro-245 in CD93) showed an excellent probability score respectively of 0.98 and 0.70, and the distances are higher than the threshold of 8 Å. On the contrary, as shown in Figure 2., there are no pairs of residues in Pose 2 with a distance below 8 Å. The closest in space co-evolved couple is Multimerin-2 Ala-585/ CD93 Pro-245 with a probability score equal to 0.70 and a distance of 11.5 Å. Differently from Pose 1, all the other selected pairs of Pose 2 presented distances larger than 20 Å. Being able to verify the presence of two co-evolved couples of residues in Pose 1, this first analysis suggested it as the most reliable pose.

CD93	**MMRN2**	**Probability**	**Distance (Å) POSE 1**	**Distance (Å) POSE 2**
E131	**E601**	0,98	13,2	24,2
E131	**E610**	0,82	5,5	31,5
Y125	**A597**	0,7	9,6	25,1
P245	**A585**	0,7	8,0	11,5

Table 1. Evolutionary coupling results summary. Probability score and distance related to Pose 1 and related to Pose 2.

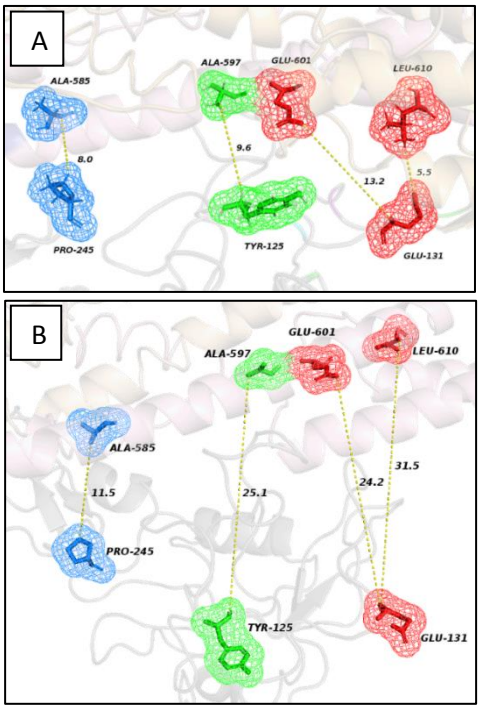


Figure 11. *Representation of the closest co-evolved residues couples (coloured sticks) in Pose 1 (A) and in Pose 2 (B). with their relative distances, resulting from EVcouplings analysis.*

The second step evaluated if the two selected poses fitted with experimental results coming from the study of Galvagni et al. 2017. In this paper, an extensive mutation analysis of the CD93/Multimerin-2 binding strength among wild type and mutated proteins was evaluated by solid phase assay. In order to characterize the interacting surface of CD93, point mutations were introduced and the CD93 mutants expressed in 293T cells were analysed using Western blots to assess the expression of the soluble recombinant proteins. Every missense mutation was evaluated with mCSM-PPI2, a novel machine learning computational tool designed to more accurately predict the effects of missense mutations on protein-protein interaction binding affinity (Rodrigues et al. 2019). mCSM-PPI2 uses graph-based structural signatures to model effects of variations on the inter-residue interaction network, evolutionary information, complex network metrics and energetic terms to generate an optimized predictor. The closer the residues pairs, the more accurate the prediction of interaction activity. The predicted binding affinity scores were compared with experimental results: the poses that best fits with experimental results through the affinity prediction made with mCSM-PPI2 could be consider the most reliable one. From observations made in Figure 11, for Pose 1 (Table 2), every binding strength prediction fitted with the experimental results apart from D249A (experimental results: increasing, prediction affinity: moderately decreasing) and H236A (experimental results: decreasing, prediction affinity: weakly increasing and with a distance to interface extremely elevated around 10 Å). For Pose 2 (Table 2), the major part of predictions fitted

with experimental results with two serious exceptions for E100R (experimental results: slightly increasing, prediction affinity: strongly decreasing and with a distance to interface extremely close around 3.2 Å) and D249A (experimental results: increasing, prediction affinity: strongly decreasing, with a distance to interface of 3.2). In conclusion to this second step, Pose 1 is the docking-pose that best fitted with experimental results.

RESIDUES MUTATED	DISTANCE TO INTERFACE		MCSM-PPI2 PREDICTION		AFFINITY	
	POSE 1	**POSE 2**	**POSE 1**	**POSE 2**	**POSE 1**	**POSE 2**
E100R	6,571	3,215	0,131	-1,225	Increasing	Decreasing
E242A	5,402	4,288	-0,45	-0,646	Decreasing	Decreasing
N249A	5,084	3,259	-0,6	-1,311	Decreasing	Decreasing
L256Q	1,67	2,403	-0,003	-0,035	Decreasing	Decreasing
F238T	9,45	7,537	-0,804	-0,714	Decreasing	Decreasing
C104S	6,855	5,392	-0,39	-0,26	Decreasing	Decreasing
N249R	5,084	3,259	-0,405	-0,738	Decreasing	Decreasing
N246R	6,896	10,085	-0,029	-0,013	Decreasing	Decreasing
F248T	7,269	7,23	-0,366	-0,236	Decreasing	Decreasing
C136S	6,305	6,262	-0,335	-0,32	Decreasing	Decreasing
H236A	9,904	9,649	0,028	-0,063	Increasing	Decreasing

Table 2. *Binding affinity predictions. Every missense mutation of the two poses were evaluated with mCSM-PPI2; results were compared with experimental outcomes.*

As a third step, we have then mapped the gnomAD missense variants to the structure (Karczewski et al. 2019). The complex structure of every pose was then analysed using the PDBePISA tool for the selection of interface residues (Krissinel and Henrick 2007). We have noticed that in Pose 1, the total amount of variants (excluding missense) seen at interface positions of CD93 is larger than in Pose 2. In Pose 1, for a total of 47 interface residues, we observed 26 amino-acid variants positions (excluding missense), while in Pose 2, for a total of 36 interface residues, no missense variants were found in only 15 positions **(Table3)**.

The fourth step was based on the exploration regional intolerance to missense variation in interface residues located in Pose 1 and Pose 2 by looking into Missense Tolerance Ratio (MTR) (Traynelis et al.2017; Silk et al 2019) scores. It was crucial in this last step to evaluate missense variant deleteriousness by examining its surrounding regional intolerance and to calculate the MTR scores at their position (Table 3). As you can see in Figure 12, there are no missense intolerant regions placed in the CD93 interfaces both in Pose 1 and in Pose 2.

Summarizing our results, Pose 1 could be the best docking pose among all the identified ones by PPI docking, as per experimental data.

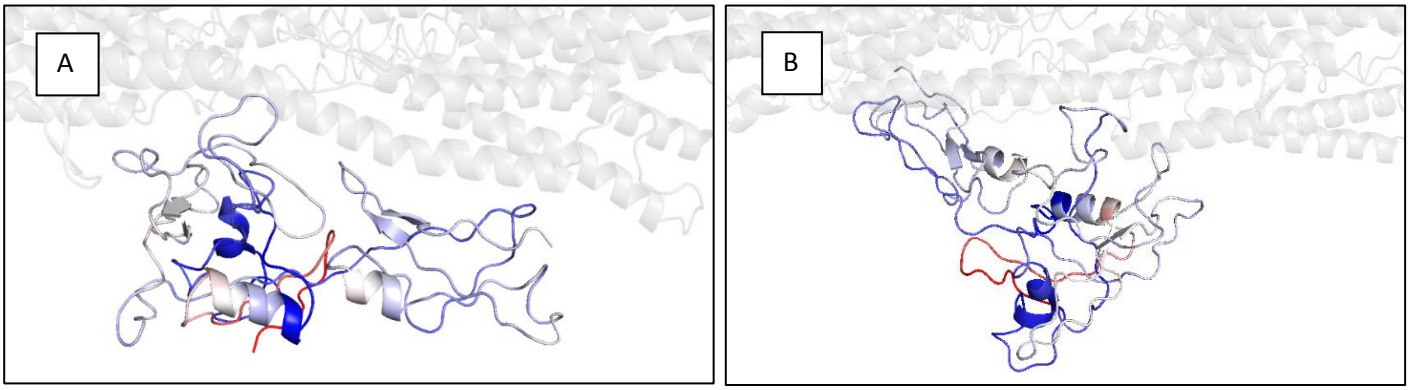


Figure 12. *Interface residues in CD93. Panel A represents Pose 1, panel B represents Pose 2. Red regions are the most intolerant, blue regions are the most tolerant.*

POSE 1				**POSE 2**		
RESIDUES	***gnomAD***	***MTR***		***RESIDUES***	***gnomAD***	***MTR***
K64	-1	1		Q98	1	1
K103	-1	0,96		R99	-1	1
D106	-1	0,84		E100	1	1
P107	1	0,83		K101	1	0,97
S108	-1	0,86		K103	-1	0,96
L109	1	0,85		D106	-1	0,84
P110	-1	0,82		P107	1	0,83
L111	-1	0,84		S108	-1	0,86
K112	1	0,88		L109	1	0,85
W116	-1	0,82		P110	-1	0,82
G120	-1	0,78		G118	-1	0,76
E121	-1	0,76		G119	1	0,82
D122	1	0,78		E121	-1	0,76
T123	1	0,76		I137	1	0,8
P124	-1	0,77		S138	1	0,77
Y125	-1	0,8		K139	-1	0,74
S126	1	0,8		R140	1	0,73
N127	-1	0,83		G166	1	0,76
W128	-1	0,82		S167	-1	0,85
H129	-1	0,82		L191	1	1
K130	-1	0,81		A192	-1	1,01
E131	-1	0,81		L193	1	1

L132	-1	0,85		G194	1	1,03
R133	1	0,85		G195	1	1,02
N134	1	0,82		P196	1	0,97
S135	1	0,82		E213	-1	0,74
S138	1	0,77		K241	1	0,85
R140	1	0,73		E242	1	0,84
K159	-1	0,74		K243	1	0,85
E162	-1	0,76		D249	1	0,81
P164	1	0,74		W250	-1	0,82
A192	-1	1,01		G251	1	0,81
L193	1	1		S252	-1	0,82
G194	1	1,03		S253	1	0,82
G195	1	1,02		L256	-1	0,79
P196	1	0,97		L257	-1	0,77
K241	1	0,85				
K243	1	0,85				
A244	-1	0,84				
P245	-1	0,85				
V247	-1	0,86				
D249	1	0,81				
W250	-1	0,82				
G251	1	0,81				
S253	1	0,82				
L256	-1	0,79				
C257	-1	0,77				

Table 3. *Results from mapping the gnomAD missense variants to CD93: b-factor=1/-1, where "1" indicated one or more missense variants found at this amino acid position and "-1" represented no missense variants seen at this amino acid position; for each ones is associated the MTR score.*

Conclusion

In silico and experimental procedures were used as a basis to determine CD93 structure-function relationship. The CD93/Multimerin-2 complex was analyzed *in vitro*, dissecting interactions occurring in specific conditions. Homology modeling and protein docking procedures were used to predict the involvement of specific amino acid residues in the CD93/Multimerin-2 interaction. Furthermore, structural analysis of regions of the protein-protein interface were conducted using bioinformatic tools, showing the key role of amino acid residues in the interaction. MD simulations and several post-dynamics analysis have been used to provide a comprehensive understanding of the impact of the inferred mutations on the protein structure and functions. The obtained results demonstrate that the CD93/Multimerin-2 interaction is involved in angiogenesis regulation, opening the possibility to develop new therapeutic tools. Furthermore, the CD93 F238 amino acid residue is instrumental for binding to Multimerin-2, as well as the presence of the so called "non-canonical" disulfide bridge. Finally, we provide an approach to evaluate the best pose of protein–protein complex structures according to new experimental data. With the application of bioinformatic tools, we have described a four-steps workflow in order to predict effects of variations on the inter-residue interaction network at the PPI, based on evolutionary information, complex network metrics and energetic affinity; in addition, it allows to map and explore regional intolerance to missense variation. These observations could provide a basis for the development of anti-angiogenic drugs therapy.

References

Abraham MJ, Murtola T., Schulz R, Pall S, Smith JC, Hess B and Lindahl E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers Software X, 1-2: 19-25.

Amadei, A, Linssen ABM, and Berendsen HJC. (1993). Essential dynamics of proteins. *Proteins*. 17:412–425

Aurell E, Ekeberg M. 2012. Inverse Ising inference using all the data. Physical Review Letters 108:090201. doi: 10.1103/PhysRevLett.108.090201.

Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. (1984). "Molecular Dynamics with Coupling to an External Bath". Journal of Chemical Physics. 81 (8): 3684–3690. Bibcode:1984JChPh..81.3684B. doi:10.1063/1.448118.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H., Shindyalov I.N., Bourne P.E.. (2000) *The Protein Data Bank Nucleic Acids Research*, 28: 235-242.

Blackburn, J. W., Lau, D. H., Liu, E. Y., Ellins, J. , Vrieze, A. M., Pawlak, E. N., Dikeakos, J. D. and Heit, B. (2019), Soluble CD93 is an apoptotic cell opsonin recognized by αxβ2. Eur. J. Immunol., 49: 600-610. doi:10.1002/eji.201847801

Braghetta P., Ferrari A., de Gemmis P., Zanetti M., Volpin D., Bonaldo P., Bressan G. M. (2004). Overlapping, complementary and site-specific expression pattern of genes of the EMILIN/multimerin family. *Matrix Biol*. 22, 549–55610.1016/j.matbio.2003.10.005

Bussi, G., Donadio, D. & Parrinello, M. (2007) Canonical sampling through velocity rescaling. *J Chem Phys*. 126, 014101.

Colladel R., Pellicani R., Andreuzzi E., Paulitti A, Tarticchio G., Todaro F., Colombatti A., Mongiat M., (2016) MULTI- MERIN2 binds VEGF-A primarily via the carbohydrate chains exerting an angiostatic function and impairing tumor growth, *Oncotarget* 7 2022–2037.

Colombatti A, Spessotto P, Doliana R, Mongiat M, Bressan GM, Esposito G. (2012) The EMILIN/Multimerin family. *Front Immunol*.; 2:93. Published 2012 Jan 6.

Darden, T., York, D. & Pedersen, L. (1993). Particle mesh Ewald: An N·log (N) method for Ewald sums in large systems. *J Chem Phys*. 98, 10089–10092.

Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys 87(1):012707.

Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A. & Caves, L. S. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*. 22, 2695–2696

Greenlee MC, Sullivan SA, Bohlson SS. (2008) CD93 and related family members: their role in innate immunity. *Curr Drug Targets*. (2):130-8.

Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* 18, 2714-2723.

Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. J Comput Chem. 18, 1463–1472.

(Hopf, Thomas A and Schärfe, Charlotta P.I. and Rodrigues, João P.G.L.M. and Green, Anna G and Kohlbacher, Oliver and Sander, Chris and Bonvin, Alexandre M.J.J. and Marks, Debora S. Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 2014;3:e03430)

Ichiye, T. & Karplus, M. C (1991) ollective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*. 11, 205–217

Ikewaki, N., Sonoda, T. & Inoko, H. (2013) Unique properties of cluster of differentiation 93 in the umbilical cord blood of neonates. *Microbiol. Immunol*. 57, 822–32.

Janson, G., Zhang, C., Prado, M. G. & Paiardini, A. (2017) PyMod 2.0: improvements in protein sequence-structure analysis and homology modeling within PyMOL. *Bioinformatics* 33, 444–446

Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci USA 2013;110:15674–9.

Kao Y.-C., Jiang S.-J., Pan W.-A., Wang K.-C., Chen P.-K, Wei H.- J., Chen W.-S., Chang B.-I, Shi G.-Y, Wu H.-L., (2012) The Epidermal Growth Factor-like domain of CD93 is a potent angiogenic factor, *PLoS ONE* 7 e51647.

(Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes, Konrad J Karczewski et al. 2019).

Kelley LA et al. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 10, 845-858

Khan KA, Naylor AJ, Khan A, et al. (2017) Multimerin-2 is a ligand for group 14 family C-type lectins CLEC14A, CD93 and CD248 spanning the endothelial pericyte interface. *Oncogene*. 36(44):6097-6108.

Krissine E. l, Henrick K., (2007) Inference of macromolecular assemblies from crystalline state, *J. Mol. Biol*. 372 774–797.

Laskowski R.A., MacArthur M.W., Moss D.S., Thornton J.M., (1993) PROCHECK: a program to check the stereochemical quality of protein structures, *J. Appl. Crystallogr*. 26 283–291.

Lindorff-larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. (2010) Improved sidechain torsion potentials for the Amber ff99SB protein force field. *Proteins.* 78(8):1950–8.

Loghmani, H., & Conway, E. M. (2018). Exploring traditional and non-traditional roles for thrombomodulin. *Blood*, (), blood-2017-12-768994.

Lugano R, Vemuri K, Yu D, et al. (2018) CD93 promotes β1 integrin activation and fibronectin fibrillogenesis during tumor angiogenesis. *J Clin Invest*. 128(8):3280-3297.

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE 6(12): e28766.

Morcos F, Pagnani ALB, Bertolinod A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue co-evolution captures native contacts across many protein families. arXiv:1110.5223v2 [q-bio.QM].

Nativel, B., et al., (2016) Soluble expression of disulfide-bonded C-type lectin like domain of human CD93 in the cytoplasm of Escherichia coli, *J. Immunol. Methods*

Orlandini, Maurizio & Nucciotti, Sara & Galvagni, Federico & Bardelli, Monia & Rocchigiani, Marina & Petraglia, Felice & Oliviero, Salvatore. (2008). Morphogenesis of human endothelial cells is inhibited by DAB2 via Src. FEBS letters. 582. 2542-8. 10.1016/j.febslet.2008.06.025.

Orlandini M., F. Galvagni, M. Bardelli, M. Rocchigiani, C. Lentucci, F. Anselmi, A. Zippo, L. Bini, S. Oliviero, (2014) The characterization of a novel monoclonal antibody against CD93 unveils a new antiangiogenic target, *Oncotarget* 5 2750–2760.

Paracuellos P, Briggs DC, Carafoli F, Lončar T, Hohenester E. (2015) Insights into Collagen Uptake by C-type Mannose Receptors from the Crystal Structure of Endo180 Domains 1–4. *Structure* (London, England:1993). 23(11):2133-2142.

Piovesan D., Minervini G., Tosatto S.C.E. (2016) The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Research*, 44 (W1), pp. W367-74

Pires DEV, Ascher DB, Blundell TL. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research.* 42(Web Server issue): W314-W319.

Pospisilova, E., Kukacka, Z., Kavan, D., Novak, P., Chmelik, J. (2017) NMR structure of human DCL-1 (CD302) extracellular domain.

(Carlos H M Rodrigues, Yoochan Myung, Douglas E V Pires, David B Ascher, mCSM-PPI2: predicting the effects of mutations on protein–protein interactions, Nucleic Acids Research (2019) , gkz383, https://doi.org/10.1093/nar/gkz383).

Schrodinger L., *The PyMOL Molecular Graphics System*, Version 1.8, 2015.

Sievers F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, (2011) Fast, scalable generation of high- quality protein multiple sequence alignments using Clustal Omega, *Mol. Syst. Biol*. 7 539.

(Michael Silk, Slavé Petrovski, David B Ascher, MTR-Viewer: identifying regions within genes under purifying selection, Nucleic Acids Research (2019) , gkz457, https://doi.org/10.1093/nar/gkz457).ù

Teicher B. A. (2019). CD248: A therapeutic target in cancer and fibrotic diseases. *Oncotarget*, 10(9), 993–1009.

The UniProt Consortium UniProt: the universal protein knowledgebase. *Nucleic Acids Res. 45*: D158-D169 (2017)

Tovchigrechko A., I.A. Vakser, (2006) GRAMM-X public web server for protein–protein docking, *Nucleic Acids Res. 34* W310–W314.

(Traynelis J.,* Silk M.,* Wang Q., Berkovic S.F., Liu L., Ascher D.B., Balding D.J., Petrovski S. (2017). Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Research (2017)),

Turner PJ. XMGrace, Version 5.1.19. Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR; 2005

Verdone G, Doliana R, Corazza A, Colebrooke SA, Spessotto P, Bot S et al. (2008) The solution structure of EMILIN1 globular C1q domain reveals a disordered insertion necessary for interaction with the α4β1 integrin. *Journal of Biological Chemistry*. Jul 4;283(27):18947-18956.

Webb, B. & Sali, A. (2014) Protein structure modeling with MODELLER. *Methods Mol Biol* 1137, 1–15

Yang J, R Yan, A Roy, D Xu, J Poisson, Y Zhang. (2015) The I-TASSER Suite: Protein structure and function prediction. *Nature Methods*, 12: 7-8

Zelensky, A. N. and Gready, J. E. (2003), Comparative analysis of structural properties of the C-type-lectin-like domain (CTLD). *Proteins*, 52: 466-477.

Zelensky, A. N. and Gready, J. E. (2005), The C-type lectin-like domain superfamily. *The FEBS Journal*, 272: 6179-6217.