UNIVERSITY OF CALIFORNIA SAN DIEGO

Biologically-interpretable machine learning for microbial genomics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioengineering

by

Erol Sincar Kavvas

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Yoav Freund
Professor Christian Metallo
Professor Victor Nizet
Professor Shankar Subramaniam

2020

The dissertation of Erol Sincar Kavvas is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To my parents, Jale and Mustafa

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work could not have been done without the help and support from countless people. First and foremost, I would like to thank my advisor Bernhard Ø. Palsson for all his guidance and inspiration throughout my graduate studies. I could not imagine a better advisor than Dr. Palsson for fostering the immense curiosity I had for biology when I first started the graduate program. His passion for systems biology was infectious and a constant source of inspiration. I'm thankful for the numerous books he lent me and for never failing to respond to my emails, even as ridiculous as some of them were. I was always amazed by Dr. Palsson's ability to balance the free-thinking mindset of a child (this is a compliment) and the get-things-done attitude of a successful CEO. Whatever I pursue in life, I'll be sure to channel as much Dr. Palsson energy as possible.

I have been fortunate to have had many inspiring mentors at the SBRG. First, I would like to thank Jonathan Monk for being a major source of support in my first couple years and for teaching me the basics of research and paper writing. I'm also thankful to Laurence Yang and David Heckmann for similar research lessons and for the insightful feedback that shaped my FBA-GWAS study. I'd like to thank Aarash Bordbar for the valuable internship opportunity at Sinopia Biosciences. I'd also like to thank Adam Feist for guiding me in my final years of researching *E. coli* evolution.

I am grateful for everyone at SBRG. There are so many colleagues and friends to thank here, but I would especially like to thank Anand, Colton, Saugat, Yara, CJ, Patrick, Bin, Muyao, James, Jared, David, Julia, Sonal, and Jacob for daily chats and relaxing outdoor lunches. I'd also like to thank my batchmates Yara, Xin, and Eddie who joined SBRG at the same time as me. I'd also like to thank Marc Abrams for helping me with countless tasks and for being a

generally tight dude.

I would like to thank the friends I made these past five years in grad school. In particular, I'd like to thank Michael Ostertag, Vish Ramesh, Ashish Manohar, and Ritvik Vasan for being tight AF and for providing terrible life advice. I'd like to thank my friends Nate Chapin, Chris Camona, Swiss Greg, Kyle, Co, Lennart and the Chemistry 2015 PhD group for all the fun.

I would also like to thank my funding sources that have supported this work. These include the National Institute of Allergy and Infectious Disease (AI124316) and the Novo Nordisk Foundation (NNF10CC1016517).

Chapter 2 is a reprint of material published in: **ES Kavvas**, E Catoui, N Mih, JT Yurkovich, Y Seif, N Dillon, D Heckmann, A Anand, L Yang, C Nizet, JM Monk and BO Palsson. 2018. "Machine learning and structural analysis of *Mycobacterium tuberculosis* pangenome identifies genetic signatures of antibiotic resistance" *Nature Communications* 9 (4306). The dissertation author is the primary author.

Chapter 3 is a reprint of the material published in: **ES Kavvas**,Y Seif, JT Yurkovich, C Norsigian, S Poudel, WW Greenwald, S Ghatak, BO. Palsson and JM Monk. 2018. "Updated and standardized genome-scale reconstruction of *Mycobacterium tuberculosis* H37Rv, iEK1011, simulates flux states indicative of physiological conditions" *BMC Systems Biology* 12 (25). The dissertation author is the primary author.

Chapter 4 is a reprint of the material in: **ES Kavvas**, L Yang, JM Monk, D Heckmann, BO Palsson. 2020. "A biochemically-interpretable machine learning classifier for microbial GWAS" *Nature Communications* 11 (2580). The dissertation author is the primary author.

Chapter 5 is a reprint of the material: **ES. Kavvas**, MR. Antoniewicz, C. Long, Y. Ding, JM. Monk, BO. Palsson, A. Feist. (2020). Laboratory evolution of multiple *E. coli*

strains reveals unifying principles of adaptation but diversity in driving genotypes. *bioRxiv* DOI:10.1101/2020.05.19.104992. The dissertation author is the primary author.

VITA

2015	Bachelor of Science in Civil and Environmental Engineering, University of California Davis
2020	Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

Bin Du, Daniel C. Zielinski, **Erol S. Kavvas**, Andreas Drager, Justin Tan, Zhen Zhang, Kayla E. Ruggiero, Garri A. Arzumanyan and Bernhard O. Palsson. 2016. Evaluation of rate law approximations in bottom-up kinetic models of metabolism. *BMC Systems Biology* 10:40.

ES Kavvas,Y Seif, JT Yurkovich, C Norsigian, S Poudel, WW Greenwald, S Ghatak, BO. Palsson and JM Monk. 2018. "Updated and standardized genome-scale reconstruction of *Mycobacterium tuberculosis* H37Rv, iEK1011, simulates flux states indicative of physiological conditions" *BMC Systems Biology* 12 (25).

Yara Seif, **Erol Kavvas**, Jean-Christophe Lachance, James T Yurkovich, Sean-Paul Nuccio, Xin Fang, Edward Catoiu, Manuela Raffatellu, Bernhard O Palsson, Jonathan M Monk. 2018. "Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits" *Nature Communications* 9:3771.

ES Kavvas, E Catoui, N Mih, JT Yurkovich, Y Seif, N Dillon, D Heckmann, A Anand, L Yang, C Nizet, JM Monk and BO Palsson. 2018. "Machine learning and structural analysis of *Mycobacterium tuberculosis* pangenome identifies genetic signatures of antibiotic resistance" *Nature Communications* 9 (4306).

Charles J Norsigian, **Erol Kavvas**, Yara Seif, Bernhard O Palsson, Jonathan M Monk. 2018. "iCN718, an updated and improved genome-scale metabolic network reconstruction of Acinetobacter baumannii AYE" *Frontiers in genetics* 9, 121.

Kumari S Choudhary, Nathan Mih, Jonathan Monk, **Erol Kavvas**, James T Yurkovich, George Sakoulas, Bernhard O Palsson. 2018. "The Staphylococcus aureus two-component system AgrAC displays four distinct genomic arrangements that delineate genomic virulence factor signatures" *Frontiers in microbiology* 9, 1082.

Nathan Mih, Elizabeth Brunk, Ke Chen, Edward Catoiu, Anand Sastry, **Erol Kavvas**, Jonathan M Monk, Zhen Zhang, Bernhard O Palsson. 2018. "ssbio: a Python framework for structural systems biology" *Bioinformatics* 34 (12), 2155-2157.

Xin Fang, Jonathan M Monk, Nathan Mih, Bin Du, Anand V Sastry, **Erol Kavvas**, Yara Seif, Larry Smarr, Bernhard O Palsson. 2018. "Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa" *BMC systems biology* 12 (1), 66.

Y Seif, JM Monk, H Machado, **E Kavvas**, BO Palsson. 2019. "Systems Biology and Pangenome of Salmonella O-Antigens" *mBio* 10 (4), e01247-19.

Jason C Hyun, **Erol S Kavvas**, Jonathan M Monk, Bernhard O Palsson. 2020. "Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens" *PLoS computational biology* 16 (3), e1007608.

ES Kavvas, L Yang, JM Monk, D Heckmann, BO Palsson. 2020. "A biochemically-interpretable machine learning classifier for microbial GWAS" *Nature Communications* 11 (2580).

ES. Kavvas, MR. Antoniewicz, C. Long, Y. Ding, JM. Monk, BO. Palsson, A. Feist. (2020). Laboratory evolution of multiple E. coli strains reveals unifying principles of adaptation but diversity in driving genotypes. *bioRxiv* DOI:10.1101/2020.05.19.104992.

ABSTRACT OF THE DISSERTATION

Biologically-interpretable machine learning for microbial genomics

by

Erol Sincar Kavvas

Doctor of Philosophy in Bioengineering

University of California San Diego, 2020

Professor Bernhard Ø. Palsson, Chair

Advancements in high-throughput biotechnology have enabled the unprecedented detailing of microbial diversity. Researchers now have the opportunity to understand evolution as a function of the genomic, transcriptomic, metabolic, and physiological variables underlying differential fitness. A comprehensive understanding of microbial evolution will help eradicate infectious disease, engineer robust synthetic circuits, and tackle environmental issues facing our planet. While the information revolution in biology has enabled researchers to simultaneously measure various biomolecules at low costs, a major bottleneck remains in translating these datasets to actionable knowledge. In this proposal, we aim to address the challenge of biological data anal-

ysis through development of computational methods that leverage both the predictive power of machine learning (ML) and the biological interpretability of mechanistic genome-scale models. First, classical ML is applied to thousands of drug-tested *Mycobacterium tuberculosis* genome sequences to recover 33 known genetic determinants of antimicrobial resistance (AMR) and 24 novel candidates. Second, a biochemically-interpretable ML model is developed and applied to the same genomics dataset to reveal metabolic mechanisms of AMR. Third, independent component analysis is applied to a multi-omics dataset of *E. coli* laboratory evolution to reveal multi-scale adaptation principles governing causal mutations. In conclusion, this dissertation broadened our understanding of microbial evolution through development and application of interpretable ML models.

Chapter 1

Introduction

Underlying microbial diversity are stories of evolutionary adaptation. From antimicrobial resistance (AMR) acquisition to increased acid tolerance, microbes are able adapt to a seemingly endless range of environmental scenarios through alteration of their genetic program by the process of evolution. All stories of adaptation are therefore written in the DNA sequences of microbial genomes. With technological advancements enabling cheap and accessible genome sequencing, there are now public databases filled with thousands of microbial genomes [1], providing researchers the unprecedented opportunity to read the evolutionary history of microbes. For the deadly pathogen *Mycobacterium tuberculosis* (TB), the thousands of publicly available drug-tested genomes sequences may be utilized towards understanding the causes of AMR that may consequently lead to better treatment regimens and assist novel drug development. However, despite the availability of genome sequences, it remains challenging to deduce the evolutionary causes underlying genetic diversity, highlighting the need for novel mathematical methods that can predicatively link genetic variants to meaningful evolutionary causes.

1.1 Causation in Biology

"Cause and effect in biology is a farcry from that in physics"

— Ernst Mayr

Unlike physics, biology is subject to both proximal and distal causation [2] [3]. Proximal causation describes physical phenomena and can be formulated with basic principles such as thermodynamics and mass conservation. Distal causation on the other hand is unique to biology and describes the process of evolution, in which natural selection favors those individuals in a genetically diverse population that have more fit functions than other members of the population. In terms of modeling, distal causation necessitates the specification of a fitness function that can distinguish the survivability of each individual in a diverse population. The challenge is that such a fitness function can not be specified *a priori* using physical principles but is instead determined by natural selection. For example, while the proximal cause of how microbes produce antibiotics can be explained using biophysical principles such as biochemistry and mass conservation, the distal cause of why microbial antibiotic production exists in the first place has no physical basis but is instead due to being able to kill other microbes in the population. Therefore, we move away from pure physical models and instead focus our modeling efforts in two categories: (1) statistics and machine learning models, and (2) constraint-based genome-scale models. The first category is based on identifying key correlations in the data while the second category is based on identifying key mechanisms.

1.2 Machine learning models for identifying predictive features

In the 1920s, the mathematician Ronald Fisher advanced the field of statistics as a means to identify underlying causes in biological data sets. Without having to account for any biologi-

cal mechanisms, these statistical models provided levels of significance for associations between biological measurements, enabling researchers to better dissect the underlying causes of observed traits. For these reasons, statistical models have been the basis for the majority of scientific discoveries. In the field of genome-wide association studies (GWAS), statistical models are used to filter millions of genetic variants for those most likely to cause an observed trait such as disease [4].

Although statistics remains as the dominant methodology for analyzing biological data sets, another mechanism-agnostic modeling approach known as machine learning (ML) has become immensely popular in recent years. In contrast to statistical models, which are designed to draw inferences about the relationship between variables, ML models are designed to find generalizable predictive patterns [5]. By adding features like L1-regularization to the ML optimization problem, the ML model learns to make accurate predictions using as few input features as possible, thereby filtering the data for key features. Application to microbial GWAS data sets has shown that ML-derived predictive features correspond to known genetic determinants of antibiotic resistance [6–11]. For transcriptomics data, patterns derived from the ML method of independent component analysis have been shown reflect known regulons and enable quantitative modeling of the transcriptional regulatory network (TRN) [12].

1.3 Constraint-based models address dual causation

The primary limitation of current ML models is their inability to infer biological mechanisms. Fundamentally, current ML models, such as the Support Vector Machine (SVM), contain no information about the function of genes or how they interact in a biomolecular system to create phenotypes. Over the past couple of decades, the computational analysis of biochemical networks

in microorganisms has been advanced through the use of genome-scale models (GEMs) [3] [13]. By computing metabolic flux states consistent with imposed biological constraints, GEMs have been shown to predict a range of cellular functions, making them a valuable tool for analyzing multi-omics datasets [14]. Significantly, GEMs represent proximal causes as constraints (e.g., reaction stoichiometry) and distal causes through the objective function (e.g., maximize biomass biosynthetic flux), thereby addressing the unique duality of biological causation.

1.4 Towards mechanistic machine learning

Although GEMs are transparent genotype-phenotype models, they are largely outperformed by machine learning models in direct comparisons of prediction accuracy. Approaches have thus been developed that integrate meaningful GEM computations with predictive "black-box" machine learning to enable "white-box" interpretations of data [15]. These approaches have worked well for endogenous metabolomics data by using the GEM to directly transform the measurements to meaningful inputs for "black box" machine learning. Approaches that integrate the interpretability of genome-scale models with the predictability of machine learning models may therefore realize the promise of big data biology.

1.5 References

1. Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y. & Sobral, B. W. PATRIC, the bacterial bioinformatics database and analysis resource. en. *Nucleic acids research* **42,** D581–91. ISSN: 0305-1048, 1362-4962 (Jan. 2014).

2. Mayr, E. *This is Biology: The Science of the Living World* ISBN: 9780674884694. `https://books.google.com/books?id=-ddVamDO-xcC` (Belknap Press of Harvard University Press, 1998).

3. Palsson, B. Ø. *Systems Biology: Constraint-based Reconstruction and Analysis* en. ISBN: 9781316239940 (Cambridge University Press, Jan. 2015).

4. Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. en. *Nature reviews. Genetics* **18,** 41–50. ISSN: 1471-0056, 1471-0064 (Jan. 2017).

5. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* **16,** 199–231. https://doi.org/10.1214/ss/1009213726 (Aug. 2001).

6. Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. & Palsson, B. O. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. en. *Nature communications* **9,** 4306. ISSN: 2041-1723, 2041-1723 (Oct. 2018).

7. Boolchandani, M., D'Souza, A. W. & Dantas, G. Sequencing-based methods and resources to study antimicrobial resistance. en. *Nature reviews. Genetics.* ISSN: 1471-0056, 1471-0064 (Mar. 2019).

8. Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A. R., Will, R., Xia, F. & Stevens, R. Antimicrobial Resistance Prediction in PATRIC and RAST. en. *Scientific reports* **6,** 27930. ISSN: 2045-2322 (June 2016).

9. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P. & Zhang, L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. en. *Microbiome* **6,** 23. ISSN: 2049-2618 (Feb. 2018).

10. Walker, T. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. *Nature reviews. Microbiology.* ISSN: 1740-1526 (2019).

11. Earle, S. G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., Spencer, C. C. A., Iqbal, Z., Clifton, D. A., Hopkins, K. L., Woodford, N., Smith, E. G., Ismail, N., Llewelyn, M. J., Peto, T. E., Crook, D. W., McVean, G., Walker, A. S. & Wilson, D. J. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. en. *Nature microbiology* **1,** 16041. ISSN: 2058-5276 (Apr. 2016).

12. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. The Escherichia coli transcriptome mostly consists of independently regulated modules. en. *Nature communications* **10,** 5536. ISSN: 2041-1723 (Dec. 2019).

13. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. en. *Cell* **161,** 971–987. ISSN: 0092-8674, 1097-4172 (May 2015).

14. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. en. *Nature reviews. Genetics* **15,** 107–120. ISSN: 1471-0056, 1471-0064 (Feb. 2014).

15. Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., Walker, G. C. & Collins, J. J. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. en. *Cell.* ISSN: 0092-8674, 1097-4172 (May 2019).

Chapter 2

Machine learning of *M. tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance

Mycobacterium tuberculosis is a serious human pathogen threat exhibiting complex evolution of antimicrobial resistance (AMR). Accordingly, the many publicly available datasets describing its AMR characteristics demand disparate data-type analyses. Here, we develop a reference strain-agnostic computational platform that uses machine learning approaches, complemented by both genetic interaction analysis and 3D structural mutation-mapping, to identify signatures of AMR evolution to 13 antibiotics. This platform is applied to 1595 sequenced strains to yield four key results. First, a pan-genome analysis shows that *M. tuberculosis* is highly conserved with sequenced variation concentrated in PE/PPE/PGRS genes. Second, the platform corroborates 33 genes known to confer resistance and identifies 24 new genetic signatures of AMR. Third,

97 epistatic interactions across 10 resistance classes are revealed. Fourth, detailed structural analysis of these genes yields mechanistic bases for their selection. The platform can be used to study other human pathogens.

2.1 Background

Advancements in genome sequencing technologies have made available thousands of drug-tested *M. tuberculosis* genomes in public databases. With available sequences expected to surpass 60,000 during the next five years [1], there is impetus for new quantitative approaches that excel at analyzing massive datasets. Methods that explicitly account for structure amongst features—such as those found in the field of machine learning—will be essential for addressing this *M. tuberculosis* data deluge [2].

To date, most approaches compare *M. tuberculosis* genome sequences against the H37Rv reference strain in order to identify single nucleotide polymorphisms (SNPs). Following SNP identification, most studies then focus on the subset of previously identified resistance-determining SNPs that have been previously determined to be key resistance-determining mutations, specifically those within a handful of genes encoding proteins targeted by drugs [3]. While such studies have proven to be powerful for diagnostics [4] and elucidating mutational steps to AMR [3], they do not account for potential genome-wide mutations reflecting positive AMR selection, such as those found to be related to cell wall permeability, efflux pumps, and compensatory mechanisms [5].

Specific genome-wide functional analyses in *M. tuberculosis* have shown that *ald* loss-of-function [6], *ubiA* gain-of-function [7], and *thyA* loss-of-function [8] mutations occur in off-target reactions and confer resistance through modulation of metabolite pools. These results exemplify

the complex interplay underlying AMR phenotypes that extends beyond the few genes currently utilized in diagnostic studies. In addition to limitations of a narrow genetic view, the identification of other types of resistance-conferring mutations, such as deletions [9, 10], suggest that SNPs are no longer comprehensive in describing the mutational landscape of *M. tuberculosis* AMR evolution.

Here, we apply a reference-agnostic machine learning approach complemented by both genetic interaction and protein structural analysis to deduce the variability in genetic content and AMR of 1,595 *M. tuberculosis* strains. The complete analysis recapitulates known AMR mechanisms and infers specific selection pressures through directed hypotheses.

2.2 Results

2.2.1 Characterizing the *M. tuberculosis* pan-genome

Our first goal was to characterize and understand the gene content of sequenced *M. tuberculosis* strains. We selected a representative set of 1,595 *M. tuberculosis* strains for which AMR testing data was available from the PATRIC database [11]. These strains come from a wide range of studies [3, 12–27]. Strains were selected for their genetic, geographic, and AMR phenotypic diversity (Supplementary Fig. 1). The geographic diversity of these strains reflects areas heavily burdened by *M. tuberculosis* (Supplementary Fig. 1a). We constructed a phylogenetic tree for the 1,595 strains using a robust set of lineage-defining SNPs [28] (Supplementary Fig. 1b and Methods). Finally, strains were selected in order to provide a distribution across commonly used *M. tuberculosis* treatment regimens (Methods). Of these 1,595 strains, 1,282 strains had AMR testing data for isoniazid, rifampicin, streptomycin, and ethambutol (Supplementary Fig. 1c)

and 946 (59%) were resistant to both isoniazid and rifampicin. Following the selection of strains, we determined the pan-genome (i.e., the union of all genes across the strains) represented by these data and analyzed the distribution of various genomic features (i.e., core genes, virulence factors, etc.). The pan-genome analysis described a general theme of high conservation (see Supplementary Note for further discussion of *M. tuberculosis* pan-genome).

2.2.2 Assessing allele frequencies in the pan-genome identifies key resistance-conferring genes

Although the *M. tuberculosis* pan-genome clusters provide an informative view of the global genetic repertoire within a species, they lack the resolution necessary to discriminate between most AMR phenotypes. To elucidate fine-grained genetic variation indicative of AMR evolution, we separated each pan-genome cluster into groups of exact amino acid sequence variants, or alleles (Supplementary Fig. 3g). In contrast to alignment-based perspectives, the allele-based pan-genome does not reduce non-H37Rv variants to a collection of SNPs, but instead represents variants in their functional protein coding form. This approach accounts for all protein-coding alleles in the *M. tuberculosis* pan-genome, thereby representing the extensive strain-to-strain variation observed in bacterial genomes without biasing the variations relative to a single reference genome.

We used mutual information [29] as an association metric to identify resistance-determining genes with this newly constructed variant pan-genome and the accompanying AMR dataset (Methods). Importantly, this approach identified primary resistance-conferring genes previously reported in the literature (Figure 2.1). In addition to mutual information, we calculated associations using a chi-squared test and an ANOVA F-test, both of which identified similar

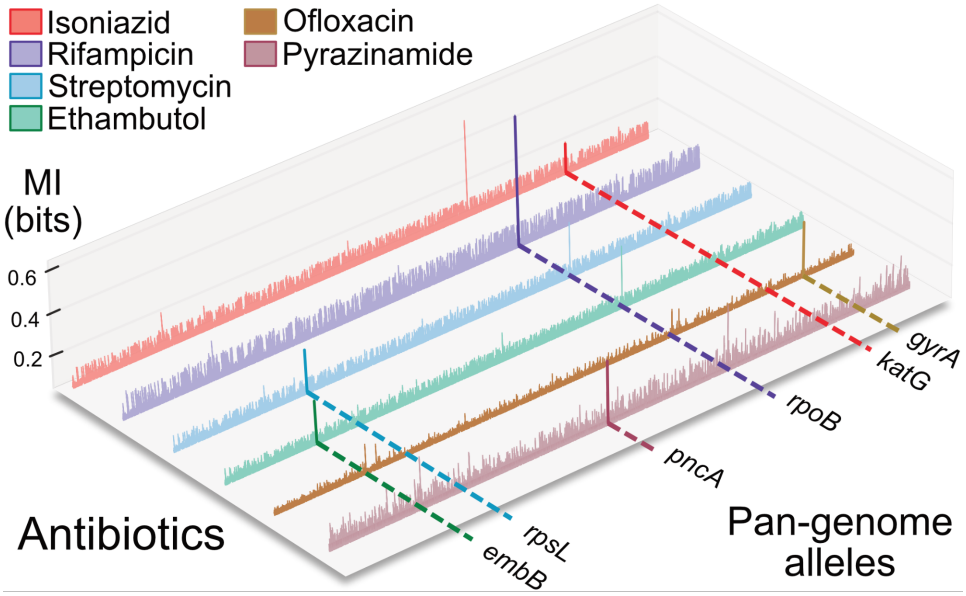

Figure 2.1: Identification of key resistance-conferring genes using mutual. The pairwise mutual information (vertical axis) between the pan-genome alleles and antibiotic resistance was calculated across all possible pairs. The listed genes correspond to the pan-genome alleles that hold the most information about the listed drug's AMR phenotype.

sets of key AMR genes ($P < 0.005$; Bonferroni correction) (Supplementary Data File 1). These results suggest that allele frequencies based on exact sequence (i.e., without a metric for genetic distance) are capable of identifying AMR genes, which has previously been shown with k-mer based approaches [30–32].

2.2.3 Machine learning identifies known resistance genes and novel candidates

Although simple and effective, pairwise association tests (i.e., mutual information, chi-squared, and ANOVA F-test) do not simultaneously account for multiple alleles because the pairwise calculations consider variants independently of one another. Thus, we tailored a support vector machine (SVM)—a method that inherently accounts for structure amongst the features—to uncover AMR-conferring genes (Methods). Using the allele presence-absence across strains as the features, the SVM identified an additional seven known AMR gene-antibiotic rela-

Table 2.1: Known AMR genes uncovered by machine learning. The eight antibiotics shown each have an AUC greater than 0.80 (Supplementary Fig. 5). *Not found in top 40 ranked alleles determined by mutual information, chi-squared, and ANOVA F-test.

Antibiotics	Known AMR genes
isoniazid	*katG* [33], *inhA** [34] , *fabG1* [35]
rifampicin	*rpoB* [36], *rpoC** [37] *Rv3239c* [38]
ethambutol	*embB* [39], *embC* [39], *ubiA** [40], *embR** [41]
pyrazinamide	*pncA* [42]
streptomycin	*rpsL* [43], *gidB* [44]
ofloxacin	*gyrA* [45]
4-aminosalicylic acid	*folC** [8], *thyA** [46]
ethionamide	*ethA* [47], *inhA** [34]
Known AMR genes associated with other antibiotics	*dprE1* [48], *ald* [6], *alr* [49], *murA* [50], *pks2* [51], *pks12* [52], *ppsA* [53], *ppsD* [53], *drrB* [54], *drrC* [54], *moeW* [48], *Rv0687* [55], *mshD* [56], *gyrB* [45], *Rv1877* [57], *Rv0194* [58]

tions absent from the top 40 ranked alleles determined by pairwise associations, including those associated with complex resistance (Table 2.1). In particular, ubiA, a resistance gene recently found to confer high level resistance to ethambutol [40], appeared as a strong signal across the ensemble of SVM simulations—despite not being accounted for in contemporary *M. tuberculosis* diagnostics (Supplementary Data File 2).

The SVM method revealed an abundance of AMR-implicated genes involved in metabolic pathways (119/317, 37.5%) (Supplementary Data File 2). In fact, the majority of the known AMR-determinants are metabolic enzymes (24/33, 73%). We found over 20 genes related to cell wall processes (26/317, 8.2%), which is consistent with previous findings of convergent AMR evolution in *M. tuberculosis* [5]. Furthermore, many high-signal AMR genes, such as pbpA and mmpS3, have recently been identified as determinants of intrinsic *M. tuberculosis* AMR [59]. The full list of identified genes for each drug is provided (Supplementary Data File 2).

2.2.4 Machine learning uncovers genetic interactions contributing to AMR

Beyond identifying AMR genes, four key attributes of our ensemble SVM learning approach enable analysis of genetic interactions underlying variable AMR phenotypes (Methods and Supplementary Fig. 4): (1) the weighting of a particular allele in a specific SVM hyperplane scales with its contribution to a particular AMR phenotype, (2) the sign of the weighting (positive or negative) corresponds to the contribution of that allele to the AMR phenotype (i.e., positive weights correspond to resistance while the negative weights correspond to susceptibility), (3) the magnitude and sign of an allele weighting is dependent upon the magnitudes and signs of other alleles within the same hyperplane, and (4) the use of bootstrapping (i.e., randomized subsampling of the population with replacement), and stochastic gradient descent ensures variability in the weights, signs, and set of alleles for each SVM hyperplane. Motivated by attributes 3 and 4, we hypothesized that two genes may interact if the weights, signs, and appearance of their alleles are significantly correlated across the ensemble of SVM hyperplanes (Methods). Therefore, to identify genetic interactions contributing to AMR in *M. tuberculosis* strains, we constructed a correlation matrix of allele weights across the ensemble of randomized SVM hyperplanes (Supplementary Data File 3) and filtered for the top 60 highest gene-gene correlations for eight AMR classifications. The resulting set of gene-gene pairs were interrogated through logistic regression modeling, selecting those gene pairs with statistically significant allele-allele interactions (P¡0.05; Benjamini-Hochberg correction) (Methods and Supplementary Fig. 4). This approach uncovered 74 potential genetic interactions (Supplementary Table 3).

We can use the evolution of ethambutol resistance as a case study to examine the output of our approach. Epistasis analysis of ethambutol AMR genes implicated interactions between embB, ubiA and embR; all genes known to contribute to ethambutol resistance [40] [60]. Although

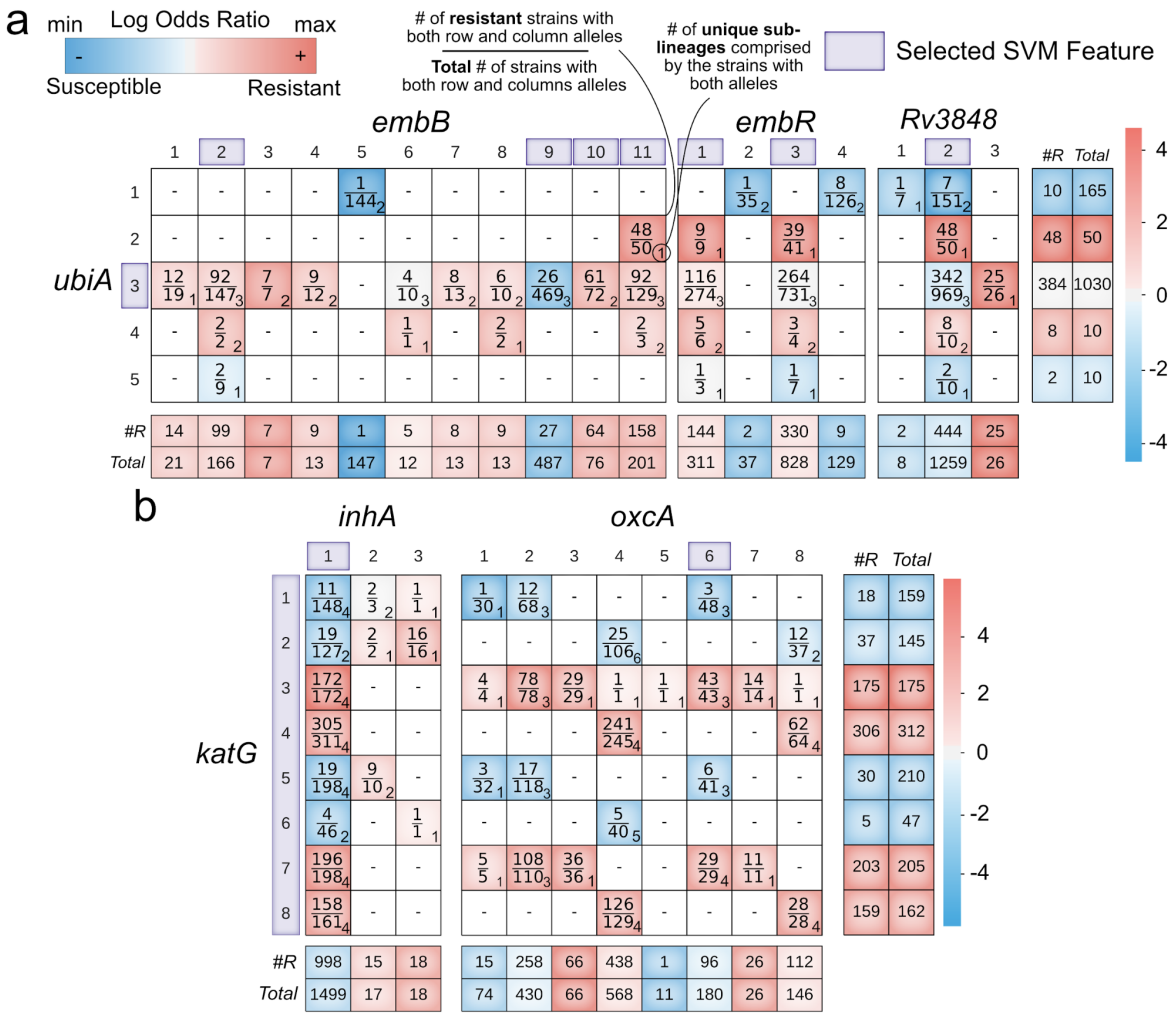

Figure 2.2: Co-occurrence of epistatic genes identified in (a) ethambutol and (b) isoniazid. For the rows on the bottom and on the far right, "#R" refers to the total number of strains that have the allele and are resistant to the specific drug. Total refers to the total number of strains that have that allele that were tested on that specific drug. Each cell is colored by the log odds ratio (LOR) with respect to the AMR phenotype. The numbers in the bottom right of each allele co-occurrence box describes the number of unique sub-lineages comprised by the strains with both alleles (Methods). The alleles enclosed by a purple box represent those chosen as features by the support vector machine (SVM). Note that in some cases the rows and columns do not sum up to the total strains due to rare cases when strains lack those alleles (Methods).

the embR alleles appeared few times across the multiple SVM simulations, their appearance was highly correlated with alterations in the sign and weight of the ubiA allele (see Supplementary Figure 6). This implies that embR is only a predictive feature within the context of ubiA, which

may result from the weak penetrance of embR alleles within *M. tuberculosis* (Figure 2.2a). Logistic regression modeling identified significant allele-allele interactions between ubiA and embR alleles (Supplementary Table 3). We investigated these interactions through a co-occurrence table of the genes, where each cell corresponds to the number of resistant strains with both alleles over the total number of strains with both alleles (Figure 2.2a). The log odds ratio (LOR)—a measurement of the association of the co-occurrence of both alleles with AMR phenotype—was used to color each cell in the co-occurrence table ((Figure 2.2a, see Methods). We observed that the resistant-dominant ubiA alleles (i.e., those with high positive LOR), 2 and 4, occurred exclusively in the background of non-susceptible-dominant embR alleles (Figure 2.2a). Interestingly, in contrast to embB and ubiA, no embR allele appeared as a clear resistance determinant (Figure 2.2a). Furthermore, neither embR nor ubiA were significantly associated with ethambutol AMR in pairwise associations tests (Table 1 and Supplementary Data File 1), showing that our ensemble-based machine learning approach uncovers *M. tuberculosis* AMR complexity. In addition to these known AMR determinants of ethambutol, our analysis implicated ubiA interactions with Rv3848 in ethambutol resistance (Table 2 and Supplementary Table 3). Interestingly, the resistant-dominant allele of Rv3848 occurs exclusively in the background of the AMR-neutral ubiA allele 3, hinting at an alternative route of high-level ethambutol resistance.

For identified isoniazid AMR genes, the co-occurrence table highlighted cases where either katG or inhA genes provide the dominant mode of resistance (Figure 2.2b). Specifically, the incidence of susceptible katG alleles 1, 2, 5, and 6 (i.e., low LOR) with the resistance inhA alleles 2 and 3 (i.e., high LOR) showed that isoniazid resistance in our dataset arose from either katG or inhA mutations, but not both. Aside from these two highly studied isoniazid AMR determinants, epistatic interactions between katG and oxcA appeared with a high signal and further displayed

an interesting co-occurrence relationship with katG (Figure 2.2b). This epistatic interaction for oxcA has not been previously described; specifically, alleles 3 and 7 of oxcA appear exclusively in isoniazid resistant strains. While the AMR phenotypes for the strains containing these alleles may be attributed to the presence of the resistance dominant katG alleles 3 and 7, as is often offered in studies to "explain resistance", the variation in AMR phenotypes across the different alleles were determined to be significant by the machine learning algorithm and thus motivated further investigation. Co-occurrence tables of epistatic AMR genes are provided for the 10 antibiotic classifications (Supplementary Data File 4).

2.2.5 Structural analysis of implicated AMR genes suggest a mechanistic driver of selection

Although the machine learning results agree with experimental literature, it remains unclear whether the uncovered genetic features are either true determinants of AMR or possible artifacts of the statistical learning algorithm. To gain additional insight into whether or not the uncovered alleles are causal in AMR evolution, we mapped the alleles of the 254 AMR genes to protein structures using both experimental crystal structures (20/254) and predicted homology models (50/254) using the ssbio Python package (Methods) [61]. Out of the 254 genes, 217 had available protein sequence annotations (i.e., binding domains, secondary structures, etc.). First, we established a positive control by mapping the alleles of known AMR genes to protein structures and verified that resistance conferring alleles were located in annotated structural regions that indicate the known mechanism of action (Supplementary Fig. 7). For example, structural mapping of the isoniazid AMR-determinant, inhA, showed that the resistance-dominant alleles of 2 and 3 are located within two NAD-binding domains (Figure 2.3a). The incidence of these two alleles

in proximal NAD-binding domains is congruent with the experimentally-derived mechanism of action, which describes the bactericidal effect of tight binding between the isoniazid-NAD adduct and inhA [62, 63]. Moreover, the resistance-conferring mutations in the NAD-binding domains explains the previously described allele co-occurrence of susceptible katG alleles 1, 2, 5, and 6, with resistant *inhA* alleles 2 and 3, because the isoniazid-NAD adduct results from binding to katG, which would only occur if the *M. tuberculosis* strain lacks the resistance-conferring katG mutation that disables the isoniazid binding opportunity. With established confidence through case-controls, we set out to analyze the implicated and uncovered AMR genes.

Revisiting the ethambutol case study, we noticed that the susceptible-dominant embR alleles shared an SNP that is 14.6 Å away from the DNA-binding domain (Figure 2.3a). Given that embR is a positive regulator of embB [64] and that the expression of embB decreases in the presence of ethambutol [40], the SNP suggests a relative increase over alleles 1 and 3 in expression of the ethambutol target, embB, through increased DNA binding. For oxcA, the resistance-dominant alleles, 3 and 7, uniquely share mutations at residue 253, which is contained in the thiamin diphosphate-dependent enzyme M-terminal domain and is 4.51 Å proximal to a mutation at residue 224 shared by most alleles (Figure 2.3). Notably, oxcA is an essential oxalyl-CoA decarboxylase enzyme that converts toxic oxalyl-CoA to CO2 and formyl-CoA, and plays a role in low pH adaptation in E. coli [65]. The totality of studies describing the poisonous effect of glyoxylate [66], significant acid stress in the macrophage environment, use of CO2 as a carbon source [67], and the importance of glyoxylate metabolism in antibiotic tolerance [68], all suggest that the uncovered resistance-conferring adaptations in oxcA increase depletion of oxalyl-CoA through increased binding affinity of the thiamin diphosphate cofactor. Without structural models, sequence annotations of structural features enabled the delineation of resistant

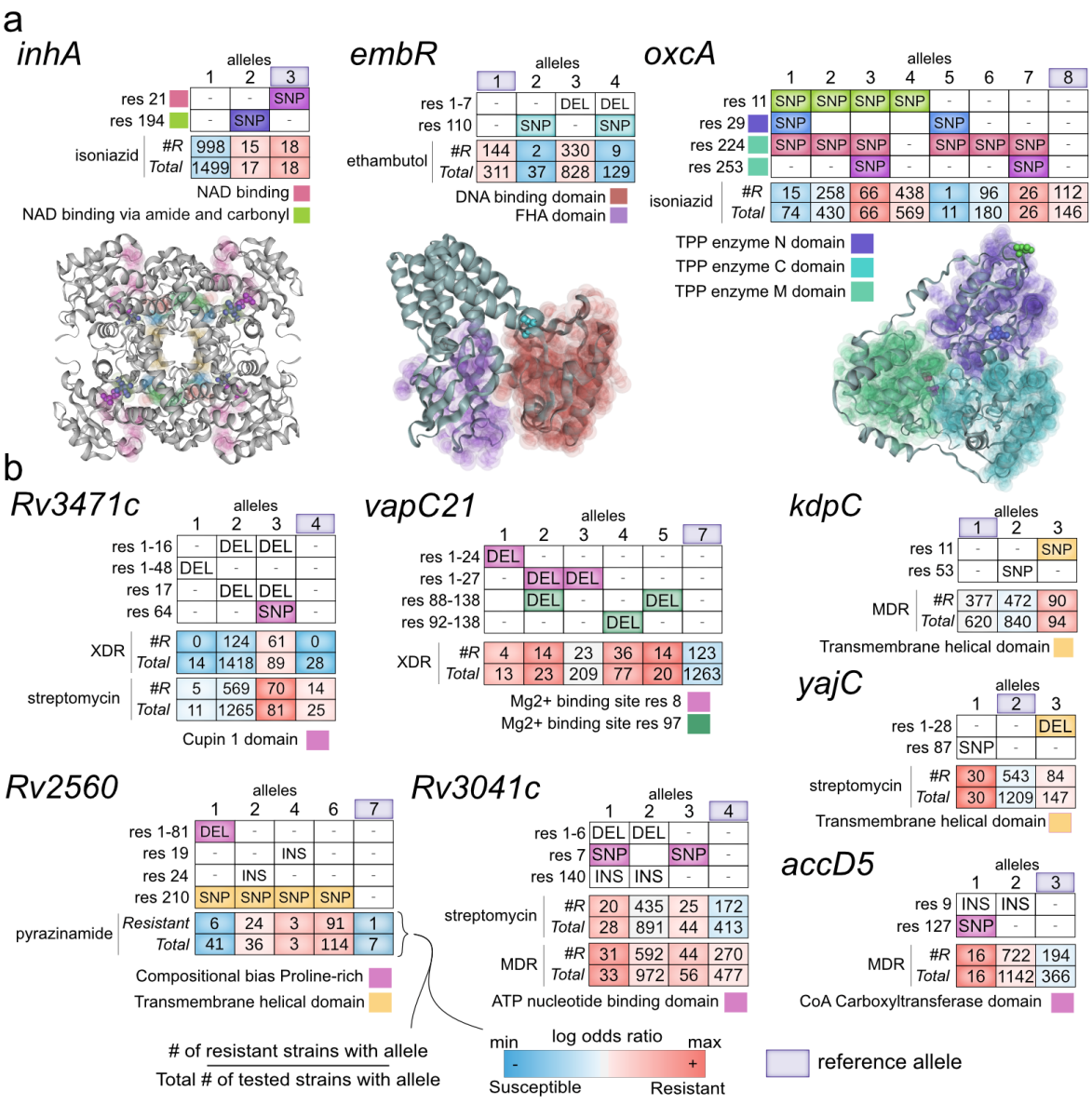

Figure 2.3: 3D and annotated protein structure mutation maps for identified AMR genes. (a) 3D protein structures with mapped mutations are shown for inhA, embR, and oxcA. The colors adjacent to and within the structural mutation table correspond to domains and mutations displayed on the protein structure, respectively. (b) Mutation tables for seven new AMR genes. The colors in the mutation table correspond to the incidence of an annotated structural feature located below the table. The two rows directly below the mutation table are colored according to the log odds ratio between the allele frequency and AMR phenotype. Two AMR classes are shown for Rv3471c and Rv3041c.

and susceptible allele mutations to unique structural domains—highlighting an advantage of our exact-variant perspective (Figure 2.3b). We provide a list of newly implicated AMR genes

Table 2.2: Newly proposed AMR genes. The mutation column represents the distinguishing mutation for the resistant or susceptible-dominant allele(s). Abbreviations: R, resistant; S, susceptible; EMB, ethambutol; PAS, para-aminosalicylic acid; INH, isoniazid; PZA, pyrazinamide; RMP, rifampicin; SM, streptomycin; OFX, ofloxacin; ETA, ethionamide; MDR, multidrug resistant; XDR, extensively-drug resistant.

Gene	**Drug**	**Dominant allele**	**Mutation**	**Structural domain feature**
Rv3848	EMB, XDR	R: (25/26)	SNP	Outside transmembrane helical domain
embR	EMB	S: (2/37, 9/129)	SNP	Proximal to DNA-binding domain
Rv3129	EMB	R: (8/11)	SNP	–
proC	EMB	S: (1/27, 11/127)	SNP	–
kdpC	EMB	R: (80/91)	SNP 11	Inside transmembrane helical domain
oxcA	INH	R: (66/66, 26/26)	SNP 253	TPP enzyme M-terminal domain
chp2	ETA	R: (29/37, 34/60)	SNP 296	DELs in mutagen and helical domain
lipD	ETA	R: (48/58, 8/12)	SNP 105	Inside beta-lactamase domain
Rv3471c	ETA, XDR, SM	R: (48/50)	SNP 64	Inside Cupin 1 domain
mmpL11	PAS	R: (35/48)	SNP 520	–
Rv0044c	PAS	R: (13/13)	DEL 137–264	BAC Luciferase
Rv0954	PAS	R: (34/46, 4/6)	SNP 223	Different mutational backgrounds
Rv2560	PZA	S: (6/41)	DEL 1–80	Compositional bias Proline-rich domain
Rv2090	RIF, INH	S: (9/67, 6/46, 5/51)	SNP 295	–
lpqZ	RIF	S: (10/91, 12/79)	SNP 119	Within opuAC signaling domain
Rv1597	RIF, MDR, INH	R: (18/19)	SNP 196	No mutation in methyltransferase domain
Rv1543	RIF, MDR	S: (10/84, 12/80)	SNP 128	Proximal to binding domain
nuoL	MDR, PAS	R: (17/17)	SNP 503	Outside transmembrane helical domain
dnaA	SM	R: (22/22)	SNP 233	Proximal to nucleotide binding domain 213
yajC	SM	R: (30/30)	SNP 87	Within transmembrane helical domain
accD5	OFX, MDR	R: (16/16)	SNP 127	Within CoA carboxyltransferase domain
Rv3041c	RIF, OFX, SM, MDR	R: (20/28, 25/44)	SNP 140	SNP in ATP binding domain
VapC21	XDR	R: (14/23, 14/20)	DEL 88–138	Within second magnesium binding domain

along with their associated antibiotic, key mutation frequency, and structural protein features (Table 2.2).

2.2.6 Geographic stratification of resistant and susceptible alleles provide insight into country-specific adaptations

Since our set of *M. tuberculosis* strains spans multiple continents, we geographically contextualized our set of SVM-derived AMR genes towards delineating possible country-specific adaptations. We observed that resistant and susceptible alleles of the identified AMR genes were stratified amongst specific countries of origin: resistant-dominant alleles were primarily located in Belarus, South Africa, and South Korea, while susceptible alleles were primarily located in India (Table 2.2). The geographic locality of ethambutol, rifampicin, and isoniazid resistant alleles suggests a genetic basis underlying the successful proliferation of *M. tuberculosis* in Belarus—a country with the highest prevalence of multidrug resistant (MDR) strains ever recorded [69]. We observed that the resistant alleles associated with para-aminosalicylic acid (PAS) were based in the high-burden MDR country of South Korea. Since PAS was a key component in the standard MDR treatment regimen of South Korea [70], these alleles may represent specific adaptations to post-MDR PAS treatment that could be leveraged to better optimize the regimen. In total, these results portray a geographic basis for *M. tuberculosis* AMR evolution and demonstrate that our phylogenetically-agnostic machine learning approach is capable of capturing population behavior, which often confounds AMR predictions [71, 72].

2.3 Discussion

The data deluge on *M. tuberculosis* and its AMR characteristics is likely to continue unabated until all *M. tuberculosis* strains isolated from patients will be sequenced with associated metadata to guide clinical management. A reference-agnostic computational platform needs to

be developed to receive, warehouse and continually analyze this data. We have taken the first step at developing a computational platform to meet this challenge. The platform was applied to 1,595 sequenced strains to yield results in four categories: pan-genome properties, identification of genes conferring antibiotic resistance, their epistatic interactions, and protein structure based mechanistic insights.

The pan-genome properties derived by our computational platform reflect the current understanding of *M. tuberculosis* genetic variability. The other three categories of results are intertwined. We recovered 33 known AMR genes and uncovered an additional 24 novel genetic targets. This demonstrates the platform's ability to generate hypotheses that may expand our knowledge of the genetic basis of AMR in *M. tuberculosis* . Some of these new targets are surprising (e.g., Rv3471c) and some are understandable (e.g., oxcA), but all provide an impetus for more detailed experimental studies (Supplementary Note).

The third and fourth categories of results are interconnected and detail intricate features underlying *M. tuberculosis* AMR evolution. The 74 epistatic interactions revealed are new but in many cases involve known gene partners (e.g., ubiA). In other cases, these new epistatic interactions involve novel gene products (e.g., Rv2090). This novelty, reinforced by structural insights, inform a new line of experimental inquiry (Supplementary Note). The larger implications of these intricacies are threefold: (1) genetic background contributes to AMR phenotypic variation, but may be subtle (e.g., embR); (2) high-level resistance mutations are prevalent in off-target genes, such as transmembrane proteins (e.g., Rv3848); and (3) high-level resistance mutations localize to countries with poor *M. tuberculosis* management (i.e., Belarus). These features point to the adverse effects of prolonged treatment [73].

While our framework successfully identifies genetic AMR signatures, there are limita-

tions to our approach that future efforts may expand upon. For one, our platform utilizes prior knowledge of known gene-antibiotic relationships and thus does not provide a means to uniquely deconvolve out an association of a region with a specific drug (Supplementary Note). In addition, while our structural analysis provided a foundation for hypothesizing potential evolutionary drivers, it did not provide further support to the causality of an allele. Novel statistical methods may leverage variations in structural features towards supporting causal alleles. Furthermore, our approach lacks the ability to understand systemic relationships connecting the alleles on a mechanistic level, such as interacting changes in biochemical flux. Future efforts may integrate genome-scale models of pathogens towards elucidating and understanding the genetic signatures of antibiotic resistance [74]

Taken together, the platform presented here meets the pressing need for disparate data type analysis enabled by rapidly growing data available for *M. tuberculosis* pathogenesis and AMR. It both recovers known AMR features (i.e., positive control) and reveals new ones. This platform utilizes a unique combination of pan-genomic analysis, machine learning, structural analysis, and geographic contextualization. These data types are likely to become available for all urgent and serious threat human prokaryotic pathogens in the near future. Similar results to those presented here are thus likely to appear on a pathogen-specific basis in the coming years.

Acknowledgements

E.K, J.M.M, and B.O.P conceived and designed the study. E.K conducted all analysis, with contributions from E.C, N.M, D.H., and J.M.M. E.K., Y.S., and J.M.M performed the pan-genome analysis. E.K. and D.H. performed the epistatic interaction analysis. E.C. and N.M.. developed the 3D protein structural analysis pipeline. E.K., J.T.Y, E.C., N.M., Y.S., N.D., A.A.,

L.Y., D.H., V.N., J.M.M., and B.O.P. provided study oversight, wrote the manuscript, and edited the manuscript. J.M.M and B.O.P managed the study. All authors reviewed and approved the final manuscript. We thank Anand Sastry for helpful discussions regarding machine learning.

This research was supported by the NIAID grant (AI124316), the NIGMS (GM102098), and the Novo Nordisk Foundation Grant Number NNF10CC1016517.

Chapter 2 is a reprint of material published in: **ES Kavvas**, E Catoui, N Mih, JT Yurkovich, Y Seif, N Dillon, D Heckmann, A Anand, L Yang, C Nizet, JM Monk and BO Palsson. 2016. "Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance" *Nature Communications* 9: 4306. The dissertation author is the primary author.

2.4 References

1. admin & Cruz, A. G. *Home* `http://www.crypticproject.org/`. Accessed: 2017-7-3.

2. Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A. R., Will, R., Xia, F. & Stevens, R. Antimicrobial Resistance Prediction in PATRIC and RAST. en. *Scientific reports* **6,** 27930. ISSN: 2045-2322 (June 2016).

3. Manson, A. L., Cohen, K. A., Abeel, T., Desjardins, C. A., Armstrong, D. T., Barry 3rd, C. E., Brand, J., TBResist Global Genome Consortium, Chapman, S. B., Cho, S.-N., Gabrielian, A., Gomez, J., Jodals, A. M., Joloba, M., Jureen, P., Lee, J. S., Malinga, L., Maiga, M., Nordenberg, D., Noroc, E., Romancenco, E., Salazar, A., Ssengooba, W., Velayati, A. A., Winglee, K., Zalutskaya, A., Via, L. E., Cassell, G. H., Dorman, S. E., Ellner, J., Farnia, P., Galagan, J. E., Rosenthal, A., Crudu, V., Homorodean, D., Hsueh, P.-R., Narayanan, S., Pym, A. S., Skrahina, A., Swaminathan, S., Van der Walt, M., Alland, D., Bishai, W. R., Cohen, T., Hoffner, S., Birren, B. W. & Earl, A. M. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. en. *Nature genetics* **49,** 395–402. ISSN: 1061-4036, 1546-1718 (Mar. 2017).

4. Walker, T. M., Kohl, T. A., Omar, S. V., Hedge, J., Del Ojo Elias, C., Bradley, P., Iqbal, Z., Feuerriegel, S., Niehaus, K. E., Wilson, D. J., Clifton, D. A., Kapatai, G., Ip, C. L. C., Bowden, R., Drobniewski, F. A., Allix-Béguec, C., Gaudin, C., Parkhill, J., Diel, R., Supply,

P., Crook, D. W., Smith, E. G., Walker, A. S., Ismail, N., Niemann, S., Peto, T. E. A. & Modernizing Medical Microbiology (MMM) Informatics Group. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. en. *The Lancet infectious diseases* **15,** 1193–1202. ISSN: 1473-3099, 1474-4457 (Oct. 2015).

5. Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., Warren, R. M., Streicher, E. M., Calver, A., Sloutsky, A., Kaur, D., Posey, J. E., Plikaytis, B., Oggioni, M. R., Gardy, J. L., Johnston, J. C., Rodrigues, M., Tang, P. K. C., Kato-Maeda, M., Borowsky, M. L., Muddukrishna, B., Kreiswirth, B. N., Kurepina, N., Galagan, J., Gagneux, S., Birren, B., Rubin, E. J., Lander, E. S., Sabeti, P. C. & Murray, M. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. en. *Nature genetics* **45,** 1183–1189. ISSN: 1061-4036, 1546-1718 (Oct. 2013).

6. Desjardins, C. A., Cohen, K. A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B. J., Shea, T. P., Almeida, D. V., Manson, A. L., Salazar, A., Padayatchi, N., O'Donnell, M. R., Mlisana, K. P., Wortman, J., Birren, B. W., Grosset, J., Earl, A. M. & Pym, A. S. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. en. *Nature genetics* **48,** 544–551. ISSN: 1061-4036, 1546-1718 (May 2016).

7. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. en. *Bioinformatics* **22,** 1658–1659. ISSN: 1367-4803 (July 2006).

8. Zheng, J., Rubin, E. J., Bifani, P., Mathys, V., Lim, V., Au, M., Jang, J., Nam, J., Dick, T., Walker, J. R., Pethe, K. & Camacho, L. R. para-Aminosalicylic acid is a prodrug targeting dihydrofolate reductase in Mycobacterium tuberculosis. en. *The Journal of biological chemistry* **288,** 23447–23456. ISSN: 0021-9258, 1083-351X (Aug. 2013).

9. Moradigaravand, D., Grandjean, L., Martinez, E., Li, H., Zheng, J., Coronel, J., Moore, D., Török, M. E., Sintchenko, V., Huang, H., Javid, B., Parkhill, J., Peacock, S. J. & Köser, C. U. dfrA thyA Double Deletion in para-Aminosalicylic Acid-Resistant Mycobacterium tuberculosis Beijing Strains. en. *Antimicrobial agents and chemotherapy* **60,** 3864–3867. ISSN: 0066-4804, 1098-6596 (June 2016).

10. Martinez, E., Holmes, N., Jelfs, P. & Sintchenko, V. Genome sequencing reveals novel deletions associated with secondary resistance to pyrazinamide in MDR Mycobacterium tuberculosis. en. *The Journal of antimicrobial chemotherapy* **70,** 2511–2514. ISSN: 0305-7453, 1460-2091 (Sept. 2015).

11. Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y. & Sobral, B. W. PATRIC, the bacterial bioinformatics database and analysis resource. en. *Nucleic acids research* **42,** D581–91. ISSN: 0305-1048, 1362-4962 (Jan. 2014).

12. Miyoshi-Akiyama, T., Matsumura, K., Iwai, H., Funatogawa, K. & Kirikae, T. Complete annotated genome sequence of Mycobacterium tuberculosis Erdman. en. *Journal of bacteriology* **194,** 2770. ISSN: 0021-9193, 1098-5530 (May 2012).

13. Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., Supply, P., Kalinowski, J. & Niemann, S. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. en. *PLoS medicine* **10,** e1001387. ISSN: 1549-1277, 1549-1676 (Feb. 2013).

14. Wu, W., Zheng, H., Zhang, L., Wen, Z., Zhang, S., Pei, H., Yu, G., Zhu, Y., Cui, Z., Hu, Z., Wang, H. & Li, Y. A genome-wide analysis of multidrug-resistant and extensively drug-resistant strains of Mycobacterium tuberculosis Beijing genotype. en. *Molecular genetics and genomics: MGG* **288,** 425–436. ISSN: 1617-4615, 1617-4623 (Sept. 2013).

15. Majid, M., Kumar, N., Qureshi, A., Yerra, P., Kumar, A., Kumar, M. K., Tiruvayipati, S., Baddam, R., Shaik, S., Srikantam, A. & Ahmed, N. Genomes of Two Clinical Isolates of Mycobacterium tuberculosis from Odisha, India. en. *Genome announcements* **2.** ISSN: 2169-8287 (Mar. 2014).

16. Ng, K. P., Yew, S. M., Chan, C. L., Chong, J., Tang, S. N., Soo-Hoo, T. S., Na, S. L., Hassan, H., Ngeow, Y. F., Hoh, C. C., Lee, K. W. & Yee, W. Y. Draft Genome Sequence of the First Isolate of Extensively Drug-Resistant (XDR) Mycobacterium tuberculosis in Malaysia. en. *Genome announcements* **1.** ISSN: 2169-8287 (Jan. 2013).

17. Lin, N., Liu, Z., Zhou, J., Wang, S. & Fleming, J. Draft genome sequences of two super-extensively drug-resistant isolates of Mycobacterium tuberculosis from China. en. *FEMS microbiology letters* **347,** 93–96. ISSN: 0378-1097, 1574-6968 (Oct. 2013).

18. Lanzas, F., Karakousis, P. C., Sacchettini, J. C. & Ioerger, T. R. Multidrug-resistant tuberculosis in panama is driven by clonal expansion of a multidrug-resistant Mycobacterium tuberculosis strain related to the KZN extensively drug-resistant M. tuberculosis strain from South Africa. en. *Journal of clinical microbiology* **51,** 3277–3285. ISSN: 0095-1137, 1098-660X (Oct. 2013).

19. Cohen, K. A., Abeel, T., Manson McGuire, A., Desjardins, C. A., Munsamy, V., Shea, T. P., Walker, B. J., Bantubani, N., Almeida, D. V., Alvarado, L., Chapman, S. B., Mvelase, N. R., Duffy, E. Y., Fitzgerald, M. G., Govender, P., Gujja, S., Hamilton, S., Howarth, C., Larimer, J. D., Maharaj, K., Pearson, M. D., Priest, M. E., Zeng, Q., Padayatchi, N., Grosset, J., Young, S. K., Wortman, J., Mlisana, K. P., O'Donnell, M. R., Birren, B. W., Bishai, W. R., Pym, A. S. & Earl, A. M. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. en. *PLoS medicine* **12,** e1001880. ISSN: 1549-1277, 1549-1676 (Sept. 2015).

20. Ismail, A., Teh, L. K., Ngeow, Y. F., Norazmi, M. N., Zainul, Z. F., Tang, T. H., Najimudin, N. & Salleh, M. Z. Draft Genome Sequence of a Clinical Isolate of Mycobacterium tuberculosis Strain PR05. en. *Genome announcements* **1.** ISSN: 2169-8287 (June 2013).

21. Karuthedath Vellarikkal, S., Patowary, A., Singh, M., Periwal, V., Singh, A. V., Singh, P. K., Garg, P., Mohan Katoch, V., Katoch, K., Jangir, P. K., Sharma, R., Open Source Drug Discovery Consortium, Chauhan, D. S., Scaria, V. & Sivasubbu, S. Draft Genome Sequence of a Clinical Isolate of Multidrug-Resistant Mycobacterium tuberculosis East African Indian Strain OSDD271. en. *Genome announcements* **1.** ISSN: 2169-8287 (Aug. 2013).

22. Al Rashdi, A. S. A., Jadhav, B. L., Deshpande, T. & Deshpande, U. Whole-Genome Sequencing and Annotation of a Clinical Isolate of Mycobacterium tuberculosis from Mumbai, India. en. *Genome announcements* **2.** ISSN: 2169-8287 (Mar. 2014).

23. Winglee, K., Manson McGuire, A., Maiga, M., Abeel, T., Shea, T., Desjardins, C. A., Diarra, B., Baya, B., Sanogo, M., Diallo, S., Earl, A. M. & Bishai, W. R. Whole Genome Sequencing of Mycobacterium africanum Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. en. *PLoS neglected tropical diseases* **10,** e0004332. ISSN: 1935-2727, 1935-2735 (Jan. 2016).

24. Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., Blum, M. G. B., Rüsch-Gerdes, S., Mokrousov, I., Aleksic, E., Allix-Béguec, C., Antierens, A., Augustynowicz-Kopeć, E., Ballif, M., Barletta, F., Beck, H. P., Barry 3rd, C. E., Bonnet, M., Borroni, E., Campos-Herrero, I., Cirillo, D., Cox, H., Crowe, S., Crudu, V., Diel, R., Drobniewski, F., Fauville-Dufaux, M., Gagneux, S., Ghebremichael, S., Hanekom, M., Hoffner, S., Jiao, W.-W., Kalon, S., Kohl, T. A., Kontsevaya, I., Lillebæk, T., Maeda, S., Nikolayevskyy, V., Rasmussen, M., Rastogi, N., Samper, S., Sanchez-Padilla, E., Savic, B., Shamputa, I. C., Shen, A., Sng, L.-H., Stakenas, P., Toit, K., Varaine, F., Vukovic, D., Wahl, C., Warren, R., Supply, P., Niemann, S. & Wirth, T. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. en. *Nature genetics* **47,** 242–249. ISSN: 1061-4036, 1546-1718 (Mar. 2015).

25. Isaza, J. P., Duque, C., Gomez, V., Robledo, J., Barrera, L. F. & Alzate, J. F. Whole genome shotgun sequencing of one Colombian clinical isolate of Mycobacterium tuberculosis reveals DosR regulon gene deletions. en. *FEMS microbiology letters* **330,** 113–120. ISSN: 0378-1097, 1574-6968 (May 2012).

26. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry 3rd, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S. & Barrell, B. G. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. en. *Nature* **393,** 537–544. ISSN: 0028-0836 (June 1998).

27. Camus, J.-C., Pryor, M. J., Médigue, C. & Cole, S. T. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. en. *Microbiology* **148,** 2967–2973. ISSN: 0026-2617, 1350-0872 (Oct. 2002).

28. Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., Portugal, I., Pain, A., Martin, N. & Clark, T. G. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. en. *Nature communications* **5,** 4812. ISSN: 2041-1723 (Sept. 2014).

29. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27,** 623–656. ISSN: 0005-8580 (Oct. 1948).

30. Earle, S. G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., Spencer, C. C. A., Iqbal, Z., Clifton, D. A., Hopkins, K. L., Woodford, N., Smith, E. G., Ismail, N., Llewelyn, M. J., Peto, T. E., Crook, D. W., McVean, G., Walker, A. S. & Wilson, D. J. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. en. *Nature microbiology* **1,** 16041. ISSN: 2058-5276 (Apr. 2016).

31. Lees, J. A., Vehkala, M., Välimäki, N., Harris, S. R., Chewapreecha, C., Croucher, N. J., Marttinen, P., Davies, M. R., Steer, A. C., Tong, S. Y. C., Honkela, A., Parkhill, J., Bentley, S. D. & Corander, J. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. en. *Nature communications* **7,** 12797. ISSN: 2041-1723 (Sept. 2016).

32. Jaillard, M., Tournoud, M., Lima, L., Lacroix, V., Veyrieras, J.-B. & Jacob, L. *Representing Genetic Determinants in Bacterial GWAS with Compacted De Bruijn Graphs* en. Mar. 2017.

33. Musser, J. M., Kapur, V., Williams, D. L., Kreiswirth, B. N., Van Soolingen, D. & Van Embden, J. D. A. Characterization of the catalase-peroxidase gene (katG) and inhA locus in isoniazid-resistant and-susceptible strains of Mycobacterium tuberculosis by automated DNA sequencing: restricted array of mutations associated with drug resistance. *The Journal of infectious diseases* **173,** 196–202. ISSN: 0022-1899 (1996).

34. Rozwarski, D. A., Grant, G. A., Barton, D. H., Jacobs Jr, W. R. & Sacchettini, J. C. Modification of the NADH of the isoniazid target (InhA) from Mycobacterium tuberculosis. en. *Science* **279,** 98–102. ISSN: 0036-8075 (Jan. 1998).

35. Torres, J. N., Paul, L. V., Rodwell, T. C., Victor, T. C., Amallraja, A. M., Elghraoui, A., Goodmanson, A. P., Ramirez-Busby, S. M., Chawla, A., Zadorozhny, V., Streicher, E. M., Sirgel, F. A., Catanzaro, D., Rodrigues, C., Gler, M. T., Crudu, V., Catanzaro, A. & Valafar, F. Novel katG mutations causing isoniazid resistance in clinical M. tuberculosis isolates. en. *Emerging microbes & infections* **4,** e42. ISSN: 2222-1751 (July 2015).

36. Taniguchi, H., Aramaki, H., Nikaido, Y., Mizuguchi, Y., Nakamura, M., Koga, T. & Yoshida, S. Rifampicin resistance and mutation of the rpoB gene in Mycobacterium tuberculosis. en. *FEMS microbiology letters* **144,** 103–108. ISSN: 0378-1097 (Oct. 1996).

37. De Vos, M., Müller, B., Borrell, S., Black, P. A., van Helden, P. D., Warren, R. M., Gagneux, S. & Victor, T. C. Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. en. *Antimicrobial agents and chemotherapy* **57,** 827–832. ISSN: 0066-4804, 1098-6596 (Feb. 2013).

38. Louw, G. E., Warren, R. M., Gey van Pittius, N. C., Leon, R., Jimenez, A., Hernandez-Pando, R., McEvoy, C. R. E., Grobbelaar, M., Murray, M., van Helden, P. D. & Victor, T. C. Rifampicin reduces susceptibility to ofloxacin in rifampicin-resistant Mycobacterium tuberculosis through efflux. en. *American journal of respiratory and critical care medicine* **184,** 269–276. ISSN: 1073-449X, 1535-4970 (July 2011).

39. Telenti, A., Philipp, W. J., Sreevatsan, S., Bernasconi, C., Stockbauer, K. E., Wieles, B., Musser, J. M. & Jacobs Jr, W. R. The emb operon, a gene cluster of Mycobacterium tuberculosis involved in resistance to ethambutol. en. *Nature medicine* **3,** 567–570. ISSN: 1078-8956 (May 1997).

40. Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M., McNeil, M., Peterson, S. N., Chatterjee, D., Fleischmann, R., *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-[beta]-D-arabinose biosynthetic and utilization pathway genes. *Nature genetics* **45,** 1190–1197. ISSN: 1061-4036 (2013).

41. Xu, Y., Jia, H., Huang, H., Sun, Z. & Zhang, Z. Mutations Found in embCAB, embR, and ubiA Genes of Ethambutol-Sensitive and -Resistant Mycobacterium tuberculosis Clinical Isolates from China. en. *BioMed research international* **2015,** 951706. ISSN: 2314-6133, 2314-6141 (Aug. 2015).

42. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. en. *Nature medicine* **2,** 662–667. ISSN: 1078-8956 (June 1996).

43. Nair, J., Rouse, D. A., Bai, G.-H. & Morris, S. L. The rpsL gene and streptomycin resistance in single and multiple drug-resistant strains of Mycobacterium tuberculosis. *Molecular microbiology* **10,** 521–527. ISSN: 0950-382X (1993).

44. Wong, S. Y., Lee, J. S., Kwak, H. K., Via, L. E., Boshoff, H. I. M. & Barry 3rd, C. E. Mutations in gidB confer low-level streptomycin resistance in Mycobacterium tuberculosis. en. *Antimicrobial agents and chemotherapy* **55,** 2515–2522. ISSN: 0066-4804, 1098-6596 (June 2011).

45. Von Groll, A., Martin, A., Jureen, P., Hoffner, S., Vandamme, P., Portaels, F., Palomino, J. C. & da Silva, P. A. Fluoroquinolone resistance in Mycobacterium tuberculosis and mutations in gyrA and gyrB. en. *Antimicrobial agents and chemotherapy* **53,** 4498–4500. ISSN: 0066-4804, 1098-6596 (Oct. 2009).

46. Fivian-Hughes, A. S., Houghton, J. & Davis, E. O. Mycobacterium tuberculosis thymidylate synthase gene thyX is essential and potentially bifunctional, while thyA deletion confers

resistance to p-aminosalicylic acid. en. *Microbiology* **158,** 308–318. ISSN: 0026-2617 (Feb. 2012).

47. Morlock, G. P., Metchock, B., Sikes, D., Crawford, J. T. & Cooksey, R. C. ethA, inhA, and katG loci of ethionamide-resistant clinical Mycobacterium tuberculosis isolates. en. *Antimicrobial agents and chemotherapy* **47,** 3799–3805. ISSN: 0066-4804 (Dec. 2003).

48. Wang, F., Sambandan, D., Halder, R., Wang, J., Batt, S. M., Weinrick, B., Ahmad, I., Yang, P., Zhang, Y., Kim, J., Hassani, M., Huszar, S., Trefzer, C., Ma, Z., Kaneko, T., Mdluli, K. E., Franzblau, S., Chatterjee, A. K., Johnsson, K., Johnson, K., Mikusova, K., Besra, G. S., Fütterer, K., Robbins, S. H., Barnes, S. W., Walker, J. R., Jacobs Jr, W. R. & Schultz, P. G. Identification of a small molecule with activity against drug-resistant and persistent tuberculosis. en. *Proceedings of the National Academy of Sciences of the United States of America* **110,** E2510–7. ISSN: 0027-8424, 1091-6490 (July 2013).

49. Nakatani, Y., Opel-Reading, H. K., Merker, M., Machado, D., Andres, S., Kumar, S. S., Moradigaravand, D., Coll, F., Perdigão, J., Portugal, I., Schön, T., Nair, D., Devi, K. R. U., Kohl, T. A., Beckert, P., Clark, T. G., Maphalala, G., Khumalo, D., Diel, R., Klaos, K., Aung, H. L., Cook, G. M., Parkhill, J., Peacock, S. J., Swaminathan, S., Viveiros, M., Niemann, S., Krause, K. L. & Köser, C. U. Role of Alanine Racemase Mutations in Mycobacterium tuberculosis d-Cycloserine Resistance. en. *Antimicrobial agents and chemotherapy* **61.** ISSN: 0066-4804, 1098-6596 (Dec. 2017).

50. Eschenburg, S., Priestman, M. & Schönbrunn, E. Evidence That the Fosfomycin Target Cys115in UDP-N-acetylglucosamine Enolpyruvyl Transferase (MurA) Is Essential for Product Release. *The Journal of biological chemistry* **280,** 3757–3763. ISSN: 0021-9258 (2004).

51. Gopal, P., Yee, M., Sarathy, J., Low, J. L., Sarathy, J. P., Kaya, F., Dartois, V., Gengenbacher, M. & Dick, T. Pyrazinamide Resistance Is Caused by Two Distinct Mechanisms: Prevention of Coenzyme A Depletion and Loss of Virulence Factor Synthesis. en. *ACS infectious diseases* **2,** 616–626. ISSN: 2373-8227 (Sept. 2016).

52. Philalay, J. S., Palermo, C. O., Hauge, K. A., Rustad, T. R. & Cangelosi, G. A. Genes required for intrinsic multidrug resistance in Mycobacterium avium. en. *Antimicrobial agents and chemotherapy* **48,** 3412–3418. ISSN: 0066-4804 (Sept. 2004).

53. Bisson, G. P., Mehaffy, C., Broeckling, C., Prenni, J., Rifat, D., Lun, D. S., Burgos, M., Weissman, D., Karakousis, P. C. & Dobos, K. Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, rpoB mutant Mycobacterium tuberculosis. en. *Journal of bacteriology* **194,** 6441–6452. ISSN: 0021-9193, 1098-5530 (Dec. 2012).

54. Li, G., Zhang, J., Guo, Q., Wei, J., Jiang, Y., Zhao, X., Zhao, L.-L., Liu, Z., Lu, J. & Wan, K. Study of efflux pump gene expression in rifampicin-monoresistant Mycobacterium tuberculosis clinical isolates. en. *The Journal of antibiotics* **68,** 431–435. ISSN: 0021-8820 (July 2015).

55. Jang, J., Kim, R., Woo, M., Jeong, J., Park, D. E., Kim, G. & Delorme, V. Efflux attenuates the anti-bacterial activity of Q203 in Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy.* ISSN: 0066-4804 (Apr. 2017).

56. Vilchèze, C., Av-Gay, Y., Attarian, R., Liu, Z., Hazbón, M. H., Colangeli, R., Chen, B., Liu, W., Alland, D., Sacchettini, J. C. & Jacobs Jr, W. R. Mycothiol biosynthesis is essential for ethionamide susceptibility in Mycobacterium tuberculosis. en. *Molecular microbiology* **69,** 1316–1329. ISSN: 0950-382X, 1365-2958 (Sept. 2008).

57. Li, X.-Z., Elkins, C. A. & Zgurskaya, H. I. *Efflux-Mediated Antimicrobial Resistance in Bacteria: Mechanisms, Regulation and Clinical Implications* en. ISBN: 9783319396583 (Springer, Nov. 2016).

58. Danilchanka, O., Mailaender, C. & Niederweis, M. Identification of a novel multidrug efflux pump of Mycobacterium tuberculosis. en. *Antimicrobial agents and chemotherapy* **52,** 2503–2511. ISSN: 0066-4804, 1098-6596 (July 2008).

59. Xu, W., DeJesus, M. A., Rücker, N., Engelhart, C. A., Wright, M. G., Healy, C., Lin, K., Wang, R., Park, S. W., Ioerger, T. R., Schnappinger, D. & Ehrt, S. Chemical Genetic Interaction Profiling Reveals Determinants of Intrinsic Antibiotic Resistance in Mycobacterium tuberculosis. en. *Antimicrobial agents and chemotherapy* **61.** ISSN: 0066-4804, 1098-6596 (Dec. 2017).

60. Brossier, F., Sougakoff, W., Bernard, C., Petrou, M., Adeyema, K., Pham, A., Amy de la Breteque, D., Vallet, M., Jarlier, V., Sola, C. & Veziris, N. Molecular Analysis of the embCAB Locus and embR Gene Involved in Ethambutol Resistance in Clinical Isolates of Mycobacterium tuberculosis in France. en. *Antimicrobial agents and chemotherapy* **59,** 4800–4808. ISSN: 0066-4804, 1098-6596 (Aug. 2015).

61. Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J. M., Zhang, Z. & Palsson, B. O. *ssbio: A Python Framework for Structural Systems Biology* en. July 2017.

62. Rozwarski, D. A., Grant, G. A., Barton, D. H., Jacobs Jr, W. R. & Sacchettini, J. C. Modification of the NADH of the isoniazid target (InhA) from Mycobacterium tuberculosis. en. *Science* **279,** 98–102. ISSN: 0036-8075 (Jan. 1998).

63. Rawat, R., Whitty, A. & Tonge, P. J. The isoniazid-NAD adduct is a slow, tight-binding inhibitor of InhA, the Mycobacterium tuberculosis enoyl reductase: adduct affinity and drug resistance. en. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 13881–13886. ISSN: 0027-8424 (Nov. 2003).

64. Sharma, K., Gupta, M., Pathak, M., Gupta, N., Koul, A., Sarangi, S., Baweja, R. & Singh, Y. Transcriptional control of the mycobacterial embCAB operon by PknH through a regulatory protein, EmbR, in vivo. en. *Journal of bacteriology* **188,** 2936–2944. ISSN: 0021-9193 (Apr. 2006).

65. Werther, T., Zimmer, A., Wille, G., Golbik, R., Weiss, M. S. & König, S. New insights into structure-function relationships of oxalyl CoA decarboxylase from Escherichia coli. en. *The FEBS journal* **277,** 2628–2640. ISSN: 1742-464X, 1742-4658 (June 2010).

66. Puckett, S., Trujillo, C., Wang, Z., Eoh, H., Ioerger, T. R., Krieger, I., Sacchettini, J., Schnappinger, D., Rhee, K. Y. & Ehrt, S. Glyoxylate detoxification is an essential function of malate synthase required for carbon assimilation in Mycobacterium tuberculosis. en. *Proceedings of the National Academy of Sciences of the United States of America* **114,** E2225–E2232. ISSN: 0027-8424, 1091-6490 (Mar. 2017).

67. Beste, D. J. V., Bonde, B., Hawkins, N., Ward, J. L., Beale, M. H., Noack, S., Nöh, K., Kruger, N. J., Ratcliffe, R. G. & McFadden, J. ^{13}C metabolic flux analysis identifies an unusual route for pyruvate dissimilation in mycobacteria which requires isocitrate lyase and carbon dioxide fixation. en. *PLoS pathogens* **7,** e1002091. ISSN: 1553-7366, 1553-7374 (July 2011).

68. Nandakumar, M., Nathan, C. & Rhee, K. Y. Isocitrate lyase mediates broad antibiotic tolerance in Mycobacterium tuberculosis. en. *Nature communications* **5,** 4306. ISSN: 2041-1723 (June 2014).

69. Skrahina, A., Hurevich, H., Zalutskaya, A., Sahalchyk, E., Astrauko, A., van Gemert, W., Hoffner, S., Rusovich, V. & Zignol, M. Alarming levels of drug-resistant tuberculosis in Belarus: results of a survey in Minsk. en. *The European respiratory journal: official journal of the European Society for Clinical Respiratory Physiology* **39,** 1425–1431. ISSN: 0903-1936, 1399-3003 (June 2012).

70. Park, J. S. Issues Related to the Updated 2014 Korean Guidelines for Tuberculosis. en. *Tuberculosis and respiratory diseases* **79,** 1–4. ISSN: 1738-3536 (Jan. 2016).

71. Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. en. *Nature reviews. Genetics* **18,** 41–50. ISSN: 1471-0056, 1471-0064 (Jan. 2017).

72. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. en. *Current opinion in microbiology* **25,** 17–24. ISSN: 1369-5274, 1879-0364 (June 2015).

73. Gagneux, S., Long, C. D., Small, P. M., Van, T., Schoolnik, G. K. & Bohannan, B. J. M. The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. en. *Science* **312,** 1944–1946. ISSN: 0036-8075, 1095-9203 (June 2006).

74. Kavvas, E. S., Seif, Y., Yurkovich, J. T., Norsigian, C., Poudel, S., Greenwald, W. W., Ghatak, S., Palsson, B. O. & Monk, J. M. Updated and standardized genome-scale reconstruction of Mycobacterium tuberculosis H37Rv, iEK1011, simulates flux states indicative of physiological conditions. *BMC systems biology* **12,** 25. ISSN: 1752-0509 (Mar. 2018).

Chapter 3

An updated genome-scale model of *M. tuberculosis* H37Rv metabolism

The efficacy of antibiotics against *M. tuberculosis* has been shown to be influenced by experimental media conditions. Investigations of *M. tuberculosis* growth in physiological conditions have described an environment that is different from common in vitro media. Thus, elucidating the interplay between available nutrient sources and antibiotic efficacy has clear medical relevance. While genome-scale reconstructions of *M. tuberculosis* have enabled the ability to interrogate media differences for the past 10 years, recent reconstructions have diverged from each other without standardization. A unified reconstruction of *M. tuberculosis* H37Rv would elucidate the impact of different nutrient conditions on antibiotic efficacy and provide new insights for therapeutic intervention.

We present a new genome-scale model of *M. tuberculosis* H37Rv, named iEK1011, that unifies and updates previous *M. tuberculosis* H37Rv genome-scale reconstructions. We func-

tionally assess iEK1011 against previous models and show that the model increases correct gene essentiality predictions on two different experimental datasets by 6% (53% to 60%) and 18% (60% to 71%), respectively. We compared simulations between in vitro and approximated in vivo media conditions to examine the predictive capabilities of iEK1011. The simulated differences recapitulated literature defined characteristics in the rewiring of TCA metabolism including succinate secretion, gluconeogenesis, and activation of both the glyoxylate shunt and the methylcitrate cycle. To assist efforts to elucidate mechanisms of antibiotic resistance development, we curated 16 metabolic genes related to antimicrobial resistance and approximated evolutionary drivers of resistance. Comparing simulations of these antibiotic resistance features between in vivo and in vitro media highlighted condition-dependent differences that may influence the efficacy of antibiotics.

3.1 Background

The success of *M. tuberculosis* as a pathogen has been partially attributed to its unique metabolic capabilities. The metabolic network of *M. tuberculosis* has evolved to withstand and navigate the harsh environment imposed by the alveolar macrophage. Most bacteria cannot thrive in this hypoxic, acidic and nutrient-limited condition, yet it is in this harsh environment that *M. tuberculosis* encounters and evolves resistance to antibiotics. Elucidating the robust properties of metabolism that enable *M. tuberculosis* pathogenicity and drug resistance evolution has become a key area of research.

Recent studies have demonstrated that the choice of experimental media conditions plays an important role in understanding physiologically-relevant phenotypes of *M. tuberculosis* [1]. Commonly used experimental media conditions such as Middlebrook 7H9 are known to be much

different from the physiological environment. For example, despite extensive research describing fatty acids as a key carbon source within the macrophage environment, most studies forgo the inclusion of fatty acids in the media, opting instead for glucose or glycerol [2]. Perhaps it is no surprise then that differences between the in vitro and in vivo environments have been shown to affect antibiotic screening results [3–11]. In particular, it has been shown that hypoxic or nutrient limited conditions alter the metabolism of *M. tuberculosis* to a nonreplicating, drug-resistant state [5–7]. Specific mechanism-changing effects between in vitro and in vivo conditions have been shown to occur for many antibiotics [9, 10].

While it is understood that differences in experimental media conditions lead to phenotypic variations, dissecting a mechanistic understanding of these different phenotypes remains challenging. Genome-scale models (GEMs) of metabolisms have emerged as powerful tools to computationally probe the effect of media composition on a cell's phenotype [12]. For the past 10 years, GEMs have provided a mechanistic basis for exploring the metabolic capabilities of *M. tuberculosis* on the systems-level. GEMs of *M. tuberculosis* have helped interrogate a variety of biological phenomena, from understanding the transcriptional regulatory network [13] to elucidating metabolic interactions between *M. tuberculosis* and the alveolar macrophage [14].

While new *M. tuberculosis* H37Rv GEMs have enabled novel insights, they have been constructed from different base models resulting in divergent representations of the metabolic network. For example, gene-protein-reaction rules (GPRs) (i.e., the Boolean relationship between a gene, or set of genes, and the corresponding reaction(s)) differ within reactions shared amongst models (e.g., The GPR of *Rv0904c* differs between iOSDD and iSM810). In addition to variation in network topology, divergent GEMs have a variety of identifiers used for metabolite and reaction names, making them difficult to compare and build from (e.g., "R" reaction identi-

fier nomenclature used in most models built primarily off of GSMN-TB). While such differences may seem insignificant, the presence of multiple divergent *M. tuberculosis* H37Rv reconstructions hinders progress and may result in future wasted efforts [15].

Here, we present iEK1011, a new GEM of *M. tuberculosis* H37Rv that unifies, standardizes, and updates previous divergent GEMs of this model organism. We assess the performance of iEK1011 to that of previous GEMs through gene essentiality prediction on two different datasets. iEK1011 is further characterized by performing simulations that examine the model's predictions in physiological conditions and interrogate differences between in vitro and in vivo media conditions. Finally, in order to provide a comprehensive platform for elucidating antibiotic resistance (AMR), we integrate knowledge derived from experimental literature into iEK1011.

3.2 Results

3.2.1 Workflow for updating, unifying, and standardizing previous reconstructions of *M. tuberculosis*

In order to ensure a comprehensive unification, we first gathered and compared available reconstructions of *M. tuberculosis* H37Rv. Since the first two *M. tuberculosis* H37Rv reconstructions released in 2007 [16, 17], a total of 9 reconstructions have been built (Table 3.1). Most models were largely based off of either iNJ661 [16] or GSMN-TB [17]. Specifically, out of the most recent *M. tuberculosis* reconstructions, sMtb [18], iSM810 [13] and gal2015 [19] were primarily built from GSMN-TB while iOSDD [20] was built from iNJ661.

Using a variety of both quantitative and qualitative criteria (e.g., standardized identifiers, gene essentiality predictions, mass balanced reactions; see Methods), iOSDD and sMtb were cho-

Table 3.1: Summary of existing genome-scale models of *M. tuberculosis* . iAB-AMØ-1410-Mt-661 has over 2000 genes because it combines an updated version of iNJ661 with a macrophage model. The model provided by Garay et al. was given the name of gal2015 because it is unnamed in the original publication.

Model	**Year**	**Genes**	**Reactions**	**Metabolites**	**Reference**
iNJ661	2007	661	1025	826	[16]
GSMN-TB	2007	726	856	645	[17]
MMF-RmwBo	2009	776	1108	???	[21]
HQMTB	2009	686	607	734	[22]
iNJ661v	2010	663	1049	838	[23]
iAB-AMØ-1410-Mt-661	2010	2071	4489	3400	[14]
MergedTBmodel	2012	917	1400	1017	[24]
GSMN-TB1.1	2013	759	876	667	[25]
iOSDD890	2014	890	1152	961	[20]
sMtb	2014	915	1192	929	[18]
gal2015	2015	760	965	754	[19]
iSM810	2015	810	938	723	[13]
iNJ661mu	2016	672	1057	846	[26]
iEK1011	2017	1011	1228	998	This study

sen as the base reconstructions for the unification process (Figure 3.1A). The recently developed *M. tuberculosis* H37Rv BioCyc Database [27] provided an additional reconstruction resource to supplement the standardized draft model. The reconstruction process was performed following a clear workflow (Figure 3.1A): the base models were mapped to standardized BiGG identifiers [28], joined into a draft model of shared reactions and unified by assessing model disagreements. The resulting unified draft model was then expanded through manual curation of new biochemical knowledge. Thus, the reconstruction process was iterative and involved constant re-evaluation of model content (see Methods).

The resulting unified and updated reconstruction of *M. tuberculosis* H37Rv, named iEK1011, contains 1011 genes, 1228 reactions, and 998 metabolites. iEK1011 encapsulates the majority of genes in the previous models based on either iNJ661 or GSMN-TB (Figure 3.1B). iEK1011 accounts for 96% of sMtb genes (874 of 915 genes) and 91% of iOSDD genes (807 of

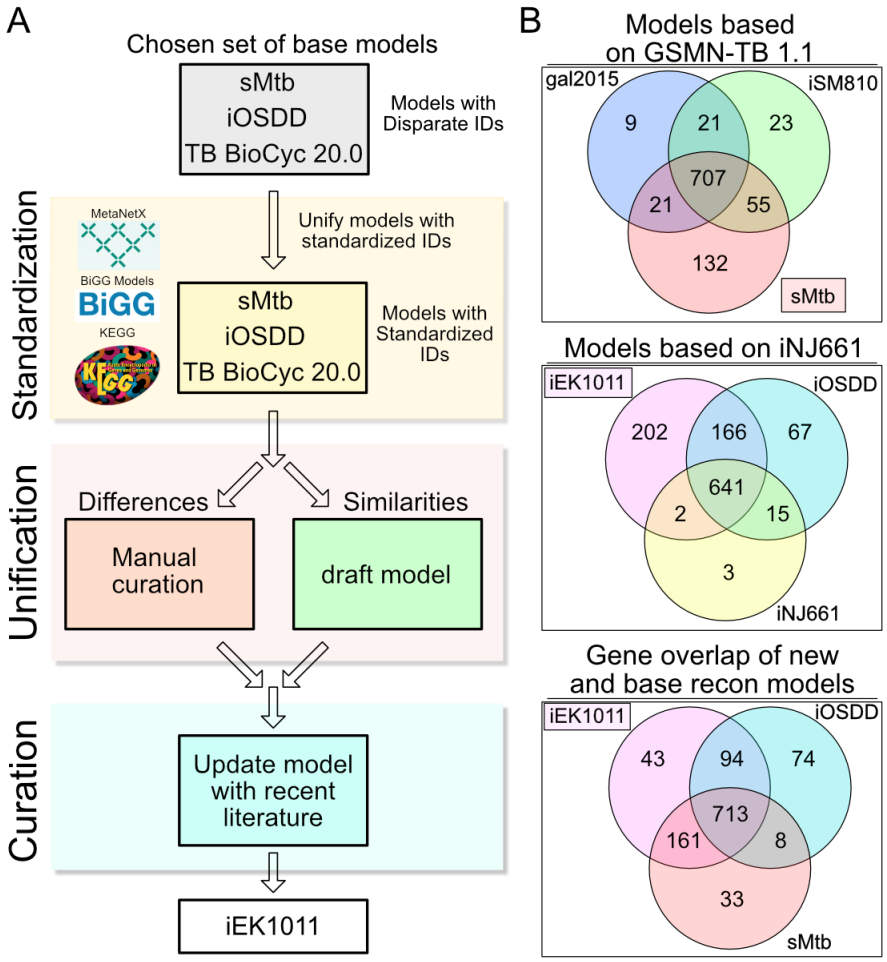

Figure 3.1: (A) Workflow of reconstruction process. A draft GEM model was built from the TB BioCyc 20.0 database and mapped to BIGGs IDs along with sMtb and iOSDD. The models were then unified by first joining the similarities between them, followed by manual curation of model differences literature and database validation. (B) Overlap of genes across different model sets. The model that covers most of the models within the particular set is enclosed by a box.

890). A total of 151 unique genes from iOSDD, iNJ661, gal2015, iSM810, and sMtb were not accounted for in iEK1011 (see Additional File 2) either due to insufficient evidence necessary to resolve major inconsistencies across models or lack of confidence in gene annotation.

In addition to unifying previous reconstructions, iEK1011 incorporated 39 new genes absent from previous models. In particular, sulfur metabolism was updated by adding the cysO-dependent biosynthesis of L-cysteine, which connects molybdenum metabolism with sulfur

metabolism through the use of moeZ in both pathways [29]. New pathways and reactions include heme uptake [30], tuberculosinol biosynthesis [31], ergothioneine biosynthesis [32, 33], and mycobilin biosynthesis [34]. The resulting unified reconstruction of *M. tuberculosis* , iEK1011, provides a biochemically-derived knowledge-base that can be functionally assessed computationally.

3.2.2 Functional assessment of iEK1011

iEK1011 was converted to a mathematical model to examine the functional capabilities of the improved reconstruction and to quantitatively compare it with previous reconstructions. The primary tool for evaluating genome-scale reconstructions of *M. tuberculosis* H37Rv has been in silico gene essentiality testing. Therefore, we used gene essentiality as a metric for evaluating and comparing the performance of iEK1011. Gene essentiality predictions across previous *M. tuberculosis* H37Rv reconstructions were determined using the same data and quantitative score used in evaluating the predictive ability of iSM810 [13]. In addition to the commonly used gene essentiality dataset by Griffin et al. [35], a recent gene essentiality dataset by DeJesus et al. [36] was also utilized in our model comparisons. The primary differentiating feature between the datasets was the media condition used to generate them (see Additional File 2). Using these gene-essentiality datasets, we evaluated and compared the ability of five models (iNJ661, iOSDD, sMtb, iSM810, and iEK1011) to predict gene essentiality.

When using the Griffin dataset, we found that iEK1011 increases the prediction of true positives (i.e., the model correctly predicts growth for the gene knockout when the gene is annotated as non-essential) by 23% (579) (Figure 3.2A) over sMtb (470), which had the largest number of true positives amongst the previous models. iEK1011 gene essentiality predictions decrease the number of false negatives (i.e., the model incorrectly predicts no growth for the

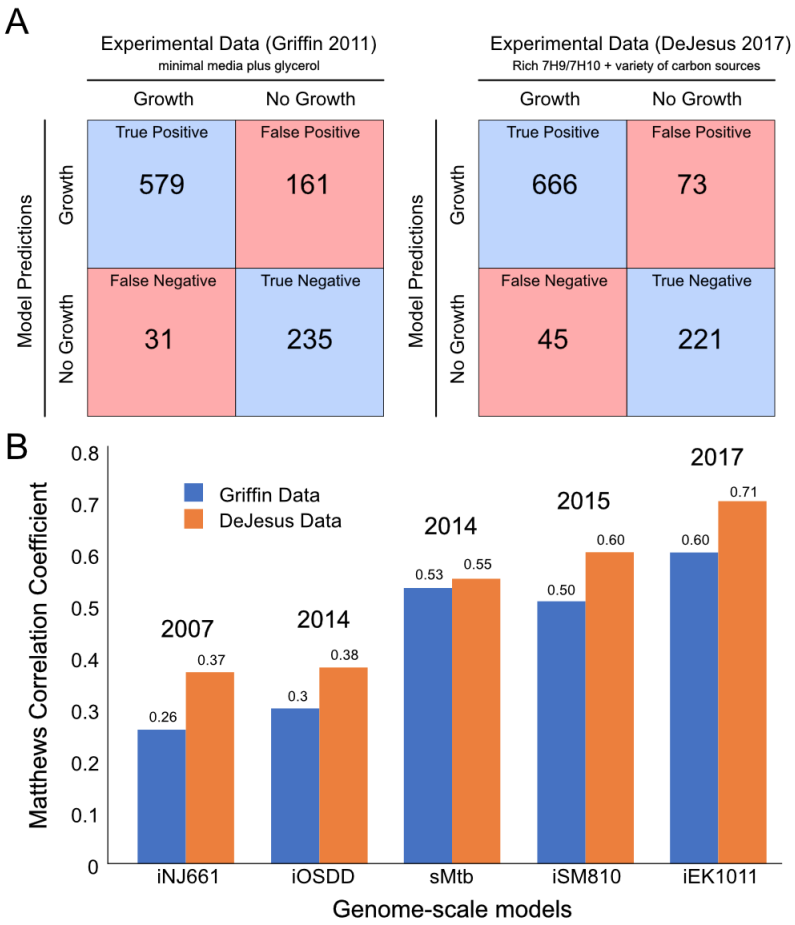

Figure 3.2: Gene Essentiality Prediction Comparisons. (A) Model-predicted gene essentiality results compared to both the Griffin et al. and deJesus et al. essentiality experimental datasets. (B) Gene essentiality performance using the Matthews Correlation Coefficient. iSM810 and sMtb, which were both built off of GSMN-TB 1.1, significantly outperform iNJ661 and iOSDD. iEK1011 outperforms all models on both gene essentiality datasets.

gene knockout when the gene is annotated as non-essential) by 11.4% (31) (Figure 3.2A) over iSM810 (35), which had the least number of false negatives amongst the previous models.

With respect to the more recent DeJesus essentiality dataset, iEK1011 increases the number of true positives by 24% (666) (Figure 3.2A) over sMtb (538), and iEK1011 increases the number of true negatives by 11% (221) (Figure 3.2A) over sMtb (199). iEK1011 decreases the number of false positives by 14% (73) over sMtb (83), and increases the number of false negatives

by one (45) over iSM810 (44). The increase in one more false positive over iSM810 is due to having 9 genes tested in the false negative category that are not contained in iSM810. Moreover, relating specific groups, such as false negatives or true positives, against other models with a different number of genes may not correctly represent the changes due to significant differences in class sizes.

In order to account for the variations in class sizes amongst models, we calculated the Matthews Correlation Coefficient (MCC) for each model's prediction on both gene essentiality datasets (Figure 3.2B). iEK1011 scores the highest on both datasets with an MCC of 0.60 and 0.71 on the Griffin and DeJesus dataset, respectively (Figure 3.2B) (see Additional File 2). These iEK1011 MCC values are a 6% and 18% increase over the previous best model MCC's of sMtb and iSM810 on the Griffin and DeJesus dataset, respectively.

Although the DeJesus essentiality dataset is more recent than the Griffin dataset by 6 years, the media condition used in determining essentiality on the DeJesus dataset was not as well defined because it utilized oleic-albumin-dextrose-catalase (OADC) in middlebrook 7H10/7H9 media supplemented with a variety of carbon sources [36]. The contents of OADC are not well defined primarily because of albumin, which may supplement amino acids to *M. tuberculosis* . The extent of OADC's impact remains unknown, which ultimately hinders the ability to rigorously define the inputs for GEMs, which are crucial components of COBRA methods [37]. Conversely, the media used in Griffin was well defined as minimal media supplemented with glycerol [35]. Therefore, the increase in MCC by 6% over sMtb on the Griffin essentiality data should be evaluated with more confidence than the significantly higher percent increase in MCC over all models on the DeJesus dataset. Thus, the gene essentiality results presented above demonstrate improved predictive capability of iEK1011 over previous *M. tuberculosis* GEMs.

3.2.3 iEK1011 qualitatively recapitulates flux states indicative of physiologically relevant media conditions

While the gene essentiality predictions are a useful metric to evaluate model quality, we prioritized the model's ability to recapitulate *M. tuberculosis* behavior described in the literature. Specifically, an emphasis was placed on central carbon metabolism given its distinctive usage in *M. tuberculosis* and recent emergence as an unexpected research frontier [38]. In addition, we focused on *M. tuberculosis* studies involving conditions relevant to pathogenicity [1]. Therefore, we compared simulations between two conditions relevant to the purpose of this study: Lowenstein-Jensen media, representing in vitro drug testing conditions; and an in vivo nutrient condition approximated from the literature that attempts to replicate the pathogenic state. We used Flux Variability Analysis (FVA) [39] and randomized sampling [40] to characterize and compare the fluxes between the two media conditions.

Taking advantage of recent studies investigating nitrogen metabolism within the context of *M. tuberculosis* pathogenicity [41–44], we set the in vivo nitrogen sources to be composed of nitrate, aspartate, asparagine, glutamate, urea, and glutamine ((Figure 3.3), see Additional File 2). Under hypoxic in vivo conditions, iEK1011 predicts use of nitrate in a respiratory role as opposed to a nitrogen source where it is taken in and reduced to nitrite by narG, and then exported out of the cell, a finding consistent with previous experiments [43]. The chosen in vivo carbon sources include fatty acids (both even and odd chain), cholesterol, CO2, and Alanine. Fatty acids were chosen as the primary source of carbon in vivo due to the vast amount of literature evidence supporting the claim that *M. tuberculosis* uses host-derived fatty acids [2, 45, 46]. iEK1011 catabolizes fatty acids through beta-oxidation, which generates acetate (even chain fatty acid catabolism), propionyl-CoA (odd chain fatty acid catabolism) and acetyl-

CoA (Figure 3.3). Although CO2 was incorporated due to evidence showing it being fixated by *M. tuberculosis* in an approximated in vivo environment [44], iEK1011 was not predicted to fixate CO2 due to a net gluconeogenic flux through phosphoenolpyruvate carboxykinase - a simulation result also found in Beste et al. [44]. Alanine was included as a nutrient due to evidence describing it to be in abundant quantities within the alveolar macrophage and being imported from the macrophage [44].

The differences in flux state simulations predicted by iEK1011 between the two conditions recapitulate key behavior described in the literature. Specifically, in the approximated in vivo condition involving hypoxia and growth on fatty acids, model-predicted flux decreases through TCA with an accompanying increase in succinate secretion (Figure 3.3). iEK1011 predicts the secretion of succinate to allow optimal growth in these conditions because it removes an intracellular proton, allowing for membrane potential related reactions such as oxidative phosphorylation to proceed. This mechanism has been previously described to be specific and essential in *M. tuberculosis* hypoxia adaptation [47]. Thus, iEK1011 can recapitulate known physiological phenomena using stoichiometry alone.

In addition to succinate secretion, iEK1011 simulates the activation of both the glyoxylate shunt and the methylcitrate cycle in response to both hypoxia and growth on fatty acids [47, 48]. Although the median flux values are low (Figure 3.3) (based on markov chain monte carlo sampling of the solution space [40]), FVA simulations show maximum flux values through methylcitrate cycle and glyoxylate shunt to have a threefold and twofold increase in in vivo media conditions relative to in vitro conditions, respectively (see Additional File 2). Furthermore, the metabolic model does not account for the toxic effect of glyoxylate and propionate which has been shown to necessitate flux through glyoxylate shunt and methylcitrate cycle. While iEK1011

simulations do not account for characteristics like toxicity, the examples outlined above show that iEK1011 is capable of qualitatively recapitulating key phenomena uncovered in recent years.

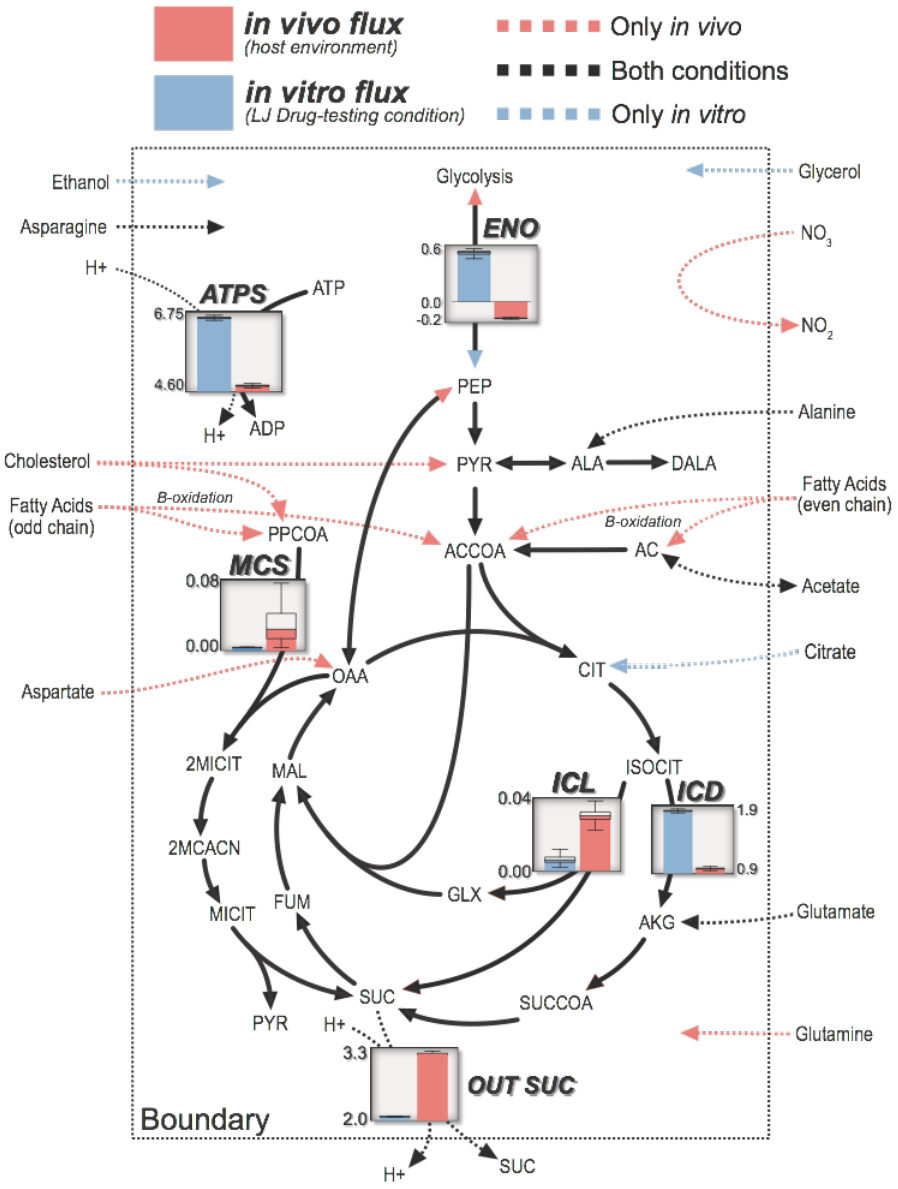

Figure 3.3: Metabolic map of flux differences through central carbon metabolism in iEK1011 between approximate in vitro and in vivo conditions. The media conditions are represented by nutrients outside of the dotted boundary line.

Table 3.2: Table of antibiotics and the associated genes whose mutations confer antibiotic resistance.

Drug	Gene	iEK1011 reaction	Reference
Ethambutol	*embABC*	EMB	[49]
	ubiA	DCPT	[50]
	aftA	AFTA	[50]
D-cycloserine	*alr*	ALAR	[51]
	ddl	ALAALAR	[52]
	ald	ALAD_L, GXRA	[53]
Isoniazid	*katG*	CAT	[54]
	inhA	FAS	[55]
	fabG1	MYCSacp56/58/50	[56]
Benzothiazinones	*dprE1*	DCPE	[57]
PAS	*thyA*	TMDS	[58]
	ribD	FOLR2, ASPRAUR, DHPPDA2	[58]
	folC	DHFS, THFGLUS	[58]
Pyrazinamide	*pncA*	NNAM	[59]
Ethionamide	*mshC*	CIGAMS	[60]
Rifampicin	*drrABC*	PDIMAT, PPDIMAT	[61]

3.2.4 iEK1011 as a computational knowledge base for interrogating features of antibiotic resistance

We have shown that iEK1011 is a valuable source of computational inquiry through gene essentiality predictions and its ability to recapitulate phenomena described in the literature. In addition to providing a computational platform, GEMs are fundamentally a knowledge-base that are capable of contextualizing a variety of concepts that extend beyond the genome-scale metabolic network [37]. Taking advantage of this ability to incorporate abstractions, we translate knowledge derived from experimental investigations of antibiotic resistance (AMR) evolution into a format that can be integrated into GEMs.

Using the extensive literature on the mechanism of AMR evolution in *M. tuberculosis* , we curated a relational table between antibiotics, genes, and metabolic reactions for eight different antibiotics (Table 3.2). The genes associated with a particular antibiotic are those known to be

central to AMR evolution (i.e., mutations in the genes that code for the reactions often confer resistance to specific drugs). Displaying AMR genes on a metabolic map of iEK1011 portrays relationships that would be difficult to comprehend without a GEM (Figure 3.4). Notably, we found that the close topological relationships between para-aminosalicylic acid, ethambutol, D-cycloserine, and pyrazinamide may hint at pleiotropic effects (i.e., mutations that affect multiple phenotypes) of resistance conferring mutations on the efficacy of different antibiotics.

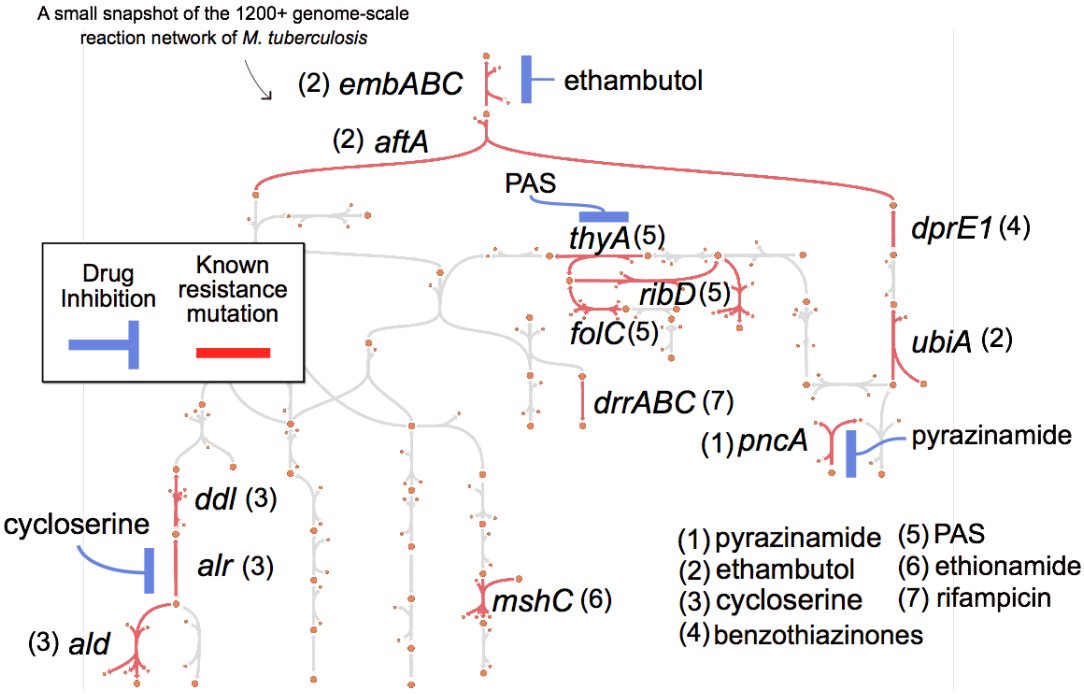

Figure 3.4: Escher map of arabinogalactan-peptidoglycan complex biosynthesis with known resistance-conferring genes mapped. The gene-antibiotic relation is indicated by the number placed proximal to the gene. The mechanistic effect by the antibiotic is indicated by the blue line. No blue line is shown for mutations in which the gene-antibiotic relation remains unclear (i.e., mshC, drrBC), Escher-usable maps were built for multiple subsystems in iEK1011 (see Additional File 4).

In order to incorporate specific antibiotic pressures into iEK1011, we evaluated each antibiotic and associated a biochemical objective function that approximates the evolutionary drivers of selection (Table 3.3). In the case of ethambutol, it has been shown that flux-increasing

Table 3.3: List of objective functions related to the evolutionary drivers of antibiotic resistance. The abbreviations are as follows: PAS (para-aminosalicylic acid), MAX (maximize), MIN (minimize).

Drug	Objective	Reaction in iEK1011	Reference
Ethambutol	MAX DPA production	decda__tb_c $\rightarrow \phi$	[50]
PAS	MAX Tetrahydrofolate production	thf_c $\rightarrow \phi$	[58]
D-cycloserine	MAX L-Alanine Production	ala__L_c $\rightarrow \phi$	[53]
Ethionamide	MIN mycothiol production	msh_c $\rightarrow \phi$	[62]

mutations in ubiA confer resistance by increasing the production of decaprenylphosphoryl-b-D-arabinose (DPA), which outcompetes ethambutol for embB binding spots [50]. Therefore, in a GEM, the evolutionary pressure imposed by ethambutol can be approximated as a metabolic objective where the production of DPA is maximized (Table 3.3). A total of four antibiotics were associated with approximated objective functions representing evolutionary forces (see Methods for further reasoning of the choice of objective function).

Taking advantage of the translation of antibiotic features to formats amenable by iEK1011, we simulated the evolutionary pressures induced by antibiotics and calculated the maximum and minimum fluxes for the AMR-associated reactions in both in vivo and in vitro conditions through FVA.

There were few differences in relative flux for a specific drug objective between these conditions. However, those that were uncovered highlighted potential impacts of environmental/media composition differences. In particular, we see major differences in the fluxes that correspond to optimizing the approximated ethambutol-induced evolutionary pressure (Figure 3.5). Furthermore, this ethambutol flux is correlated with fluxes induced by the approximated d-cycloserine objective. Closer inspection of the uptake differences driving these differential flux states points to L-alanine as a key environmental influence. In particular, the differential fluxes within the cases of ethambutol resistance-conferring genes ubiA (DCPT) and embB (EMB), as well as d-

cycloserine resistance-conferring genes of alr (ALAR), ald (ALAD_L), and ddlA (ALAALAr), exemplify the differential effect of environmental L-alanine availability. Notably, L-alanine has been shown to be an important substrate in the macrophage environment (Beste 2013). While L-alanine and other amino acids may be available in LJ drug-testing media due to utilization of egg base or bovine serum, our analysis only accounted for metabolites that were explicitly stated in defined quantities within the media conditions. With respect to the efficacy of antibiotics, these results suggest that d-cycloserine and ethambutol may be less effective in vivo due to increased availability of L-alanine, which is a key precursor reaction catalyzed by AMR genes targeted by d-cycloserine and ethambutol, whereas in vitro conditions may increase susceptibility to ethambutol. In both cases, the significant decrease in model-predicted maximum ALAD_L (ald) flux is in line with studies describing the deleterious mutations in ald that confer resistance to D-cycloserine [53]. Altogether, iEK1011 provides a knowledge base for relating antibiotic resistance features through genome-scale metabolic network analysis.

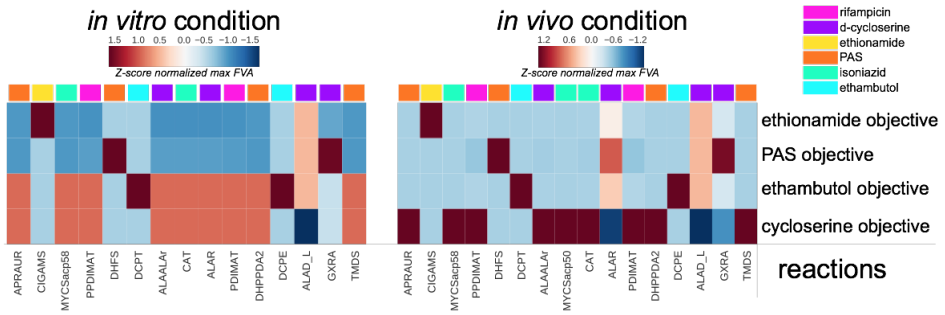

Figure 3.5: Heatmaps of maximum FVA values for for a matrix representing FVA values for the curated AMR reactions across simulations of different drug-specific objective functions (see (Table 3.2) for curated list of AMR genes and their associated iEK1011 reactions, see (Table 3.3) for drug-specific objectives).

3.3 Discussion

The divergence of *M. tuberculosis* H37Rv reconstructions has created an unnecessary obstacle in contextualizing the increasing growth of biochemical data for this troublesome pathogen. In order to address experimental insights to pathogenic conditions and alleviate roadblocks for future reconstruction efforts of *M. tuberculosis* H37Rv, we built a unified and updated GEM of *M. tuberculosis* , iEK1011. We tested the predictive potential of iEK1011 by comparing gene essentiality predictions with previous models and showed that iEK1011 outperforms previous models. We further assessed the predictive capabilities of iEK1011 by comparing simulated flux states between in vitro drug testing and approximate in vivo media conditions. Comparisons recapitulate specific phenomena indicative of biochemical flux states seen in physiological conditions. We incorporated antibiotic resistance knowledge in iEK1011, which enabled a network-based perspective of multi-antibiotic resistance evolution.

iEK1011 unified previous *M. tuberculosis* H37Rv reconstructions and encompassed the majority of genes within the two divergent groups of reconstructions. Additionally, iEK1011 incorporates new pathways including the incorporation of ergothioneine biosynthesis. This addition will aid a quantitative elucidation of the relationships between sulfur metabolism, bioenergetic homeostasis, and redox balance [33]. As a unified, standardized, and updated model, iEK1011 provides a base for future models of *M. tuberculosis* H37Rv.

Functional assessment of previous *M. tuberculosis* H37Rv reconstructions through gene essentiality predictions showed that iEK1011 achieves a higher MCC than previous models on two different datasets. While the two datasets were crucial in both assessing and driving iEK1011 reconstruction, experimental gene essentiality datasets derived from physiologically-relevant conditions are warranted for understanding the human-restricted lifestyle of *M. tuberculosis* .

Using iEK1011, we qualitatively determined differences in biochemical states between in vitro and approximated in vivo conditions. We showed that iEK1011 successfully recapitulates specific phenomena described in physiologically-relevant studies of *M. tuberculosis* . Future reconstruction efforts may target iEK1011's lack of predicted CO2 fixation [44] and account for compartmentalized co-metabolism of multiple substrates [63, 64]. iEK1011 may provide a base for future host-pathogen integrated reconstructions that leverage valuable experimental data.

An integrated knowledge-base of genome-scale metabolism and antibiotic resistance components may enable new perspectives for understanding and combating *M. tuberculosis* H37Rv. We translated experimental knowledge of AMR genes and specific adaptation mechanisms to formats amenable to iEK1011. Comparing simulations of these AMR features between in vitro and in vivo conditions emphasized the potential impact of hypoxia and L-alanine availability on the pressures induced by antibiotics. Future constraint-based analysis of *M. tuberculosis* AMR may leverage new experimental approaches, such as those that have analyzed changes in essentiality under antibiotic exposure [65].

Taken together, iEK1011 is a new, comprehensive and predictive constraint-based model of *M. tuberculosis* H37Rv. In this study, we computationally demonstrate that in vivo nutrient sources absent from in vitro media significantly alter the flux state of central carbon metabolism. As experimental insights to *M. tuberculosis* pathogenicity and antibiotic resistance continue to grow, this GEM will provide a foundation to connect disparate data types and knowledge.

Acknowledgements

EK reconstructed the network, performed the analysis, and drafted the manuscript. BOP and JM conceived the study and revised the manuscript. YS, JTY, CN, SP, SG, and BG revised

the manuscript.

This research was supported by the NIH NIAID grant (1-UO1-AI124316-01).

Chapter 3 is a reprint of material published in: **ES Kavvas**, Y Seif, JT Yurkovich, C Norsigian, S Poudel, WW Greenwald, S Ghatak, BO Palsson, and JM Monk. "Updated and standardized genome-scale reconstruction of Mycobacterium tuberculosis H37Rv, iEK1011, simulates flux states indicative of physiological conditions." *BMC Systems Biology* 12: 25. The dissertation author is the primary author.

3.4 References

1. Cumming, B. M. & Steyn, A. J. C. Metabolic plasticity of central carbon metabolism protects mycobacteria. en. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 13135–13136. ISSN: 0027-8424, 1091-6490 (Oct. 2015).

2. Lee, W., VanderVen, B. C., Fahey, R. J. & Russell, D. G. Intracellular Mycobacterium tuberculosis exploits host-derived fatty acids to limit metabolic stress. en. *The Journal of biological chemistry* **288,** 6788–6800. ISSN: 0021-9258, 1083-351X (Mar. 2013).

3. Sakoulas, G., Okumura, C. Y., Thienphrapa, W., Olson, J., Nonejuie, P., Dam, Q., Dhand, A., Pogliano, J., Yeaman, M. R., Hensler, M. E., Bayer, A. S. & Nizet, V. Nafcillin enhances innate immune-mediated killing of methicillin-resistant Staphylococcus aureus. en. *Journal of molecular medicine* **92,** 139–149. ISSN: 0946-2716, 1432-1440 (Feb. 2014).

4. Russell, D. G., Barry, C. E. & Flynn, J. L. Tuberculosis: what we don't know can, and does, hurt us. *Science* **328,** 852–856. ISSN: 0036-8075 (2010).

5. Wang, F., Sambandan, D., Halder, R., Wang, J., Batt, S. M., Weinrick, B., Ahmad, I., Yang, P., Zhang, Y., Kim, J., Hassani, M., Huszar, S., Trefzer, C., Ma, Z., Kaneko, T., Mdluli, K. E., Franzblau, S., Chatterjee, A. K., Johnsson, K., Johnson, K., Mikusova, K., Besra, G. S., Fütterer, K., Robbins, S. H., Barnes, S. W., Walker, J. R., Jacobs Jr, W. R. & Schultz, P. G. Identification of a small molecule with activity against drug-resistant and persistent tuberculosis. en. *Proceedings of the National Academy of Sciences of the United States of America* **110,** E2510–7. ISSN: 0027-8424, 1091-6490 (July 2013).

6. Wayne, L. G. & Hayes, L. G. An in vitro model for sequential study of shiftdown of Mycobacterium tuberculosis through two stages of nonreplicating persistence. en. *Infection and immunity* **64,** 2062–2069. ISSN: 0019-9567 (June 1996).

7. Wayne, L. G. & Sohaskey, C. D. Nonreplicating persistence of Mycobacterium tuberculosis 1. *Annual Reviews in Microbiology* **55,** 139–163 (2001).

8. Mitchison, D. A. & Coates, A. R. M. Predictive in vitro models of the sterilizing activity of anti-tuberculosis drugs. en. *Current pharmaceutical design* **10,** 3285–3295. ISSN: 1381-6128 (2004).

9. Zhang, Y. & Mitchison, D. The curious characteristics of pyrazinamide: a review. en. *The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease* **7,** 6–21. ISSN: 1027-3719 (Jan. 2003).

10. Prosser, G. A. & de Carvalho, L. P. S. Metabolomics Reveal d-Alanine:d-Alanine Ligase As the Target of d-Cycloserine in Mycobacterium tuberculosis. en. *ACS medicinal chemistry letters* **4,** 1233–1237. ISSN: 1948-5875 (Dec. 2013).

11. Pethe, K., Sequeira, P. C., Agarwalla, S., Rhee, K., Kuhen, K., Phong, W. Y., Patel, V., Beer, D., Walker, J. R., Duraiswamy, J., Jiricek, J., Keller, T. H., Chatterjee, A., Tan, M. P., Ujjini, M., Rao, S. P. S., Camacho, L., Bifani, P., Mak, P. A., Ma, I., Barnes, S. W., Chen, Z., Plouffe, D., Thayalan, P., Ng, S. H., Au, M., Lee, B. H., Tan, B. H., Ravindran, S., Nanjundappa, M., Lin, X., Goh, A., Lakshminarayana, S. B., Shoen, C., Cynamon, M., Kreiswirth, B., Dartois, V., Peters, E. C., Glynne, R., Brenner, S. & Dick, T. A chemical genetic screen in Mycobacterium tuberculosis identifies carbon-source-dependent growth inhibitors devoid of in vivo efficacy. en. *Nature communications* **1,** 57. ISSN: 2041-1723 (Aug. 2010).

12. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. en. *Cell* **161,** 971–987. ISSN: 0092-8674, 1097-4172 (May 2015).

13. Ma, S., Minch, K. J., Rustad, T. R., Hobbs, S., Zhou, S.-L., Sherman, D. R. & Price, N. D. Integrated Modeling of Gene Regulatory and Metabolic Networks in Mycobacterium tuberculosis. en. *PLoS computational biology* **11,** e1004543. ISSN: 1553-734X, 1553-7358 (Nov. 2015).

14. Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø. & Jamshidi, N. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. en. *Molecular systems biology* **6,** 422. ISSN: 1744-4292 (Oct. 2010).

15. Monk, J., Nogales, J. & Palsson, B. O. Optimizing genome-scale network reconstructions. en. *Nature biotechnology* **32,** 447–452. ISSN: 1087-0156, 1546-1696 (May 2014).

16. Jamshidi, N. & Palsson, B. Ø. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. en. *BMC systems biology* **1,** 26. ISSN: 1752-0509 (June 2007).

17. Beste, D. J. V., Hooper, T., Stewart, G., Bonde, B., Avignone-Rossa, C., Bushell, M. E., Wheeler, P., Klamt, S., Kierzek, A. M. & McFadden, J. GSMN-TB: a web-based genome-

scale network model of Mycobacterium tuberculosis metabolism. en. *Genome biology* **8,** R89. ISSN: 1465-6906 (2007).

18. Rienksma, R. A., Suarez-Diez, M., Spina, L., Schaap, P. J. & Martins dos Santos, V. A. P. Systems-level modeling of mycobacterial metabolism for the identification of new (multi-)drug targets. *Seminars in immunology* **26,** 610–622. ISSN: 1044-5323 (Dec. 2014).

19. Garay, C. D., Dreyfuss, J. M. & Galagan, J. E. Metabolic modeling predicts metabolite changes in Mycobacterium tuberculosis. en. *BMC systems biology* **9,** 57. ISSN: 1752-0509 (Sept. 2015).

20. Vashisht, R., Bhat, A. G., Kushwaha, S., Bhardwaj, A., OSDD Consortium & Brahmachari, S. K. Systems level mapping of metabolic complexity in Mycobacterium tuberculosis to identify high-value drug targets. en. *Journal of translational medicine* **12,** 263. ISSN: 1479-5876 (Oct. 2014).

21. Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., Cheng, T.-Y., Moody, D. B., Murray, M. & Galagan, J. E. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. en. *PLoS computational biology* **5,** e1000489. ISSN: 1553-734X, 1553-7358 (Aug. 2009).

22. Kalapanulak, S. High quality genome-scale metabolic network reconstruction of mycobacterium tuberculosis and comparison with human metabolic network: application for drug ... (2009).

23. Fang, X., Wallqvist, A. & Reifman, J. Development and analysis of an in vivo-compatible metabolic network of Mycobacterium tuberculosis. en. *BMC systems biology* **4,** 160. ISSN: 1752-0509 (Nov. 2010).

24. Chindelevitch, L., Stanley, S., Hung, D., Regev, A. & Berger, B. MetaMerge: scaling up genome-scale metabolic reconstructions with application to Mycobacterium tuberculosis. en. *Genome biology* **13,** r6. ISSN: 1465-6906 (Jan. 2012).

25. Lofthouse, E. K., Wheeler, P. R., Beste, D. J. V., Khatri, B. L., Wu, H., Mendum, T. A., Kierzek, A. M. & McFadden, J. Systems-based approaches to probing metabolic variation within the Mycobacterium tuberculosis complex. en. *PloS one* **8,** e75913. ISSN: 1932-6203 (Sept. 2013).

26. Puniya, B. L., Kulshreshtha, D., Mittal, I., Mobeen, A. & Ramachandran, S. Corrigendum: Integration of Metabolic Modeling with Gene Co-expression Reveals Transcriptionally Programmed Reactions Explaining Robustness in Mycobacterium tuberculosis. en. *Scientific reports* **6,** 24916. ISSN: 2045-2322 (Apr. 2016).

27. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S. & Karp, P. D. The MetaCyc database of metabolic pathways and enzymes

and the BioCyc collection of pathway/genome databases. en. *Nucleic acids research* **44,** D471–80. ISSN: 0305-1048, 1362-4962 (Jan. 2016).

28. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. en. *Nucleic acids research* **44,** D515–22. ISSN: 0305-1048, 1362-4962 (Jan. 2016).

29. Hatzios, S. K. & Bertozzi, C. R. The regulation of sulfur metabolism in Mycobacterium tuberculosis. en. *PLoS pathogens* **7,** e1002036. ISSN: 1553-7366, 1553-7374 (July 2011).

30. Tullius, M. V., Harmston, C. A., Owens, C. P., Chim, N., Morse, R. P., McMath, L. M., Iniguez, A., Kimmey, J. M., Sawaya, M. R., Whitelegge, J. P., Horwitz, M. A. & Goulding, C. W. Discovery and characterization of a unique mycobacterial heme acquisition system. en. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 5051–5056. ISSN: 0027-8424, 1091-6490 (Mar. 2011).

31. Layre, E., Lee, H. J., Young, D. C., Martinot, A. J., Buter, J., Minnaard, A. J., Annand, J. W., Fortune, S. M., Snider, B. B., Matsunaga, I., Rubin, E. J., Alber, T. & Moody, D. B. Molecular profiling of Mycobacterium tuberculosis identifies tuberculosinyl nucleoside products of the virulence-associated enzyme Rv3378c. en. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 2978–2983. ISSN: 0027-8424, 1091-6490 (Feb. 2014).

32. Richard-Greenblatt, M., Bach, H., Adamson, J., Peña-Diaz, S., Li, W., Steyn, A. J. C. & Av-Gay, Y. Regulation of Ergothioneine Biosynthesis and Its Effect on Mycobacterium tuberculosis Growth and Infectivity. en. *The Journal of biological chemistry* **290,** 23064–23076. ISSN: 0021-9258, 1083-351X (Sept. 2015).

33. Saini, V., Cumming, B. M., Guidry, L., Lamprecht, D. A., Adamson, J. H., Reddy, V. P., Chinta, K. C., Mazorodze, J. H., Glasgow, J. N., Richard-Greenblatt, M., Gomez-Velasco, A., Bach, H., Av-Gay, Y., Eoh, H., Rhee, K. & Steyn, A. J. C. Ergothioneine Maintains Redox and Bioenergetic Homeostasis Essential for Drug Susceptibility and Virulence of Mycobacterium tuberculosis. en. *Cell reports* **14,** 572–585. ISSN: 2211-1247 (Jan. 2016).

34. Nambu, S., Matsui, T., Goulding, C. W., Takahashi, S. & Ikeda-Saito, M. A new way to degrade heme: the Mycobacterium tuberculosis enzyme MhuD catalyzes heme degradation without generating CO. en. *The Journal of biological chemistry* **288,** 10101–10109. ISSN: 0021-9258, 1083-351X (Apr. 2013).

35. Griffin, J. E., Gawronski, J. D., Dejesus, M. A., Ioerger, T. R., Akerley, B. J. & Sassetti, C. M. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. en. *PLoS pathogens* **7,** e1002251. ISSN: 1553-7366, 1553-7374 (Sept. 2011).

36. DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., Rubin, E. J., Schnappinger, D., Ehrt, S., Fortune, S. M., Sassetti, C. M. & Ioerger, T. R. Comprehensive

Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. en. *mBio* **8.** ISSN: 2150-7511 (Jan. 2017).

37. Palsson, B. Ø. *Systems Biology: Constraint-based Reconstruction and Analysis* en. ISBN: 9781316239940 (Cambridge University Press, Jan. 2015).

38. Rhee, K. Y., de Carvalho, L. P. S., Bryk, R., Ehrt, S., Marrero, J., Park, S. W., Schnappinger, D., Venugopal, A. & Nathan, C. Central carbon metabolism in Mycobacterium tuberculosis: an unexpected frontier. en. *Trends in microbiology* **19,** 307–314. ISSN: 0966-842X, 1878-4380 (July 2011).

39. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. en. *Metabolic engineering* **5,** 264–276. ISSN: 1096-7176 (Oct. 2003).

40. Megchelenbrink, W., Huynen, M. & Marchiori, E. optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. en. *PloS one* **9,** e86587. ISSN: 1932-6203 (Feb. 2014).

41. Gouzy, A., Larrouy-Maumus, G., Wu, T.-D., Peixoto, A., Levillain, F., Lugo-Villarino, G., Guerquin-Kern, J.-L., Gerquin-Kern, J.-L., de Carvalho, L. P. S., Poquet, Y. & Neyrolles, O. Mycobacterium tuberculosis nitrogen assimilation and host colonization require aspartate. en. *Nature chemical biology* **9,** 674–676. ISSN: 1552-4450, 1552-4469 (Nov. 2013).

42. Gouzy, A., Larrouy-Maumus, G., Bottai, D., Levillain, F., Dumas, A., Wallach, J. B., Caire-Brandli, I., de Chastellier, C., Wu, T.-D., Poincloux, R., Brosch, R., Guerquin-Kern, J.-L., Schnappinger, D., Sório de Carvalho, L. P., Poquet, Y. & Neyrolles, O. Mycobacterium tuberculosis exploits asparagine to assimilate nitrogen and resist acid stress during infection. en. *PLoS pathogens* **10,** e1003928. ISSN: 1553-7366, 1553-7374 (Feb. 2014).

43. Gouzy, A., Poquet, Y. & Neyrolles, O. Nitrogen metabolism in Mycobacterium tuberculosis physiology and virulence. en. *Nature reviews. Microbiology* **12,** 729–737. ISSN: 1740-1526, 1740-1534 (Nov. 2014).

44. Beste, D. J. V., Nöh, K., Niedenführ, S., Mendum, T. A., Hawkins, N. D., Ward, J. L., Beale, M. H., Wiechert, W. & McFadden, J. 13C-flux spectral analysis of host-pathogen metabolism reveals a mixed diet for intracellular Mycobacterium tuberculosis. en. *Chemistry & biology* **20,** 1012–1021. ISSN: 1074-5521, 1879-1301 (Aug. 2013).

45. Daniel, J., Maamar, H., Deb, C., Sirakova, T. D. & Kolattukudy, P. E. Mycobacterium tuberculosis uses host triacylglycerol to accumulate lipid droplets and acquires a dormancy-like phenotype in lipid-loaded macrophages. en. *PLoS pathogens* **7,** e1002093. ISSN: 1553-7366, 1553-7374 (June 2011).

46. Muñoz-Elías, E. J. & McKinney, J. D. Carbon metabolism of intracellular bacteria. en. *Cellular microbiology* **8,** 10–22. ISSN: 1462-5814 (Jan. 2006).

47. Eoh, H. & Rhee, K. Y. Multifunctional essentiality of succinate metabolism in adaptation to hypoxia in Mycobacterium tuberculosis. en. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 6554–6559. ISSN: 0027-8424, 1091-6490 (Apr. 2013).

48. Eoh, H. & Rhee, K. Y. Methylcitrate cycle defines the bactericidal essentiality of isocitrate lyase for survival of Mycobacterium tuberculosis on fatty acids. en. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 4976–4981. ISSN: 0027-8424, 1091-6490 (Apr. 2014).

49. Sreevatsan, S., Stockbauer, K. E., Pan, X., Kreiswirth, B. N., Moghazeh, S. L., Jacobs Jr, W. R., Telenti, A. & Musser, J. M. Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations. en. *Antimicrobial agents and chemotherapy* **41,** 1677–1681. ISSN: 0066-4804 (Aug. 1997).

50. Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M., McNeil, M., Peterson, S. N., Chatterjee, D., Fleischmann, R., *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-[beta]-D-arabinose biosynthetic and utilization pathway genes. *Nature genetics* **45,** 1190–1197. ISSN: 1061-4036 (2013).

51. Caceres, N. E., Harris, N. B., Wellehan, J. F., Feng, Z., Kapur, V. & Barletta, R. G. Overexpression of the D-alanine racemase gene confers resistance to D-cycloserine in Mycobacterium smegmatis. *Journal of bacteriology* **179,** 5046–5055. ISSN: 0021-9193 (1997).

52. Neuhaus, F. C. & Lynch, J. L. THE ENZYMATIC SYNTHESIS OF D-ALANYL-D-ALANINE. 3. ON THE INHIBITION OF D-ALANYL-D-ALANINE SYNTHETASE BY THE ANTIBIOTIC D-CYCLOSERINE. en. *Biochemistry* **3,** 471–480. ISSN: 0006-2960 (Apr. 1964).

53. Desjardins, C. A., Cohen, K. A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B. J., Shea, T. P., Almeida, D. V., Manson, A. L., Salazar, A., Padayatchi, N., O'Donnell, M. R., Mlisana, K. P., Wortman, J., Birren, B. W., Grosset, J., Earl, A. M. & Pym, A. S. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. en. *Nature genetics* **48,** 544–551. ISSN: 1061-4036, 1546-1718 (May 2016).

54. Musser, J. M., Kapur, V., Williams, D. L., Kreiswirth, B. N., Van Soolingen, D. & Van Embden, J. D. A. Characterization of the catalase-peroxidase gene (katG) and inhA locus in isoniazid-resistant and-susceptible strains of Mycobacterium tuberculosis by automated DNA sequencing: restricted array of mutations associated with drug resistance. *The Journal of infectious diseases* **173,** 196–202. ISSN: 0022-1899 (1996).

55. Rozwarski, D. A., Grant, G. A., Barton, D. H., Jacobs Jr, W. R. & Sacchettini, J. C. Modification of the NADH of the isoniazid target (InhA) from Mycobacterium tuberculosis. en. *Science* **279,** 98–102. ISSN: 0036-8075 (Jan. 1998).

56. Torres, J. N., Paul, L. V., Rodwell, T. C., Victor, T. C., Amallraja, A. M., Elghraoui, A., Goodmanson, A. P., Ramirez-Busby, S. M., Chawla, A., Zadorozhny, V., Streicher, E. M., Sirgel, F. A., Catanzaro, D., Rodrigues, C., Gler, M. T., Crudu, V., Catanzaro, A. & Valafar, F. Novel katG mutations causing isoniazid resistance in clinical M. tuberculosis isolates. en. *Emerging microbes & infections* **4,** e42. ISSN: 2222-1751 (July 2015).

57. Makarov, V., Manina, G., Mikusova, K., Möllmann, U., Ryabova, O., Saint-Joanis, B., Dhar, N., Pasca, M. R., Buroni, S., Lucarelli, A. P., Milano, A., De Rossi, E., Belanova, M., Bobovska, A., Dianiskova, P., Kordulakova, J., Sala, C., Fullam, E., Schneider, P., McKinney, J. D., Brodin, P., Christophe, T., Waddell, S., Butcher, P., Albrethsen, J., Rosenkrands, I., Brosch, R., Nandi, V., Bharath, S., Gaonkar, S., Shandil, R. K., Balasubramanian, V., Balganesh, T., Tyagi, S., Grosset, J., Riccardi, G. & Cole, S. T. Benzothiazinones kill Mycobacterium tuberculosis by blocking arabinan synthesis. en. *Science* **324,** 801–804. ISSN: 0036-8075, 1095-9203 (May 2009).

58. Zheng, J., Rubin, E. J., Bifani, P., Mathys, V., Lim, V., Au, M., Jang, J., Nam, J., Dick, T., Walker, J. R., Pethe, K. & Camacho, L. R. para-Aminosalicylic acid is a prodrug targeting dihydrofolate reductase in Mycobacterium tuberculosis. en. *The Journal of biological chemistry* **288,** 23447–23456. ISSN: 0021-9258, 1083-351X (Aug. 2013).

59. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. en. *Nature medicine* **2,** 662–667. ISSN: 1078-8956 (June 1996).

60. Vilchèze, C., Av-Gay, Y., Attarian, R., Liu, Z., Hazbón, M. H., Colangeli, R., Chen, B., Liu, W., Alland, D., Sacchettini, J. C. & Jacobs Jr, W. R. Mycothiol biosynthesis is essential for ethionamide susceptibility in Mycobacterium tuberculosis. en. *Molecular microbiology* **69,** 1316–1329. ISSN: 0950-382X, 1365-2958 (Sept. 2008).

61. Li, G., Zhang, J., Guo, Q., Wei, J., Jiang, Y., Zhao, X., Zhao, L.-L., Liu, Z., Lu, J. & Wan, K. Study of efflux pump gene expression in rifampicin-monoresistant Mycobacterium tuberculosis clinical isolates. en. *The Journal of antibiotics* **68,** 431–435. ISSN: 0021-8820 (July 2015).

62. Vilchèze, C., Av-Gay, Y., Attarian, R., Liu, Z., Hazbón, M. H., Colangeli, R., Chen, B., Liu, W., Alland, D., Sacchettini, J. C. & Jacobs Jr, W. R. Mycothiol biosynthesis is essential for ethionamide susceptibility in Mycobacterium tuberculosis. en. *Molecular microbiology* **69,** 1316–1329. ISSN: 0950-382X, 1365-2958 (Sept. 2008).

63. De Carvalho, L. P. S., Fischer, S. M., Marrero, J., Nathan, C., Ehrt, S. & Rhee, K. Y. Metabolomics of Mycobacterium tuberculosis reveals compartmentalized co-catabolism of carbon substrates. en. *Chemistry & biology* **17,** 1122–1131. ISSN: 1074-5521, 1879-1301 (Oct. 2010).

64. Zimmermann, M., Kogadeeva, M., Gengenbacher, M., McEwen, G., Mollenkopf, H.-J., Zamboni, N., Kaufmann, S. H. E. & Sauer, U. Integration of Metabolomics and Transcriptomics

Reveals a Complex Diet of Mycobacterium tuberculosis during Early Macrophage Infection. en. *mSystems* **2.** ISSN: 2379-5077 (July 2017).

65. Xu, W., DeJesus, M. A., Rücker, N., Engelhart, C. A., Wright, M. G., Healy, C., Lin, K., Wang, R., Park, S. W., Ioerger, T. R., Schnappinger, D. & Ehrt, S. Chemical Genetic Interaction Profiling Reveals Determinants of Intrinsic Antibiotic Resistance in Mycobacterium tuberculosis. en. *Antimicrobial agents and chemotherapy* **61.** ISSN: 0066-4804, 1098-6596 (Dec. 2017).

Chapter 4

A biochemically-interpretable machine learning classifier for microbial GWAS

Current machine learning classifiers have successfully been applied to whole-genome sequencing data to identify genetic determinants of antimicrobial resistance (AMR), but they lack causal interpretation. Here we present a metabolic model-based machine learning classifier, named Metabolic Allele Classifier (MAC), that uses flux balance analysis to predict binary phenotypes of microbial genomes. We apply the MAC to a dataset of 1,595 drug-tested Mycobacterium tuberculosis strains and show that MACs achieve prediction accuracy on par with mechanism-agnostic machine learning models (isoniazid AUC=0.93) while enabling a biochemical interpretation of the genotype-phenotype map. Interpretation of MACs for three antibiotics (pyrazinamide, para-aminosalicylic acid, and isoniazid) recapitulates known AMR mechanisms

and suggest a biochemical basis for how the identified alleles cause AMR. Extending flux balance analysis to identify accurate sequence classifiers thus contributes mechanistic insights to GWAS, a field thus far dominated by mechanism-agnostic results.

4.1 Background

Mycobacterium tuberculosis (TB) claims 1.6 million lives annually and resists eradication through evolution of antimicrobial resistance (AMR) [1]. To elucidate AMR mechanisms, researchers have applied machine learning approaches to large-scale genome sequencing and drug-testing datasets for identifying genetic determinants of AMR [2–7]. While current machine learning approaches have provided a predictive tool for microbial genome-wide association studies (GWAS), such "black-box" models are incapable of mechanistically interpreting genetic associations. Such a limitation has become increasingly apparent in TB, where numerous experimental studies have shown that AMR-associated genetic variants often reflect network-level metabolic adaptations to antibiotic-induced selection pressures (Supplementary Figure 1, Supplementary Table 1) [8–12]. These studies show that identified genetic associations have corresponding network-level associations that are highly informative of AMR mechanisms. However, current GWAS results only provide predictions for which alleles are most important, not their functional effects. Therefore, machine learning models that incorporate biochemical network structure may naturally extend GWAS results by estimating functional effects of identified alleles, leading to an enhanced understanding of AMR [13–15].

Over the past couple of decades, the computational analysis of biochemical networks in microorganisms has been advanced through the use of genome-scale models (GEMs) [16] [17]. By computing metabolic flux states (see Glossary for definition of terms) consistent with imposed

biological constraints, GEMs have been shown to predict a range of cellular functions, making them a valuable tool for analyzing multi-omics datasets [18]. Although GEMs are transparent genotype-phenotype models, they are largely outperformed by machine learning models in direct comparisons of prediction accuracy. Approaches have thus been developed that integrate meaningful GEM computations with predictive "black-box" machine learning to enable "white-box" interpretations of data [19]. These approaches have worked well for endogenous metabolomics data by using the GEM to directly transform the measurements to meaningful inputs for "black box" machine learning.

This approach, however, may not be amenable to analyzing microbial GWAS data, in which the genetic parameters of the GEM are not directly observed (see Supplementary Notes). GEMs have previously modeled genetic variation at the resolution of gene presence-absence [20–23], but have not yet been used to link nucleotide-level genetic variation (i.e., alleles) to observed phenotypes (i.e., AMR) in a predictive manner [24]. Since alleles are the primary forms of causal variation identified in GWAS, an approach for mechanistically integrating information about alleles is of major interest [25].

Here we develop a GEM-based machine learning framework for modeling datasets used in GWAS and apply it to a sequencing dataset of drug-tested TB strains. We show that our framework achieves high performance in accurately classifying AMR phenotypes of TB strains. We then characterize the identified classifiers for pyrazinamide, isoniazid, and para-aminosalicylic acid AMR and show that they identify key genetic determinants and pathway activity discriminating between resistant and susceptible TB strains. This work demonstrates how GEMs can be used directly as an input-output machine learning model to extract both genetic and biochemical network-level insights from microbial GWAS datasets.

4.2 Results

4.2.1 Assessing genes implicated in AMR mechanisms motivates the use of a genome-scale metabolic model for data analysis

We first set out to assess the scope of a potential mechanism-based genotype-phenotype map using a dataset of 1,595 drug-tested TB strains [2, 26] and a GEM of TB H37Rv, named iEK1011 [27]. The acquired genetic variant matrix (G) of the 1,595 strains describes 3,739 protein-coding genes and their 12,762 allelic variants, where each variant is defined as a unique amino acid sequence for the protein coding gene. Our analysis therefore does not account for synonymous amino acid changes and intergentic genetic variants. The corresponding drug susceptibility status for a strain is described by a binary 'susceptibility' or 'resistance' phenotype to a particular antibiotic. iEK1011 accounts for 1,011 genes (26% of H37Rv) and comprises a metabolic network of 1,229 reactions and 998 metabolites.

Comparing the gene list between iEK1011 and the genomics dataset, we found that 26% (981/3,739) of the total genes and 25% (3,310/12,762) of the total variants described by the genetic variant matrix were accounted for by the GEM. To evaluate iEK1011's potential to model causal variants, we compiled a list of AMR genes and compared this list to the gene list of iEK1011 (Supplementary File 1; Methods). We found that 72% (32/44) of known AMR genes are accounted for in iEK1011 (Supplementary Table 1). In the case of six drugs (ethambutol, isoniazid, d-cycloserine, para-aminosalicylic acid, ethionamide, and pyrazinamide), 87% (20/23) of their AMR genes were accounted for in iEK1011. AMR genes not explicitly accounted for in iEK1011 were primarily related to DNA transcription (e.g., rpoB) and transcriptional regulation (e.g., embR). The antibiotics rifampicin, ofloxacin, and streptomycin do not have AMR genes

accounted for in iEK1011 and are therefore out of scope for our study. Taken together, the abundance of AMR genes accounted for in iEK1011 motivated a GEM-driven analysis of the TB AMR dataset.

4.2.2 A metabolic model-based framework for classifying microbial genomes

While we have shown that a GEM accounts for the majority of known genetic determinants of AMR in TB, computational methods do not exist for integrating a fine-grained description of allelic variation with GEMs to directly predict binary phenotypes (i.e., AMR susceptible/resistant classification). We thus set out to develop a GEM-based machine learning framework for analyzing the TB dataset. The developed method, named Metabolic Allele Classifier (MAC), takes the genome sequence of a particular TB strain as its input and classifies strains as either resistant or susceptible to a specific antibiotic (Figure 4.1a). Specifically, the MAC is an allele-parameterized form of flux balance analysis [28, 29] that represents a strain as a set of allele-specific flux capacity constraints and classifies AMR according to the optimum value attained by optimizing an antibiotic-specific objective.

We formulate the MAC within the flux balance analysis framework as follows,

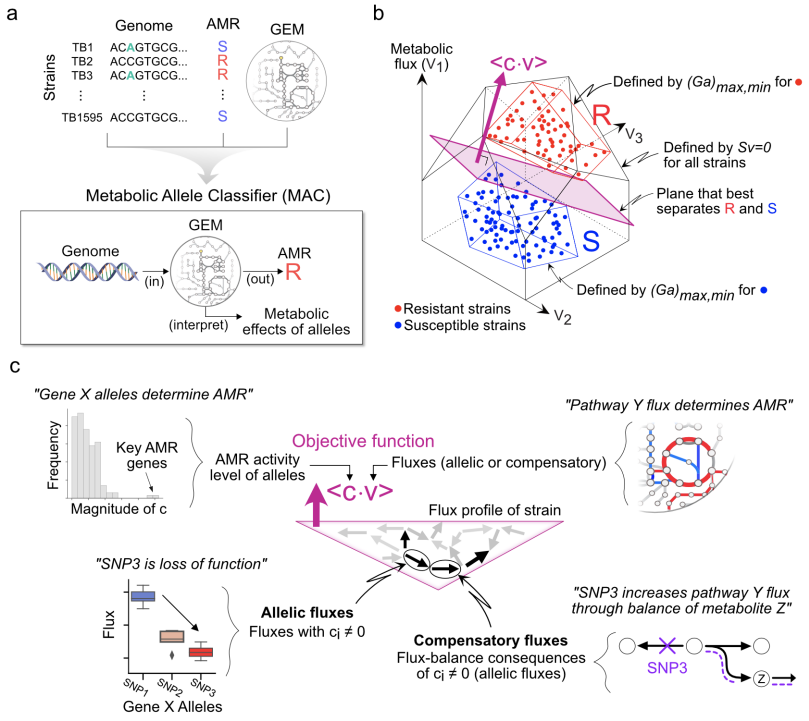

Figure 4.1: A metabolic systems approach for genetic associations. (a) In this study, data describing TB genome sequences and AMR data types are integrated with a metabolic model to learn a biochemically-interpretable classifier, named Metabolic Allele Classifier (MAC). The MAC parameters consist of allele-specific flux capacity constraints, a, and an antibiotic-specific metabolic objective, c, both of which are inferred from the data. (b) The optimal MAC describes strain-specific polytopes in flux space that separate into resistant (R) and susceptible (S) regions. The MAC objective function, $c^T v$, is identified as normal to the plane that best separates R and S. (c) The learned MAC provides a biochemically-based hypothesis of AMR mechanisms and allele-specific effects through interpretation of c and v. The genome-scale flux state of a strain, v, consists of fluxes that are directly activated by alleles (allelic fluxes) and those that are flux-balance consequences of the allele-activated fluxes (compensatory fluxes). Abbreviations: S, susceptible; R, resistant; AMR, antimicrobial resistance.

$$H_{MAC} = \text{sign}(\max_{v} c^T \boldsymbol{v} + b) \qquad \text{(Antibiotic-specific objective)}$$

$$s.t.$$

$$\boldsymbol{Sv} = \boldsymbol{0} \qquad \text{(Flux balance constraint)}$$

$$\boldsymbol{v}^{lb} \leq \boldsymbol{v} \leq \boldsymbol{v}^{ub} \qquad \text{(Over-all min/max flux constraints)}$$

$$Ga^{lb} = v^{lb} \leq v \leq v^{ub} = Ga^{ub} \qquad \text{(Allele-specific min/max flux constraints)}$$

Where each line of the MAC formulation is briefly described with plain text to the right, and further detailed by the correspondingly ordered bullet points below;

- $H_{y,k}$ is the sign of the MAC optimum value that classifies a strain, k, as either resistant (R) or susceptible (S) to a specific antibiotic, y (see Supplementary Notes for comparison between the MAC and the Support Vector Machine). The optimum value is determined by optimizing the objective function, $\max c_y^T v_k$, which describes a linear combination of the metabolic fluxes, v_k, and is specific to an antibiotic, y. The antibiotic-specific objective coefficients, c_y^T, are unknown a-priori and inferred from the data as a normal to the plane that best separates resistant and susceptible strains (Figure 4.1b).

- The classical flux-balance constraints, $S_v k = 0$, ensure that for each strain, k, the net mass flux through each of their metabolites is balanced to 0 (i.e., steady internal homeostatic state), where S is the stoichiometric matrix with 998 metabolites (rows) and 1229 reactions (columns).

- The constraints on the fluxes (reaction rates) through the metabolic reactions, $v^{lb,ub}$, describe the overall min/max flux constraints not changed by allelic variation and are thus the same for all strains. Geometrically, the constraints $v^{lb,ub}$ and $S_v k = 0$ define a polytope in which all strain-specific fluxes must reside (Figure 4.1b).

- The binary genetic variant matrix, $G_{k,i}$, is the primary data type used in GWAS and describes the presence/absence of i alleles (columns) across k strains (rows).

- The constraints, $G_{k,i} a_{i,j}^{lb,ub} = v_{k,j}^{lb,ub}$, represent the genome sequence of each strain (represented as a row in G) as a set of allele-specific flux constraints, $v_{k,j}^{lb,ub}$. The allele-constraint matrix, $a^{lb,ub}$, describes the allele-specific flux constraint values of i alleles (rows) that

encode for enzymes catalyzing j reactions (columns) (see Supplementary Notes for further explanation on the biological relationship between alleles and flux constraints). The allele-constraint matrix is unknown a-priori and inferred from the data. Geometrically, Ga describes strain-specific polytopes that represent the best separation of resistant and susceptible strains within the overall flux space (Figure 4.1b).

Importantly, the MAC was formulated such that for each strain-antibiotic classification, $H_{y,k}$, there exists a corresponding flux state, v_k, thereby providing a biochemical network explanation of the classification. Geometrically, the flux state of the metabolic network of a particular strain is described by the intersection of the objective function with its genome-specific polytope (Figure 4.1c).

The objective function corresponds to the fluxes through a set of metabolic reactions that form the basis for the MAC. By the fundamental nature of flux balancing, these reactions identify activity levels of discriminating pathways. The objective function that best separates the two polytopes formed by the spaces of resistant and sensitive phenotypes is a plane that describes a critical level of pathway activity that discriminates between the R and S phenotypes. Thus, the separating plane consists of fluxes that are directly activated by alleles ($c_i \neq 0$) and those that result from flux-balance consequences of $c_i \neq 0$. Statistical tests can then be performed using the set of all strain-specific intersections to identify both significant flux states discriminating between resistant and susceptible strains (Supplementary Figure 2a) as well as their underlying allele-specific flux effects (Supplementary Figure 2b). The MAC is therefore a biochemically interpretable machine learning classifier.

4.2.3 Validation of Metabolic Allele Classifiers

We utilized randomized sampling, machine learning, and model selection to identify predictive MACs (see Supplementary Figures 4-5, Methods, and Supplementary Notes for further details of the process outlined below). Specifically, the MACs were trained on the same 375 strains to predict antibiotic phenotypes with 1,220 strains set aside for testing. Since the computational cost of estimating MACs scales poorly with the number of alleles utilized, we limited the set of alleles modeled by the MAC to 237, describing 107 genes consisting of both known and unknown relations to AMR (Supplementary File 1). The known AMR genes provide validation cases while the unknown genes enable novel insights.

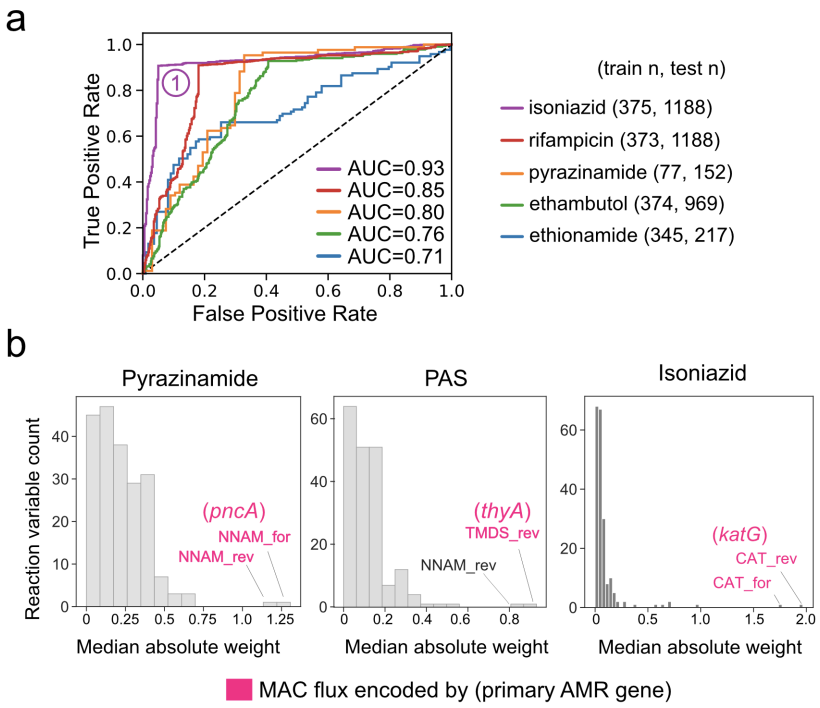

Figure 4.2: Validation of Metabolic Allele Classifiers. (a) Receiver operator characteristic (ROC) curves for MAC AMR predictions determined using a test set of 1,188 isoniazid-tested strains. (b) Histogram of median absolute MAC objective function coefficients (cyT) for pyrazinamide, para-aminosalicylic acid, and isoniazid MACs. The reaction variables corresponding to the two largest coefficients are noted in text. The reaction variable corresponding to the primary genetic determinant is colored pink. Abbreviations: AUC, area under the curve.

We assessed MACs for isoniazid, rifampicin, pyrazinamide, ethambutol, and ethionamide using held out test sets and find that the MACs generally achieve high classification performance (Figure 4.2a), with scores similar to our previous mechanism-agnostic machine learning models [2]. The MACs were further validated by assessing their ability to recover the primary AMR genes. We find that the largest objective weights for pyrazinamide, para-aminosalicylic acid, and isoniazid MACs correspond to the primary known AMR genes of antibiotics (Figure 4.2b). These results show that the MAC performs on par with state-of-the-art machine learning approaches in AMR classification and identification of primary AMR genes.

4.2.4 MACs reveal known and new antibiotic resistance determinants

The ability of MACs to efficiently predict AMR phenotypes (i.e., high accuracy, low complexity) suggests that the model parameters have biological relevance. Furthermore, in contrast to "black-box" machine learning models, the genotype-phenotype map of a MAC was designed to satisfy known biological constraints on metabolism e.g., reaction stoichiometry, mass conservation, gene-product-reaction encoding, nutrient environment. Therefore, we hypothesized that MACs should not only identify genetic determinants of AMR, but also provide metabolic systems explanations of their predictions.

Below, we focus our analysis on three case studies: pyrazinamide, para-aminosalicylic acid, and isoniazid AMR. These three antibiotics were chosen due to having both characterized and uncharacterized mechanisms underlying their associated alleles, allowing for both test cases and novel insights for the MAC. We analyze the best MACs for each antibiotic through four steps: (i) identification of significant fluxes discriminating between resistant and susceptible strains (i.e., "flux GWAS"), (ii) pathway enrichments of significant fluxes, (iii) identification of key allelic flux

effects, and (iv) network-level flux tracing of allelic effects (Methods).

4.2.5 Pyrazinamide resistance

To identify key flux states discriminating between resistant and susceptible strains, we performed statistical associations between the strain-specific MAC fluxes, v_k, and pyrazinamide AMR phenotypes using the training set of 77 strains (52 resistant, 25 susceptible) (we refer to this as "Flux GWAS", see Figure 4.1d). Flux GWAS identified 25 significant reaction fluxes (Bonferroni corrected $P < 4.66x10 - 5$, 0.05/1073 reactions) whose gene-protein-reaction rules overlapped with 8 genes modeled by MAC alleles (pncA, ansP2, fadD26, ppsA, and drrABC) (Supplementary Figure 7a; Supplementary File 3).

To gain a coarse systems view of the 25 significant fluxes, we performed pathway enrichment tests using a curated gene-pathway annotation list consisting of both BioCyc [30] and KEGG pathways [31] that accounts for 32% of protein-coding genes in the H37Rv genome (1,254/3906) (Supplementary File 2; Methods). Of the 245 total pathways, 5 were enriched with significant fluxes with less than 5% false discovery rate (FDR¡0.05) [32] and were primarily described by "phthiocerol biosynthesis" and "nicotinate and nicotinamide metabolism" (Figure 4.3a). These results recapitulate two pyrazinamide features describing flux variation in nicotinamidase activity [33] and phthiocerol dimycocerosate (PDIM) biosynthesis [12].

We then set out to understand the genetic basis for the flux associations by identifying loci in which the AMR association of each allele was correlated with their flux distribution ("LOR-flux correlation") (see Methods). The idea here is that resistant alleles have different metabolic effects than susceptible alleles for key genes. These allele-specific flux differences underlie the AMR classification accuracy of the MAC. We identified significant LOR-flux correlations at pncA,

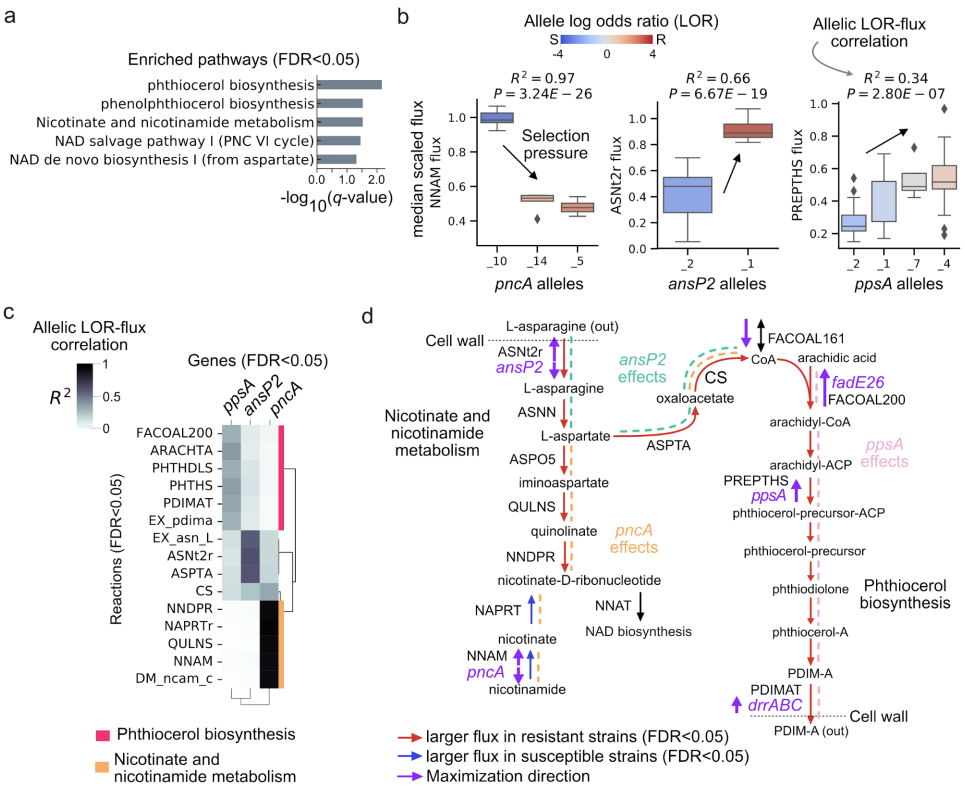

Figure 4.3: Characterization of pyrazinamide MACs. (a) Horizontal bar plots of pathways enriched with significant pyrazinamide-associated fluxes with FDR¡0.05. (b) Boxplots of pncA, ansP2, and ppsA allele-specific fluxes for the reactions catalyzed by their gene-products. Alleles are rank ordered from least to greatest by their log odds ratio (LOR), from left to right. The boxes are colored according to the allele LOR, where positive corresponds to resistant (R) dominant while negative corresponds to susceptible (S) dominant. Regression between allele LOR and flux is plotted. See Supplementary Data File 3 for list of mutations per allele. (c) Clustered heatmap of allele LOR-flux correlations between genes (y-axis) and significant reactions fluxes (x-axis). (d) Pathway depiction of "nicotinate and nicotinamide metabolism" and "phthiocerol biosynthesis" with objective variables plotted. Coenzyme-A generation from L-asparagine through aspartate decarboxylase (ASPTA) and citrate synthase (CS) is also depicted. Traced allelic effects are shown as dashed lines and colored for pncA, ansP2, and ppsA.

ansP2, and ppsA loci (FDR¡0.05) (Figure 4.3b). Specifically, the MACs infer a flux decreasing selection pressure at the pncA locus and flux increasing selection pressures at the ansP2 and ppsA loci. The estimated decreased enzymatic activity of pncA is consistent with studies describing resistant pncA mutants as loss of function [34]. Mutations in ppsA have previously been linked to pyrazinamide AMR [12] and convergent AMR evolution [35] while ansP2 mutants have not

yet been associated with AMR.

To understand the global effects of pncA, ppsA, and ansP2 alleles on the metabolic network, we traced out their LOR-flux correlation through the 25 significant reactions (Figure 4.3c). For ansP2, we observe that the increased generation of L-asparagine by the resistant ansP2 allele was utilized to generate coenzyme A (CoA) through aspartate aminotransferase (ASPTA) and citrate synthase (CS) (Figure 4.3d), which recapitulates experimental studies describing L-aspartate-based modulation of CoA as a pyrazinamide resistance mechanism [12]. However, our results differ from that of the proposed panD-based pantothenate route for CoA generation [36–38]. The lack of pyrazinamide-associated panD alleles in our dataset may underlie this discrepancy.

In summary, pyrazinamide MACs correctly identify pncA and ppsA alleles as major genetic determinants and recapitulate nicotinamide metabolism, CoA biosynthesis, and phthiocerol metabolism as key metabolic associations [12, 34]. As for new hypothesis, the MACs implicate ansP2 mutants in resistance through L-aspartate-based modulation of the coenzyme-A pool.

4.2.6 Para-aminosalicylic resistance

We performed flux GWAS using the para-aminosalicylic acid training set of 375 strains (80 resistant, 295 susceptible) and identified 52 fluxes discriminating between resistant and susceptible strains (Bonferroni corrected P ¡ 4.66x10-5, 0.05/1073 reactions) (Supplementary Figure 7b, Supplementary File 4). Of these 52 reactions, 10 were directly encoded by MAC alleles of 8 genes (thyA, katG, pncA, alar, cysK2, ald, fadE26, aspB, kdg, and inhA). Pathway enrichment tests of these 52 reactions identified “S-adenosyl-L-methionine cycle II”, “NAD de novo biosynthesis I (from aspartate)”, and “cysteine and methionine metabolism” as key para-aminosalicylic acid

pathways (FDR¡0.05) (Figure 4.4a). The identification of "cysteine and methionine metabolism" recapitulates known metabolic effects of para-aminosalicylic acid [39].

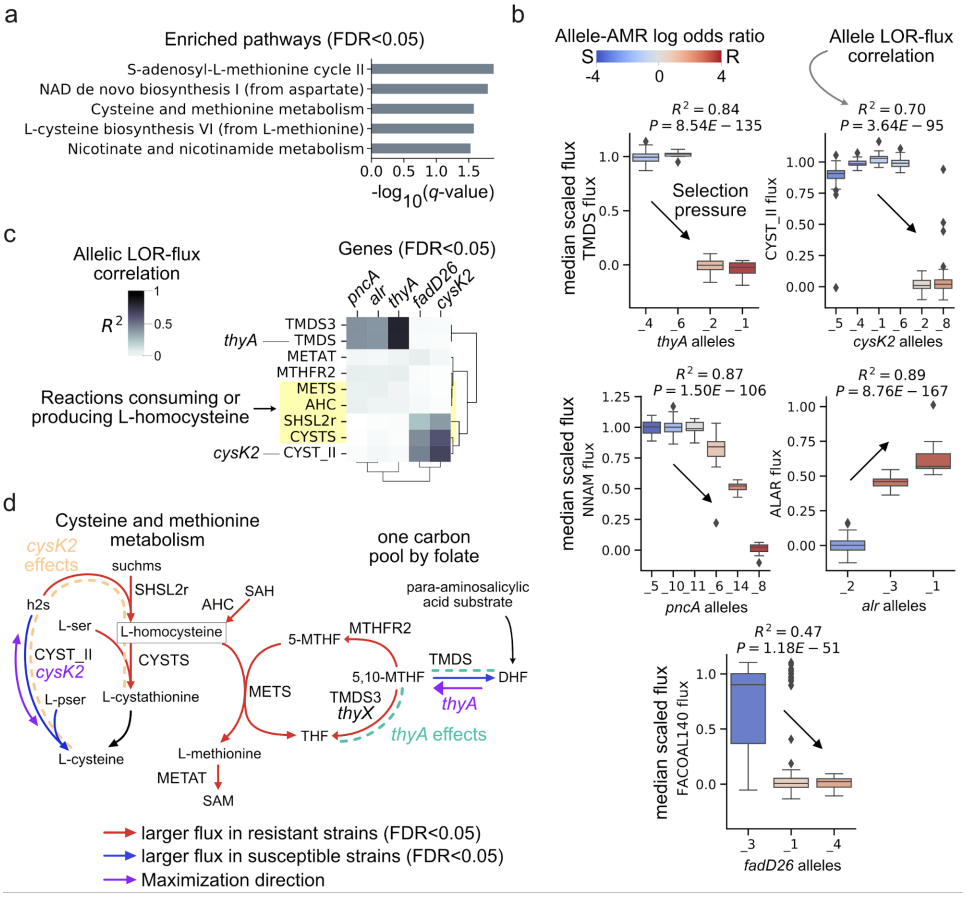

Figure 4.4: Characterization of para-aminosalicylic acid MACs. (a) Horizontal bar plots of pathways enriched with significant para-aminosalicylic acid-associated fluxes with FDR¡0.05. (b) Boxplots of thyA, cysK2, alr, pncA, and fadD26 allele-specific fluxes for the reactions catalyzed by their gene-products. Alleles are rank ordered from least to greatest by their log odds ratio (LOR), from left to right. The boxes are colored according to the allele LOR, where positive corresponds to resistant (R) dominant while negative corresponds to susceptible (S) dominant. Regression between allele LOR and flux is plotted. See Supplementary Data File 4 for list of mutations per allele. (c) Clustered heatmap of allele LOR-flux correlations for significant reactions in cysteine and methionine metabolism. (d) Pathway depiction of "cysteine and methionine metabolism" and "one carbon pool by folate". Significant allelic effects are shown by dashed lines and colored for thyA and cysK2.

We tested these genes for allelic LOR-flux correlations and identified selection pressures at thyA, cysK2, alr, pncA, and fadD26 loci (FDR¡0.05, R2¿0.1) (Figure 4.4b). Specifically, the

MACs infer flux decreasing selection pressures at the thyA, cysK2, pncA, and fadD26 loci and a flux increasing selection pressure at the alr locus. The estimated decreased enzymatic activity of thyA resistant alleles is consistent with experimental studies describing thyA resistant mutants as loss of function [8, 40]. The identification of alr and pncA—known determinants of cycloserine and pyrazinamide, respectively—reflect the co-resistance of these strains and are not known to have selective pressure in para-aminosalicylic acid treatment. Of these genes, only cysK2 encodes an enzyme in "cysteine and methionine pathway" and has not been previously linked to AMR.

We traced out the allelic LOR-flux correlation of cysK2 through cysteine and methionine pathway flux and found that their effects positively correlated with fadD26 alleles and negatively with thyA, alr, and pncA alleles (Figure 4.4c). Resistant cysK2 alleles are estimated to lead to increased flux through O-succinylhomoserine (SHSL2r) and cystathionine beta-synthase (CYSTS). The effect of cysK2 decreases from SHSL2r to CYSTS at the L-homocysteine flux balance node, which implicates L-homocysteine modulation as the cysK2 selection pressure (Figure 4.4d). Notably, L-homocysteine was experimentally identified as the most differentially perturbed metabolite resulting from para-aminosalicylic acid treatment [39].

In summary, para-aminosalicylic acid MACs recover thyA as the primary genetic determinant and recapitulate cysteine and methionine metabolism as a major pathway induced by the drug. As for novel hypothesis, the MACs implicate deleterious cysK2 mutants in resistance through modulation of L-homocysteine that may either arise from deleterious thyA mutants or para-aminosalicylic acid treatment.

4.2.7 Isoniazid resistance

We performed flux GWAS using the isoniazid training set of 375 strains (248 resistant, 127 susceptible) and identified 160 significant fluxes (Bonferroni corrected P ¡ 4.66x10-5, 0.05/1073 reactions) (Supplementary Figure 7c, Supplementary File 5). We find that only 11.3% (18/160) of the significant fluxes were catalyzed by gene-products of the MAC alleles. Pathway enrichments of the 160 significant fluxes identified "TCA cycle V", "oxidative phosphorylation", "superpathway of mycolate biosynthesis", and "gluconeogenesis I" as key isoniazid pathways (FDR¡0.05) (Figure 4.5a). These results are consistent with numerous studies demonstrating TCA and oxidative phosphorylation as key TB pathways altered by isoniazid treatment [41–43] and studies generally linking antibiotic efficacy to these pathways [44]. In general, we found that resistant strains were characterized by decreased respiratory activity, which is consistent with studies connecting decreased respiration to increased isoniazid resistance [42]. The genes encoding enzymes in these enriched pathways correspond to known (inhA, fabD, kasA, accD6, fadE24, ndh) and unknown (accD5, nuoL, gpdA2) genetic determinants of isoniazid resistance; however, none of these encoded for reactions annotated with "TCA cycle V".

We tested the significant fluxes for allelic LOR-flux correlations and identified selection pressures at katG, ndh, nuoL, accD6, gpdA2, fabD, kasA, and accD5 loci (FDR¡0.05) (Figure 4.5b). Specifically, the MACs infer flux decreasing selection pressures at the ndh, nuoL, fabD, gpdA2, and kasA loci and a flux increasing selection pressure at the katG, accD6, and accD5 locus (MCOATA is depicted in reverse direction). The resulting increased CAT flux observed in resistant strains is consistent with studies describing the majority of resistance-conferring katG alleles in clinical isolates as preserving catalase-peroxidase activity while disabling isoniazid binding (i.e., strains carrying susceptible-dominant katG alleles have low catalase-peroxidase flux due

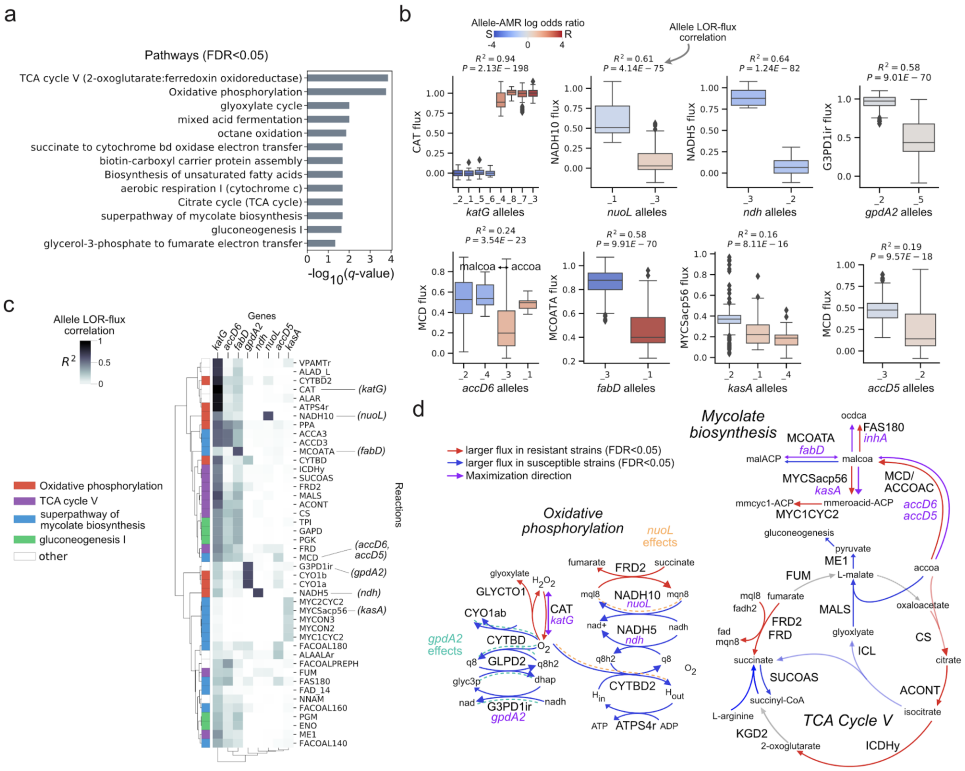

Figure 4.5: Characterization of isoniazid MACs. (a) Horizontal bar plots of pathways enriched with significant isoniazid-associated fluxes with FDR¡0.05. (b) Boxplots of katG, nuoL, ndh, gpdA2, accD6, fabD, kasA, and accD5 allele-specific fluxes for the reactions catalyzed by their gene-products. Alleles are rank ordered from least to greatest by their log odds ratio (LOR), from left to right. The flux (y-axis) is the median scaled flux across the high-quality isoniazid MACs. The boxes are colored according to the allele LOR, where positive corresponds to resistant (R) dominant while negative corresponds to susceptible (S) dominant. See Supplementary Data File 5 for list of mutations per allele. (c) Clustered heatmap of allele LOR-flux correlations for significant reactions in "TCA Cycle V", "Oxidative phosphorylation", and "Mycolate biosynthesis". (d) Pathway depiction of "TCA Cycle V", "Oxidative phosphorylation", and "Mycolate biosynthesis". Significant allelic effects are shown by dashed lines and colored for gpdA2 and nuoL.

to isoniazid binding) [45, 46]. The increased flux towards mycolic acid biosynthesis in resistant strains by fabD, accD6, and kasA is consistent with studies showing increased expression of these genes resulting from isoniazid treatment [47]. Furthermore, the metabolite acted on by these genes, malonyl-CoA, has recently been shown to have a significant fold change in response to 16 antibiotics in TB [48].

We traced out significant LOR-flux correlations of these genes through the enriched pathways to elucidate their global network effects (Figure 4.5c). For the novel genetic determinants, nuoL and gpdA2, we find that their alleles have significant flux effects in cytochrome bd oxidase reactions (CYTBD, CYTBD2) traced through menaquinone and ubiquinone flux balance nodes, respectively (Figure 4.5d). The allelic effects of the primary genetic determinant, katG, are similarly traced through cytochrome bd oxidase flux by oxygen. The importance of cytochrome bd oxidase has recently been linked to isoniazid [41]. These results implicate gpdA2 and nuoL mutants in isoniazid AMR through modulation of quinone/menaquinone pools.

In summary, isoniazid MACs recover the primary (katG) and secondary (inhA, fabD, kasA, accD6, fadE24, ndh) genetic determinant and recapitulate oxidative phosphorylation, TCA, and mycolic acid biosynthesis as major pathways induced by the drug [41–43]. As for novel genetic hypothesis, the MACs implicate gpdA2 and nuoL mutants in resistance through modulation of menaquinone and ubiquinone that may either arise from katG mutants or isoniazid-induced oxidative stress.

4.2.8 Conventional pathway analyses do not recapitulate network-level AMR mechanisms

To assess how MAC results compare to mechanism-agnostic approaches, we performed conventional pathway analysis of the 197 alleles (Supplementary File 6, Methods). Comparison of pathway-based analysis showed that results derived from conventional pathway enrichments do not recapitulate the antibiotic mechanisms for isoniazid, pyrazinamide, and para-aminosalicylic acid. For isoniazid, a total of five pathways were enriched (FDR¡0.05); however, the significant allelic associations enriched in pathways were simply those annotated for katG, such as “superoxide

radicals degradation" and "tryptophan metabolism". For para-aminosalicylic acid, "L-alanine biosynthesis I" was the only enriched pathway while no pathway was enriched for pyrazinamide alleles (FDR¡0.05).

These results show that flux balance constraints are required to generate meaningful network-level hypotheses for identified genetic associations. The basis for this advancement is that flux balances represent how the entirety of metabolic gene products come together to produce balanced homeostatic states.

4.3 Discussion

We have developed a computational framework for analyzing data sets (comprised of genotypes and binary phenotypes) using a genome-scale model (GEM) to identify the genetic and metabolic basis for TB AMR (Fig. 1a). The identification of the underlying biochemical mechanisms is reflected in the MAC. We first discuss our approach, emphasizing key design choices, and then describe the results it generates when applied to the TB dataset.

The outcome of the MAC depends on two major design choices: the set of alleles and the objective function that optimally separates strains into resistant and sensitive strain cohorts in the overall metabolic flux space. Although our approach does not explicitly require prior knowledge of key AMR genes, we chose a set of alleles with just over 100 genes with known and implicated AMR relations in order to both provide test cases and to address the combinatorial explosion of sampling possible allelic effects. Relaxing the current computational bottleneck in identifying MACs will enable the utilization of all alleles. For determining the objective function, our approach was based on the key insight that a linear program may behave as a machine learning classifier if its objective optimizes in the direction normal to a predictive classification

plane. While we utilized PCA, L1-logistic regression, and the BIC metric to identify sparse linear objectives, there are potentially alternative avenues that could be taken. The major concept that should sustain in any model selection strategy is that a good model is simple (in structure) yet accurate (in its predictions). Application of the MAC to other GWAS datasets may therefore benefit from tuning these parameters appropriately.

The MAC advances current GWAS machine learning approaches by enabling a biochemical interpretation of genetic associations. Although advancements have been made to increase the "explainability" of black-box machine learning models [49–51], such interpretations are limited by the lack of mechanistic knowledge incorporated in the model. We show that causal biochemical explanations for classifications can be derived by constraining a machine learning classifier to satisfy knowledge-based biological constraints (gene function, reaction stoichiometry, flux balance, etc).

Our interpretation of MACs for pyrazinamide, para-aminosalicylic acid, and isoniazid AMR identified genome-scale flux states and key pathways discriminating resistant and susceptible strains. Notably, we found the MAC-identified pathways to be consistent with known antibiotic mechanisms. In contrast, conventional pathway analysis using only alleles was unable to recapitulate known pathway mechanisms. The MAC therefore provides a mechanistic approach for pathway-based analysis of genome-wide associations [52].

Dissection of the allele-specific fluxes underlying the significant fluxes further clarified the genotype-phenotype map and provided hypotheses regarding specific allelic effects. For example, pyrazinamide MACs implicate an ansP2 allele as a novel resistance determinant through increased uptake of asparagine towards L-aspartate-based CoA generation. The MAC thus extends allele-phenotype associations (i.e., LOR) by estimating allele-specific flux effects and their network

interactions.

Taken together, the framework presented here meets the pressing need to integrate comprehensive biochemical mechanisms for the analysis of genomics-phenomics datasets. Our framework both recovers known gene-AMR relations and provides novel insights regarding their metabolic basis. As genome sequences, phenotypes, and genome-scale network reconstructions of microbes continue to grow in size and scope, similar results to those presented here are likely to appear in the coming years. This initial development of an FBA based GWAS analysis (FBA-GWAS) is likely to continue the development of a mechanistic basis into future GWAS methods.

Acknowledgements

We would like to thank Anand Sastry, Jean-Christophe Lachance, Yara Seif, and Jason Hyun for helpful discussions and Marc Abrams for editing the manuscript.

This research was supported by the NIAID grant (AI124316), the NIGMS (GM102098), and the Novo Nordisk Foundation Grant Number NNF10CC1016517.

Chapter 4 is a reprint of the material in: **ES Kavvas**, L Yang, JM Monk, D Heckmann, BO Palsson. 2020. "A biochemically-interpretable machine learning classifier for microbial GWAS" *Nature Communications* 11 (2580). The dissertation author is the primary author.

4.4 References

1. Organization, W. H. *et al. Global tuberculosis report 2018* (World Health Organization, 2018).

2. Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. & Palsson, B. O. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. en. *Nature communications* **9,** 4306. ISSN: 2041-1723, 2041-1723 (Oct. 2018).

3. Boolchandani, M., D'Souza, A. W. & Dantas, G. Sequencing-based methods and resources to study antimicrobial resistance. en. *Nature reviews. Genetics.* ISSN: 1471-0056, 1471-0064 (Mar. 2019).

4. Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A. R., Will, R., Xia, F. & Stevens, R. Antimicrobial Resistance Prediction in PATRIC and RAST. en. *Scientific reports* **6,** 27930. ISSN: 2045-2322 (June 2016).

5. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P. & Zhang, L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. en. *Microbiome* **6,** 23. ISSN: 2049-2618 (Feb. 2018).

6. Walker, T. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. *Nature reviews. Microbiology.* ISSN: 1740-1526 (2019).

7. Earle, S. G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., Spencer, C. C. A., Iqbal, Z., Clifton, D. A., Hopkins, K. L., Woodford, N., Smith, E. G., Ismail, N., Llewelyn, M. J., Peto, T. E., Crook, D. W., McVean, G., Walker, A. S. & Wilson, D. J. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. en. *Nature microbiology* **1,** 16041. ISSN: 2058-5276 (Apr. 2016).

8. Zheng, J., Rubin, E. J., Bifani, P., Mathys, V., Lim, V., Au, M., Jang, J., Nam, J., Dick, T., Walker, J. R., Pethe, K. & Camacho, L. R. para-Aminosalicylic acid is a prodrug targeting dihydrofolate reductase in Mycobacterium tuberculosis. en. *The Journal of biological chemistry* **288,** 23447–23456. ISSN: 0021-9258, 1083-351X (Aug. 2013).

9. Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M., McNeil, M., Peterson, S. N., Chatterjee, D., Fleischmann, R., *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-[beta]-D-arabinose biosynthetic and utilization pathway genes. *Nature genetics* **45,** 1190–1197. ISSN: 1061-4036 (2013).

10. Desjardins, C. A., Cohen, K. A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B. J., Shea, T. P., Almeida, D. V., Manson, A. L., Salazar, A., Padayatchi, N., O'Donnell, M. R., Mlisana, K. P., Wortman, J., Birren, B. W., Grosset, J., Earl, A. M. & Pym, A. S. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. en. *Nature genetics* **48,** 544–551. ISSN: 1061-4036, 1546-1718 (May 2016).

11. Hicks, N. D., Yang, J., Zhang, X., Zhao, B., Grad, Y. H., Liu, L., Ou, X., Chang, Z., Xia, H., Zhou, Y., Wang, S., Dong, J., Sun, L., Zhu, Y., Zhao, Y., Jin, Q. & Fortune, S. M. Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. en. *Nature microbiology* **3,** 1032–1042. ISSN: 2058-5276 (Sept. 2018).

12. Gopal, P., Yee, M., Sarathy, J., Low, J. L., Sarathy, J. P., Kaya, F., Dartois, V., Gengenbacher, M. & Dick, T. Pyrazinamide Resistance Is Caused by Two Distinct Mechanisms: Prevention of Coenzyme A Depletion and Loss of Virulence Factor Synthesis. en. *ACS infectious diseases* **2,** 616–626. ISSN: 2373-8227 (Sept. 2016).

13. Yu, M. K., Ma, J., Fisher, J., Kreisberg, J. F., Raphael, B. J. & Ideker, T. Visible Machine Learning for Biomedicine. en. *Cell* **173,** 1562–1565. ISSN: 0092-8674, 1097-4172 (June 2018).

14. Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R. & Ideker, T. Using deep learning to model the hierarchical structure and function of a cell. en. *Nature methods* **15,** 290–298. ISSN: 1548-7091, 1548-7105 (Apr. 2018).

15. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-Generation Machine Learning for Biological Networks. en. *Cell* **173,** 1581–1592. ISSN: 0092-8674, 1097-4172 (June 2018).

16. Palsson, B. Ø. *Systems Biology: Constraint-based Reconstruction and Analysis* en. ISBN: 9781316239940 (Cambridge University Press, Jan. 2015).

17. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. en. *Cell* **161,** 971–987. ISSN: 0092-8674, 1097-4172 (May 2015).

18. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. en. *Nature reviews. Genetics* **15,** 107–120. ISSN: 1471-0056, 1471-0064 (Feb. 2014).

19. Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., Walker, G. C. & Collins, J. J. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. en. *Cell.* ISSN: 0092-8674, 1097-4172 (May 2019).

20. Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. & Palsson, B. Ø. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. en. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 20338–20343. ISSN: 0027-8424, 1091-6490 (Dec. 2013).

21. Bosi, E., Monk, J. M., Aziz, R. K., Fondi, M., Nizet, V. & Palsson, B. Ø. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. en. *Proceedings of the National Academy of Sciences of the United States of America* **113,** E3801–9. ISSN: 0027-8424, 1091-6490 (June 2016).

22. Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., Catoiu, E., Raffatellu, M., Palsson, B. O. & Monk, J. M. Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. en. *Nature communications* **9,** 3771. ISSN: 2041-1723 (Sept. 2018).

23. Fang, X., Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., Seif, Y., Smarr, L. & Palsson, B. O. Escherichia coli B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. en. *BMC systems biology* **12,** 66. ISSN: 1752-0509 (June 2018).

24. Cardoso, J. G. R., Andersen, M. R., Herrgård, M. J. & Sonnenschein, N. Analysis of genetic variation and potential applications in genome-scale metabolic modeling. en. *Frontiers in bioengineering and biotechnology* **3,** 13. ISSN: 2296-4185 (Feb. 2015).

25. Lees, J. A. & Bentley, S. D. Bacterial GWAS: not just gilding the lily. en. *Nature reviews. Microbiology* **14,** 406. ISSN: 1740-1526, 1740-1534 (July 2016).

26. Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y. & Sobral, B. W. PATRIC, the bacterial bioinformatics database and analysis resource. en. *Nucleic acids research* **42,** D581–91. ISSN: 0305-1048, 1362-4962 (Jan. 2014).

27. Kavvas, E. S., Seif, Y., Yurkovich, J. T., Norsigian, C., Poudel, S., Greenwald, W. W., Ghatak, S., Palsson, B. O. & Monk, J. M. Updated and standardized genome-scale reconstruction of Mycobacterium tuberculosis H37Rv, iEK1011, simulates flux states indicative of physiological conditions. *BMC systems biology* **12,** 25. ISSN: 1752-0509 (Mar. 2018).

28. Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. en. *Applied and environmental microbiology* **60,** 3724–3731. ISSN: 0099-2240 (Oct. 1994).

29. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nature biotechnology* **28,** 245–248. ISSN: 1087-0156, 1546-1696 (Mar. 2010).

30. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P. & Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. en. *Nucleic acids research* **42,** D459–71. ISSN: 0305-1048, 1362-4962 (Jan. 2014).

31. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. en. *Nucleic acids research* **47,** D590–D595. ISSN: 0305-1048, 1362-4962 (Jan. 2019).

32. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. en. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 9440–9445. ISSN: 0027-8424 (Aug. 2003).

33. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. en. *Nature medicine* **2,** 662–667. ISSN: 1078-8956 (June 1996).

34. Zhang, H., Deng, J.-Y., Bi, L.-J., Zhou, Y.-F., Zhang, Z.-P., Zhang, C.-G., Zhang, Y. & Zhang, X.-E. Characterization of Mycobacterium tuberculosis nicotinamidase/pyrazinamidase. en. *The FEBS journal* **275,** 753–762. ISSN: 1742-464X (Feb. 2008).

35. Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., Warren, R. M., Streicher, E. M., Calver, A., Sloutsky, A., Kaur, D., Posey, J. E., Plikaytis, B., Oggioni, M. R., Gardy, J. L., Johnston, J. C., Rodrigues, M., Tang, P. K. C., Kato-Maeda, M., Borowsky, M. L., Muddukrishna, B., Kreiswirth, B. N., Kurepina, N., Galagan, J., Gagneux, S., Birren, B., Rubin, E. J., Lander, E. S., Sabeti, P. C. & Murray, M. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. en. *Nature genetics* **45,** 1183–1189. ISSN: 1061-4036, 1546-1718 (Oct. 2013).

36. Gopal, P., Nartey, W., Ragunathan, P., Sarathy, J., Kaya, F., Yee, M., Setzer, C., Manimekalai, M. S. S., Dartois, V., Grüber, G. & Dick, T. Pyrazinoic Acid Inhibits Mycobacterial Coenzyme A Biosynthesis by Binding to Aspartate Decarboxylase PanD. en. *ACS infectious diseases* **3,** 807–819. ISSN: 2373-8227 (Nov. 2017).

37. Dillon, N. A., Peterson, N. D., Rosen, B. C. & Baughn, A. D. Pantothenate and pantetheine antagonize the antitubercular activity of pyrazinamide. en. *Antimicrobial agents and chemotherapy* **58,** 7258–7263. ISSN: 0066-4804, 1098-6596 (Dec. 2014).

38. Zhang, S., Chen, J., Shi, W., Liu, W., Zhang, W. & Zhang, Y. Mutations in panD encoding aspartate decarboxylase are associated with pyrazinamide resistance in Mycobacterium tuberculosis. en. *Emerging microbes & infections* **2,** e34. ISSN: 2222-1751 (June 2013).

39. Chakraborty, S., Gruber, T., Barry 3rd, C. E., Boshoff, H. I. & Rhee, K. Y. Para-aminosalicylic acid acts as an alternative substrate of folate metabolism in Mycobacterium tuberculosis. en. *Science* **339,** 88–91. ISSN: 0036-8075, 1095-9203 (Jan. 2013).

40. Moradigaravand, D., Grandjean, L., Martinez, E., Li, H., Zheng, J., Coronel, J., Moore, D., Török, M. E., Sintchenko, V., Huang, H., Javid, B., Parkhill, J., Peacock, S. J. & Köser, C. U. dfrA thyA Double Deletion in para-Aminosalicylic Acid-Resistant Mycobacterium tuberculosis Beijing Strains. en. *Antimicrobial agents and chemotherapy* **60,** 3864–3867. ISSN: 0066-4804, 1098-6596 (June 2016).

41. Zeng, S., Soetaert, K., Ravon, F., Vandeput, M., Bald, D., Kauffmann, J.-M., Mathys, V., Wattiez, R. & Fontaine, V. Isoniazid Bactericidal Activity Involves Electron Transport Chain Perturbation. en. *Antimicrobial agents and chemotherapy* **63.** ISSN: 0066-4804, 1098-6596 (Mar. 2019).

42. Vilchèze, C., Hartman, T., Weinrick, B., Jain, P., Weisbrod, T. R., Leung, L. W., Freundlich, J. S. & Jacobs Jr, W. R. Enhanced respiration prevents drug tolerance and drug resistance

in Mycobacterium tuberculosis. en. *Proceedings of the National Academy of Sciences of the United States of America* **114,** 4495–4500. ISSN: 0027-8424, 1091-6490 (Apr. 2017).

43. Nandakumar, M., Nathan, C. & Rhee, K. Y. Isocitrate lyase mediates broad antibiotic tolerance in Mycobacterium tuberculosis. en. *Nature communications* **5,** 4306. ISSN: 2041-1723 (June 2014).

44. Lobritz, M. A., Belenky, P., Porter, C. B. M., Gutierrez, A., Yang, J. H., Schwarz, E. G., Dwyer, D. J., Khalil, A. S. & Collins, J. J. Antibiotic efficacy is linked to bacterial cellular respiration. en. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 8173–8180. ISSN: 0027-8424, 1091-6490 (July 2015).

45. Wengenack, N. L., Uhl, J. R., St Amand, A. L., Tomlinson, A. J., Benson, L. M., Naylor, S., Kline, B. C., Cockerill 3rd, F. R. & Rusnak, F. Recombinant Mycobacterium tuberculosis KatG(S315T) is a competent catalase-peroxidase with reduced activity toward isoniazid. en. *The Journal of infectious diseases* **176,** 722–727. ISSN: 0022-1899 (Sept. 1997).

46. Pym, A. S., Saint-Joanis, B. & Cole, S. T. Effect of katG mutations on the virulence of Mycobacterium tuberculosis and the implication for transmission in humans. en. *Infection and immunity* **70,** 4955–4960. ISSN: 0019-9567 (Sept. 2002).

47. Wilson, M., DeRisi, J., Kristensen, H. H., Imboden, P., Rane, S., Brown, P. O. & Schoolnik, G. K. Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. en. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 12833–12838. ISSN: 0027-8424 (Oct. 1999).

48. Zampieri, M., Szappanos, B., Buchieri, M. V., Trauner, A., Piazza, I., Picotti, P., Gagneux, S., Borrell, S., Gicquel, B., Lelievre, J., Papp, B. & Sauer, U. High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. en. *Science translational medicine* **10.** ISSN: 1946-6234, 1946-6242 (Feb. 2018).

49. Ribeiro, M. T., Singh, S. & Guestrin, C. *Why Should I Trust You?: Explaining the Predictions of Any Classifier* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, Aug. 2016), 1135–1144. ISBN: 9781450342322.

50. Adadi, A. & Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **6,** 52138–52160. ISSN: 2169-3536 (2018).

51. Gunning, D. & Aha, D. *DARPA's Explainable Artificial Intelligence (XAI) Program* 2019.

52. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. en. *Nature reviews. Genetics* **11,** 843–854. ISSN: 1471-0056, 1471-0064 (Dec. 2010).

Chapter 5

Laboratory evolution of multiple *E. coli* strains reveals unifying principles of adaptation but diversity in driving genotypes

Fitness landscapes are a central concept in evolutionary biology and have been thoroughly detailed in terms of genotypes. However, our understanding of the selected metabolic and gene expression adaptations, and their dependence on genetic background, remains limited. Here, we reveal multi-scale adaptation principles in the *E. coli* species by taking multi-omics measurements of six different strains throughout their adaptive evolution to glucose minimal media. Statistics and matrix factorization is applied to yield four key results. First, analysis of the metabolic and physiological data shows evolutionary convergence in growth rate, glucose uptake

rate, glycolytic ATP and NADH production but divergence in NADPH production strategies. Second, factorization-based analysis of the transcriptome revealed six conserved transcriptomic adaptations describing increased expression of ribosome and amino acid biosynthetic genes and decreased expression of stress response and structural genes. Third, correlation analysis identifies five tradeoffs underlying the transcriptomic profiles. Fourth, statistical tests leveraging ALE design identify four mutation-flux correlates and eight mutation-transcriptomic correlates that link mutations to systems level adaptation principles. Our total results reveal the dominant metabolic and regulatory constraints governing *E. coli* growth adaptation that either distinguish strains or are conserved principles.

5.1 Introduction

Advancements in biotechnology have enabled the unprecedented detailing of microbial evolution. The process of evolution can now be studied in a controlled laboratory environment, where genome sequencing and phenotypic measurements are routine [1, 2]. Although studies utilizing genome sequences and fitness measurements have provided valuable insights ranging from the dynamics of evolution on long time-scales [3–5] to general features of epistasis [6], evolutionary principles at the levels of gene regulation and metabolism remain unelucidated. Moreover, the generality of principles identified in experimental evolution studies is ambiguous since studies often focus on a single strain, not a species. For example, different strains of *E. coli* have been shown to exhibit diverse regulatory and metabolic functions and thus may have different constraints governing their evolutionary trajectories [7]. A fundamental multi-scale description of evolutionary landscapes may therefore be deduced through multi-omic measurements of different strain-specific experimental evolutions. Towards revealing multi-scale features of evolutionary

landscapes, researchers have taken transcriptomic and fluxomic measurements in their experimental evolution studies [8]. However, it remains challenging to extract insights from these data types due to a lack of effective data analysis methods, especially for gene expression data sets. To date, no statistical correlation has been made between selected mutations and these multi-omics measurements. Our lab has recently shown the effectiveness of independent component analysis (ICA) to quantitatively interpret transcriptomic datasets in terms of transcription factors [9]. Therefore, ICA and novel statistical approaches may reveal fundamental regulatory principles and provide links between mutations and transcriptomic changes analogous to those seen in genetic association studies.

Here, we reveal multi-scale adaptation principles in the *E. coli* species by taking multi-omics measurements of six different strains throughout their adaptive evolution to glucose minimal media. Our total results reveal the dominant metabolic and regulatory constraints governing *E. coli* growth adaptation that either distinguish strains or are conserved principles.

5.2 Results

5.2.1 Consistent genetics in evolution of multiple *E. coli* strains

Six different *E. coli* wild-type strains exhibiting diverse genetics (K-12 MG1655, K-12 W3110, BL21, C, and Crooks) (Figure 5.1a). were evolved to select for rapid growth. Independent triplicates of each strain were evolved under a strict selection pressure for growth in that the cultures never left the exponential phase under batch culture 37 C and M9 glucose (see Methods). Whole genome sequencing was performed for clones of all replicate lineages while 13-C fluxomics, RNA-seq, and physiological measurements were performed for a single replicate

lineage (Figure 5.1b). We find that all strains start with different growth rates but evolve to rates ranging between 0.98 and 1.11 hr-1 (Dt = 42 mins) (Figure 5.1c). Some strains (W and Crooks) operate near this optimal in their wild-type state, while others require genetic mutations to achieve the observed optimal (MG1655, W3110, BL21, and C). We observed striking consistency in mutated genes, where each strain had at least one gene with a selected mutation in all replicate lineages (Figure 5.1d). A total of seven genes (pykF, zwf, spoT, mrdA, hns/tdk, rpoC, rpoB) had selected mutations appear both in multiple strains and in more than one replicate lineage. The commonality of selected mutations indicated similar evolutionary constraints facing these strains and motivated inquiry of their metabolic and gene expression profiles.

5.2.2 Characteristics of physiological and metabolic adaptations

Since a total of 8 selected mutations were in genes encoding metabolic enzymes—two of which appear multiple strains (zwf, pykF)—we hypothesized that the strains may be evolving towards similar metabolic states. We thus set out to examine convergent and divergent phenotypes along the ALE trajectory by performing statistical tests for each physiological and fluxomic measurement between the wild-type (WT) and end-point (EP) flasks for each strain (see Methods). Of the 187 total phenotypes, 64 were identified as convergent (i.e., points became closer together) and 6 were identified as divergent (i.e., points became further apart) with false discovery rate (FDR) less than 5% (Figure 5.2a). Of the convergent phenotypes, we find that 86% (55/64) were growth-correlated (spearman rho<0.05, FDR<0.05) (see Supplementary File 1).

We find that the convergent features are related to glucose uptake, glycolysis and oxidative phosphorylation while the top ranked divergent features relate to NADPH production through malic enzyme (ME2) and pentose phosphate pathway (PPP) (Fig. 2A). Inspection of

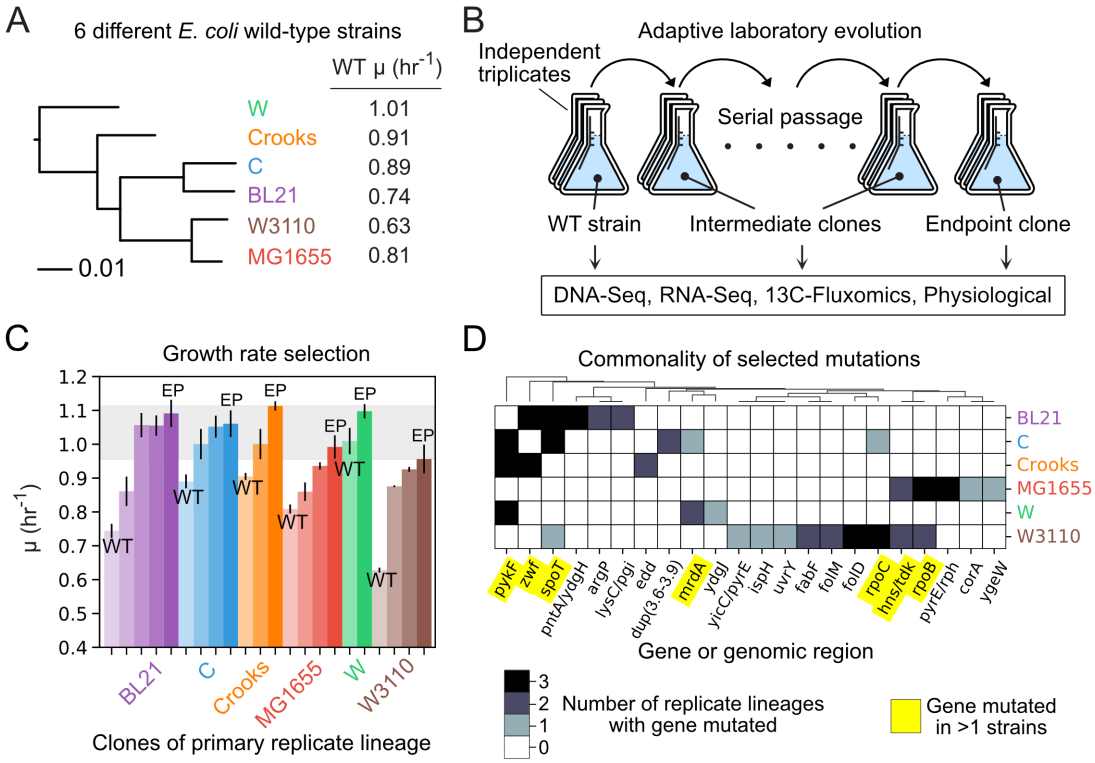

Figure 5.1: Overview of selected *E. coli* strains, experimental design, and key adaptation trends (A) Phylogenetic tree of six different *E. coli* wild-type starting strains utilized in this study. The wild-type (WT) growth rates (u) of the strains are noted. (B) Adaptive laboratory evolution was performed for each strain using independent triplicates. The wild-type (WT), evolved intermediate, and evolved end-point clones underwent multi-omics measurements. (C) Bar Plot of measured growth rates for wild type (WT), intermediate, and end point (EP) flasks for each strain. Clones are ordered left to right by trajectory. (D) Heatmap of gene-level mutation frequency across replicate lineages of each strain. The intergenic region between two genes is noted by a dash "/".

the ALE trajectories for the most convergent (Mann-Whitney U¿169, P<5.7x10-5) and divergent (Mann-Whitney U=19, P=5.7x10-5) phenotypes showed that phenotypes do not monotonically increase/decrease along the ALE (i.e., not always increasing or decreasing along trajectory) (Figure 5.2b). For example, although the glucose uptake rate has a significant net increase between WT and EP strains, four of the strains have one ALE jump where glucose uptake decreases. Principal component analysis of the fluxes showed that two components explain 93% of the variation and correspond to ATP production through oxidative phosphorylation and glycolysis

(80%), and NADPH balance through pentose phosphate pathway and transhydrogenases (13%) (Supplementary Figure 1).

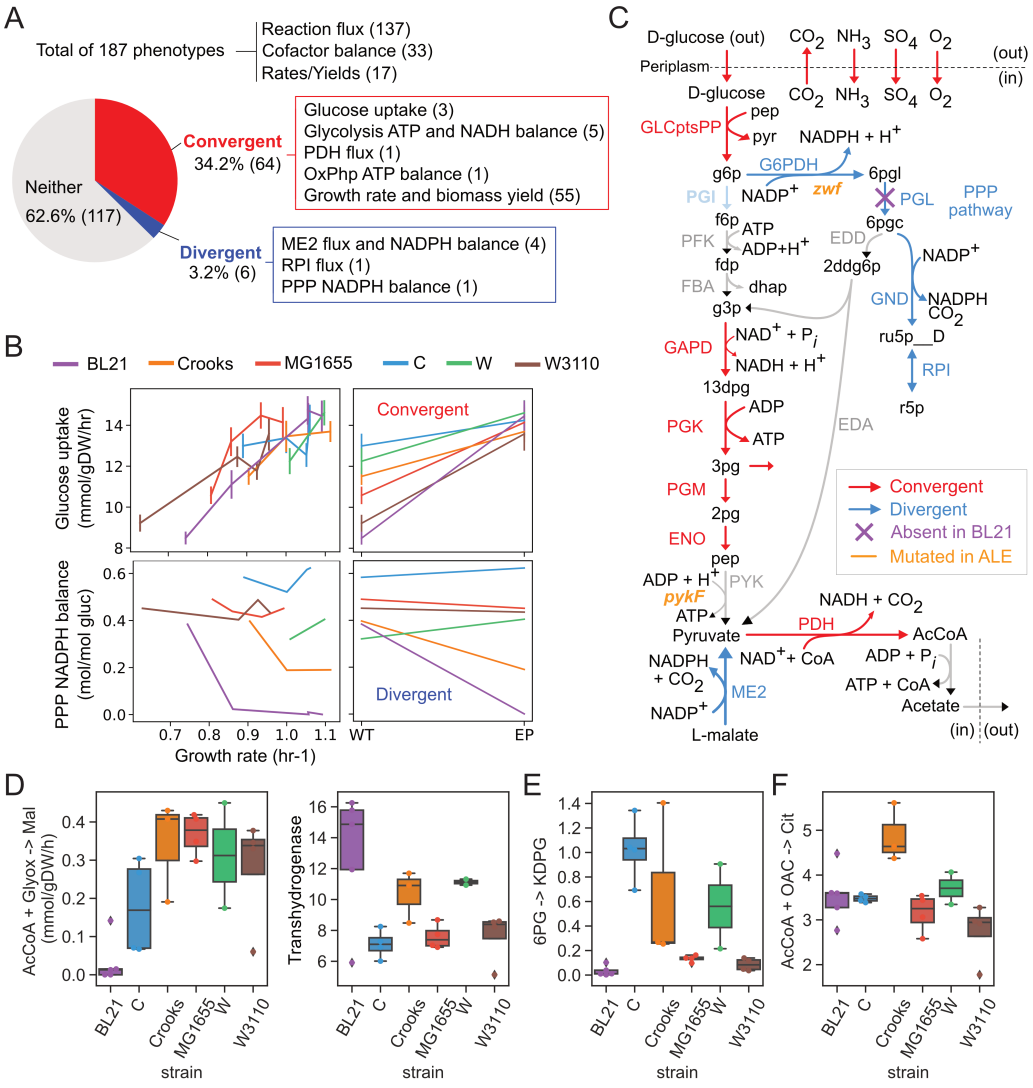

Figure 5.2: Adaptation in physiology and metabolism. (A) Pie chart describing the fraction of phenotypes that converge or diverge. Numbers in parentheses describe the number of related phenotypes. (B) Line plots of glucose uptake (top) and PPP NADPH balance (bottom) vs growth rate. Line plots and frequency distributions for WT and EP are plotted to the right for both cases. (C) Metabolic map of reactions in glycolysis, PPP, and exchange reactions colored according to whether they diverge or converge. Blue describes divergence and red describes convergence. (D-F) Bar plots of four reaction fluxes (absolute) that have strain-specific distributions. Abbreviations: abs, absolute flux (mmol/gDW/hr); rel, relative flux (mol/mol glucose); TCA, citric acid cycle; PPP, pentose phosphate pathway; ME2, malic enzyme; OxPhp, Oxidative phosphorylation; PDH, pyruvate dehydrogenase.

To determine whether specific reaction fluxes distinguish specific strains, we tested all fluxes for strain-specific distributions and found four subsystems specific to BL21, Crooks and C (ANOVA F-test, FDR<0.05). The BL21 strain uniquely had no flux through glyoxylate shunt while having the highest flux through transhydrogenase (Figure 5.2d). Since BL21 can't regenerate NADPH through PPP due to lacking the pgl gene encoding 6-phosphogluconolactonase (PGL) reaction activity (Meier, Jensen and Duus, 2012), the high transhydrogenase flux likely compensates to regenerate NADPH. Furthermore, we find that all BL21 flask lineages select for mutations in the intergenic region of a transhydrogenase (pntA/ydgH) (we test for mutation correlates later in this study) (Figure 5.2d). C strain uniquely had high flux through the Entner-Doudoroff (ED) pathway while BL21, MG1655, and W3110 had almost none (Figure 5.2e). Crooks uniquely had the highest flux through TCA (Figure 5.2f). In total, these results describe convergent and divergent phenotypes that are either conserved or distinguish strains.

5.2.3 Characteristics of transcriptome adaptation in *E. coli*

Underlying the phenotypic differences of these strains are differences in gene expression strategies. We thus set out to analyze the transcriptome of these strains by performing both differential expression analysis and a matrix factorization approach. Differential expression analysis showed that the number of differentially expressed genes (DEGs) generally decreases along the trajectory, with the exception of the last BL21 flask (Figure 5.3a). To make sense of these expression changes, we applied an alternative RNA-seq analysis workflow that was shown to enable quantitative analysis of the *E. coli* transcriptome from the perspective of transcription factors [9]. The authors showed that independent component analysis (ICA) deconvolved a large compendium of *E. coli* MG1655 RNA-seq data into a linear combination of independent sources

that reflect known regulons ("iModulons"), and source weightings ("iModulon activities"), which describe the global regulatory state [10]. Using the previous set of 92 iModulons, we transformed the flask-specific gene expression profiles into flask-specific iModulon activities (see Methods, Supplementary Figure 2).

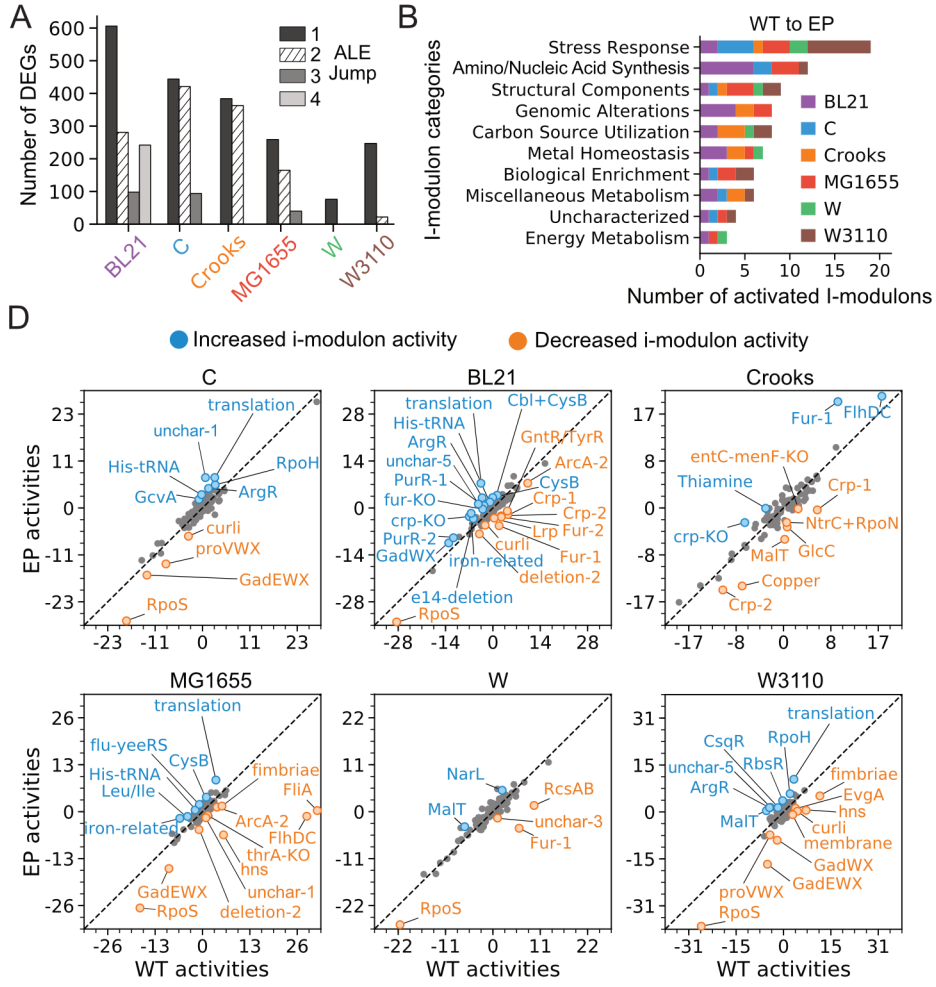

Figure 5.3: Characterization of gene expression adaptations. (A) Number of differentially expressed genes (DEG) for each strain-specific jump in growth rate during ALE. (B) Bar plot of total iModulon activation count in terms of iModulon functional category. The count is summed across the 6 strains activated ranked by the total number of times they were differentially activated between WT and EP flasks. (C) Bar plot of iModulons ranked by the total number of times they were differentially activated in an ALE jump. (D) Differential iModulon activity plots (DIMA). Comparison of iModulon activities between wild-type (WT) and end-point (EP) flasks for each strain. Significant altered iModulons are colored red and noted with text.

In order to first understand the different starting points of the strains, we tested for iModulons that distinguish WT expression profiles and identify a total of 38 iModulons (Table 1, FDR<0.005). For BL21, the iModulons imply an original environment that was cold (cspA), anaerobic and nitrate rich (ArcA-2), with gluconate (GntR/TyrR), allantoin, fructose, and arabinose (AllR/AraC/FucR) as possible carbon sources. For C, the identified iModulons hint at a background with high acidity and osmotic stress (EvgA, proVWX). The low OxyR activity in Crooks implies a WT environment facing low oxidative stress while high FliA activity in MG1655 implies that high motility was advantageous to its original environment. The relatively high GadEWX in W3110 implies an original environment with high acid stress.

To understand what iModulons changed the most throughout the ALEs, we performed differential activity analysis between the WT and EP flasks of each strain (see Methods). We find a total of 57 iModulons that were differentially activated at least once amongst the different strains (P<0.05, FC¿2). The most commonly activated iModulons corresponded to stress response and amino/nucleic acid biosynthesis (Figure 5.3b). The W3110 strain had the largest number of differentially activated stress response iModulons while BL21 had the most activated amino/nucleic acid biosynthesis iModulons. With respect to the total number of differentially activated iModulons, we find that BL21 has the most while W has the least (Figure 5.3d), which reflects their respective change in growth rate. Of those activated, we find decreased activity in iModulons describing stress response (rpoS, gadEWX, rpoH, hns-related, proVWX) and motility (FlhDC, FliA, curli, fimbriae, RcsAB) while increased activity in iModulons describing translation machinery (translation), amino acid biosynthesis (ArgR, His-tRNA).

5.2.4 Linear growth-dependent transcriptome adaptations conserved in *E. coli*

While differential activity analysis identifies general regulatory trends along the trajectory, it does not directly account for changes in quantitative growth rates or similarity between strains. We thus tested for iModulons that exhibit linear growth-dependence in all strains and identify six iModulons (Figure 5.4, median Pearson —R—¿0.75, median P-value<0.05). Of the six, three are positively correlated with growth-rate and describe the expression of ribosomal genes (translation), arginine biosynthetic genes (ArgR), and nutrient response (ppGpp). The other three iModulons are negatively correlated with growth-rate and describe stress response (RpoS, GadEWX) and structural assembly (curli). These results describe growth-dependent transcriptome adaptations that are mostly conserved in the *E. coli* species.

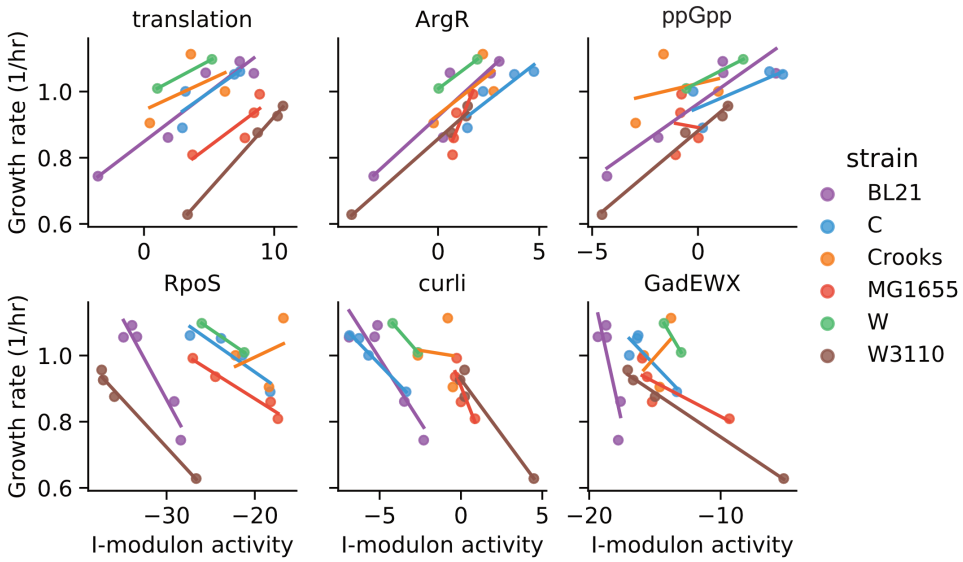

Figure 5.4: Conserved growth-dependent transcriptome. Strain-specific line plots of growth rate vs iModulon activity for six iModulons (median Pearson abs(R)¿0.75, median P-value 0.05).

5.2.5 Regulatory trade-offs governing *E. coli* adaptation

The identification of both positively and negatively growth-correlated iModulons imply the existence of regulatory tradeoffs, and thereby a lower dimensionality of the iModulon activities (i.e., increased expression of certain genes requires decreased expression of others). We thus used principal component analysis to further decompose iModulon activities. Prior to PCA, we first transform the activity matrix (flask-specific) to the difference in flask activity along the trajectory (jump-specific) in order to identify components describing general adaptation trends as opposed to strain differences (see Supplementary Figure XX for PCA of flask-specific iModulon activities). We find that the first three PCA components explain the majority of the variance and have an explained variance ratio of 40%, 28%, and 12%, respectively (Figure 5.5a). The first component describes activation of flagella machinery and is owed to the large deviation in FlhDC and FliA activity seen in the first MG1655 jump. The second component describes metal-related iModulons (Fur-1, Fur-2, iron-related, efuR-repair, Copper) and growth-correlated iModulons (RpoS, translation, ppGpp). The third PCA component primarily describes carbon metabolism iModulons (Crp-1, Crp-2, MalT,) with positive weight and stress-response and structural iModulons with opposite weight (RpoS, GadWX, GadEWX, hns-related, CspA, curli). We test for negative correlations and identify a total of six potential tradeoffs (RpoS vs translation/ppGpp, Fur-2 vs translation/ppGpp, and Fur-1 vs Copper) in component 1 and (Crp-KO vs Crp-1, Crp-2) in component 2 (Figure 5.5b-d).

5.2.6 Statistical tests leveraging ALE design reveal key mutational effects

Comparing mutations is challenging due to the large number of genetic differences between strains. We therefore leveraged the directionality of the ALE data by transforming the flask-

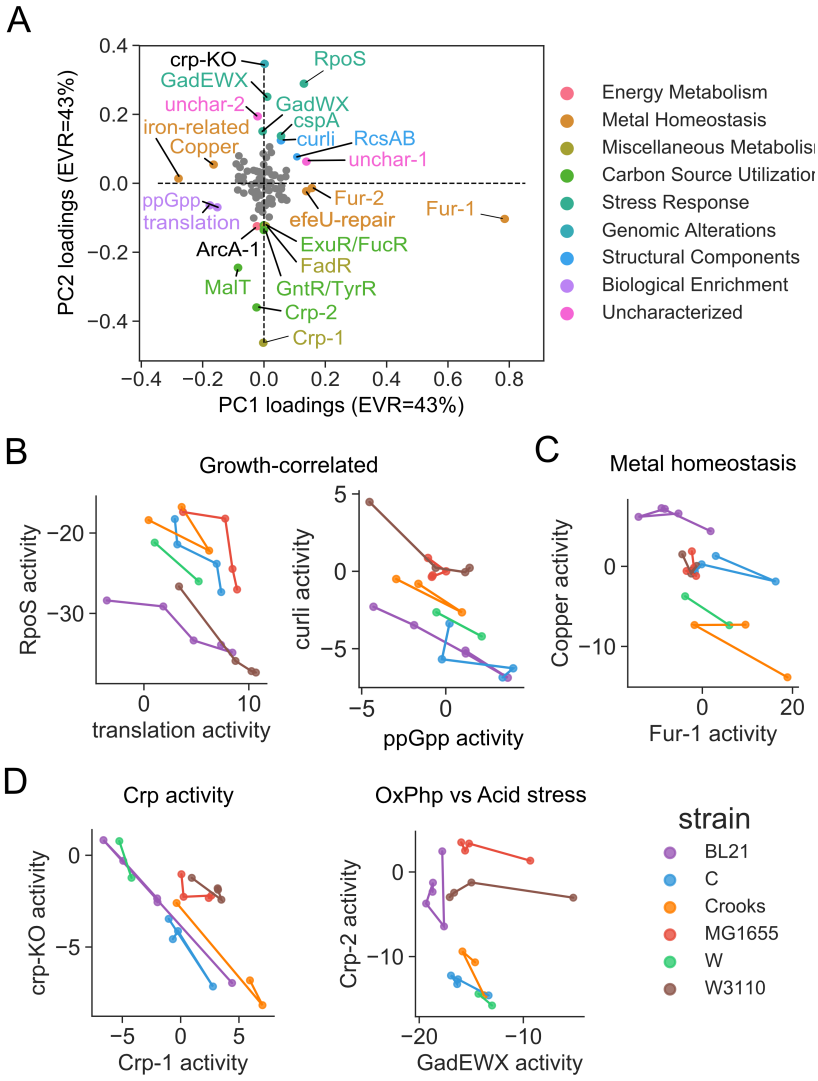

Figure 5.5: Regulatory trade-offs governing *E. coli* adaptations. (A) Plot of PCA loadings for components 1 and 2. (B-D) Strain-specific line plots for iModulon activities for trade-offs reflecting growth-correlated iModulons, metal homeostasis, crp activity, proton motive force.

specific reaction fluxes and iModulon activities to jump-specific differences in flux and activity, thereby narrowing the view of genetic differences to those selected in ALE. Using the jump-oriented perspective of the data, we then tested for significant associations between jump-specific differences in reaction flux and iModulon activity with the selection of mutations at both the nucleotide and gene levels (i.e., gene level groups two different ALE mutations together if they

appear in the same gene).

For the metabolic fluxes, we find four flux correlations primarily describing reactions involved in co-factor balancing (FDR<5%) (Figure 5.6a). Specifically, zwf mutations are correlated with ΔG6PDH flux (NADPH balance through PP pathway), pykF mutations with ΔME2 flux (NADPH balance through Malic Enzyme), and lysC with ΔSUCCOAS flux (ATP and NADPH through TCA cycle). We find that the zwf mutation in Crooks is uniquely associated with delta ΔED pathway flux. For the iModulon activities, we identify eight mutation correlates that fall into four different iModulon functional categories describing stress response, motility, structural components, and carbohydrate metabolism (FDR<5%) (Figure 5.6b-e).

We find that similar statistical tests using DE fold changes instead of iModulon activities did not uncover any significant correlations. Notably, there are only 7 cases where the selection of a mutation coincided with significant differential expression of gene (Supplementary Figure 2). Factorization-based analysis therefore enables statistical associations between the transcriptome and selected mutations.

5.3 Discussion

Taken together, our total analysis of the multi-strain ALEs revealed metabolic and transcriptomic adaptations principles of the *E. coli* species. Characterization of the phenotypic data showed specific convergent and divergent features between the WT and EP flasks of these strains. It remains open how many peripheral phenotypes change with the core genes. Since the experimental condition was glucose minimal media, it remains unclear what principles are specific to glucose minimal media and which ones are not. Future studies may gain deeper insight by diversifying the measured phenotypes of these strains through high throughput approaches such

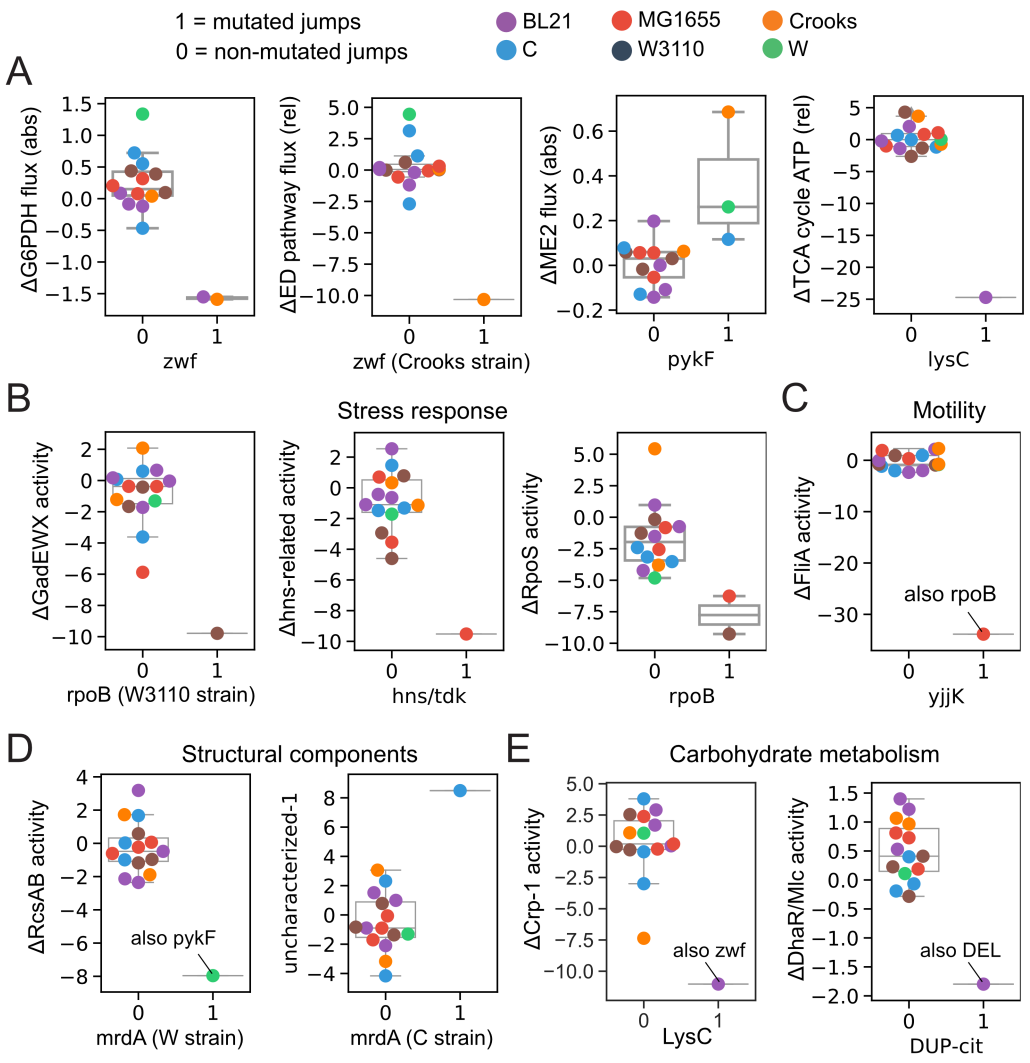

Figure 5.6: (A) Boxplots of significant correlations between mutations and changes in metabolic fluxes. The terms "abs" and "rel" in parenthesis refer to absolute flux (mmol/gDw/h) and relative flux (mol/mol gluc), respectively. (B-E) Boxplots of significant correlations between mutations and changes in iModulon activities. The boxplots are grouped by iModulon functional category. Genes with strains in parenthesis note a strain-specific mutation correlation. Mutations are grouped at the gene-level unless otherwise.

as biolog plates. Our ICA-based analysis of the transcriptome revealed key growth-correlated gene sets and tradeoffs governing *E. coli* adaptation. By leveraging ALE design and the ICA-determined iModulon weights, we identified . Many of these associations make sense (i.e., zwf with Δ G6PDH flux, hns/tdk with Δhns-related iModulon activity) while others provide novel

insights. Together, our results point to energy balance and proteome allocation (stress response, structural components, motility) as the dominant constraints governing *E. coli* adaptation. Including more samples would increase the identification of metabolic and regulatory features associated with mutations, providing a more clear picture of the logic underlying evolutionary selection. Our results show that fluxomics and transcriptomics data types are valuable data types for characterizing adaptive landscapes.

Acknowledgements

We are grateful to the Novo Nordisk Foundation (NNF10CC1016517) and the NIH NIAID (Grant U01AI124316) for their support.

Chapter 5 is a reprint of the material: **ES. Kavvas**, MR. Antoniewicz, C. Long, Y. Ding, JM. Monk, BO. Palsson, A. Feist. (2020). Laboratory evolution of multiple *E. coli* strains reveals unifying principles of adaptation but diversity in driving genotypes. *bioRxiv* DOI:10.1101/2020.05.19.104992. The dissertation author is the primary author.

5.4 References

1. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental Escherichia coli population. en. *Nature* **489,** 513–518. ISSN: 0028-0836, 1476-4687 (Sept. 2012).

2. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Applied and environmental microbiology* **81,** 17–30. ISSN: 0099-2240, 1098-5336 (Jan. 2015).

3. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. en. *Nature* **551,** 45–50. ISSN: 0028-0836, 1476-4687 (Nov. 2017).

4. Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Médigue, C., Schneider, D. & Lenski, R. E. Tempo and mode of genome evolution in a 50,000-generation experiment. en. *Nature* **536,** 165–170. ISSN: 0028-0836, 1476-4687 (Aug. 2016).

5. Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E. & Kim, J. F. Genome evolution and adaptation in a long-term experiment with Escherichia coli. en. *Nature* **461,** 1243–1247. ISSN: 0028-0836, 1476-4687 (Oct. 2009).

6. Kryazhimskiy, S., Rice, D. P., Jerison, E. R. & Desai, M. M. Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. en. *Science* **344,** 1519–1522. ISSN: 0036-8075, 1095-9203 (June 2014).

7. Monk, J. M., Koza, A., Campodonico, M. A., Machado, D., Seoane, J. M., Palsson, B. O., Herrgård, M. J. & Feist, A. M. Multi-omics Quantification of Species Variation of Escherichia coli Links Molecular Features with Strain Phenotypes. en. *Cell systems* **3,** 238–251.e12. ISSN: 2405-4712 (Sept. 2016).

8. Long, C. P., Gonzalez, J. E., Feist, A. M., Palsson, B. O. & Antoniewicz, M. R. Dissecting the genetic and metabolic mechanisms of adaptation to the knockout of a major metabolic enzyme in Escherichia coli. en. *Proceedings of the National Academy of Sciences of the United States of America* **115,** 222–227. ISSN: 0027-8424, 1091-6490 (Jan. 2018).

9. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. The Escherichia coli transcriptome mostly consists of independently regulated modules. en. *Nature communications* **10,** 5536. ISSN: 2041-1723 (Dec. 2019).

10. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. *The Escherichia coli Transcriptome Consists of Independently Regulated Modules* en. Apr. 2019.

Chapter 6

Conclusions

The advent of high throughput next-generation sequencing has ushered in a new era in biology, allowing quantitative understanding of processes within the cell on an unprecedented scale. The development of data modeling approaches that are both predictive and interpretable are therefore required for transforming the explosion of biological datasets to valuable knowledge. Such data-driven insights will have a broad impact, ranging from aiding drug development to microbial engineering. In this dissertation, we develop and apply biologically-interpretable models to various omics data types describing microbial diversity to improve our understanding of *M. tuberculosis* drug resistance evolution and the multi-scale *E. coli* adaptive landscape.

In the first chapter of this dissertation "Machine learning of *M. tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance", we applied classical machine learning to a large genomics dataset in order to gain insight into the genetic basis of AMR. Our reference-agnostic pan-genome approach was sufficient in capturing AMR variants through common univariate statistical tests. Our machine learning approach levering L1-regularization, bootstrapping, and model averaging was able to identify 33 known AMR genes and 24 novel candidates.

Further analysis of these genes through epistatic and structural analysis provided hypothesis for how these mutations may enable AMR.

The second chapter of this dissertation "An updated genome-scale model of *M. tuberculosis* H37Rv metabolism" describes an updated and standardized metabolic reconstruction of *M. tuberculosis* strain H37Rv. The new GEM improves gene essentiality predictions over previous models, computes biologically meaningful flux states, and captures 72% of known AMR genes. This study provided a stepping stone for the third thesis chapter.

The third chapter of this dissertation "A biochemically-interpretable machine learning classifier for microbial GWAS" describes a novel mechanism-based machine learning model, named Metabolic Allele Classifier (MAC), that is able to not only identify key genetic variants, but also identifies mechanistic explanations for the predictive genotype-phenotype mapping. We show that FBA—and any linear program— can behave as a ML classifier through design of the objective function. For validation, we show that the MAC achieves prediction accuracy on par with mechanism-agnostic machine learning models (isoniazid AUC=0.93). Application of the MAC to three antibiotics recovers the primary and second genetic determinants of AMR, and also estimates detailed metabolic explanations for the accuracy predictions that elucidate allele-specific effects and pathway-level antibiotic effects. Our inferred mechanisms are consistent with the literature and provide novel hypothesis for tackling AMR. We expect our method to have a major impact in the field of genetic associations due to the value of providing mechanistic biological explanations.

In the last chapter of this dissertation "Laboratory evolution of multiple *E. coli* strains reveals unifying principles of adaptation but diversity in driving genotypes" we examine multi-scale principles of the *E. coli* adaptive landscape that are either conserved or strain-specific.

Application of independent component analysis enables the identification of 6 conserved transcriptomic adaptations that describe decreasing stress response and structural component genes and increasing ribosome and amino acid biosynthesis genes. Our ALE-based statistical tests were able to associate causal mutations with specific changes in metabolic states and transcriptomic profiles, which reveal co-factor metabolism and the stress response proteome as dominant evolutionary constraints.

Multi-omics datasets of microbial adaptation have the potential to revolutionize industrial biotechnology and medicine if meaningful insights are extracted from them. In this dissertation we develop novel mathematical methods to cover three aspects of microbial adaptive landscapes: genetic determinants of AMR evolution, metabolic determinants of AMR evolution, and lastly the multi-scale genetic, metabolic, and transcriptomic adaptation principles conserved in *E. coli* . These findings improve our understanding of microbial adaptation and pave the way for better treatment regiments and more effective engineering of *E. coli* for industrial biotechnology.

Appendix A

Machine learning of *M. tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance - Supplementary Information

A.1 Methods

A.1.1 *M. tuberculosis* strain dataset

The selected set of *M. tuberculosis* strains are representative of various antimicrobial resistance phenotypes, geographic isolation sites, and genetic diversity. References for the published and unpublished data sets can be found in Supplementary Table 5. The sequencing data for the TB Antibiotic Resistance Catalog (TB-ARC) projects (Supplementary Table 5) were generated

at the Broad institute. Additional information for each of these unpublished projects can be found at the Broad Institute website. Because Africa exhibits the most diverse set of *M. tuberculosis* strains in the world [1], a third of our strains were isolated there (Supplementary Fig. 1). Furthermore, the chosen dataset constitutes a wide spectrum of isolation hotspots, ranging from 144 strains in Sweden to 141 strains in Belarus. Notably, 78 strains were isolated from South Korea, a country that has endured a significant increase in *M. tuberculosis* incidence since 2005 [2]. In total, 70% of the selected strains were in "high burden countries" [2].

A.1.2 *M. tuberculosis* pan-genome construction and QA/QC

We employed QA/QC of the constructed 1595 pan-genome by initially filtering out outlier strains. The initial selection of 1603 strains was reduced to 1595 upon review of both the cluster size distribution and the number of unique clusters across the set of all strains (Supplementary Fig. 3a-b). We found only 4 strains in the PATRIC database that had either a very low (¡2000) or high number (¿5500) of clusters. The final selection of 1595 strains has a cluster size distribution between 3900 and 4400, and a reasonable unique cluster distribution where the number of unique clusters did not exceed 160 (note that unique is defined here as being in only one strain). The pan-genome of all 1595 strains was constructed by clustering protein sequences based on their sequence homology using the CD-hit package (v4.6). CD-hit clusters protein sequences based on their sequence identity [3]. CD-hit clustering was performed with 0.8 threshold for sequence identity and a word length of 5.

A.1.3 Pan-genome core and unique cutoff determination

We determined the core and unique pan-genome through sensitivity analysis by plotting the change in core and unique cutoff values by the change in percentage. The cutoffs were chosen to be at the point where the second derivative of the curve is the largest. The curve represents the change in pan-genome core percentage to changes in the number of strains a gene must be found in to be defined as core (Supplementary Fig. 3c-d).

A.1.4 Phylogenetic Tree and categorization of lineages

We created a robust phylogenetic tree of the 1,595 strains using SNPs at the core genome. Specifically, we chose a set of 2,803 core genes that appeared in at least 1,593 strains, included the H37Rv reference strain (83332.12). We used needle [4] to align sequences within the 2803 pan-genome clusters (a cluster is representative of a particular loci) to the H37Rv reference allele. We built a binary SNP matrix using all of the SNPs identified from the 2803 genes (21,206 SNPs in total), and then estimated a maximum-likelihood phylogeny using RaXML version 8 [5]. The tree was visualized using iTOL [6].

We used an existing SNP typing scheme [7] for categorizing the strains into lineages and sublineages. Specifically, we used a total of 141 SNPs for identifying lineages and sublineages for our 1595 TB strains. These SNPs were previously determined to be sufficient for categorizing lineages [7]. Of these SNPs, 61 were in non-synonymous sites and the other 70 were SNPs found in drug resistance genes. These 141 SNPs comprised a total of 74 genes. The presence of SNPs were then used to categorize the strains into the defined lineages [7]. Of the 1595 strains, 1366 strains were categorized and 229 were uncategorized. The remaining 229 strains were categorized according to their proximity to strains with lineage-defining SNPs, with proximity

defined according to our core genome SNP phylogeny. We have included the frequency of lineage variants in order to help readers discern between epistatic alleles and those in tight linkage (Supplementary Table 6). Implicated co-occurring alleles that span different lineages are unlikely to be in tight linkage (i.e., hitchhikers). We determined the lineages of our set of *M. tuberculosis* strains using previously defined lineage/sub-lineage SNPs [7].

For the numeric subscripts shown in Figure 2—describing the number of unique sublineages for each allele-allele pair—were determined as the maximum number of unique sub-lineages at a single branch amongst all lineage/sublineage branches.. For example, an allele co-occurrence which has strains in both lineage 1.1 and 1.2 counts as two sublineages. An allele co-occurrence which has strains in both lineage 1.1, 1.1.2, 1.1.3, 1.1.3.1, 1.1.3.2, and 1.1.3.3 counts as three sublineages (1.1.3.1, 1.1.3.2, and 1.1.3.3). If an allele co-occurrence has strains in sublineages 4.1, 4.1.2, and 4.1.2.1, then only one sublineage is counted, since the strains can be traced through a single lineage (4.1 to 4.1.2 to 4.1.2.1).

A.1.5 Identification of key resistance-conferring genes with mutual information, chi-squared, and ANOVA

Mutual information (MI) has many statistical benefits which include being a nonparametric method that can quantify nonlinear relationships, unlike Pearson's correlation which measures a linear relationship. MI has proven to be a natural and powerful means to equitably quantify statistical associations in large datasets [8]. The pairwise mutual information was calculated for each column vector in the unique variant pan-genome with each drug susceptibility vector (Supplementary Fig. 3g). The discrete entropy calculations were carried out using the Non-Parametric Entropy Estimation Toolbox (NPEET, https://github.com/gregversteeg/NPEET).

Since both vectors are binary, the "naive" implementation of discrete entropy estimation used in NPEET is sufficient. The formula for calculating MI is as follows (2):

where X is the presence(1)/absence(0) distribution of a specific allele across the 1595 strains and Y is the resistant(1)/susceptible(0) distribution of a specific drug across the 1595 strains, and $x \in [0, 1], y \in [0, 1]$. The top 40 MI associations for 11 drugs are recorded (Supplementary Data File 1). Associations were similarly calculated with chi-squared and ANOVA tests. P-values were adjusted using the Bonferroni multiple-hypothesis testing correction. Theses statistical tests and corrections were implemented using the python package, statsmodels [9]. The top 40 associations determined by chi-squared and ANOVA F-test were recorded for 10 AMR classifications are recorded (Supplementary Data File 1).

A.1.6 Allele feature selection through ensemble Support Vector Machine

The Support Vector Machine (SVM) attempts to account for all variants together by learning a multidimensional hyperplane that best separates the susceptible and resistant strains. The resulting hyperplane is a function of all exact-variant vectors in the pan-genome. Since the goal is not to predict resistance with high accuracy, but to instead extract key insights from the data, we take a more "loose" approach by gearing the linear SVM with an L1-norm penalty and stochastic gradient descent optimization algorithm. It is known that there is a tradeoff between accuracy and feature selection. The L1-norm enforces sparsity in the decision function, which is ideal for feature selection. The stochastic gradient descent algorithm, in conjunction with the L1-norm, returns a different set of significant features each run. Since the chosen SVM does not reach the same solution, we look at the ensemble of 200 SVM feature selection simulations. Furthermore, we performed bootstrapping by randomly selecting a subpopulation representing

80% of the training data for each SVM simulation.

Prior to simulation, we took out the primary resistance-conferring gene of an antibiotic from the machine learning analysis of other antibiotics in order to amplify the signal of other genes - a pre-processing step previously utilized in AMR gene identification studies [10] (Supplementary Table 2). For example, all katG alleles were only accounted for as features in the machine learning analysis for isoniazid. Furthermore, we removed all mobile element proteins, PE/PPE/PE-PGRS proteins, transposases, and hypothetical proteins from consideration in the machine learning analysis due to primarily appearing in the accessory and unique pan-genome of *M. tuberculosis* which may confound the results, as previously discussed in the pan-genome analysis. Finally, we balanced the class weight in the SVM algorithm in order to account for the imbalance of resistant and susceptible strains seen for each drug in our dataset.

Features were selected from the SVM based on a threshold value. The value was determined through 10-fold cross-validation where the threshold value was optimized through grid search (Supplementary Table 2). The use of bootstrapping in the machine learning algorithm may account for biased subpopulations in the data, which often confounds GWAS analysis for *M. tuberculosis* [11, 12].

A.1.7 Determination of potential epistatic genes from SVM ensemble correlations

Leveraging machine learning towards identification of genetic interactions, we constructed a correlation matrix of allele weights across the ensemble of randomized SVM hyperplanes for each antibiotic (Supplementary File 3). We limited our machine learning analysis to AMR classifications that achieved an average AUC (i.e., average area under ensemble of receiver-

operator curves) greater than 0.80 (Supplementary Fig. 5). We selected the top 100 gene-gene correlations that include genes in the top 25 ranked SVM alleles for each antibiotic. We limited the correlations to in the top 25 ranked alleles in order to avoid the case when low weighted alleles appear sparsely with other low weighted alleles which lead to significant correlations. The resulting set of gene-gene pairs were then analyzed using a logistic regression model in order to determine statistically significant interactions. The filtering of potential gene-gene pairs prior to classical quantitative epistasis analysis addresses the problem of combinatorial explosion of pairwise interaction terms in conventional techniques. Identification of significant epistatic interactions using logistic regression models. We utilized logistic regression to identify significant epistatic interactions. A logistic regression model was built for each potential gene-gene pair previously determined by the ensemble SVM correlation analysis. The variables of the gene-gene logistic regression model were composed of both alleles and allele-allele interaction variables (3).

The interaction variables, aibj, were limited to those in which the two alleles co-occur in at least one strain. The interaction variable was the dot product of the two allele presence-absence vectors. In order to account for collinearity in the variables, we applied the following three filtering criteria (note that ai is interchangeable with bj):

1. If the allele ai presence-absence is the same as the interaction aibj presence-absence, remove the aibj interaction variable from the logistic regression model

2. If the allele ai presence-absence is equal to allele bj presence-absence, remove both variables as well as the allele-allele interaction variable, aibj.

3. If the allele ai presence-absence is equal to the sum of all interaction variables involving that allele (i.e., aibj for all j), remove the allele variable but keep the interaction variables.

We filtered for allele-allele interactions with P-value ¡ 0.05 after Benjamini-Hochberg multiple-testing corrections. The resulting set of gene-gene interactions encompassing significant allele-allele interactions were portrayed through allele co-occurrence tables (Supplementary Data File 4). Logistic regression and statistical tests were implemented using the python package statsmodels [9].

A.1.8 Calculation of log odds ratio visualized in allele co-occurrence tables

The odds ratio of each cell in the allele co-occurrence tables was determined by the following equation (4),

where BR is the number of strains that have both alleles and are resistant to the specified antibiotic, NR is the number of strains that do not have both alleles and are resistant to the specified antibiotic, BS is the of strains that have both alleles and are susceptible to the specified antibiotic, NS is the number of strains that do not have both alleles and are susceptible to the specified antibiotic. For a single allele, the odds ratio was calculated the same way with each variable representing the single allele case. If any of the four values (BR, BS, NR, NS) were zero, 0.5 was added to each value in order to ensure a value when computing the logarithm of the odds ratio.

A.1.9 Missing alleles in allele co-occurrence tables counts

The lack of specific alleles shown in the allele co-occurrence table is due to strains missing some alleles. For example, embB allele 5 is found in 147 strains but only 144 strains have both embB allele 5 and ubiA allele 2 (Fig. 2). Specifically, the three strains missing the three ubiA alleles are the following PATRIC strains as described by their genome identifiers: 1423432.3,

1448794.3, 1448824.3. Searching on the PATRIC database for either "ubiA" or "Rv3806c" results in 0 hits for these organisms. While it is unlikely that the strain is missing this allele, these limitations are not due to the analysis but instead results from the selection of strains. These events happen quite rarely and were accounted for in the partitioning of pan-genome portions. The large sample size was able to recapitulate the key genes due to large sample size.

A.1.10 Structural protein analysis of identified AMR genes

For identified AMR genes, the ssbio software was used to gain gene-specific, protein sequence and structure based information about residue-level changes (SNPs and deletions) present in the *M. tuberculosis* alleles [13]. Each AMR gene was mapped to a reference protein sequence file obtained from UniProt [14] and sequence-based metadata identifying protein-specific features (e.g. active sites, secondary structures, and mutations in studied wild-type strains) was used to determine the occurrence of allele-specific AMR mutations within the gene feature set (Supplementary Table 4). When available, AMR genes were additionally mapped to experimentally obtained protein structures from the RCSB Protein Data Bank or to homology structures generated using the Iterative Threading ASSEmbly Refinement (I-TASSER) platform [15, 16]. To help elucidate the mechanistic effects of AMR mutations, both AMR mutations and the residue-level feature set were mapped to these structures and visualized using the NGLview Jupyter notebook plugin [17]. The structural information was utilized to calculate distances between each mutation and annotated protein feature (Supplementary Table 4).

A.2 Supplementary Notes

A.2.1 Characteristics of 1,595 Strain *M. tuberculosis* dataset

The chosen strains come from a wide range of studies [18–34]. Because Africa exhibits the most diverse set of *M. tuberculosis* strains in the world [1], a third of our strains were isolated in South Africa (Supplementary Fig. 1a). Furthermore, the chosen dataset constitutes a wide spectrum of isolation hotspots, ranging from 144 strains in Sweden to 141 strains in Belarus. Notably, 78 strains were isolated from South Korea, a country that has endured a significant increase in *M. tuberculosis* incidence since 2005 [2]. In total, 70% of the selected strains were in "high burden countries" [2].

A.2.2 Characterizing the *M. tuberculosis* pan-genome

Following selection of the representative set of *M. tuberculosis* genome sequences, we determined the pan-genome (i.e., the union of all genes across all strains) represented by these data (Methods). We categorized the genome content across all 1,595 strains as "core" (the set of genes shared by at least 1590 strains), "accessory" (the set of genes present in some, but not all, strains), or "unique" (the set of genes found in at most 5 strains) [35, 36]; the cutoffs for each of these categories were evaluated using sensitivity analyses (Methods). The resulting pan-genome consisted of 11,039 clusters, where each cluster represents a grouping of protein variants determined to be sufficiently similar to each other (i.e., ¿80% sequence similarity). Using these partitioning criteria, the core, accessory, and unique genomes were composed of 3,419 genes (31%), 2,402 genes (21.8%), and 5,218 genes (47.3%), respectively (Supplementary Fig. 2a). The core genome made up 80% of the average genome in our dataset, a result in agreement with the hypothesis that *M. tuberculosis* is a clonal species [37]. This diversity is in stark contrast

to that of Escherichia coli, which has a core genome percentage estimated to be between 20% and 50% of the average full E. coli genome [38], and Staphylococcus aureus, where we recently calculated the core genome to comprise 56% of the average genome [36]. Furthermore, we found that virulence factors were highly conserved in the *M. tuberculosis* core genome (93%, 414/445 genes) (Supplementary Table 1 and Supplementary Note).

The remaining 7,620 genes that comprise the accessory and unique genomes represent the genetic variability across *M. tuberculosis* strains. A significant portion of the unique and accessory genome was attributed to Pro-Glu (PE)-related proteins and hypothetical proteins (Supplementary Fig. 2b). Specifically, PE-related proteins represent products that contain the characteristic motifs Pro-Glu (PE), Pro-Pro-Glu (PPE), or polymorphic GC-rich sequence motifs (PE-PGRS) [39] and make up approximately 10% of the average *M. tuberculosis* coding capacity [40]. Because of significant variation in both PE-related proteins and hypothetical proteins, we computed the shape of the pan-genome by filtering out PE/PPE genes and genes with lengths that were significantly longer (¿1 standard deviation) than the mean gene length of 1000 bp, which are likely result of sequencing or annotation errors. In total, this led to the removal of 1,335 genes clusters from the pan-genome. The majority of these genes (826) were PE/PPE genes. Following the removal of these genes we find that the pan-genome is closed for our 1595 strains of *M. tuberculosis* (Supplementary Fig. 2c).

A.2.3 Pan-genome COG Categories

We used eggNog with the eggNog-mapper tool [41] to functionally categorize the pan-genome into Clusters of Orthologous Groups (COGs) [42] (Supplementary Fig. 3f). We filtered out clusters annotated as PE genes or those marked as hypothetical proteins in order to focus on

the functionally characterized pan-genome. The core genome made up less than 50% of the clusters annotated with defense mechanisms (V), signal transduction mechanisms (T), inorganic ion transport and metabolism (P), and secondary metabolism (Q) COGs. In contrast, the core made up more than 70% of clusters annotated with intracellular trafficking, secretion, and vesicular transport (U), and translation, ribosomal structure and biogenesis (J).

A.2.4 Virulence factors are highly conserved in the core genome

The pathogenicity of *M. tuberculosis* can be partly attributed to its unique set of virulence factors, whose variable distribution may provide further insight into pathogenic requirements. Thus, we determined the distribution of 445 virulence factors, curated by the PATRIC database [43], across the constructed pan-genome. Of the 445 virulence factors, 7.0% (31 genes) were in the accessory genome and 93.0% (414 genes) were in the core genome (Supplementary Table 1). Of the 31 accessory virulence genes, 17 were PPE/PE/PGRS genes (Supplementary Table 1). Also partitioned in the accessory genome was a set of six virulence factors composed of genes encoding the phospholipases C (plcC, plcD, plcA, and plcB) [44], and lipR (a lipolytic esterase). The remaining eight virulence factors found in the accessory genome were kdpD, mceC, rpfA, trpD, aceAa, ribA1, Rv0969, and ctpV, and two ESAT-6 like proteins, esxG and esxH. esxG and esxH comprise part of the ESX-3 secretion system involved in mycobactin-mediated iron acquisition but may play an additional role in virulence [45]. The isocitrate lyase subunit (aceAa) is a nonessential gene within the glyoxylate shunt and is downregulated in antibiotic conditions [46].

In addition to virulence factors, we investigated the "CD4 counteractome"—defined as the specific set of genes necessary for coping with the immune environment generated by CD4 T

cells [47]. We found that all of the genes were partitioned in the core genome with the exception of a trpD, Rv1053, and three adenylate cyclases (Rv1358, Rv1359, and Rv1319c) (Supplementary Table 1). Interestingly, the existence of an alternative tryptophan biosynthesis pathway suggested by [48, 49] is supported by the partitioning of trpD in the accessory genome.

Among the accessory genes found in the virulome and counteractome, trpD (anthranilate phosphoribosyltransferase) stood out as it is an essential tryptophan biosynthesis gene. Interestingly, in a study comparing trpE and trpD deleted strains, it was found that the trpE deleted strains had a 100,000 fold loss of viability after 2 weeks in contrast to the trpD deleted ones which could not achieve such a level after 13 weeks[48, 49]. Zang et al. hypothesized that such a difference could either be due to either "an accumulation of intermediary metabolites or an as of yet undescribed alternative tryptophan biosynthesis pathway"[49]. In our case, the partitioning of trpD to the accessory genome could either be due to the absence of trpD in 1000+ strains or due to trpD having significant sequence variability. A quick check on the PATRIC database corroborates our findings in that many strains lack an annotated trpD. Given the drastic experimental differences between trpD and trpE deletions and the rare occurrence of accessory virulence factors, we believe that the significant absence of trpD in the constructed pan-genome supports the claim that there is an alternative Tryptophan biosynthesis pathway in *M. tuberculosis* .

A.2.5 Motivation for using mutual information and observation of shared AMR signals across multiple antibiotics

In our study, mutual information (MI) was used to quantify the dependence between the labeled phenotype distribution of a specific drug (resistant or susceptible) and the distribution of a specific variant (presence or absence), across all tested strains (Supplementary Figure 3g). MI

was chosen due to having many statistical benefits, which include being a nonparametric method that can quantify nonlinear relationships unlike Pearson's correlation which measures a linear relationship. MI has proven to be a natural and powerful means to equitably quantify statistical associations in large datasets [8]. In addition to key AMR genes (Fig. 1), mutual information picks up a other known resistance-conferring genes including ethA (Rv3854) [50], papA2 (Rv1182) [51], drrA (Rv2936) [52], drrB (Rv2937) [52], gidB (Rv3919c) [53], moeW (Rv2338c) [54] and ubiA (Rv3806c) [55, 56] (Supplementary Data File 1).

MI showed that the variants associated with the highest signals are often those representative of susceptible rather than resistant phenotypes, thus indicating that knowledge of the presence of a susceptible variants in *M. tuberculosis* holds more informational value in determining the AMR phenotype.

It is important to note that *M. tuberculosis* treatment consists of the combined use of multiple drugs, which in turn make many *M. tuberculosis* strains (reflected in their genomes) resistant to multiple antibiotics. Therefore, it comes as no surprise that key resistance-determining genes showed up as tall peaks with other drugs (Fig. 1). These multi-antibiotic resistant *M. tuberculosis* strains make relating a specific variant to a AMR challenging [10, 57].

A.2.6 Motivation of ensemble support vector machine and limitations

Although simple and effective, mutual information does not account for the relationship between interacting alleles since the pairwise calculations consider variants independently of one another. In order to uncover possible structures in our dataset related to AMR, we used a Support Vector Machine (SVM) to select AMR-associated alleles. We introduced both unstable and randomized behavior in the SVM by using an L1-norm penalty and stochastic gradient

descent. A "noisier" SVM was used in order to address the following two inherent biases in the AMR data: (1) that the binary AMR phenotype (resistant or susceptible) is biased towards in vitro drug testing conditions, and (2) that the binary AMR phenotype does not account for varying levels of drug efficacy which may determine high level resistance. We looked at an ensemble of noisy SVM simulations for each drug in order to get a notion of significance (genes that pop out in many simulations are more likely to be significant) (Methods).

The unstable and randomized SVM method may slightly relieve the bias introduced by the AMR phenotypes (resistant or susceptible) experimentally determined from in vitro testing conditions. As noted earlier, the host environment of *M. tuberculosis* is drastically different from the one encountered in the petri dish, and such differences influence the efficacy of drugs [58]. Moreover, the AMR phenotype is binary and does not consider variation in the drug concentration profiles. Therefore, "explaining resistance" by finding a minimal set of mutations that best explains the in vitro AMR phenotypes may not capture subtle genetic adaptations. Other possible influential adaptations, however, such as those under the complex resistance category that have been shown to result in varying levels of resistance [55], may be hidden within the genomic data. Thus, this "loose" machine learning method extracts features from suboptimal peaks as well as from areas surrounding the global optima. Furthermore, it is important to note that current treatments of *M. tuberculosis* infection consists of the combined use of multiple drugs, which in turn make many *M. tuberculosis* strains resistant to multiple antibiotics.

A key biomarker that was not uncovered was the streptomycin AMR-determinant, rrs, because only protein coding genes were taken into account in our analysis. We find many cell wall genes implicated in the analysis as well including pks12 [59], pks9, pks2, dprE1, pks7, pks1 , pks6, ltp1, and ddpX. Furthermore, many implicated alleles occur in sulfur metabolism including

cysK2, serA1, moaE2, mec, and metZ. The presence of cydC as an implicated gene was interesting because studies have shown that it is important for host immune response and that disruptions in cydC affect antibiotic efficacy [60, 61].

A.2.7 Detailed perspective of the presented platform-derived results.

Defining SNPs is not required for identification of AMR genes. Defining SNPs relative to the *M. tuberculosis* H37Rv reference strain has provided the foundation both for diagnostics and for identifying novel resistance-conferring mutations but has limited a comprehensive and unbiased analysis of the *M. tuberculosis* AMR mutational landscape [10, 62–64]. Our representation of genetic variation and subsequent identification of key AMR genes demonstrates that reference-based genetic variation is not required for comprehensively identifying AMR genes. Rather, by representing genetic features as exact allele sequences, each strain in our dataset contains a single genetic feature for each of its genes, which removes potential confounding effects that may arise when multiple genetic features appear in a single gene.

A.2.8 Limitations of our view of genetic variation

The primary limitation in our view of genetic variation is that we do not account for non-protein coding genes. Therefore, our analysis is unable to identify known non-protein coding genes that confer resistance such as eis and rrs. Furthermore, by only looking at protein sequences, we do not account for synonymous SNPs, which have been shown to confer resistance [55]. While we focused our view on protein-coding genes and their protein sequences, there is no limitation in the ability of our computational platform to account for non-protein coding genes and synonymous SNPs.

A.2.9 Machine learning enables increased identification of known AMR genes over GWAS.

Our results suggest that a machine learning approach that accounts for multi-dimensional correlations is more powerful than a typical GWAS-based approach that tests positions on the genome individually for association with a phenotype [65]. Implementing an ensemble SVM identified 33 known AMR genes, including an additional 7 gene-to-antibiotic relations absent from our lists derived from pairwise statistical associations. Our observation of significant correlations between embB, ubiA, and embR implied that machine learning may provide a base for the quantitative analysis of epistatic interactions. In particular, our pipeline identified an optimal mapping between multiple genetic features and AMR phenotypes. This mapping elucidates complex relations underlying AMR evolution that are hidden from simple GWAS analysis. While we utilized an SVM for its clarity, future efforts may implement machine learning methods capable of capturing more complexity, or integrate phylogenetic constraints in the optimization problem.

A.2.10 Adaptations in toxins are associated with XDR in *M. tuberculosis* .

In addition to analyzing the resistance to individual antibiotics, we looked at AMR genes predicted to contribute to MDR (multidrug-resistant, AUC: 0.96) and XDR (extensively drug-resistant, AUC: 0.92) strains of *M. tuberculosis* . In XDR cases, mazF3 (Rv1102c) appeared as the top 5th allele and vapC21 (Rv2757c) appeared as the 10th ranked allele, both of which ranked higher than alleles of known AMR determinants such as gyrA, embB, ethA, katG, thyA, ppsA, and pncA (Supplementary Data File 2). Notably, mRNA levels of mazF3 have been shown to be induced 6.0-, 8.9-, and 8-fold by isoniazid, gentamycin, and rifampicin, respectively, when grown in a non-replicating, starved state [66]. The hyperplane weights for mazF3 and vapC21

showed that mazF3 allele 6 and vapC21 allele 7 were selected as determinants for resistance and susceptibility, respectively. In addition to the mentioned XDR-associated toxins, other implicated AMR toxins that appeared across the antibiotics include mazF5 (8th rank, PAS), higA (30th rank, PAS), vapC2 (21th rank, ETH), higB (49th rank, EMB). In particular, mazF5 is part of a toxin-antitoxin module (mazF5-mazE5) that has been shown to be in the top five most differentially expressed genes in a XDR *M. tuberculosis* strain [67]. The uncovering of toxins by our machine learning approach complements and extends recent experimental studies by relating toxin variation to host-relevant AMR evolution.

A.2.11 Epistatic and protein-structure-guided generation of experimental hypothesis

Extending our sequence-based view of these implicated AMR genes by mapping alleles to protein structures provides a basis for inferring the causal driver of adaptation. We found that the two resistant-dominant alleles of oxcA uniquely share a SNP A253S located within the thiamin diphosphate-dependent enzyme M-terminal domain, which led us to hypothesize that the SNP A253S promotes acid stress resistance through increased enzyme efficiency. Observation that oxcA SNP A253S occurs in the background of katG S315T suggests the use of acidic stress and *M. tuberculosis* strains carrying the S315T harbinger mutation [30] in experimental interrogation of oxcA in high-level isoniazid resistance.

Given the difficulty of experimenting with *M. tuberculosis* —where slow growth rate, host-irrelevant media conditions, and biosafety level 3 requirements burden experimentalists—our results demonstrate that an additional interpretation of computationally-derived mutations by analysing protein structures may accelerate experimental investigation of this deadly pathogen.

Beyond mutation proximality and feature incidence, future efforts may better utilize protein structures by estimating changes in biochemical properties due to mutations, such as changes in metabolite or cofactor binding affinities [68].

A.2.12 Geographic contextualization suggests modulation of antibiotic treatment.

Our geographic contextualization of the implicated AMR genes identifies novel genetic adaptations specific to Belarus—a country that had the highest rate of MDR *M. tuberculosis* strains in the world between 2015-2016 [2]. While studies have described the genomic composition of Belarus strains in terms of the commonly used AMR genes [69], our identification of resistant-dominant alleles within Rv3848, oxcA, kdpC, dnaA, and vapC21 demonstrates that the focused view of genetic variation is limiting. Modulation of treatment regimens may reflect these genetic adaptations by removing isoniazid, streptomycin, and ethambutol. Furthermore, observation that susceptible dominant alleles of thyA, mmpL11, and ald are localized in Belarus suggests that a combinatorial antibiotic regimen based on PAS and d-cycloserine may increase the likelihood of effective MDR *M. tuberculosis* treatment. We believe that additional epidemiological perspectives should enable actionable insight to the problem of poor *M. tuberculosis* management.

A.3 Supplementary Figures

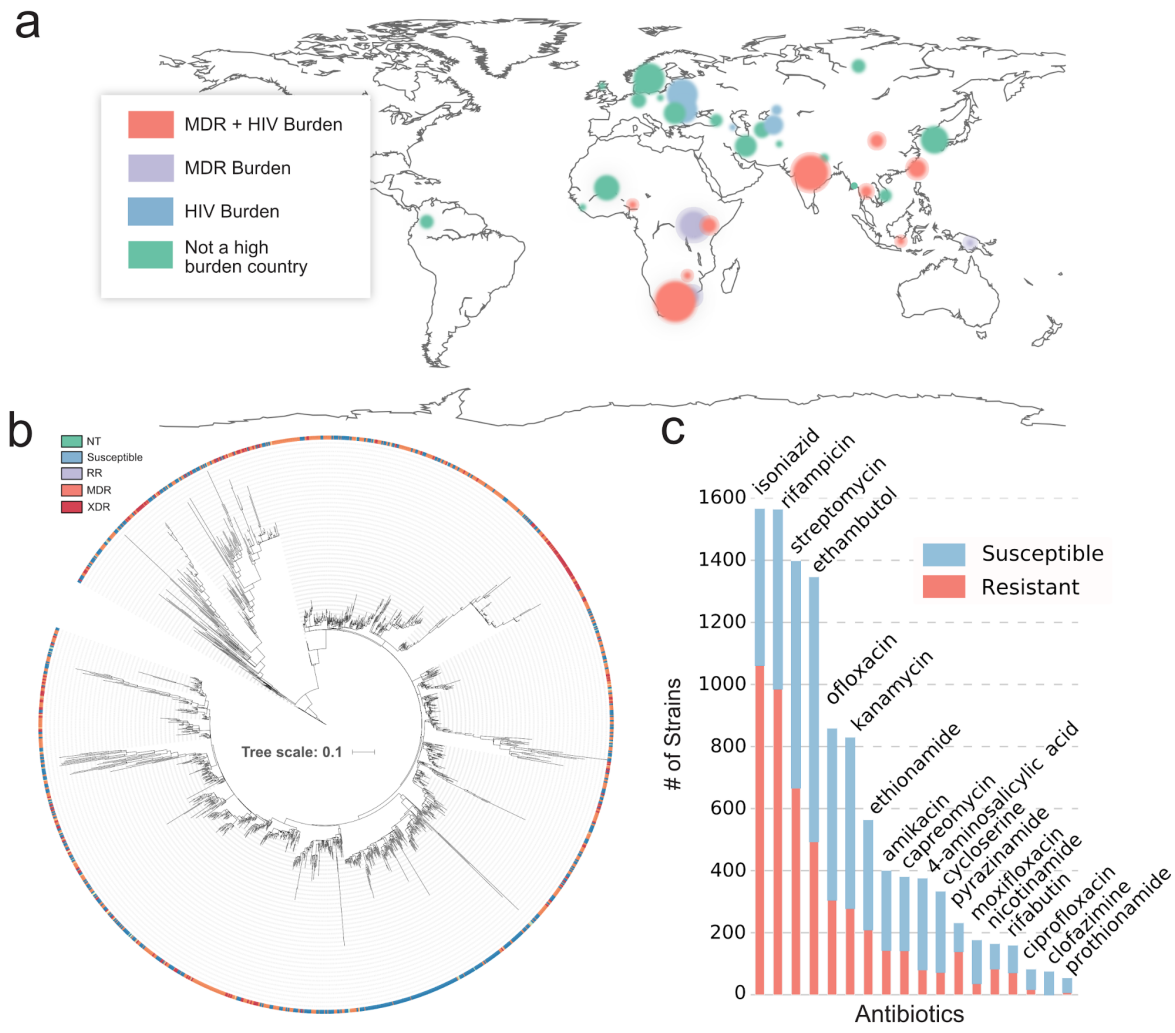

Figure A.1: Characteristics of 1595 strain dataset. *M. tuberculosis* strains were selected to span geography, resistances and phylogenetic space. (a) Geographic locations of strain isolation sites. The locations are colored according to the "high burden countries" 2016-2020 watchlist categories [2]. The size of the circles scale logarithmically with the number of strains found in that location. (b) Phylogenetic tree of the 1595 strains (Methods). (c) Specific drug characteristics tested across all 1595 strains. Abbreviations: RR, Rifampicin Resistant; MDR, Multidrug resistant; XDR, Extensively Drug Resistant; NT, Not Tested.

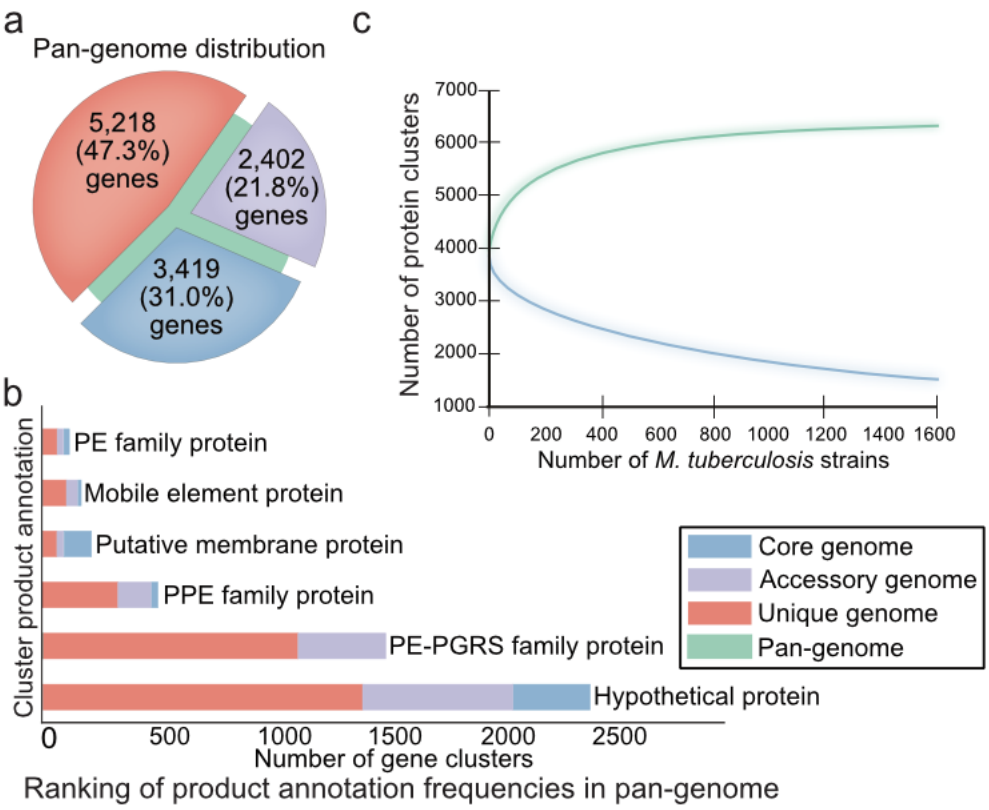

Figure A.2: *M. tuberculosis* pan-genome characteristics (a) Distribution of the core, unique, and accessory genes across the pan-genome. (b) Products annotated across the pan-genome clusters in ranked order. (c) The number of protein clusters in the pan-genome against the number of *M. tuberculosis* strains. The green line indicates the size of the pan-genome as *M. tuberculosis* strains are added to the pan-genome. The blue line indicates the size of the core genome with addition of new strains.

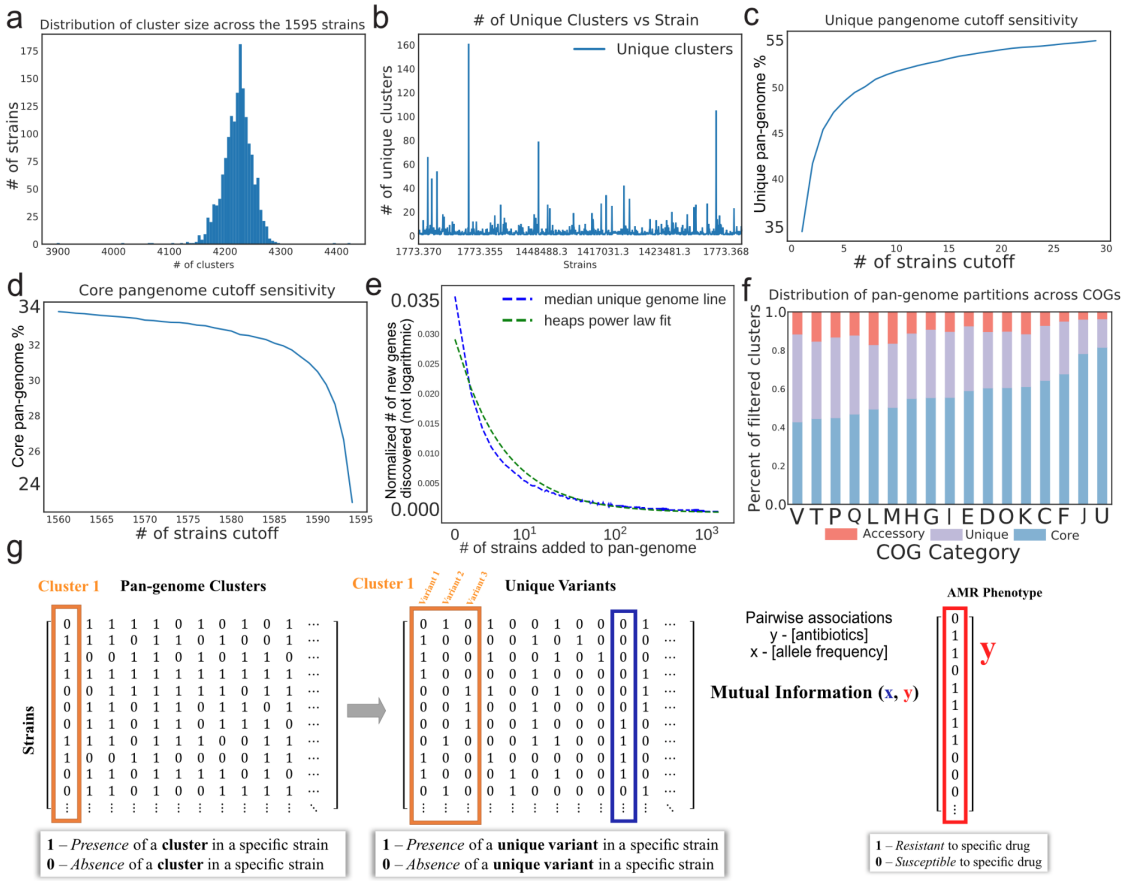

Figure A.3: Pan-genome quality check, characteristics, and allele-centric view. (a) Distribution of *M. tuberculosis* luster size across the 1595 strains. (b) Number of unique clusters per strain in our dataset. (c) Change in unique pan-genome percentage according to change in strain cutoff values. (d) Change in core pan-genome percentage according to strain cutoff values. (e) Fit of median unique genome line on Heap's power law. The y-axis is the normalized number of new genes discovered (note that this axis is not logarithmic). The x-axis is a logarithmic number of strains added to pan-genome. (f) Distribution of the functional characterized pan-genome across COG categories. (g) Higher resolution view of genetic variation and subsequent calculation of pairwise associations. The allele pan-genome was constructed by separating out sequences of exact similarity (i.e. 100% amino acid conservation) into separate columns. Therefore, each column in the allele pan-genome matrix corresponds to the frequency of a unique allele across the 1595 strains. Alleles that were found in less than 5 strains were taken out of the analysis. The mutual information between each binary absence/presence allele vector (blue) and each AMR phenotype vector (red) was taken.

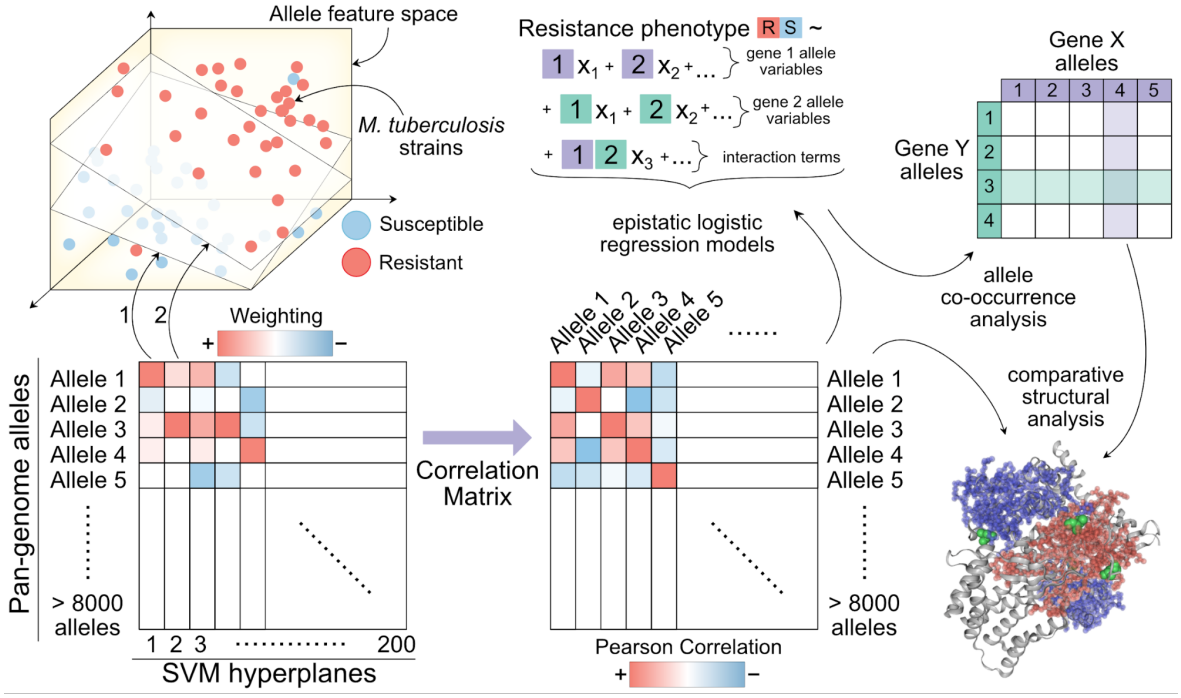

Figure A.4: Illustration of multi-layered analysis workflow. A support vector machine (SVM) was trained on random subsets of the total population with equal size (i.e., bootstrapping). The SVM utilized an L1-norm and stochastic gradient descent (SGD). Due to the randomness and L1-norm, the SVM may choose different features with different weights for each subset. Correlation matrix between the alleles was determined from the ensemble of SVMs. Large positive correlations correspond to alleles whose weights often appear together and are of the same sign (i.e., positive and positive, or negative and negative). Large negative correlations correspond to alleles whose weights often appear together but are of different signs (i.e., positive and negative). Significant correlations were evaluated using logistic regression models and visualized using allele co-occurrence tables. Mapping alleles of both high ranked genes and correlated genes concluded the quantitative analysis.

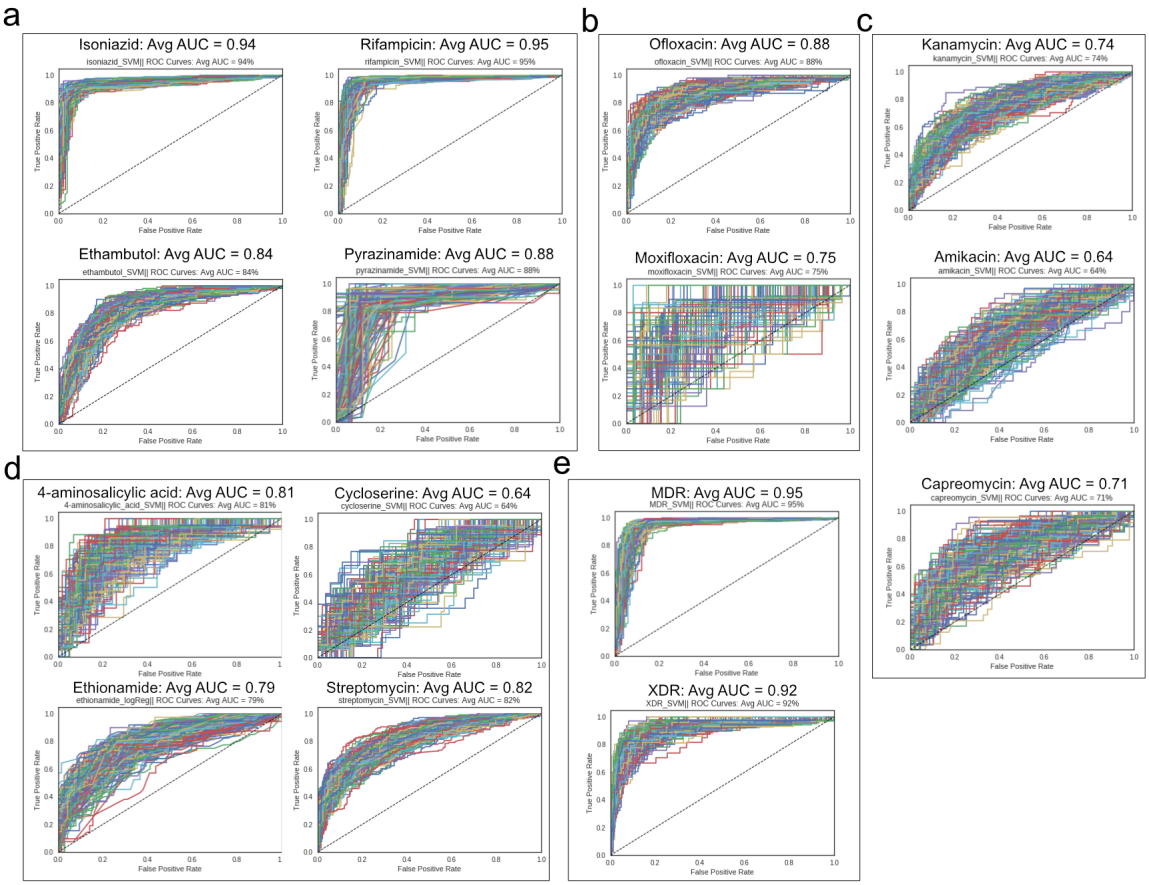

Figure A.5: Ensemble ROC curves for SGD-SVM predictions Ensemble ROC curves for SGD-SVM (stochastic gradient descent support vector machine) predictions of different AMR classifications. (a) First-line drugs: isoniazid, rifampicin, ethambutol, and pyrazinamide. (b) Second-line drugs of fluoroquinolones: ofloxacin and moxifloxacin, and (c) aminoglycosides: kanamycin, amikacin, capreomycin. (d) Other antibiotics: 4-aminosalicylic acid, cycloserine, ethionamide, streptomycin. (e) MDR (multidrug resistant) and XDR (extensively drug resistant) classification. MDR is defined as *M. tuberculosis* strains that are resistant to at least Isoniazid and Rifampicin. XDR is defined as *M. tuberculosis* strains that are MDR and resistant to at least one second line aminoglycoside (i.e., amikacin, kanamycin, or capreomycin) and resistant to at least one second line fluoroquinolones (i.e. ciprofloxacin, ofloxacin, moxifloxacin). The average AUC was calculated by averaging over AUCs for the 200 independent SGD-SVM ROC curves. The y-axis is the true positive rate and the x-axis is the false positive rate. For ethionamide, a logistic regression estimator using both an L1-norm and SGD was used instead of the SVM due to have a significantly larger AUC (0.79) than the SVM (0.71).

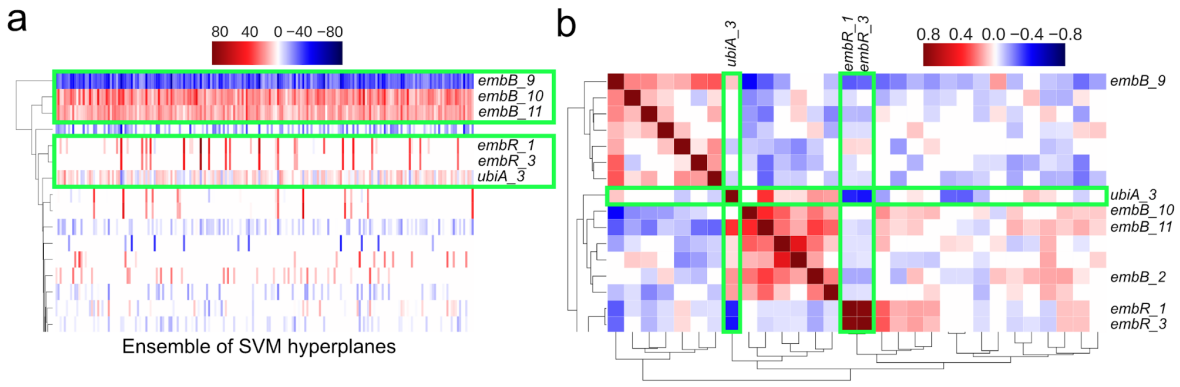

Figure A.6: Pairwise correlation of ethambutol genetic features across ensemble of SGD-SVM simulations. (a) SVM weightings across the hyperplane ensemble. The x-axis represents the iterations for each unique SVM simulation. The y-axis represents the alleles selected by each SVM. Red corresponds to a positive weighting while blue corresponds to a strong negative weighting. The alleles of embB, ubiA, and embR are highlighted in green. (b) Clustering of ethambutol allele correlation matrix. The color blue corresponds to a negative correlation while a blue color corresponds to a positive correlation. The y-axis is shown since the figure since the x-axis is the mirror of the y-axis. The alleles of embB, ubiA, and embR are highlighted in green.

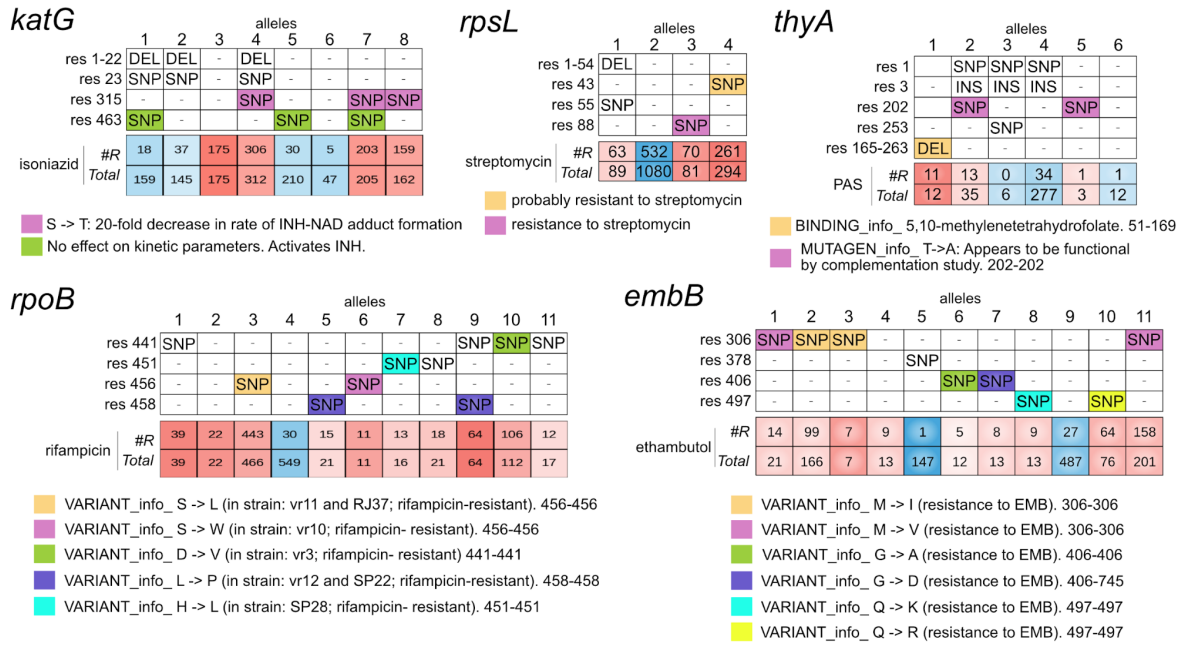

Figure A.7: Case-controls for relating MoA with uniprot annotated protein structural features. Mutation tables and uniprot color annotations are shown for katG, rpsL, thyA, rpoB, and embB.

A.4 References

1. Gagneux, S. & Small, P. M. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. en. *The Lancet infectious diseases* **7,** 328–337. ISSN: 1473-3099 (May 2007).

2. Organization, W. H. *et al.* Global tuberculosis report 2016 (2016).

3. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. en. *Bioinformatics* **22,** 1658–1659. ISSN: 1367-4803 (July 2006).

4. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. en. *Trends in genetics: TIG* **16,** 276–277. ISSN: 0168-9525 (June 2000).

5. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. en. *Bioinformatics* **30,** 1312–1313. ISSN: 1367-4803, 1367-4811 (May 2014).

6. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. en. *Nucleic acids research* **44,** W242–5. ISSN: 0305-1048, 1362-4962 (July 2016).

7. Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., Portugal, I., Pain, A., Martin, N. & Clark, T. G. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. en. *Nature communications* **5,** 4812. ISSN: 2041-1723 (Sept. 2014).

8. Kinney, J. B. & Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. en. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 3354–3359. ISSN: 0027-8424, 1091-6490 (Mar. 2014).

9. Seabold, S. & Perktold, J. *Statsmodels: Econometric and statistical modeling with python* in *Proceedings of the 9th Python in Science Conference* **57** (2010), 61.

10. Desjardins, C. A., Cohen, K. A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B. J., Shea, T. P., Almeida, D. V., Manson, A. L., Salazar, A., Padayatchi, N., O'Donnell, M. R., Mlisana, K. P., Wortman, J., Birren, B. W., Grosset, J., Earl, A. M. & Pym, A. S. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. en. *Nature genetics* **48,** 544–551. ISSN: 1061-4036, 1546-1718 (May 2016).

11. Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. en. *Nature reviews. Genetics* **18,** 41–50. ISSN: 1471-0056, 1471-0064 (Jan. 2017).

12. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. en. *Current opinion in microbiology* **25,** 17–24. ISSN: 1369-5274, 1879-0364 (June 2015).

13. Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J. M., Zhang, Z. & Palsson, B. O. *ssbio: A Python Framework for Structural Systems Biology* en. July 2017.

14. The UniProt Consortium. UniProt: the universal protein knowledgebase. en. *Nucleic acids research* **45,** D158–D169. ISSN: 0305-1048, 1362-4962 (Jan. 2017).

15. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic acids research* **28,** 235–242. ISSN: 0305-1048 (Jan. 2000).

16. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. The I-TASSER Suite: protein structure and function prediction. en. *Nature methods* **12,** 7–8. ISSN: 1548-7091, 1548-7105 (Jan. 2015).

17. Nguyen, H., Case, D. A. & Rose, A. S. NGLview - Interactive molecular graphics for Jupyter notebooks. en. *Bioinformatics.* ISSN: 1367-4803, 1367-4811 (Dec. 2017).

18. Miyoshi-Akiyama, T., Matsumura, K., Iwai, H., Funatogawa, K. & Kirikae, T. Complete annotated genome sequence of Mycobacterium tuberculosis Erdman. en. *Journal of bacteriology* **194,** 2770. ISSN: 0021-9193, 1098-5530 (May 2012).

19. Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., Supply, P., Kalinowski, J. & Niemann, S. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. en. *PLoS medicine* **10,** e1001387. ISSN: 1549-1277, 1549-1676 (Feb. 2013).

20. Wu, W., Zheng, H., Zhang, L., Wen, Z., Zhang, S., Pei, H., Yu, G., Zhu, Y., Cui, Z., Hu, Z., Wang, H. & Li, Y. A genome-wide analysis of multidrug-resistant and extensively drug-resistant strains of Mycobacterium tuberculosis Beijing genotype. en. *Molecular genetics and genomics: MGG* **288,** 425–436. ISSN: 1617-4615, 1617-4623 (Sept. 2013).

21. Majid, M., Kumar, N., Qureshi, A., Yerra, P., Kumar, A., Kumar, M. K., Tiruvayipati, S., Baddam, R., Shaik, S., Srikantam, A. & Ahmed, N. Genomes of Two Clinical Isolates of Mycobacterium tuberculosis from Odisha, India. en. *Genome announcements* **2.** ISSN: 2169-8287 (Mar. 2014).

22. Ng, K. P., Yew, S. M., Chan, C. L., Chong, J., Tang, S. N., Soo-Hoo, T. S., Na, S. L., Hassan, H., Ngeow, Y. F., Hoh, C. C., Lee, K. W. & Yee, W. Y. Draft Genome Sequence of the First Isolate of Extensively Drug-Resistant (XDR) Mycobacterium tuberculosis in Malaysia. en. *Genome announcements* **1.** ISSN: 2169-8287 (Jan. 2013).

23. Lin, N., Liu, Z., Zhou, J., Wang, S. & Fleming, J. Draft genome sequences of two super-extensively drug-resistant isolates of Mycobacterium tuberculosis from China. en. *FEMS microbiology letters* **347,** 93–96. ISSN: 0378-1097, 1574-6968 (Oct. 2013).

24. Lanzas, F., Karakousis, P. C., Sacchettini, J. C. & Ioerger, T. R. Multidrug-resistant tuberculosis in panama is driven by clonal expansion of a multidrug-resistant Mycobacterium tuberculosis strain related to the KZN extensively drug-resistant M. tuberculosis strain from South Africa. en. *Journal of clinical microbiology* **51,** 3277–3285. ISSN: 0095-1137, 1098-660X (Oct. 2013).

25. Cohen, K. A., Abeel, T., Manson McGuire, A., Desjardins, C. A., Munsamy, V., Shea, T. P., Walker, B. J., Bantubani, N., Almeida, D. V., Alvarado, L., Chapman, S. B., Mvelase, N. R., Duffy, E. Y., Fitzgerald, M. G., Govender, P., Gujja, S., Hamilton, S., Howarth, C., Larimer, J. D., Maharaj, K., Pearson, M. D., Priest, M. E., Zeng, Q., Padayatchi, N., Grosset, J., Young, S. K., Wortman, J., Mlisana, K. P., O'Donnell, M. R., Birren, B. W., Bishai, W. R., Pym, A. S. & Earl, A. M. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. en. *PLoS medicine* **12,** e1001880. ISSN: 1549-1277, 1549-1676 (Sept. 2015).

26. Ismail, A., Teh, L. K., Ngeow, Y. F., Norazmi, M. N., Zainul, Z. F., Tang, T. H., Najimudin, N. & Salleh, M. Z. Draft Genome Sequence of a Clinical Isolate of Mycobacterium tuberculosis Strain PR05. en. *Genome announcements* **1.** ISSN: 2169-8287 (June 2013).

27. Karuthedath Vellarikkal, S., Patowary, A., Singh, M., Periwal, V., Singh, A. V., Singh, P. K., Garg, P., Mohan Katoch, V., Katoch, K., Jangir, P. K., Sharma, R., Open Source Drug Discovery Consortium, Chauhan, D. S., Scaria, V. & Sivasubbu, S. Draft Genome Sequence of a Clinical Isolate of Multidrug-Resistant Mycobacterium tuberculosis East African Indian Strain OSDD271. en. *Genome announcements* **1.** ISSN: 2169-8287 (Aug. 2013).

28. Al Rashdi, A. S. A., Jadhav, B. L., Deshpande, T. & Deshpande, U. Whole-Genome Sequencing and Annotation of a Clinical Isolate of Mycobacterium tuberculosis from Mumbai, India. en. *Genome announcements* **2.** ISSN: 2169-8287 (Mar. 2014).

29. Winglee, K., Manson McGuire, A., Maiga, M., Abeel, T., Shea, T., Desjardins, C. A., Diarra, B., Baya, B., Sanogo, M., Diallo, S., Earl, A. M. & Bishai, W. R. Whole Genome Sequencing of Mycobacterium africanum Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. en. *PLoS neglected tropical diseases* **10,** e0004332. ISSN: 1935-2727, 1935-2735 (Jan. 2016).

30. Manson, A. L., Cohen, K. A., Abeel, T., Desjardins, C. A., Armstrong, D. T., Barry 3rd, C. E., Brand, J., TBResist Global Genome Consortium, Chapman, S. B., Cho, S.-N., Gabrielian, A., Gomez, J., Jodals, A. M., Joloba, M., Jureen, P., Lee, J. S., Malinga, L., Maiga, M., Nordenberg, D., Noroc, E., Romancenco, E., Salazar, A., Ssengooba, W., Velayati, A. A., Winglee, K., Zalutskaya, A., Via, L. E., Cassell, G. H., Dorman, S. E., Ellner, J., Farnia, P., Galagan, J. E., Rosenthal, A., Crudu, V., Homorodean, D., Hsueh, P.-R., Narayanan, S., Pym, A. S., Skrahina, A., Swaminathan, S., Van der Walt, M., Alland, D., Bishai, W. R., Cohen, T., Hoffner, S., Birren, B. W. & Earl, A. M. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and

spread of multidrug resistance. en. *Nature genetics* **49,** 395–402. ISSN: 1061-4036, 1546-1718 (Mar. 2017).

31. Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., Blum, M. G. B., Rüsch-Gerdes, S., Mokrousov, I., Aleksic, E., Allix-Béguec, C., Antierens, A., Augustynowicz-Kopeć, E., Ballif, M., Barletta, F., Beck, H. P., Barry 3rd, C. E., Bonnet, M., Borroni, E., Campos-Herrero, I., Cirillo, D., Cox, H., Crowe, S., Crudu, V., Diel, R., Drobniewski, F., Fauville-Dufaux, M., Gagneux, S., Ghebremichael, S., Hanekom, M., Hoffner, S., Jiao, W.-W., Kalon, S., Kohl, T. A., Kontsevaya, I., Lillebæk, T., Maeda, S., Nikolayevskyy, V., Rasmussen, M., Rastogi, N., Samper, S., Sanchez-Padilla, E., Savic, B., Shamputa, I. C., Shen, A., Sng, L.-H., Stakenas, P., Toit, K., Varaine, F., Vukovic, D., Wahl, C., Warren, R., Supply, P., Niemann, S. & Wirth, T. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. en. *Nature genetics* **47,** 242–249. ISSN: 1061-4036, 1546-1718 (Mar. 2015).

32. Isaza, J. P., Duque, C., Gomez, V., Robledo, J., Barrera, L. F. & Alzate, J. F. Whole genome shotgun sequencing of one Colombian clinical isolate of Mycobacterium tuberculosis reveals DosR regulon gene deletions. en. *FEMS microbiology letters* **330,** 113–120. ISSN: 0378-1097, 1574-6968 (May 2012).

33. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry 3rd, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S. & Barrell, B. G. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. en. *Nature* **393,** 537–544. ISSN: 0028-0836 (June 1998).

34. Camus, J.-C., Pryor, M. J., Médigue, C. & Cole, S. T. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. en. *Microbiology* **148,** 2967–2973. ISSN: 0026-2617, 1350-0872 (Oct. 2002).

35. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. en. *Current opinion in genetics & development* **15,** 589–594. ISSN: 0959-437X (Dec. 2005).

36. Bosi, E., Monk, J. M., Aziz, R. K., Fondi, M., Nizet, V. & Palsson, B. Ø. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. en. *Proceedings of the National Academy of Sciences of the United States of America* **113,** E3801–9. ISSN: 0027-8424, 1091-6490 (June 2016).

37. Supply, P., Warren, R. M., Bañuls, A.-L., Lesjean, S., Van Der Spuy, G. D., Lewis, L.-A., Tibayrenc, M., Van Helden, P. D. & Locht, C. Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area. en. *Molecular microbiology* **47,** 529–538. ISSN: 0950-382X (Jan. 2003).

38. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced Escherichia coli genomes. en. *Microbial ecology* **60,** 708–720. ISSN: 0095-3628, 1432-184X (Nov. 2010).

39. Bottai, D. & Brosch, R. Mycobacterial PE, PPE and ESX clusters: novel insights into the secretion of these most unusual protein families. en. *Molecular microbiology* **73,** 325–328. ISSN: 0950-382X, 1365-2958 (Aug. 2009).

40. Glickman, M. S. & Jacobs Jr, W. R. Microbial pathogenesis of Mycobacterium tuberculosis: dawn of a discipline. en. *Cell* **104,** 477–485. ISSN: 0092-8674 (Feb. 2001).

41. Huerta-Cepas, J., Forslund, K., Szklarczyk, D., Jensen, L. J., von Mering, C. & Bork, P. *Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper* en. Jan. 2016.

42. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. en. *Nucleic acids research* **28,** 33–36. ISSN: 0305-1048 (Jan. 2000).

43. Mao, C., Abraham, D., Wattam, A. R., Wilson, M. J. C., Shukla, M., Yoo, H. S. & Sobral, B. W. Curation, integration and visualization of bacterial virulence factors in PATRIC. en. *Bioinformatics* **31,** 252–258. ISSN: 1367-4803, 1367-4811 (Jan. 2015).

44. Raynaud, C., Guilhot, C., Rauzier, J., Bordat, Y., Pelicic, V., Manganelli, R., Smith, I., Gicquel, B. & Jackson, M. Phospholipases C are involved in the virulence of Mycobacterium tuberculosis. en. *Molecular microbiology* **45,** 203–217. ISSN: 0950-382X (July 2002).

45. Tufariello, J. M., Chapman, J. R., Kerantzas, C. A., Wong, K.-W., Vilchèze, C., Jones, C. M., Cole, L. E., Tinaztepe, E., Thompson, V., Fenyö, D., Niederweis, M., Ueberheide, B., Philips, J. A. & Jacobs Jr, W. R. Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence. en. *Proceedings of the National Academy of Sciences of the United States of America* **113,** E348–57. ISSN: 0027-8424, 1091-6490 (Jan. 2016).

46. Nandakumar, M., Nathan, C. & Rhee, K. Y. Isocitrate lyase mediates broad antibiotic tolerance in Mycobacterium tuberculosis. en. *Nature communications* **5,** 4306. ISSN: 2041-1723 (June 2014).

47. Russell, D. G. Trp'ing tuberculosis. en. *Cell* **155,** 1209–1210. ISSN: 0092-8674, 1097-4172 (Dec. 2013).

48. Parish, T. Starvation survival response of Mycobacterium tuberculosis. en. *Journal of bacteriology* **185,** 6702–6706. ISSN: 0021-9193 (Nov. 2003).

49. Zhang, Y. J., Reddy, M. C., Ioerger, T. R., Rothchild, A. C., Dartois, V., Schuster, B. M., Trauner, A., Wallis, D., Galaviz, S., Huttenhower, C., Sacchettini, J. C., Behar, S. M. &

Rubin, E. J. Tryptophan biosynthesis protects mycobacteria from CD4 T-cell-mediated killing. en. *Cell* **155,** 1296–1308. ISSN: 0092-8674, 1097-4172 (Dec. 2013).

50. Morlock, G. P., Metchock, B., Sikes, D., Crawford, J. T. & Cooksey, R. C. ethA, inhA, and katG loci of ethionamide-resistant clinical Mycobacterium tuberculosis isolates. en. *Antimicrobial agents and chemotherapy* **47,** 3799–3805. ISSN: 0066-4804 (Dec. 2003).

51. Danilchanka, O., Mailaender, C. & Niederweis, M. Identification of a novel multidrug efflux pump of Mycobacterium tuberculosis. en. *Antimicrobial agents and chemotherapy* **52,** 2503–2511. ISSN: 0066-4804, 1098-6596 (July 2008).

52. Li, G., Zhang, J., Guo, Q., Wei, J., Jiang, Y., Zhao, X., Zhao, L.-L., Liu, Z., Lu, J. & Wan, K. Study of efflux pump gene expression in rifampicin-monoresistant Mycobacterium tuberculosis clinical isolates. en. *The Journal of antibiotics* **68,** 431–435. ISSN: 0021-8820 (July 2015).

53. Wong, S. Y., Lee, J. S., Kwak, H. K., Via, L. E., Boshoff, H. I. M. & Barry 3rd, C. E. Mutations in gidB confer low-level streptomycin resistance in Mycobacterium tuberculosis. en. *Antimicrobial agents and chemotherapy* **55,** 2515–2522. ISSN: 0066-4804, 1098-6596 (June 2011).

54. Wang, F., Sambandan, D., Halder, R., Wang, J., Batt, S. M., Weinrick, B., Ahmad, I., Yang, P., Zhang, Y., Kim, J., Hassani, M., Huszar, S., Trefzer, C., Ma, Z., Kaneko, T., Mdluli, K. E., Franzblau, S., Chatterjee, A. K., Johnsson, K., Johnson, K., Mikusova, K., Besra, G. S., Fütterer, K., Robbins, S. H., Barnes, S. W., Walker, J. R., Jacobs Jr, W. R. & Schultz, P. G. Identification of a small molecule with activity against drug-resistant and persistent tuberculosis. en. *Proceedings of the National Academy of Sciences of the United States of America* **110,** E2510–7. ISSN: 0027-8424, 1091-6490 (July 2013).

55. Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M., McNeil, M., Peterson, S. N., Chatterjee, D., Fleischmann, R., *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-[beta]-D-arabinose biosynthetic and utilization pathway genes. *Nature genetics* **45,** 1190–1197. ISSN: 1061-4036 (2013).

56. Lingaraju, S., Rigouts, L., Gupta, A., Lee, J., Umubyeyi, A. N., Davidow, A. L., German, S., Cho, E., Lee, J.-I., Cho, S.-N., Kim, C. T., Alland, D. & Safi, H. Geographic Differences in the Contribution of ubiA Mutations to High-Level Ethambutol Resistance in Mycobacterium tuberculosis. en. *Antimicrobial agents and chemotherapy* **60,** 4101–4105. ISSN: 0066-4804, 1098-6596 (July 2016).

57. Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A. R., Will, R., Xia, F. & Stevens, R. Antimicrobial Resistance Prediction in PATRIC and RAST. en. *Scientific reports* **6,** 27930. ISSN: 2045-2322 (June 2016).

58. Sakoulas, G., Okumura, C. Y., Thienphrapa, W., Olson, J., Nonejuie, P., Dam, Q., Dhand, A., Pogliano, J., Yeaman, M. R., Hensler, M. E., Bayer, A. S. & Nizet, V. Nafcillin enhances innate immune-mediated killing of methicillin-resistant Staphylococcus aureus. en. *Journal of molecular medicine* **92,** 139–149. ISSN: 0946-2716, 1432-1440 (Feb. 2014).

59. Philalay, J. S., Palermo, C. O., Hauge, K. A., Rustad, T. R. & Cangelosi, G. A. Genes required for intrinsic multidrug resistance in Mycobacterium avium. en. *Antimicrobial agents and chemotherapy* **48,** 3412–3418. ISSN: 0066-4804 (Sept. 2004).

60. Shi, L., Sohaskey, C. D., Kana, B. D., Dawes, S., North, R. J., Mizrahi, V. & Gennaro, M. L. Changes in energy metabolism of Mycobacterium tuberculosis in mouse lung and under in vitro conditions affecting aerobic respiration. en. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15629–15634. ISSN: 0027-8424 (Oct. 2005).

61. Dhar, N. & McKinney, J. D. Mycobacterium tuberculosis persistence mutants identified by screening in isoniazid-treated mice. en. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 12275–12280. ISSN: 0027-8424, 1091-6490 (July 2010).

62. Moradigaravand, D., Grandjean, L., Martinez, E., Li, H., Zheng, J., Coronel, J., Moore, D., Török, M. E., Sintchenko, V., Huang, H., Javid, B., Parkhill, J., Peacock, S. J. & Köser, C. U. dfrA thyA Double Deletion in para-Aminosalicylic Acid-Resistant Mycobacterium tuberculosis Beijing Strains. en. *Antimicrobial agents and chemotherapy* **60,** 3864–3867. ISSN: 0066-4804, 1098-6596 (June 2016).

63. Martinez, E., Holmes, N., Jelfs, P. & Sintchenko, V. Genome sequencing reveals novel deletions associated with secondary resistance to pyrazinamide in MDR Mycobacterium tuberculosis. en. *The Journal of antimicrobial chemotherapy* **70,** 2511–2514. ISSN: 0305-7453, 1460-2091 (Sept. 2015).

64. Pearson, T., Busch, J. D., Ravel, J., Read, T. D., Rhoton, S. D., U'Ren, J. M., Simonson, T. S., Kachur, S. M., Leadem, R. R., Cardon, M. L., Van Ert, M. N., Huynh, L. Y., Fraser, C. M. & Keim, P. Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. en. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 13536–13541. ISSN: 0027-8424 (Sept. 2004).

65. Mieth, B., Kloft, M., Rodríguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suárez, C., Farré, X., Marigorta, U. M., Fehr, E., Dickhaus, T., Blanchard, G., Schunk, D., Navarro, A. & Müller, K.-R. Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. en. *Scientific reports* **6,** 36671. ISSN: 2045-2322 (Nov. 2016).

66. Tiwari, P., Arora, G., Singh, M., Kidwai, S., Narayan, O. P. & Singh, R. MazF ribonucleases promote *Mycobacterium tuberculosis* drug tolerance and virulence in guinea pigs. en. *Nature communications* **6,** 6059. ISSN: 2041-1723 (Jan. 2015).

67. De Welzen, L., Eldholm, V., Maharaj, K., Manson, A. L., Earl, A. M. & Pym, A. S. Whole transcriptome and genomic analysis of extensively drug-resistant Mycobacterium tuber-

culosis clinical isolates identifies downregulation of ethA as a mechanism of ethionamide resistance. en. *Antimicrobial agents and chemotherapy.* ISSN: 0066-4804, 1098-6596 (Oct. 2017).

68. Mih, N., Brunk, E., Bordbar, A. & Palsson, B. O. A Multi-scale Computational Platform to Mechanistically Assess the Effect of Genetic Variation on Drug Responses in Human Erythrocyte Metabolism. en. *PLoS computational biology* **12,** e1005039. ISSN: 1553-734X, 1553-7358 (July 2016).

69. Wollenberg, K. R., Desjardins, C. A., Zalutskaya, A., Slodovnikova, V., Oler, A. J., Quiñones, M., Abeel, T., Chapman, S. B., Tartakovsky, M., Gabrielian, A., Hoffner, S., Skrahin, A., Birren, B. W., Rosenthal, A., Skrahina, A. & Earl, A. M. Whole-Genome Sequencing of Mycobacterium tuberculosis Provides Insight into the Evolution and Genetic Composition of Drug-Resistant Tuberculosis in Belarus. en. *Journal of clinical microbiology* **55,** 457–469. ISSN: 0095-1137, 1098-660X (Feb. 2017).

Appendix B

An updated genome-scale metabolic model of *M. tuberculosis* - Supplementary Information

B.1 Methods

B.1.1 Choosing a base reconstruction

A variety of both quantitative and qualitative criteria was considered in determining which model would provide the base for the new model. The determining criterion included the amount of curated data, extent of previous model unification, gene essentiality predictions, standardized identifiers, cross-references to databases, and quality of physical representations, such as the use of an extracellular compartment and mass balanced reactions. Based on this criterion, the reconstruction of iEK1011 was based on the unification of iOSDD, sMtb, and

portions of the M. tuberculosis H37Rv BioCyc Database [1] (Figure 1A).

With regards to the selection criterion, sMtb was chosen as the primary base model. Notably, sMtb performed the best amongst the previous models in gene essentiality predictions (Figure 2B). In addition, sMtb included metabolite formulas, an extracellular compartment, and cross-references to databases. Both iSM810 and gal2015, which were both built off of GSMN-TB 1.1, lacked standardized identifiers (i.e., reactions identifiers were arbitrarily named R1, R2, etc.), metabolite formulas, and an extracellular compartment (i.e. inputs into the model could be utilized without being transported across the membrane). The lack of chemical formulas disables the assessment of mass conservation, which is a defining feature of constraint-based modeling. Furthermore, an extracellular compartment is key in distinguishing between what goes into the media and what is being transported across the membrane. While iOSDD performed well in categories related to component descriptions, it was based on iNJ661 and thus had a lower gene essentiality score, as previously shown [2]. Despite the low gene essentiality score, we utilized iOSDD as a representative for the models based on iNJ661. In this study, we show through gene essentiality predictions that the integration of iOSDD with sMtb results in a 6

The reconstruction process was straightforward (Figure 1A). The base models were first algorithmically mapped to standardized BiGG identifiers [3]. Identifiers that could not be mapped by the algorithm were manually assigned identifiers that follow the BiGGs format. Importantly, BiGGs was chosen as the standardization platform due to being a centralized repository for high-quality models. Once a standardized basis for identifiers was established, a draft reconstruction was built from the set of reactions shared across the standardized models. The differences between reactions across the models were manually assessed through literature references and added to the draft reconstruction. Once the models were unified into the draft reconstruction,

manual curation of new biochemical knowledge was incorporated in the reconstruction. The reconstruction process was iterative and involved constant re-evaluation of model components.

B.1.2 Updating the reconstruction

The model was updated with newly characterized metabolic processes, standardized identifications, and mass balanced reactions. In addition, detailed and designable metabolic maps of M. tuberculosis metabolism were manually built and provided in the supplement in order to help in silico simulation and reconstruction efforts as well as provide access to systems biology research for non-computational biologists. Specifics on using the escher maps are described in the section titled "Escher Flux Maps".

Before any updating took place, sMtb identifiers for metabolites and reactions were mapped to standardized identifiers in the BiGG Models database [3]. In addition to sMtb, the BioCyc M. tuberculosis H37Rv version 20.0 database was approximately converted to a cobra model - standardizing it first to BiGGs IDs, then MetaNetx, and then BioCyc identifiers if no BiGGs or Metanetx reference mapping was available. When an sMtb component had no equivalent BiGGs identification, a new identifier was created that followed the BiGG's nomenclature. The updated reconstruction utilizes data from Tuberculist, 2016 TB BioCyc database, and manually curated literature sources. New pathways and major GPR updates include Tuberculosinol biosynthesis, oxidized GTP and dGTP detoxification, Heme uptake and degradation, GlgE pathway update, glucosylglycerate biosynthesis I, included essential genes Rv3805c and Rv2673 in MAP complex biosynthesis, and others. In addition to incorporating updates from the new BioCyc database, we re-curated pathways that had inconsistencies across divergent models.

B.1.3 Description of GAM and NGAM parameters

Our model includes both growth-associated (GAM) and non-growth (NGAM) associated ATP maintenance parameters. NGAM quantifies the energy required by Mtb to maintain itself in a given environment while GAM quantifies growth energy requirements not accounted for in the metabolic model. For iEK1011, the GAM was chosen to be 60 mmol gDw-1, which was the same as the GAM used in previous M. tuberculosis H37Rv reconstructions of iNJ661, iAB-AMØ-1410-Mt-661, and iOSDD. For comparison, the GAM used for sMtb—a model built from the GSMN-TB line of reconstructions—was 57 mmol gDw-1. For the NGAM, iEK1011 uses a value of 3.15 mmol gDw-1h-1, which was taken from the E. coli model [4]. For comparison, the NGAM used in sMtb was set to 0.1 mmol gDw-1h-1, while the NGAM used in iSM810 was 1 mmol gDw-1h-1. We are not aware of any datasets available for M. tuberculosis that enables a rigorous evaluation of the NGAM parameter, such as those used for E. coli [5] (i.e., quantitative substrate uptake rates for different substrates).

In order to assess the sensitivity of our chosen NGAM, we recomputed the gene essentiality using an NGAM value of 1.0 and 0.01. With respect to our previous NGAM of 3.15, the NGAMs of 1.0 and 0.1 result in very similar values (Table B.1). We hope that future experimental efforts will enable a better parameterization of GAM and NGAM in genome-scale reconstructions of M. tuberculosis.

B.1.4 Flux Variability Analysis and Sampling of in vitro and in vivo conditions

All constraint-based simulations of iEK1011 were done using the python constraint-based modeling package, COBRApy [6]. While the linear program is guaranteed to find the global

Table B.1: Newly proposed AMR genes. The mutation column represents the distinguishing mutation for the resistant or susceptible-dominant allele(s). Abbreviations: R, resistant; S, susceptible; EMB, ethambutol; PAS, para-aminosalicylic acid; INH, isoniazid; PZA, pyrazinamide; RMP, rifampicin; SM, streptomycin; OFX, ofloxacin; ETA, ethionamide; MDR, multidrug resistant; XDR, extensively-drug resistant.

	iEK1011	iSM810	sMTb
NGAM	3.15	1.00	0.10
Griffin essentiality MCC	0.6	0.6	0.59
DeJesus essentiality MCC	0.71	0.7	0.7

optimum, the flux state solution to this optimization problem may not be unique, leading to the alternate optimal flux states. To account for this, we ran Flux Variability Analysis (FVA) in both the Lowenstein-Jensen media and approximated in vivo conditions using the "biomass" objective function. FVA gives the maximum and minimum amount of flux a reaction can take on. In addition to FVA, we sampled the solution space of iEK1011 on both media conditions using markov-chain monte-carlo sampling (MCMC) package available in cobrapy.

Furthermore, for both FVA and MCMC sampling, we allowed for solutions within 95% of the optimal value. The growth rate for both simulations were approximately the same to allow for a meaningful quantitative flux value comparisons.

B.1.5 Comparison of FVA across different drug objective simulations

For both in vivo and in vitro media conditions, we simulated each of the drug objectives and compared the maximum and minimum fluxes of the reactions catalyzed by the curated antibiotic resistance genes (Table 3.3). The maximum and minimum fluxes for each reaction were determined by FVA (described above) allowing for solutions within 95% of optimum. Furthermore, iEK1011 was constrained to produce at least 20% of biomass growth (i.e., the lower bound of the "biomass" reaction was set to frac*max_biomass_growth, where max_biomass_growth is the optimum value of iEK1011 when maximizing biomass on either in vivo or in vitro conditions),

and frac is the percentage of biomass to maintain while optimizing the other objective functions.

B.1.6 Gene Essentiality predictions

Gene essentiality predictions were determined using the same data and quantitative score used in evaluating the predictive ability of iSM810 [2]. The gene essentiality dataset was acquired from Griffin et al. [7]. If the Griffin essentiality confidence score was less than 0.1, the gene was determined to be essential. A growth cutoff of 20% of optimal growth was chosen to determine whether the in silico knockout was essential or not (i.e. if it was less than 20% of optimal growth, the gene was determined to be essential).

In addition to the Griffin gene essentiality dataset, we also evaluated the performance of the models in using a recent gene essentiality dataset acquired from DeJesus et al. [8]. A cutoff of 20% was used for the DeJesus dataset for the gene annotations of GD (growth defect), ES (essential), and ESD (essential domain). If growth was above 20% of optimal growth, the gene was said to be NE (non-essential) and GA (growth advantage). The matthews Correlation Coefficient was used to score the quality of each model's prediction, given by the following equation:

where TP (True Positive) represents the event where the model correctly simulates growth when a gene is nonessential. TN (True Negative) represents the event where the model correctly simulates no-growth when a gene is essential. FP (False Positive) represents the event where the model simulates no growth with the gene knockout when the gene is in fact non-essential. FN (False Negative) represents the event where the model simulates growth when the gene is in fact essential. While the Griffin et al. essentiality dataset is older, we utilized it due to having a more defined media conditions and was previous used in previous M. tuberculosis H37Rv reconstruction studies. The default objective function ("biomass") was used across all models. Differences in

MCC values between this study and that in Ma et al. [2] are due to differences in growth cutoff thresholds and media conditions. Despite inconsistencies, the values remained similar and did not change the resulting 6% increase in gene essentiality by iEK1011.

B.1.7 Approximation of literature-derived evolutionary forces of antibiotic-resistance evolution

A more in depth reasoning for the choice of objective function is described below for each antibiotic.

Ethambutol: It has been shown that flux-increasing mutations in ubiA confer resistance to ethambutol by increasing the production of decaprenylphosphoryl-b-D-arabinose (DPA), which outcompetes ethambutol for embB bindings spots [9]. Therefore, we approximate the evolutionary force of adaptation as maximizing the production of DPA.

D-cycloserine: Analogous to the mechanism of ethambutol, it has been shown that loss-of-function mutations in ald confer resistance to the d-cycloserine by increasing the pool of Alanine (i.e. ald no longer converts Alanine to pyruvate), thereby competitively inhibiting d-cycloserine [10].

Para-aminosalicylic acid (PAS): Mutations in folC, ribD, and thyA have been shown to confer resistance to PAS [11]. It was suggested that thyA mutations are selected in order to decrease the utilization of folates. In addition, it was suggested that the up-regulation of ribD occurs as an alternative when dfrA is inhibited.

Ethionamide: It has been shown that mycothiol biosynthesis is essential for ethionamide susceptibility [12]. We approximate ethionamide resistance is minimizing the production of mycothiol. It is worth noting that the objective defined for ethionamide is a much looser

approximation than the other objectives defined before.

B.2 References

1. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S. & Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. en. *Nucleic acids research* **44,** D471–80. ISSN: 0305-1048, 1362-4962 (Jan. 2016).

2. Ma, S., Minch, K. J., Rustad, T. R., Hobbs, S., Zhou, S.-L., Sherman, D. R. & Price, N. D. Integrated Modeling of Gene Regulatory and Metabolic Networks in Mycobacterium tuberculosis. en. *PLoS computational biology* **11,** e1004543. ISSN: 1553-734X, 1553-7358 (Nov. 2015).

3. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. en. *Nucleic acids research* **44,** D515–22. ISSN: 0305-1048, 1362-4962 (Jan. 2016).

4. Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M. & Palsson, B. Ø. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. en. *Molecular systems biology* **7,** 535. ISSN: 1744-4292 (Oct. 2011).

5. Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M. & Palsson, B. O. iML1515, a knowledgebase that computes Escherichia coli traits. en. *Nature biotechnology* **35,** 904–908. ISSN: 1087-0156, 1546-1696 (Oct. 2017).

6. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. en. *BMC systems biology* **7,** 74. ISSN: 1752-0509 (Aug. 2013).

7. Griffin, J. E., Gawronski, J. D., Dejesus, M. A., Ioerger, T. R., Akerley, B. J. & Sassetti, C. M. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. en. *PLoS pathogens* **7,** e1002251. ISSN: 1553-7366, 1553-7374 (Sept. 2011).

8. DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., Rubin, E. J., Schnappinger, D., Ehrt, S., Fortune, S. M., Sassetti, C. M. & Ioerger, T. R. Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. en. *mBio* **8.** ISSN: 2150-7511 (Jan. 2017).

9. Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M., McNeil, M., Peterson, S. N., Chatterjee, D., Fleischmann, R., *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-[beta]-D-arabinose biosynthetic and utilization pathway genes. *Nature genetics* **45,** 1190–1197. ISSN: 1061-4036 (2013).

10. Desjardins, C. A., Cohen, K. A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B. J., Shea, T. P., Almeida, D. V., Manson, A. L., Salazar, A., Padayatchi, N., O'Donnell, M. R., Mlisana, K. P., Wortman, J., Birren, B. W., Grosset, J., Earl, A. M. & Pym, A. S. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. en. *Nature genetics* **48,** 544–551. ISSN: 1061-4036, 1546-1718 (May 2016).

11. Zheng, J., Rubin, E. J., Bifani, P., Mathys, V., Lim, V., Au, M., Jang, J., Nam, J., Dick, T., Walker, J. R., Pethe, K. & Camacho, L. R. para-Aminosalicylic acid is a prodrug targeting dihydrofolate reductase in Mycobacterium tuberculosis. en. *The Journal of biological chemistry* **288,** 23447–23456. ISSN: 0021-9258, 1083-351X (Aug. 2013).

12. Vilchèze, C., Av-Gay, Y., Attarian, R., Liu, Z., Hazbón, M. H., Colangeli, R., Chen, B., Liu, W., Alland, D., Sacchettini, J. C. & Jacobs Jr, W. R. Mycothiol biosynthesis is essential for ethionamide susceptibility in Mycobacterium tuberculosis. en. *Molecular microbiology* **69,** 1316–1329. ISSN: 0950-382X, 1365-2958 (Sept. 2008).

Appendix C

A biochemically-interpretable machine learning classifier for microbial GWAS - Supplementary Information

C.1 Methods

C.1.1 Characteristics of utilized datasets.

The TB AMR datasets utilized in this study were acquired from a previous study that performed machine learning and protein structure analysis. References describing this data set are provided in the supplementary information of the previous study [1]. The dataset was initially acquired from the PATRIC database [2]. The sequencing and phenotypic testing data for

these strains were generated at the Broad Institute. Additional information for these sequencing projects can be found at the Broad Institute website for the TB Antibiotic Resistance Catalog (TB-ARC).

C.1.2 Curation and functional assessment of TB AMR genes

A list of known and implicated TB AMR genes was curated for 8 antibiotics (isoniazid, rifampicin, ethambutol, pyrazinamide, ofloxacin, d-cycloserine, para-aminosalicylic acid) using a combination of databases [3], experimental studies, and computational studies [1, 4–6]. Experimental studies on allele-specific effects for these AMR genes were curated utilizing a previous study performing 3D structural mutation mapping [1] and functional annotation from UNIPROT [7]. The lists of known and implicated TB AMR genes and mutational effects are provided (Supplementary File 1).

C.1.3 Modification of base genome-scale model

We performed minor modifications to the base genome-scale model, iEK1011, in order to use it for the MAC. Specifically, we performed quality-assurance and quality check (QA/QC) by removing blocked reactions (i.e., cannot carry any flux) and imposing maximum and minimum allowable flux constraints on the model determined by Loopless Flux Variability Analysis (LFVA) [8, 9]. Before FVA-derived constraints were imposed, we parameterized the exchange reactions according to the experimental nutrient media for testing AMR phenotypes, Middlebrook 7H10 (m7H10). Specifically, the LFVA simulations were constrained to have a biomass flux of at least 10% of its maximum value, and the total flux was bounded from above by 1.5 times the minimum total flux determined by parsimonious flux balance analysis [10]. The code for initializing the

base genome-scale model is provided in the code repository.

C.1.4 Generation of allele-constraint map ensemble through randomized sampling

Since knowledge of allele-specific effects are unavailable, we generated an ensemble of landscapes through randomized sampling of the allele-constraint map. Specifically, we generated an allele-constraint sample by sampling from each allele's discretized constraint set. The constraint set per allele includes the "no change" option and has a uniform probability distribution (i.e., each constraint has equal probability). An allele-constraint map sample is thus derived from sampling each allele's constraint distribution for all alleles.

C.1.5 Statistical tests for allelic AMR and flux stratification

We tested the AMR-based flux stratification of alleles by fitting a linear regression line between the allele log odds ratio (LOR) and fluxes. Linear regression was implemented using the linregress function in the scipy package. The LOR for each allele with respect to a specific antibiotic was quantified as LOR=log10((PR/PS)/(AR/AS)). PR, PS, AR, and AS denote number of strains that have the allele and are resistant (PR), have the allele and are susceptible (PS), do not have the allele and are resistant (AR), and do not have the allele and are susceptible (AS), respectively. If any of the values were 0, then 0.5 was added to each value to ensure a value when computing the logarithm. The fluxes for each allele were defined as the set of fluxes in strains containing that allele. We identified significant allelic LOR-flux correlations as having less than 5% FDR by the Benjamini-Hochberg method.

Conventional GWAS and pathway analysis of allelic variants Genome-wide association

analysis was performed to identify significant associations between allele frequencies and AMR phenotypes in our dataset using an ANOVA F-test, carried out using scikit-learn [11]. The set of significant alleles was determined by the Bonferroni-corrected significance threshold set at P ¡ 0.05/195 = 2.56x10-4. We identified metabolic pathways enriched in significant alleles through hypergeometric enrichment tests using the scipy function hypergeom and the gene-pathway annotation list described above. We identified significant pathways as having less than 5% false discovery rate (FDR) correction by the Benjamini-Hochberg method.

C.2 References

1. Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. & Palsson, B. O. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. en. *Nature communications* **9,** 4306. ISSN: 2041-1723, 2041-1723 (Oct. 2018).

2. Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y. & Sobral, B. W. PATRIC, the bacterial bioinformatics database and analysis resource. en. *Nucleic acids research* **42,** D581–91. ISSN: 0305-1048, 1362-4962 (Jan. 2014).

3. Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B. K., Church, G. M. & Murray, M. B. Tuberculosis drug resistance mutation database. en. *PLoS medicine* **6,** e2. ISSN: 1549-1277, 1549-1676 (Feb. 2009).

4. Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., Warren, R. M., Streicher, E. M., Calver, A., Sloutsky, A., Kaur, D., Posey, J. E., Plikaytis, B., Oggioni, M. R., Gardy, J. L., Johnston, J. C., Rodrigues, M., Tang, P. K. C., Kato-Maeda, M., Borowsky, M. L., Muddukrishna, B., Kreiswirth, B. N., Kurepina, N., Galagan, J., Gagneux, S., Birren, B., Rubin, E. J., Lander, E. S., Sabeti, P. C. & Murray, M. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. en. *Nature genetics* **45,** 1183–1189. ISSN: 1061-4036, 1546-1718 (Oct. 2013).

5. Coll, F., Phelan, J., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., Abdallah, A. M., Alghamdi, S., Alsomali, M., Ahmed, A. O., Portelli, S., Oppong, Y., Alves, A., Bessa, T. B., Campino, S., Caws, M., Chatterjee, A., Crampin, A. C., Dheda, K., Furnham, N., Glynn, J. R., Grandjean, L., Minh Ha, D., Hasan, R., Hasan, Z., Hibberd, M. L., Joloba, M., Jones-López, E. C., Matsumoto, T., Miranda, A., Moore, D. J., Mocillo, N., Panaiotov, S., Parkhill, J., Penha, C., Perdigão, J., Portugal, I., Rchiad, Z., Robledo, J., Sheen, P., Shesha, N. T., Sirgel, F. A., Sola, C., Oliveira Sousa, E., Streicher, E. M., Van Helden, P., Viveiros, M., Warren, R. M., McNerney, R., Pain, A. & Clark, T. G. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. en. *Nature genetics* **50,** 307–316. ISSN: 1061-4036, 1546-1718 (Feb. 2018).

6. Walker, T. M., Kohl, T. A., Omar, S. V., Hedge, J., Del Ojo Elias, C., Bradley, P., Iqbal, Z., Feuerriegel, S., Niehaus, K. E., Wilson, D. J., Clifton, D. A., Kapatai, G., Ip, C. L. C., Bowden, R., Drobniewski, F. A., Allix-Béguec, C., Gaudin, C., Parkhill, J., Diel, R., Supply, P., Crook, D. W., Smith, E. G., Walker, A. S., Ismail, N., Niemann, S., Peto, T. E. A. & Modernizing Medical Microbiology (MMM) Informatics Group. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. en. *The Lancet infectious diseases* **15,** 1193–1202. ISSN: 1473-3099, 1474-4457 (Oct. 2015).

7. The UniProt Consortium. UniProt: the universal protein knowledgebase. en. *Nucleic acids research* **45,** D158–D169. ISSN: 0305-1048, 1362-4962 (Jan. 2017).

8. Desouki, A. A., Jarre, F., Gelius-Dietrich, G. & Lercher, M. J. CycleFreeFlux: efficient removal of thermodynamically infeasible loops from flux distributions. en. *Bioinformatics* **31,** 2159–2165. ISSN: 1367-4803, 1367-4811 (July 2015).

9. Gudmundsson, S. & Thiele, I. Computationally efficient flux variability analysis. en. *BMC bioinformatics* **11,** 489. ISSN: 1471-2105 (Sept. 2010).

10. Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D. & Palsson, B. Ø. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. en. *Molecular systems biology* **6,** 390. ISSN: 1744-4292 (July 2010).

11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *Journal of machine learning research: JMLR* **12,** 2825–2830. ISSN: 1532-4435 (2011).

Appendix D

Laboratory evolution of multiple *E. coli* strains reveals unifying principles of adaptation but diversity in driving genotypes - Supplementary Information

D.1 Methods

D.1.1 Adaptive laboratory evolution and DNA sequencing

ALE was performed using 3 independent replicates of each strain. Cultures were serially propagated on M9 minimal medium [1] with 2 g/L glucose at 37C and well-mixed for proper

aeration using an automated system[1, 2] that periodically passed 150 µL of the cultures to a new fresh 30 mL flasks with a total working volume of 15 mL M9 medium (i.e., a 1:100 ratio) once they had reached an optical density (OD600) of 0.3 (Tecan Sunrise plate reader, equivalent to an OD600 of 1.3 on a traditional spectrophotometer with a 1 cm path length). Such a routine to pass at the late exponential phase of growth, was to keep cells under constant selection pressure for higher fitness, i.e. growth rate. Cultures were always maintained in excess nutrient conditions assessed by non-tapering exponential growth. The laboratory evolution was performed for a sufficient time interval to allow the cells to reach its fitness plateau. Periodically, glycerol cryogenic stocks were prepared and stored at -80C for any culture restarting. The fitness jump was observed in about 200 generations; however, the experiment was continued for approximately 900 generations to explore the possibility of any secondary fitness jump. Further passaging was stopped due to the absence of any appreciable growth rate increase in about 700 generations. The slope of ln(OD600) vs. time of four OD600 measurements from each flask was used to determine the growth rate. A cubic interpolating spline constrained to be monotonically increasing was fit to these growth rates to obtain the smoothed fitness trajectory curves. DNA resequencing was performed on a clone from the end points of evolved strains as described earlier by Lacroix et al., 2015 [2]. The ALE mutation data is provided for all replicate lineages (Supplementary Table 7).

D.1.2 RNA-sequencing and processing

Total RNA was sampled from duplicate cultures. Growth curve analysis was performed using a Bioscreen C Reader system with 200μ L culture volume per well. Two biological replicates were used in the assay. Media components were purchased from Sigma-Aldrich (St. Louis, MO). After inoculation and growth, 3mL of cell broth (OD600) was immediately added to two volumes

Qiagen RNA-protect Bacteria Reagent (6mL), vortexed for 5s, incubated at room temperature for 5min, and immediately centrifuged for 10min at 11,000g. The supernatant was decanted, and the cell pellet was stored in the -80C. Cell pellets were thawed and incubated with ReadyLyse Lysozyme, SuperaseIn, Protease K, and 20% SDS for 20min at 37C. Total RNA was isolated and purified using the Qiagen RNeasy Mini Kit columns and following vendor procedures. An on-column DNase-treatment was performed for 30min at room temperature. RNA was quantified using a Nanodrop and quality assessed by running an RNA-nano chip on a bioanalyzer. The rRNA was removed using Illumina Ribo-Zero rRNA removal kit for Gram-negative bacteria. A KAPA stranded RNA-Seq Kit (Kapa Biosystems KK8401) was used following the manufacturer's protocol to create sequencing libraries with an average insert length of around 300bp. Libraries were run on a HiSeq4000 (Illumina). All RNA-seq experiments were performed in biological duplicates from distinct samples. Raw-sequencing reads were deposited to GEO.

Raw-sequencing reads were mapped to the reference genomes using bowtie (v1.1.2)[3] with the following options "-X 1000 -n 2 -3 3". Transcript abundance was quantified using summarizeOverlaps from the R GenomicAlignments package (v1.18.0)[4]. To ensure the quality of the compendium, genes shorter than 100 nucleotides and genes with under 10 fragments per million-mapped reads across all samples were removed before further analysis. Transcripts per million (TPM) were calculated by DESeq2 (v1.22.1) [5]. The final expression compendium was log-transformed log2(TPM+1) before analysis, referred to as log-TPM. Biological replicates with R2¡0.9 between log-TPM were removed to reduce technical noise.

D.1.3 Fluxomics

Metabolic characterization by 13C metabolic flux analysis was performed as described in [6, 7]. Briefly, for 13C-tracer experiments, strains were cultured aerobically in glucose M9 minimal medium at 37C in mini-bioreactors with 10 mL working volume. Pre-cultures were grown overnight and then used to inoculate the experimental culture at an OD600 of 0.01, in which 2 g/L of [1,6-13C]glucose was present. Cells were harvested for GC-MS analysis at mid-exponential growth when OD600 was approximately 0.7. [1,6-13C]glucose was previously identified as an optimal tracer for global flux resolution [8].

Chemicals and M9 minimal medium were purchased from Sigma-Aldrich (St. Louis, MO). Isotopic tracers were purchased from Cambridge Isotope Laboratories (Tewksbury, MA): [1,6–13C]glucose (99.2% 13C) (99.7%). The isotopic purity and enrichment of all tracers were validated by GC-MS analysis. All solutions were sterilized by filtration. Samples were collected during the exponential growth phase to monitor cell growth, glucose consumption and acetate production. Cell growth was monitored by measuring the optical density at 600 nm (OD600) using a spectrophotometer (Eppendorf BioPhotometer). The OD600 values were converted to cell dry weight concentrations using a predetermined OD600-dry cell weight relationship for E. coli (1.0 OD600 = 0.32 gDW/L) [9]. After centrifugation, the supernatant was separated from the biomass pellet and glucose concentration was measured with a YSI 2700 biochemistry analyzer (YSI, Yellow Springs, OH). Acetate was measured by HPLC.

GC-MS analysis was performed on an Agilent 7890B GC system equipped with a DB-5MS capillary column (30 m, 0.25 mm i.d., 0.25 µm-phase thickness; Agilent J&W Scientific), connected to an Agilent 5977A Mass Spectrometer operating under ionization by electron impact (EI) at 70 eV. Helium flow was maintained at 1 mL/min. The source temperature was

maintained at 230°C, the MS quad temperature at 150C, the interface temperature at 280C, and the inlet temperature at 250C. GC-MS analysis of tert-butyldimethylsilyl (TBDMS) derivatized proteinogenic amino acids was performed as described [6]. Labeling of glucose (derived from glycogen) and ribose (from RNA) were determined as described. In all cases, mass isotopomer distributions were obtained by integration [10] and corrected for natural isotope abundances [11]. All mass isotopomer data are provided.

The metabolic network model used for 13C-MFA is provided. The model [7] includes all major metabolic pathways of central carbon metabolism, lumped amino acid biosynthesis reactions, and a lumped biomass formation reaction. 13C-MFA calculations were performed using the Metran software [12], which is based on the elementary metabolite units (EMU) framework [13]. Fluxes were estimated by minimizing the variance-weighted sum of squared residuals (SSR) between the measured and model predicted mass isotopomer distributions and acetate yield using non-linear least-squares regression. Flux estimation was repeated 10 times starting with random initial values for all fluxes to find a global solution. At convergence, accurate 95% confidence intervals were computed for all estimated fluxes by evaluating the sensitivity of the minimized SSR to flux variations. Precision of estimated fluxes was determined as follows60: Flux precision (stdev) = [(flux upper bound 95%) – (flux lower bound 95%)] / 4.

To describe fractional labeling of biomass amino acids G-value parameters were included in 13C-MFA. As described previously [6], the G-value represents the fraction of a metabolite pool that is produced during the labeling experiment, while 1-G represents the fraction that is naturally labeled (e.g., from inoculum). By default, one G-value parameter was included for each measured amino acid in each data set. Reversible reactions were modeled as separate forward and backward fluxes. Net and exchange fluxes were determined as follows: vnet = vf-vb; vexch

= min(vf, vb). To determine the goodness-of-fit, 13C-MFA fitting results were subjected to a 2-statistical test [14].

D.1.4 Mann-Whitney U tests for identifying convergent and divergent phenotypes

To perform statistical tests for convergent and divergent features, we transformed the data vectors describing the mean physiological and fluxomics values for the size WT and EP flasks to vectors containing the pairwise distances amongst the points. The conversion resulted in a total of 15 points for each the WT and EP flasks. The transformation to pairwise distances accounts for how close the strains were at each point (i.e., convergence describes points coming closer together). Mann-Whitney U tests were then carried out to test whether the EP pairwise distances are smaller than the WT pairwise distances (i.e., the EP values are closer together than the WT values). We calculated the p-values using both a normal approximation implemented with the mannwhitneyu function in scipy stats. Both of the statistic estimates captured the general behavior, but the normal approximation was utilized due to the lack of table p-values for U statistics less than 36. We then selected the convergent and divergent features as those with a false discovery rate (FDR) less than 5% using the Benjamini Hochberg correction, implemented in the statsmodels package version 0.9.0 [15].

D.1.5 Differential expression analysis of RNA-seq

We performed differential expression analysis of the RNA-seq profiles between consecutive ALE flasks (i.e., ALE evolution stages) using the R package DESeq2 [5]. Specifically, differential expression was performed for each pair of flasks describing the before and after of an ALE

experiment. We utilized an adaptive t prior shrinkage estimator [16] to transform the log fold changes for better ranking and visualization of the differential expression results. We performed a sensitivity analysis of the p-value and Log2 fold change thresholds on determining sets of significantly expressed genes.

D.1.6 iModulon analysis of RNA-seq data

We previously showed that Independent Component Analysis (ICA) deconvolved a large compendium of E. coli MG1655 RNA-seq data into a linear combination of independent sources ("iModulons"), that reflect known regulons, and source weightings ("iModulon activities"), which describe the global regulatory state [17]. The resulting matrix decomposition by ICA in Anand et al [17] is formulated as follows, XPRECISE = MPRECISE * APRECIS.

Where XPRECISE is the previously utilized PRECSE RNA-seq data described in transcripts per million (TPM), MPRECISE is the matrix describing the iModulon gene sets (genes as rows and iModulons as columns), and APRECISE is the sample-specific iModulon activities (iModulons as rows and samples as columns). Using the previous set of 92 iModulons (MPRECISE), we transformed the flask-specific gene expression profiles of our six E. coli strain ALEs (X6strain) into flask-specific iModulon activities (APRECISE), formulated as follows, A6strain = MPRECISE-1 * X6strain.

Where A6strain and X6strain describe the flask specific iModulon activities and flask-specific gene expression TPM profiles, respectively. The previously uncharacterized iModulons Uncharacterized-6, Uncharacterized-5, and Uncharacterized-3 were characterized in this study as hns-related, ppGpp, and CspA, respectively. Together, the 92 iModulons explained 52% of the expression variance of the multi-strain core genome, where they explained the most expression

for MG1655 (67.78%) and the least for C (44.23%).

D.1.7 Differential expression analysis of RNA-seq

Distribution of differences in iModulon activities between biological replicates were first calculated and a log-norm distribution was fit to the differences [18]. In order to test statistical significance, absolute value of difference in activity level of each iModulon between the two samples were calculated. This difference in activity was compared to the log-normal distribution from above to get a p-value. Because differences and p-value for all iModulons were calculated, the p-value was further adjusted with Benjamini-Hochberg correction to account for multiple hypothesis testing problem. Only iModulons with change in activity levels greater than 5 were considered significant. Differential activity analysis was performed for all ALE jumps as well as between the WT and EP flask for each strain.

D.1.8 Data transformation to jump-specific perspective

We utilize a jump-specific perspective of the data was taken for our iModulon PCA and mutation correlation analysis. Specifically, we transform the activity matrix (flask-specific) to the difference in flask activity along the trajectory (jump-specific) in order to identify components describing general adaptation trends as opposed to strain differences. We formulate this as follows, Xjump i, strain j = Xflask i+1, strain j - Xflask i, strain j.

Where δX describes the jump-specific dataset with 16 rows (jumps) and X describes the original flask-specific dataset with 22 rows (flasks).

D.1.9 Trade-off analysis through PCA and ANCOVA

In order to avoid harsh statistical corrections when testing all possible iModulon pairs, we performed PCA using the jump-specific iModulon activities in order to filter out a candidate set of iModulons for downstream correlation tests. Since our initial run of PCA resulted in the first component (explaining 40% of the variation) describing large FlhDC and FliA activity unique to the first MG1655 jump, we filtered out the FlhDC and FliA iModulon outliers. We then performed both analysis of covariance (ANCOVA) and pearson correlation tests for iModulons that had PCA weights greater than 0.10 in components explaining at least 5% of the variation. ANCOVA was performed to test the similarity of the strain-specific regression lines (dependence on strain-specific categorization). Tradeoffs were identified as iModulon pairs with ANCOVA R2 greater than 0.90.

D.2 References

1. Mohamed, E. T., Mundhada, H., Landberg, J., Cann, I., Mackie, R. I., Nielsen, A. T., Herrgård, M. J. & Feist, A. M. Generation of an E. coli platform strain for improved sucrose utilization using adaptive laboratory evolution. en. *Microb. Cell Fact.* **18,** 116 (June 2019).

2. LaCroix, R. A., Sandberg, T. E., O'Brien, E. J., Utrilla, J., Ebrahim, A., Guzman, G. I., Szubin, R., Palsson, B. O. & Feist, A. M. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. en. *Applied and environmental microbiology* **81,** 17–30. ISSN: 0099-2240, 1098-5336 (Jan. 2015).

3. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. en. *Genome Biol.* **10,** R25 (Mar. 2009).

4. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. & Carey, V. J. Software for computing and annotating genomic ranges. en. *PLoS Comput. Biol.* **9,** e1003118 (Aug. 2013).

5. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. en. *Genome biology* **15,** 550. ISSN: 1465-6906 (2014).

6. Long, C. P. & Antoniewicz, M. R. High-resolution 13C metabolic flux analysis. en. *Nat. Protoc.* **14,** 2856–2877 (Oct. 2019).

7. Long, C. P. & Antoniewicz, M. R. Metabolic flux responses to deletion of 20 core enzymes reveal flexibility and limits of E. coli metabolism. en. *Metab. Eng.* **55,** 249–257 (Sept. 2019).

8. Crown, S. B., Long, C. P. & Antoniewicz, M. R. Optimal tracers for parallel labeling experiments and 13C metabolic flux analysis: a new precision and synergy scoring system. *Metab. Eng.* (2016).

9. Long, C. P., Gonzalez, J. E., Sandoval, N. R. & Antoniewicz, M. R. Characterization of physiological responses to 22 gene knockouts in Escherichia coli central carbon metabolism. en. *Metab. Eng.* **37,** 102–113 (Sept. 2016).

10. Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. Accurate assessment of amino acid mass isotopomer distributions for metabolic flux analysis. en. *Anal. Chem.* **79,** 7554–7559 (Oct. 2007).

11. Fernandez, C. A., Des Rosiers, C., Previs, S. F., David, F. & Brunengraber, H. Correction of 13C mass isotopomer distributions for natural stable isotope abundance. en. *J. Mass Spectrom.* **31,** 255–262 (Mar. 1996).

12. Yoo, H., Stephanopoulos, G. & Kelleher, J. K. Quantifying carbon sources for de novo lipogenesis in wild-type and IRS-1 knockout brown adipocytes. en. *J. Lipid Res.* **45,** 1324–1332 (July 2004).

13. Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. en. *Metab. Eng.* **9,** 68–86 (Jan. 2007).

14. Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. en. *Metab. Eng.* **8,** 324–337 (July 2006).

15. Seabold, S. & Perktold, J. *Statsmodels: Econometric and statistical modeling with python* in *Proceedings of the 9th Python in Science Conference* **57** (2010), 61.

16. Zhu, A., Ibrahim, J. G. & Love, M. I. *Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences* en. Apr. 2018.

17. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A. & Palsson, B. O. The Escherichia coli transcriptome mostly consists of independently regulated modules. en. *Nature communications* **10,** 5536. ISSN: 2041-1723 (Dec. 2019).

18. Poudel, S., Tsunemoto, H., Seif, Y., Sastry, A., Szubin, R., Xu, S., Machado, H., Olson, C., Anand, A., Pogliano, J., Nizet, V. & Palsson, B. O. *Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators and role in key physiological responses* en. Mar. 2020.