A STATISTICAL METHOD FOR IDENTIFYING CHEMICAL-GENETIC INTERACTIONS USING LINEAR MIXED MODELS

A Thesis

by

ESHA DUTTA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Thomas R. Ioerger
Committee Members,	Bobak Mortazavi
	Junjie Zhang
Head of Department,	Scott Schaefer

May 2021

Major Subject: Computer Science

ABSTRACT

This research is focused on building statistical solutions to identify chemical-genomic interactions. We have used a linear mixed model to study the trend in the abundances of the genes in a population when exposed to varying concentrations of drugs. In this model, each influence of the drug on each individual gene is treated as random effects. For every gene, our model yields a gene-specific slope for the abundance vs the concentration of the drugs. These slopes are then subjected to 1-sided test to determine the genes with the most significantly outlying slopes, i.e., giving us an insight on the potential target of the drug under consideration. To gain further insights, we have used the GSEA analysis on the ranks of the slopes to understand the impact of the drugs on a pathway of genes. The developed model is validated on the publicly available chemical-genomics dataset published by the Broad Institute and multiple hypomorph libraries created by our collaborators.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Ioerger for his constant support and guidance throughout the course of this research. I would also like to thank my committee members Dr. Mortazavi and Dr. Zhang for their support.

Thanks also goes to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my family for their encouragement, patience and believing in me amidst the tough times.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Dr. Thomas Ioerger and Dr. Bobak Mortazavi of the Department of Computer Science and Professor Dr. Junjie Zhang of the Department of Biochemistry and Biophysics.

I would like to thank Nadine Ruecker (Schnappinger lab, Weil Cornell Medical College, New York) and Eun-Ik Koh (Sassetti lab, University of Massachusetts Medical School, MA) for making the chemical genomics libraries in M. tuberculosis and collecting the sequencing data, which was analyzed in this thesis.

All other work conducted for the thesis (or) dissertation was completed by the student independently.

Funding Sources

This work was also made possible in part by the Bill & Melinda Gates Foundation under Grant Number OPP0124065. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Bill & Melinda Gates Foundation.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Recently, a new group of experimental methodologies has been developed for detecting interactions between modulatory compounds (that elicit a phenotype) and genes in a bacterial genome, also known as chemical-genomic (C-G) interactions. The most common application of C-G analysis is for drug discovery, in which the goal is to discover the protein target of an inhibitor, which can yield insight about its mechanism of action.

A general approach to C-G experiments is to construct a knock-down (or depletion) library of essential genes in the organism (or “hypomorphs”). This consists of a pool (mixture) of mutant bacteria where the level of individual genes can be controlled. For example, expression of target genes can be decreased using a tetracycline-inducible (Tet) promoter (cloned onto a plasmid, with the native gene knocked out) (Schnappinger and Ehrt 2014, Evans and Mizrahi 2015). More sophisticated systems employing the ClpXP protease have been used to target genes (cloned with a C-terminal DAS tag) for proteolytic degradation, again under tetracycline control (Kim, O'Brien et al. 2013). Finally, CRISPRi can be used to express short RNA sequence guides (sgRNAs) complimentary to the target genes to inhibit their transcription (Rock, Hopkins et al. 2017).

If the genes in the library are barcoded with unique nucleotide tags, then the relative abundance of clones in the library can be efficiently profiled using deep sequencing. Furthermore, the results of an experiment (e.g., treatment with drug) can be quantified as changes in abundance (estimated via barcode counts) of different clones in the library.

Regardless of the methodology, the expression levels of the library members are knocked-down in parallel (each in a separate clone), resulting in overall growth-impairment of the culture (since these are essential genes). Then the culture is treated with an inhibitor, typically at concentrations just below the MIC. In theory, this will provide some challenge (or stress) to the population, and most clones in the library are expected to respond similarly (e.g., they all experience a similar degree of inhibition due to the drug treatment). However, for the member of the library with depletion of the specific target of the inhibitor, excess growth impairment is expected. C-G analysis relies on synergy between the stress of drug pressure combined with the stress caused by depletion of the target gene. In contrast, although absolute abundance of other members of the library would be expected to decrease due to presence of drug, because of normalization (e.g. dividing individual barcode counts by the total number of reads collected for the sample), the *relative* abundances of most genes should stay (approximately) the same. Thus, the gene(s) interacting with the compound (CGIs) can be identified as those genes that exhibit excessive (or even complete) depletion (compared to the other members of the library) at higher concentrations of the inhibitor. (Johnson, LaVerriere et al. 2019).

The objective of this research is to formulate the statistical analysis of a C-G experiment to identify the gene(s) in the library that interact with a drug by looking for those with excess depletion of barcode counts compared to the rest of the population. Quantifying statistical significance is important, because the genes can always be ranked by their apparent level of depletion, and there will always be a "most-depleted" gene, but this does not necessarily mean it is the true target; the target gene might not even be in the knock-down library. The challenge of analyzing C-G data is that there are many sources of noise in the data (barcode counts). The original abundances in the library are only approximately known, and the growth under treatments and DNA extraction/preparation for sequencing are stochastic, resulting in variance between both biological and technical replicates. More importantly, there are natural (but unpredictable, though possibly biological) reasons that the abundance of a particular library member might increase or decrease between different drug concentrations, as the population experiences different levels of stress. Although each depletion mutant might experience a distinct level of impairment, and the drug treatment should theoretically affect them all equally (except for the target gene with which it interacts), there is inevitably going to be some variance in the apparent responses to depletion of genes resulting from variance in levels of gene abundance between drug concentrations. Finally, not all genes might be represented in the library at equivalent levels, and genes that are low-abundance to begin with (or as growth-impairment increases) become more difficult to reliably estimate, as counts approach 0. Collecting sufficient sequencing data (reads) and sequencing multiple replicates greatly facilitates the statistical analysis.

In previous work, Johnson et al (Johnson, LaVerriere et al. 2019) described an approach to statistical analysis of C-G data based on a generalized linear model (GLM), called ConCensusGLM (or PROSPECT). Specifically, they fit gene abundances (normalized barcode counts) to a linear model (using the Negative Binomial distribution with a log-link function as a likelihood), with drugs (at different concentrations) as covariates. The GLM approach captures the dependence of gene abundances on drug treatments through coefficients in the linear model. Additional offsets for other covariates such as batch, plate, lane, and instrument were also estimated and subtracted out. As validation, they demonstrated that known drug targets like DNA gyrase A and RNA polymerase B exhibited among the most extreme depletion in the presence of known anti-mycobacterial drugs like moxifloxacin and rifampicin, respectively. They then ran the experiment on a large library of 50,000 compounds generated through combinatorial chemistry. To determine which interactions are statistically significant, the authors proposed using a Wald test, employing strain-wise dispersion estimates (from growth in DMSO controls). A Wald test tests whether a coefficient is significantly different from zero (Draper & Smith, 1998). However, the linear model in the ConCensusGLM pipeline treats each drug concentration independently. The only dependence on drug concentration is between individual concentrations and the DMSO control (effectively, log-fold-changes relative to no drug), and thus is more susceptible to random fluctuations. Indeed, in their experiments on the combi-chem library, they observed 95,685 "significant" interactions between a chemical library of 50,000 compounds and a knock-down library of 152 genes (1.3% of 7.2M combinations tested), using a criterion of p-value<10^{-10}. This is probably an overestimate of the number of true chemical-genetic interactions (since this implies ~2 hits for every compound, but it is highly unlikely that the true target for every compound will be represented in the library). The fact that they had to lower the p-value threshold to such an extreme level (10^{-10}) and still wound up with so many hits (2 interacting genes per compound) suggest their test is too loose and probably admits many false positives; our objective is to develop a more conservative statistical model where only true positives meet a rigorous statistical criterion.

In this thesis, we propose a new approach in which drug concentration is treated as a quantitative variable in the linear model, and the effect of concentration on gene abundance is captured by a single coefficient (a slope) for each gene that incorporates information across multiple concentrations. Interesting genes should be those where the depletion shows a consistent concentration-dependent effect. Genes that are targets of drugs are expected to show a synergistic effect, such that, at low drug concentrations, they are not more depleted than the rest of the population, but the abundance should systematically decrease as drug concentration increases (approximately around the MIC). Our approach to assessing significance is to calculate a slope of the log of gene abundance with respect to drug concentration, integrating information across a range of concentrations and capturing the trend of the counts. This is more robust because it depends on trends exhibited over multiple conditions (and hence is less sensitive to random fluctuations of gene abundance in a single condition). Not all genes follow a perfect S-curve as is typical of an ideal dose-response effect. Furthermore, it is hard to predict what concentration the drop-off will occur, since knock-down of different genes might cause different degrees of impairment (Wei, Krishnamoorthy et al. 2011), which affects the degree of synergy with the drug. But if the abundance drops off at some point within the concentration range evaluated, the overall decrease in abundance would be characterized by a negative slope.

Our approach is implemented as a linear mixed model (LMM), where the slope of each gene is represented as a random effect (conditional on drug concentration), so each gene can have its unique slope parameter. The aim of the approach is to identify the genes that exhibit the most negative slopes, that is, genes exhibiting the greatest degree of depletion as concentration increases. To determine whether the effect (slope) is statistically significant for a given gene, we compare it to the rest of the population, essentially looking for outliers. This is another fundamental difference from ConCensusGLM. In ConCensusGLM, the significance of individual gene-drug-concentration combinations is assessed using a Wald test, which aims to test only whether the coefficient is significantly different from zero. But due to the multiple unaccounted-for sources of noise in C-G experiments, there are various reasons why the slopes of some genes might be different from zero, resulting in dispersion in the distribution of slopes. Failing to account for these sources of dispersion is likely to produce many false positives. We take an empirical view that, whatever the reasons, the average gene (which is assumed not to interact with the drug), might have a slightly positive or negative slope, and we determine this variance post-hoc. True interactors (CGIs) must stand out from the spread of the population as outliers. Thus, our test for statistical significance takes advantage of the distribution of random effects to evaluate which genes have (negative) slopes that are outliers. Our implementation also incorporates an adjustment of the slopes to account for uncertainty in the slope estimates themselves. Thus, unlike a Wald test, which evaluates coefficients in isolation, our method identifies genes as significant only in the context of all the other genes. It is generally a more conservative approach that will detect fewer

CGIs, but hopefully a greater proportion of them will be genuine, by eliminating false positives.

Below is a summary of our technical contributions in this research:

- Developing a novel linear mixed model-based approach to study the influence of the abundances of the library of genes treated at various drug concentrations.
- Identifying the most significantly depleted genes in a population of genes using statistical tests for various anti-tubercular drugs such as levofloxacin, rifampin, copper etc.

The following chapters describe the details of the proposed model and the results from the experiments on the various datasets.

CHAPTER II

STATISTICAL ANALYSIS OF CHEMICAL GENOMICS DATA

Overview

The classic way of analyzing chemical genomics data is to select resistance mutants and identify its location by sequencing. However, a new approach (C-G) is to knock-out or knock-down genes one at a time and test whether the sensitivity to the inhibitor is affected. Recent technology developments (ClpXP proteolysis, Illumina sequencing) (Schnappinger & Ehrt, 2013, Wei et. al, 2011) have made it possible to scale this up to do it in parallel for many mutants (each with depletion of a different gene) at the same time in a pool (bacterial culture, “hypomorph library”). The experiment produces counts of genes (through counts of nucleotide barcodes through sequencing), and the goal of this thesis is to develop a statistical analysis method for analyzing these read counts and determining, with statistical confidence, which genes interact with which drugs, based on synergy between depletion effects of knock-down and drug treatment.

The challenges are: 1) there is a lot of noise in these experiments (stochastic variation of counts), and 2) many genes show relative changes in abundance but not always in a consistent and drug-depending way. We will show how linear mixed-models can be used to address these problems and identify the strongest candidates for C-G interactions.

The typical experimental setup for our analysis is designed as follows. There is a hypomorph library consisting of G genes whose expression can be controlled (depleted). Each clone has a knock-down of a different gene, and they are tagged with unique nucleotide barcodes that can be amplified by PCR. The library is grown in standard culture conditions (e.g., 7H9 medium), typically on 96- or 384-well plates, for a fixed number of days calibrated so that there is visible growth without inhibitors. In parallel wells, drugs are added at concentrations that span a range around the MIC (typically 2-fold dilutions), keeping in mind that growth might be severely depleted or eliminated above the MIC. After incubation of the plates, DNA is extracted from each of the wells (possibly facilitated by robotics), the barcodes are amplified using PCR, and the samples are prepared for high-throughput sequences (e.g., adapters with unique barcodes for multiplexing samples on an Illumina sequencer). It is recommended to collect multiple replicates of each culture condition (drug treatment).

After the sequencing data is obtained, the reads are de-multiplexed and formatted into a matrix containing counts $C_{g,s}$ for each gene of the hypomorph library in each sample (where S is the total number of samples). The metadata for each sample includes drug, concentration, and possibly other data that could be used as covariates such as number of days of incubation, carbon source in medium, strength of knock-down (*sspB, tet*). Finally, the counts are **normalized** to produce **relative abundances** by dividing each individual count by the total counts for that sample. This step is done to adjust for the different numbers of reads sequenced for each sample, represented as another matrix $A_{g,s}$, where the values range between 0 and 1 representing the percentages of the population. An alternative representation useful for modeling is to melt the data into a column matrix of all individual abundances Y of dimensions $n \times 1$, where $n = G \times S$.

Linear Model Selection to Analyze Chemical-Genomics Data

We tried multiple mathematical models based on simple linear regression with interaction and linear mixed models. A comparative study of these models (Zewotir et. al, 2007, Wright, 2017) helped us identify the best suited model for our data, which was then used for further analysis. We try to fit these models to capture the relationship between the (log-transformed) gene abundances and the log-transformed drug concentration. We begin our experimentation with fitting this data to a linear regression model with interaction effects between the gene and the concentration.

Linear Regression with Interactions

We start with OLS based linear regression between the log-abundances and log-concentration of the drugs and consider interactions between the genes and the concentrations of the drug. This interaction-effects generates unique slopes and intercepts corresponding to each of the genes in the population. Our model in R looks like:

log-abundances ~ 0 + (gene)+ (log-concentration) + (gene)*(log-concentration)

The second term in this equation corresponds to gene-specific intercepts. The third term corresponds to overall slope for log-concentration of the drug. The last term corresponds to gene-specific slopes for the increasing concentrations of the drug. We fit this model to the chemical-genomics analysis data published by Broad Institute (downloaded from www.chemicalgenomicsoftb.com) and compare its outcome with the alternative linear mixed model-based approach. The outcome of the model is captured in the following sections.

A fundamental problem in this approach is that for fitting a linear model, we assume that the data corresponding to every gene is independent from one other. However, as our model works on the relative abundances of the genes for every experimental condition, the relative counts of the genes are not exactly independent of each other which violates a basic assumption in fitting linear models (Winter, 2013). To overcome this problem, we use linear mixed models such that the gene-specific effects are captured as the “random effects” in the model. The following section describes further on the linear mixed model.

Linear Mixed Model

A linear mixed model (LMM) captures the relationship between (log-transformed) gene abundances and (log-transformed) drug concentrations. We fit the model separately for each drug (subscript d is implicit). The model is expressed as:

$$y = \ log_{10}Y \sim N(XB + ZU, \sigma_e^2)$$

where*,* Y is the log-abundance matrix. X is a $n \times 2$ design matrix, with a column encoding the $\log_{10}$ of the concentration and a constant (1) for the intercept. In controls or no-drug treatments, the value used is minimum of the non-zero concentrations divided by 10, so that, on a log scale, it is one less than the lowest concentration evaluated. The coefficients B are the **fixed effects** which will be fit in the model, representing an average slope and intercept for each drug (independent of gene). The ZU term represents **random effects**, for capturing the gene-specific effects. Z is a $n \times 2g$ matrix of covariates with g binary columns that encodes the information about which gene each observation represents, and the associated drug concentrations. U is a $2g \times 1$ matrix of random effects, which includes a slope and intercept for each of the gene.

The variance of observations σ_e^2 can be decomposed over the fixed and random effects by introducing the variable γ_i which represents the gene-specific errors (Zewotir & Galpin, 2007). Thus, each individual random effect U_i is assumed to be sampled from a normal distribution with 0 mean and independent variance of σ_i^2 such that,

$$\gamma_i = \frac{\sigma_i^2}{\sigma_e^2}$$

The null hypothesis is that the random effects U are drawn from a multivariate-Normal (MVN) distribution with 0 mean and unknown covariance matrix $\sum$:

$$U \sim MVN(0, \Sigma)$$

where $\Sigma = \sigma_e^2 D$, and D is a block diagonal matrix with each block corresponding to the independent variance of each of the covariates of the random effects, i.e., intercept and slope corresponding to each gene.

$$D_i = \gamma_i I_{2g}$$

γ_i represents the gene-specific component of the errors. γ_i is assumed to be sampled from a Normal distribution with 0 mean and independent variance of σ_i^2.

Our model can thus be decomposed with respect to the fixed and random effects by considering that $E(Y) = XB$ and the $var(Y) = \sigma_e^2 H$ where $H = I_n + ZDZ'$. For this model, the maximum likelihood estimates of B and U, referred to as $\hat{B}$ and $\hat{U}$ respectively, can be computed as:

$$\hat{B} = (X'H^{-1}X)^{-1}(X'H^{-1}y)$$

$$\text{and } \widehat{U} = DZ'H^{-1}(y - X\widehat{B})$$

However, since γ_i, and thus D and H, are unknown, we use restricted maximum likelihood (REML) to solve the system by an iterative procedure (Lindstrom & Bates, 1990), as implemented in the *lmer* function in the lme4 package in R. The resulting model then has estimates of the covariance matrix of the random effects, and the variance of the slopes as a population can be recovered from the diagonal elements in Σ.

We have explored two formulas for the implementation purpose.

F1: logY ~ 1+conc + (1+conc|gene)

F2: logY ~ 1 + conc + (1|gene) + (0 + conc|gene))

The primary difference between F1 and F2 is that F1 accounts for a random slope and intercept for each of the genes, whereas F2 decouples the random intercept and random slope for each of the genes. Both models consider a fixed slope and intercept for the overall population. The subsequent portion of this section compares the two linear mixed models and the linear regression with interaction effects on the chemical-genomics data published by the Broad Institute.

Comparing Linear Models

We evaluated the three models on data for trimethoprim published by the Broad Institute (Johnson, LaVerriere et al. 2019). Trimethoprim is a widely used anti-tuberculosis drug which targets a gene (dihydrofolate reductase, *dfrA*) in the folate pathway. This data comprises abundances of 155 genes profiled at various concentrations of trimethoprim. As evident from the plots below (Fig. (1)), *trpG* stands out as a significantly depleted gene in the linear regression model with interactions and the linear mixed model with F2. This is an expected outcome for trimethoprim because, although *dfrA* itself was not in the hypomorph library, *trpG* is also in the folate pathway. *TrpG* converts chorismate into PABA as the first step in the pathway. This is consistent with the analysis of this data with ConCensusGLM, which also detected *trpG* as a C-G interaction with trimethoprim. Additionally, these two models generate highly correlated slopes for all the genes in the library (Fig. (2)). Moving ahead, we use the linear mixed model (F2) for further analysis.

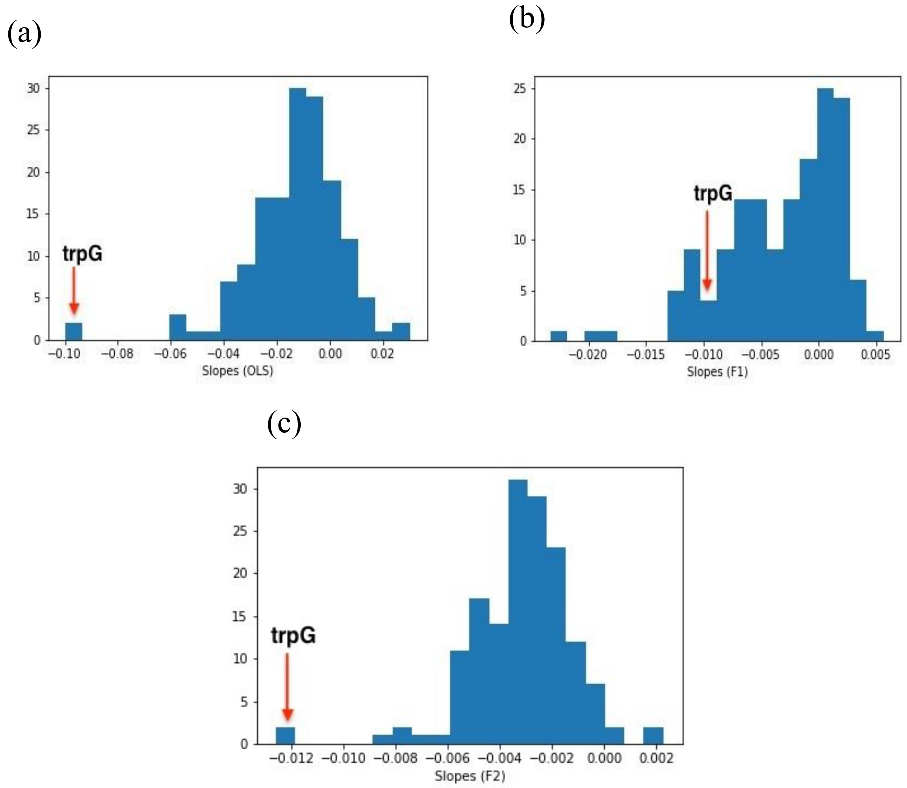

Figure 1 Histogram of the slopes from (a) OLS model (b) Linear Mixed Model (F1) (c) Linear Mixed Model (F2)

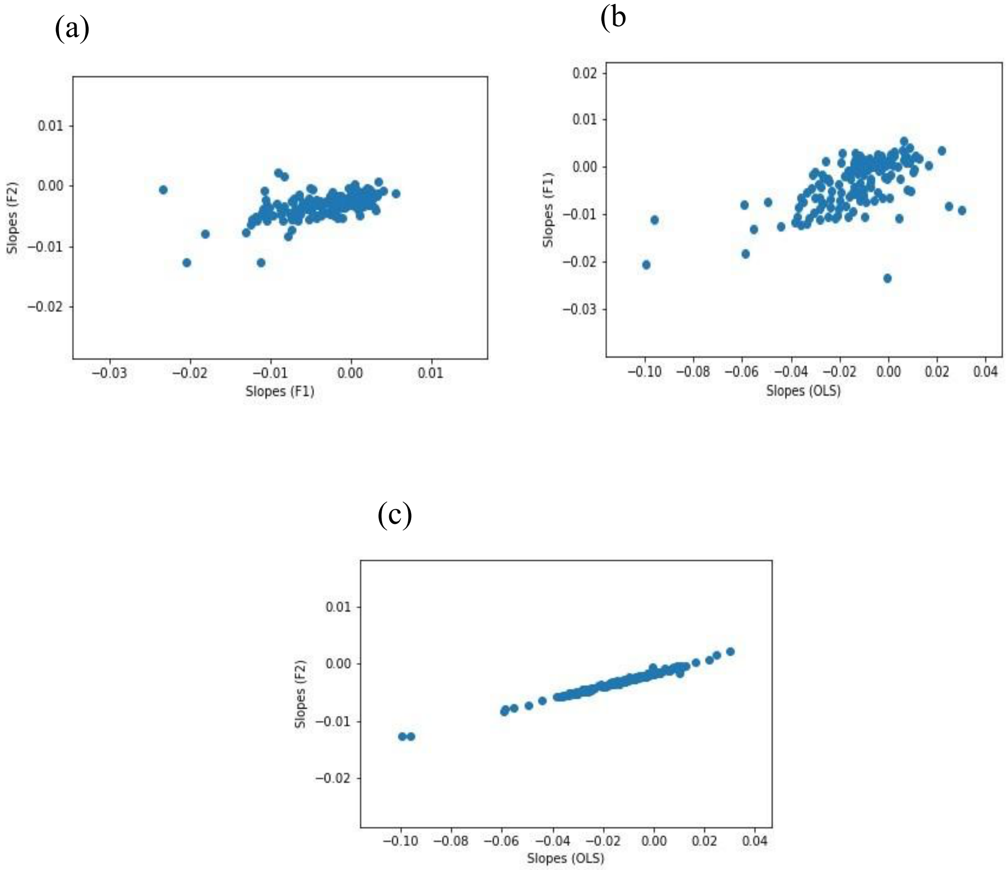

Figure 2 Scatter plots of the slopes from models (a) F1 vs F2 (b) OLS vs F1 (c) OLS vs F2. We observe that the slopes from OLS and F2 are highly correlated.

Statistical Significance

To determine which genes, have significantly negative slopes representing excess depletion, we compute Z-scores and look for outliers with respect to the overall population of genes. Our null hypothesis is that slopes, as random effects, are normally distributed with zero mean and a variance of Σ. Thus, the Z-scores of the slopes for each gene g can be computed as:

$$z_g = \frac{s_g - \mu}{\sigma_s}$$

where s_g is the slope (concentration-dependent random effect) for the gene estimated in the model, μ is the mean of these slopes and σ_s is the standard deviation of the population of slopes. For each gene, we can compute a p-value from the Z-score using a 1-sided test, $p_i = \phi(Z_i)$, which is the cumulative of the standard-Normal distribution.

This approach, however, does not account for certainty in the slope estimates themselves. For example, two genes might have very similar slopes, but the individual data points (log-abundances at log-concentrations) for one might be much more variable than the other. One such example is captured in Fig. 9. Slopes from data with high noise should be less confident. To quantify this, we compute the variance of the residuals (differences between predicted log-abundances and observations) with respect to the fitted model.

$$R = XB + ZU - \ logY$$

R corresponds to a $n \times 1$ vector of residuals for all observations. To compute the residual for a given gene, we project the residual vector in the gene space by multiplying with G which is a $n \times g$ binary matrix indicating which gene corresponds to a given observation. We then compute the covariance matrix as: $\Sigma_r \ = N^{-1}(G^T R)(G^T R)^T$, such that, $N = G^T G$ is a $g \times g$ diagonal matrix, which gives the number of observations corresponding to each gene. Thus, the diagonal elements of Σ_r give the variance of the residuals for each gene $\sigma_r^2(g)$ (relative to the regression fit).

Finally, we compute an **adjusted slope** as a mixture of the estimated slope for the gene, s_g, and the mean over the whole population, μ, weighted inversely by the variances:

$$s_g' = \frac{\frac{s_g}{\sigma_r^2(g)}+\frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma_r^2(g)}+\frac{1}{\sigma_s 2}}$$

Effectively, this reduces the magnitude of the slope for each gene estimated by the model (BLUPs) toward the mean slope in cases where the noise is high (among observations for the gene) but maintains the slope if the regression is well-supported by the observations, reflecting a consistent trend in abundance as concentration increases. It is over the population of adjusted slopes that we re-calculate the mean and standard deviation and use them to compute the Z-score and p-value for each gene.

CHAPTER III

MODEL EVALUATION

Experimental Setup

To evaluate our statistical method, we generated a hypomorph library of 162 essential genes in *M. tuberculosis* H37Rv. Degradation of target proteins was facilitated by appending a C-terminal DAS tag (a 15-amino-acid sequence recognized by SspB and ClpX) (Kim et al., 2011).

The *sspB* gene needed for initiation of the proteolytic degradation through ClpX was introduced into the *M. tuberculosis* cells on a plasmid controlled by an anhydrotetracycline (Atc) repressor. Removal of Atc, allowed expression of *sspB,* which led to the degradation of target protein through the native caseinolytic protease ClpX. To achieve non-lethal doses of protein degradation, the levels of *sspB* expression were regulated through promoter variations (Johnson et al., 2019; Kim et al., 2013; Kim et al., 2011; Lin et al., 2016). While constructing the mutant for each gene by add the DAS tag, a 10bp nucleotide barcode was inserted that was unique to each strain. For each drug, the library was grown in the presence of varying concentrations of inhibitor, the DNA was extracted from the culture, and the barcodes were amplified by PCR (polymerase chain reaction) and sequenced by next-generation sequencing (using an Illumina NovaSeq), producing millions of short (~50bp) reads for each sample. These were demultiplexed, and the barcodes for each strain were counted.

We also evaluated our model for the chemical-genomics analysis data published online by the Broad Institute (https://www.broadinstitute.org/chemical-biology/initiative-chemical-genetics) (Johnson et al., 2019). This dataset comprises the raw read counts of ~155 mutants knocked down by inhibitors – rifampicin, trimethoprim, methotrexate and BRD-4592. Johnson et.al builds a generalized linear model based statistical analysis on this dataset (Johnson et al., 2019). We tried our approach on this dataset and found that results are similar, yet our solution is more conservative and selective in terms of identifying the top hits of a drug.

Preliminary Analysis

We begin our analyses by normalizing the data which involves dividing the raw read counts of each gene with the total abundance in the sample. The relative abundances of the genes in the library varied from 0.0001% to 4.9%, with a median of 0.4%. The variances in the abundances among biological replicates represents the noise in the data which, in turn, greatly aids our statistical modeling. To bound the variances, we apply a log-transform. For convenient modeling, if a gene has an abundance of '0', we map to 1e-6, such that the lower limit of our log-transform is –6. Fig (3) shows the distribution of the log-transformed abundances when the library of genes is treated with no drug.

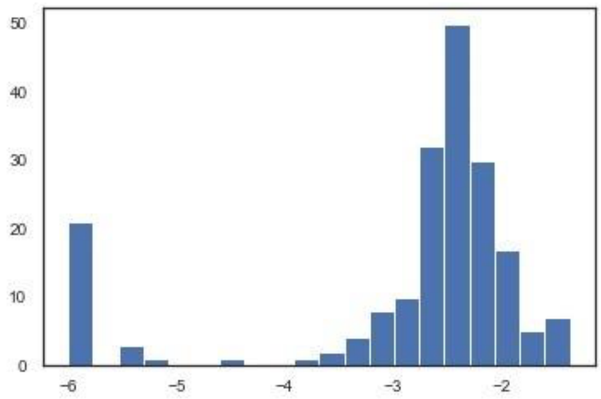

Figure 3 Distribution of the log-abundances of the genes when treated with no-drug concentration in glycerol

The following step involves a preliminary cleanup such that we discard genes with relative abundance greater than 20% of the overall population of genes. These genes are removed from our analysis because the high raw abundance of such genes is responsible for a significant reduction in the relative abundances of all the other genes in the library, which in turn, induces deleterious effects in the statistical modeling. For a robust modeling of the data, we clean our data of such outliers and then proceed to fitting the model.

Fig. (4) represents the scatter plot of the same two replicates when normalized after removing the outlying mutants if any. Here, the normalized abundances of the data look almost correlated to each other. However, there is still some variances in the abundances in the genes between the biological replicates. To stabilize the variances, we take a $\log_{10}$ transform on the same. Thus, the abundances and their variances are stabilized (more or less constant over the dependent variable) and hence fit for statistical modeling. Fig (5) represents the scatter plot of the log-abundances of the two replicates.

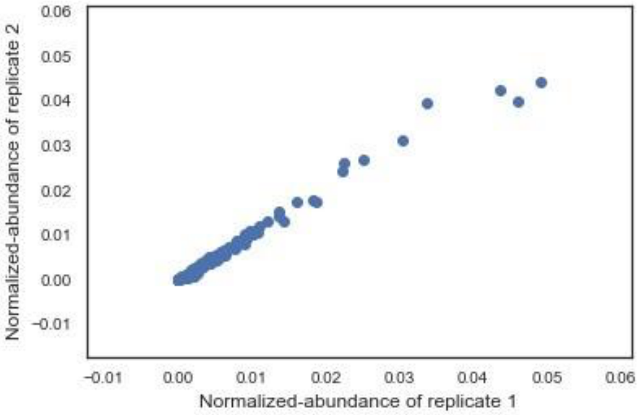

Figure 4 Scatter plot of normalized abundances of replicate 1 vs replicate 2 of the libraries of genes in no drug after removing outliers.

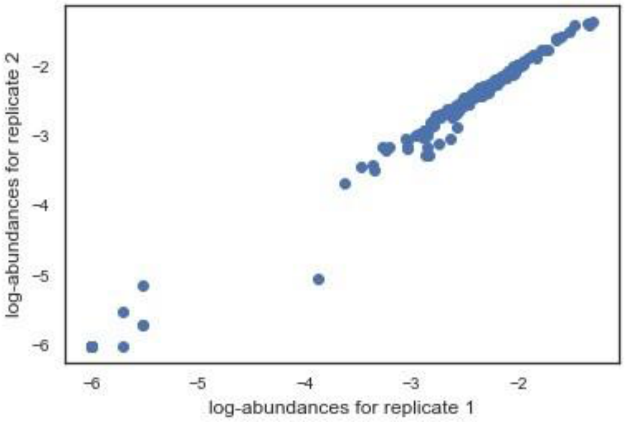

Figure 5 Scatter plot of log abundances of replicate 1 vs replicate 2 of the libraries of genes in glycerol as carbon source (no drug) after removing outliers.

Fig. (6) represents the trend of the log abundances of the genes with respect to the log-folded concentrations when the library of mutants in grown levofloxacin. The X-axis of the plot represents the samples with increasing drug concentrations. In the figure, we observe a significant gradient in the log-abundances of *Rv0006.gyrA* which is denoted by the red line in the graph.

Fig. (7) represents the boxplots of the log abundances of 2 of the mutants selected randomly from the population of genes treated with levofloxacin. *LeuS* does not play a role in levofloxacin activity, and hence its slope is flat. In contrast, *GyrA*, the target, exhibits a clear negative slope (excess depletion, indicating synergy with increasing drug concentrations.

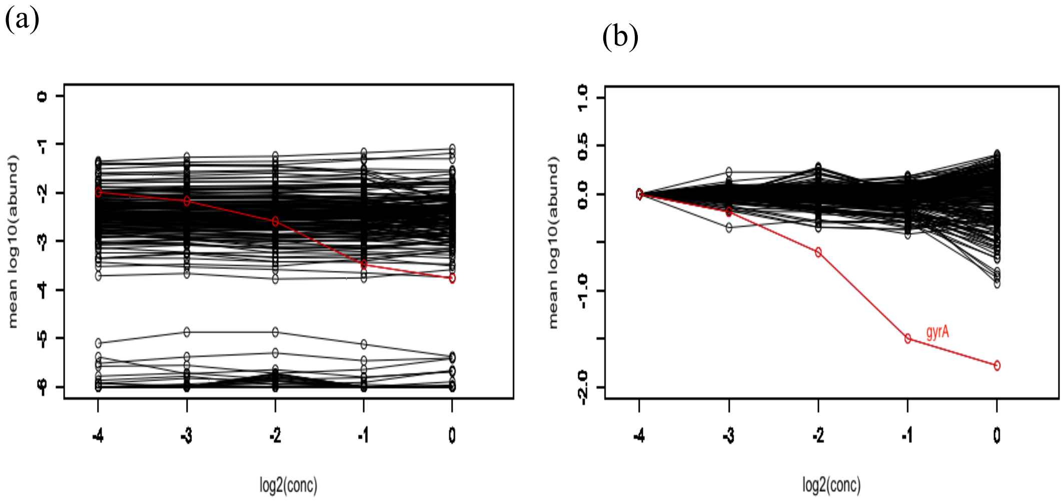

Figure 6 (a) Represents the trend of the log-abundances when a library of genes is treated with levofloxacin. (b) Represents trend of the log-abundances normalized with respect to the counts at 0xMIC. The red highlighted line refers to *gyrA*.

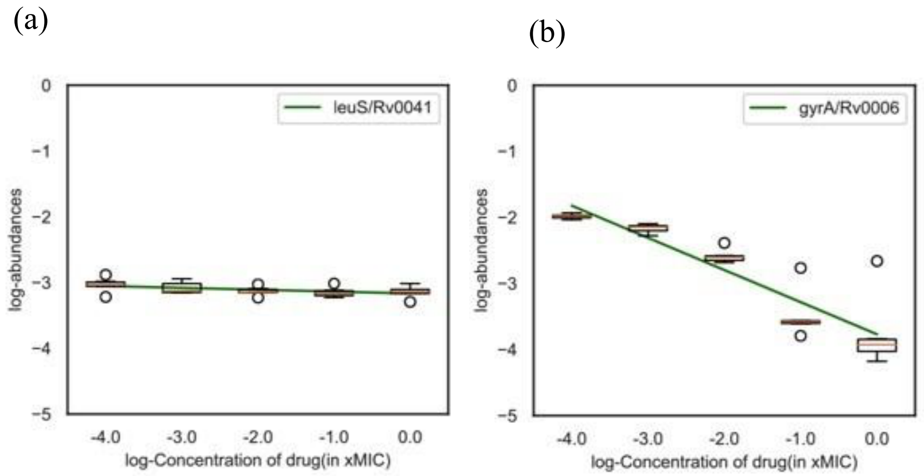

Figure 7 Boxplots of (a) *leuS* and (b) *gyrA* treated in various concentrations of levofloxacin

Identifying Targets using Linear Models

Linear mixed models are used to capture the relationship between the log-abundances and the log-concentrations of the drugs for all the genes in the library. Each gene contributes a random effect in terms of a gene-specific slope and intercept. We also have the fixed effect corresponding to the overall slope and intercept of the entire population. We fit the models separately for each of the drugs. Our data includes treating the pool strains with 7 drugs -- levofloxacin, moxifloxacin, bedaquiline, isoniazid, fidaxomicin, sulfamethoxazole, fusidic acid. We fit a separate linear mixed model for every drug, which yields us 7 different models, each with independent slope estimates for each gene. We conclude on the target hits of the conditions based on the genes that have a significantly negative slope as compared to the rest of the population.

Linear Mixed Model Output Summary

We fit the linear mixed model on the pre-processed data and extract the slopes corresponding to the random effects introduced by the genes. The target hits are the ones having outlier negative slopes. Fig. (8) represents a histogram of the distribution of the slopes when the cultures are treated with various concentrations of levofloxacin. The genes corresponding to the left ends of the histogram are the top hits for this inhibitor. In this case, the expected target is *Rv0006.gyrA*, which also comes up as the one with the most negative slope of –0.44 in our analysis and is effectively an outlier with respect to the rest of the population, thus, consistent with our expectations.

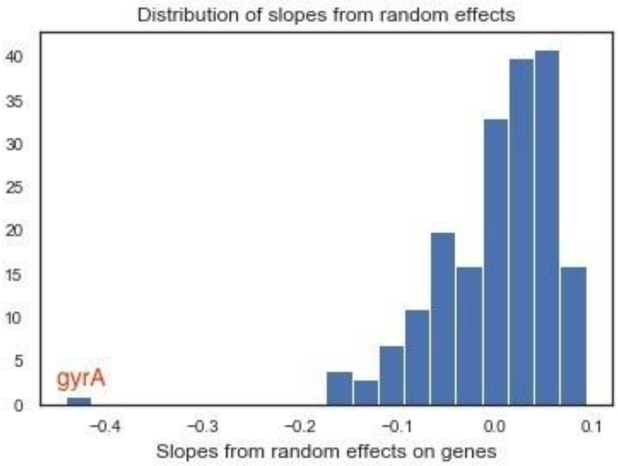

Figure 8 Distribution of (unadjusted) slopes of genes treated with various concentrations of Moxifloxacin in glycerol

Table 1 summarizes the top hits of levofloxacin. The first column represents the raw slopes of the genes as random effects in the linear mixed model. As expected, *Rv0006.gyrA* comes up as the most significantly depleted gene with the lowest slope and a very low p-value which is calculated by performing a one-sided test on the Z-scores of the slopes.

Table 1 Summary of top hits for levofloxacin before and after slope adjustment

gene	Slope(random)	Slope adj.	p-val(adj.)	q-value(adj.)
gyrA/Rv0006	-0.44	-0.36	0	2.99E-11
thyA/Rv2764c	-0.16	-0.16	0	0.13
dapB/Rv2773c	-0.17	-0.15	0	0.15
dnaE1/Rv1547	-0.14	-0.13	0	0.47
asnB/Rv2201	-0.16	-0.1	0.02	0.96
murD/Rv2155c	-0.1	-0.1	0.02	0.99
proB/Rv2439c	-0.1	-0.09	0.03	0.99

Linear Mixed Model with Slope Adjustment

Abundances of the genes across replicates often vary significantly which introduces uncertainty in the slope estimates itself. One such example is represented in Fig. (9). Thus, instead of relying on the random slopes extracted from the linear mixed model, we do a variance-based slope adjustment in such a way that, magnitude of the slopes of the genes having higher variances among replicates is reduced towards the mean slope of the overall population. Fig. (10). represents a scatter plot of the adjusted vs the unadjusted slopes of the genes present in the library. As evident from this figure, the adjusted and unadjusted slopes are closely correlated and *rpoB* stands as an outlier when the library of genes is treated with fidaxomicin.

To gain further insights on the slope adjustment approach, we plotted scatter plots of the slopes of the genes before and after adjustment. This is shown in Fig. (11) and Fig. (12). We can see that after slope adjustment *rpoB* comes out as the one with a significantly negative slope. As this adjustment accounts for the uncertainty in the slope estimates itself, it is more reliable to base our analysis on the adjusted slopes to identify the most significantly depleted gene.

The overall summary of the results after slope adjustment is summarized in Table 1.

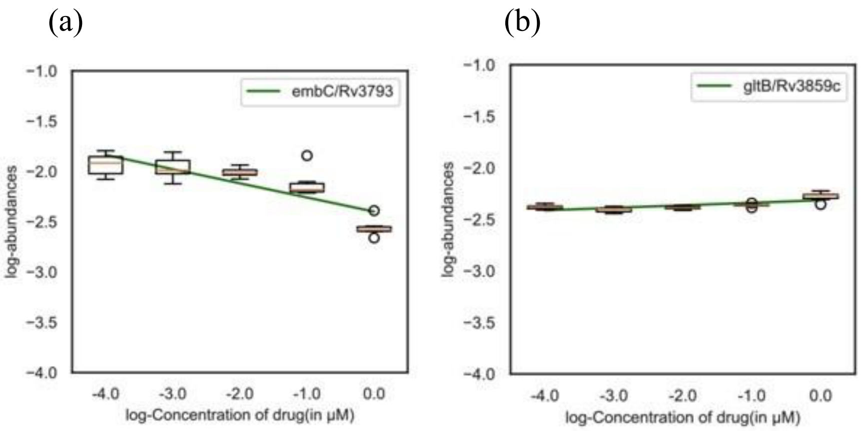

Figure 9 Boxplot of the log-abundances vs log-concentrations for (a) *embC* and (b) *gltB* when treated with levofloxacin. The slope of the green line are the unadjusted slopes for these genes which are nearly equal to –0.1 for *embC* and 0.03 for *gltB*. Slope adjustment has a higher influence on the slopes of *embC* because of the higher variance in the data itself.

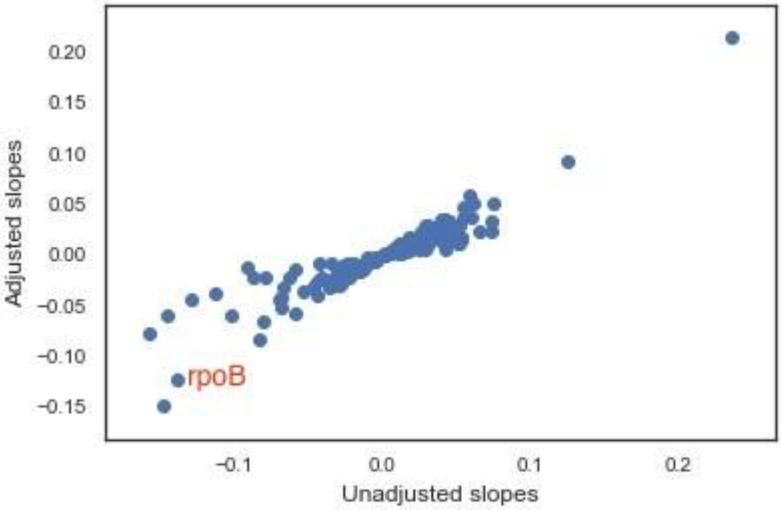

Figure 10 Scatter plot of adjusted vs unadjusted slopes of genes treated with various concentrations of fidaxomicin. *RpoB* comes out as an outlier with a negative slope.

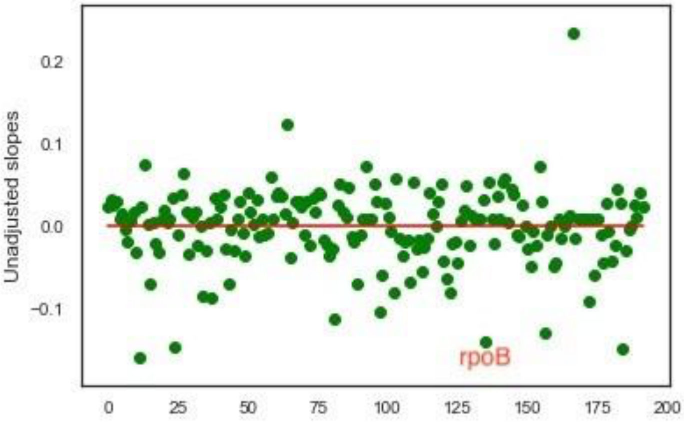

Figure 11 Scatter plot of slopes of genes treated with various concentrations of fidaxomicin. The red Line represents the mean slope of the population of genes.

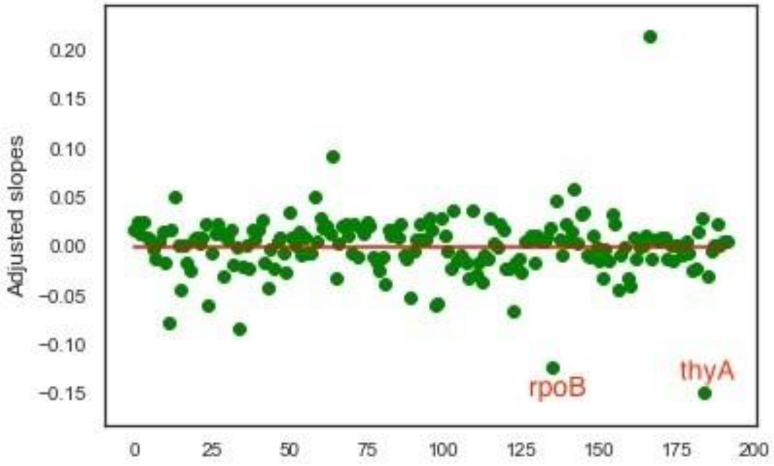

Figure 12 Scatter plot of the adjusted slopes of genes treated with various concentrations of fidaxomicin. The red line represents the mean slope of the population of genes. *RpoB* is the expected target of this drug, and it comes out as an outlier.

Fig. (13) represents the histogram of the distribution of the adjusted slopes of the genes when treated with levofloxacin. As evident from the graph, after variance-based slope adjustment, *Rv0006.gyrA* clearly comes out even more as an outlier with the most negative slope in the whole population (slope = -0.44), with an adjusted p-value of 3E-11 (based on a Normal distribution and FDR correction by the Benjamini-Hochberg procedure). Note that no other genes have adjusted p-value<0.05; our method rejects all other genes as potential false positives. This is represented by the leftmost bar in the plot. Table 2 gives an overall summary of the top hits after the slope adjustment.

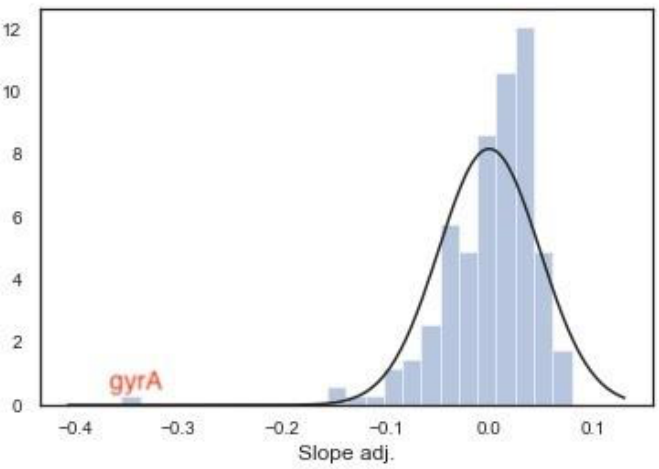

Figure 13 Distribution of adjusted slopes of genes treated with various concentrations of levofloxacin.

CHAPTER IV

RESULTS

Table (2) summarizes the targets of the various drugs used to treat our hypomorph library.

Table 2 Summary of top-ranked hit for various drugs.

Hit type	Drug	Expected Target	Relevant hits (rank)	Adjusted p-value
Protein Hits	Levofloxacin	DNA Gyrase (gyrA)	gyrA* (1)	3E-11
	Moxifloxacin	DNA Gyrase (gyrA)	gyrA (2)	0.55
	Fidaxomicin	Translation (rpoB)	rpoB*(2)	0.01
Pathway Hits	Isoniazid	Lipid biosynthesis	Lipid pathway --- KasB(8), desA1(15), desA2 (21), fabD(48), kasA(144)	0.32
	Bedaquiline	ATP synthase	ATP proton motive force pathway --- atpF(2), atpH(4), atpG(9), atpB(48)	0.13
	Sulfamethaxazole	Folate synthesis (folP1)	Folate pathway --- trpG(7), folB(29), folP1(52), dfrA (64)	0.30

Asterisks mark those that are significant (adjusted p-value < 0.05).

Levofloxacin

For levofloxacin, *GyrA* (the expected target of fluoroquinolones) was the top hit and only significant gene. The depletion effect with increasing inhibitor concentration is shown in Fig (6b) above, reflecting the chemical-genetic interaction between levofloxacin and *GyrA*. We also observed that *gyrA* has the lowest Z-score for levofloxacin as compared to the other drugs as evident in Fig. (14). This further confirms the strong interaction between gyrase genes with levofloxacin.

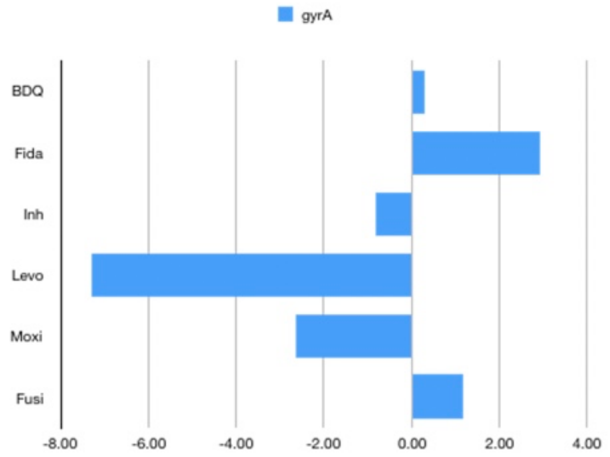

Figure 14 Barplot of the slopes of *gyrA* in various drugs.

Moxifloxacin

Moxifloxacin is also a drug of the fluoroquinolone family and the target of this drug is also *gyrA*. However, with the current data, *thyA* comes up as the most significant hit. *GyrA* comes up as the second most depleted gene. The boxplots for the log-abundances vs the log-concentrations of these two genes when treated with moxifloxacin is shown in Fig. (16). But an intriguing observation is that *thyA* comes up as the top hit for many of the drugs and thus can be ignored as a non-specific artifact. This is evident from the bar plot shown in Fig. (15).

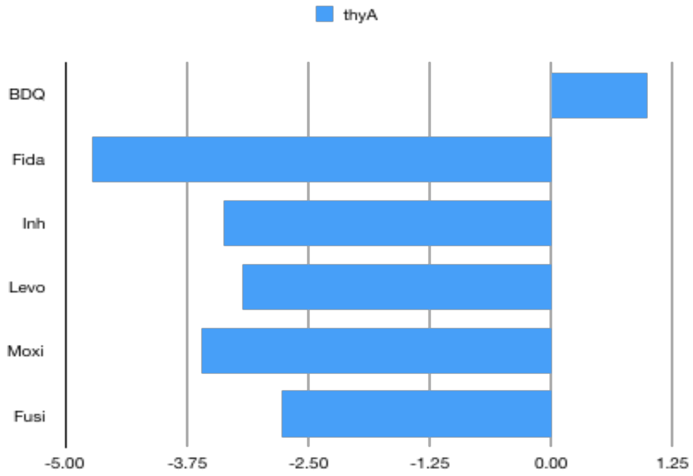

Figure 15 Barplot of the slopes of *thyA* in various drugs.

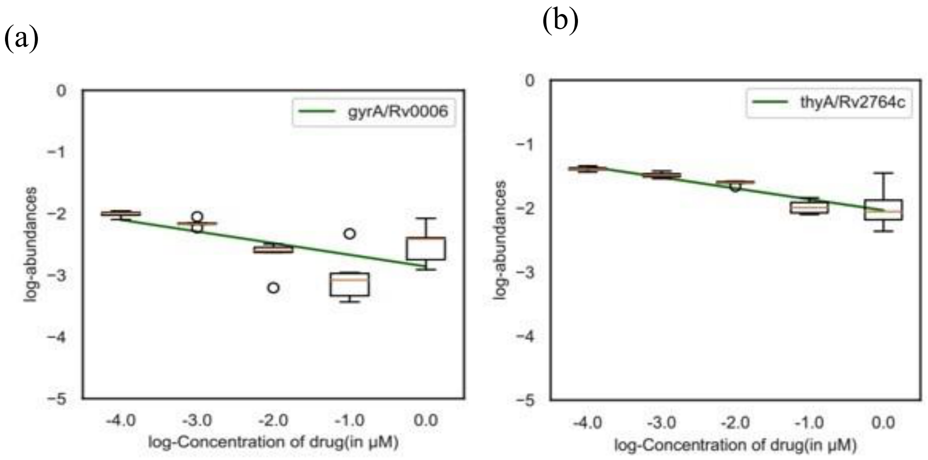

Figure 16 Boxplots of the slopes of (a) *gyrA* and (b) *thyA* when treated with moxifloxacin.

Bedaquiline

Bedaquiline targets genes of the ATP synthase pathway (*atpBDGH* are in the library), specifically subunit C of the membrane complex (Andries et. al, 2005). Though our analysis ranks *atpF* as the second lowest slope, in this case, there are no statistically significant hits. However, pathway analysis gives us more insight on how the organisms respond to bedaquiline.

It is possible that, even if none of the genes in each pathway has a significantly negative slope, the genes as a group might exhibit a systematic bias, showing depletion as a group with increasing drug concentration (Table 3). We applied GSEA analysis (Subramaniam, 2005), which ranks the genes by slope, calculates an enrichment score (ES) reflecting whether the mean rank of a subset (e.g., genes in a pathway) is above or below average, and then determines the statistical significance of the ES using Monte

Carlo sampling. We ran GSEA on the 24 COG categories (clusters of orthologous genes; (Galperin, Makarova, Wolf, & Koonin, 2015) on all the drugs. For both COG and Sanger pathways (Ashburner et al., 2000; Cole et al., 1998), we observed that ATP synthase/energy production pathway comes up as the most significant pathway when treated with bedaquiline (Table 4). The 4 *atp* genes present in this library show a systematic depletion when considered as a group (Fig (17)).

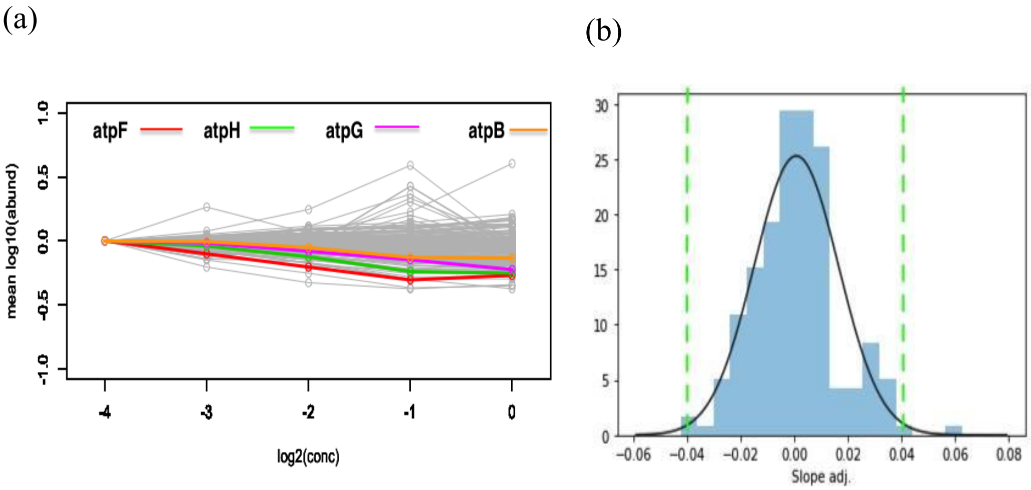

Figure 17 (a) This figure represents the trend of the log-abundances relative to the abundance at 0xMIC with respect to log-concentrations of bedaquiline. The highlighted lines represent *atpH, atpG, atpF and atpB*. (b) This plot represents the histogram of the distribution of the slopes of all the genes treated with bedaquiline.

Table 3 Summary of top hits of Bedaquiline

gene	Rank	Slope(random)	Slope adj.	p-value (adj.)	q-value(adj.)
atpF/Rv1306	2	-0.039	-0.0362	0.0096	0.844
atpH/Rv1307	4	-0.037	-0.0269	0.04	0.999
atpG/Rv1309	9	-0.029	-0.0240	0.0587	0.999
atpB/Rv1304	49	-0.019	-0.009	0.2621	1

Table 4 Summary of analysis of top Sanger pathways for bedaquiline (out of 152 pathways).

Pathway	Mean rank	ES	P-value	q-value	Description	genes
I.B.8	15	0.76	0.003	0.147	ATP-proton motive force	atpF/Rv1306(1) atpH/Rv1307(3) atpG/Rv1309(8) atpB/Rv1304(48)
I.A.1	29	0.74	0.088	0.527	Carbon compounds	manA/Rv3255c(9) adoK/Rv2202c(49)
I.B.7	33.5	0.71	0.09	0.527	Miscellaneous oxidoreductases and oxygenases	ndhA/Rv0392c(11) ccsX/Rv3673c(56)
I.J.1	34	0.76	0.034	0.367	Repressors	Rv2017/Rv2017(23) moxR1/Rv1479(45)

Isoniazid

The target for Isoniazid (INH) is *inhA* (enoyl-ACP reductase), which is in the FAS II pathway for synthesis of long-chain fatty acid, and ultimately mycolic acid. INH is a pro-drug that must be activated first to a radical by the *KatG* catalase. Thus, one might expect that depletion of *inhA* would be synergistic with INH treatment (causing barcode counts to decrease), and depletion of *KatG* would be antagonistic (causing barcode counts to increase due to enhanced survival). However, neither *inhA* nor *katG* is in the hypomorph library. Using our analysis, we observed *ino1* as the topmost significant hit. This is evident from the plots in Fig (18). The depletion of this gene can be attributed to the fact that inositol-1-phosphate synthase is used in mycothiol, which has been connected to INH MOA via redox homeostasis (Vilcheze & Jacobs, 2019). The boxplot for this gene is shown in the figure below.

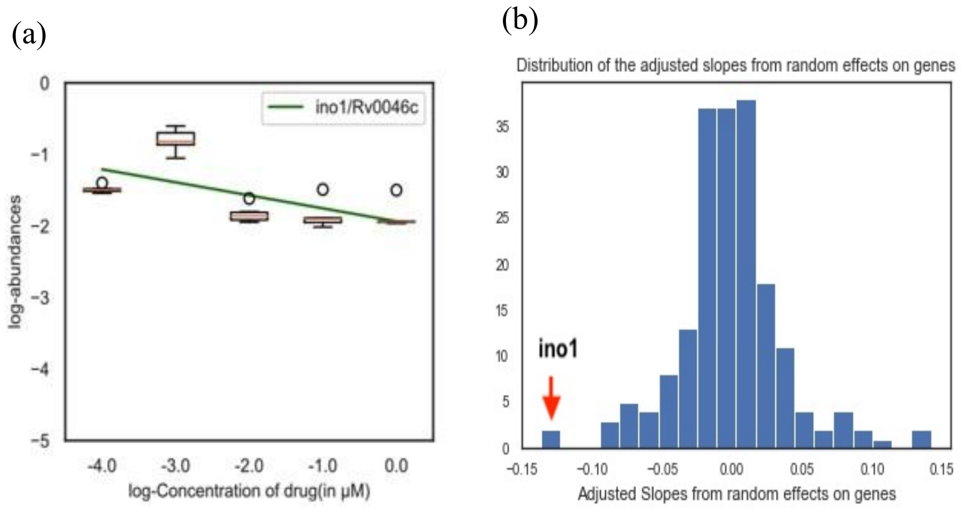

Figure 18 (a) Boxplot of the slopes of ino1 (top hit) (b) Distribution of slopes of all genes when treated with Isoniazid.

Pathway analysis yields some interesting results for isoniazid. We observe that of the various sanger pathways, fatty acid synthesis related pathways such as lipid biosynthesis, synthesis of fatty and mycolic acid come up as significant pathways. Isoniazid impacts the fatty acid related pathways, which agrees with our observation. These results are summarized in the Table (5) below.

Table 5 Summary of analysis of top Sanger pathways for isoniazid (out of 152 pathways).

Pathway	Mean rank	ES	P-value	q-value	Description	genes
V	59.2	0.502	0	0	Conserved hypotheticals	Rv1836c/Rv1836c(11) Rv0289/Rv0289(12) Rv3194c/Rv3194c(20)
I.H	46.4	0.597	0.001	0.0663	Lipid Biosynthesis	desA1/Rv0824c(5) desA2/Rv1094(10) Rv0904c/Rv0904c(15)
I.H.2	7.5	0.952	0.003	0.110	Modification of fatty and mycolic acids	desA1/Rv0824c(5) desA2/Rv1094(10)
I.B.8	40.8	0.675	0.01	0.237	ATP-proton motive force	atpH/Rv1307(24) atpB/Rv1304(32) atpF/Rv1306(43) atpG/Rv1309(64)
I.H.1	57.6	0.525	0.011	0.256	Synthesis of fatty and mycolic acids	Rv0904c/Rv0904c(15) fabD/Rv2243(21) Rv2247/Rv2247(27)
I.A.3	36.5	0.702	0.013	0.276	Fatty acids	fadD30/Rv0404(1) fadD32/Rv3801c(36) accD2/Rv0974c(50) accA2/Rv0973c(59)

Fidaxomicin

Fidaxomicin targets *rpoB* and inhibits transcription initiation (Boyaci et. al, 2018). Using our analysis, we see *rpoB* (slope = -0.13, adjusted p-value = 0.01) as the second most significantly depleted gene. *ThyA* comes up as the topmost hit but that is likely to be a false positive as it comes up as a top hit for a lot of other drugs as well. Fig. (19) indicates the boxplot of *rpoB* and the distribution of the slopes of all the other genes when treated with fidaxomicin.

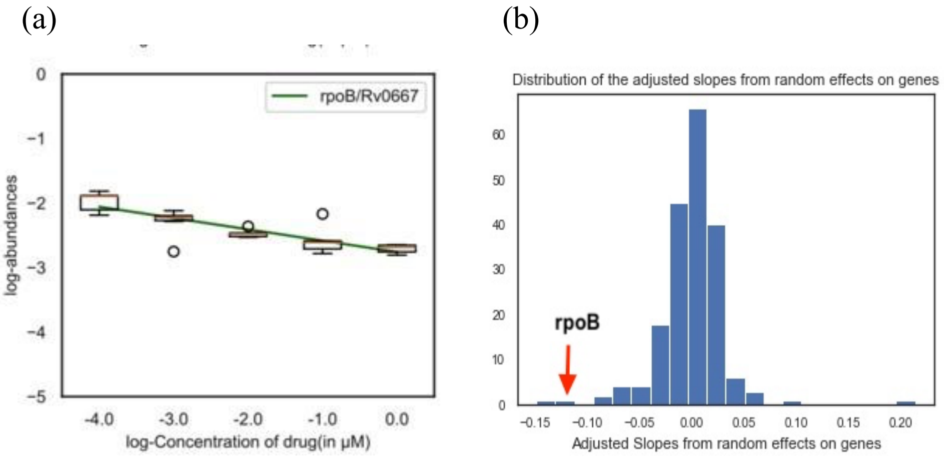

Figure 19 (a) Boxplot of the slopes of rpoB (top hit) (b) Distribution of slopes of all genes when treated with fidaxomicin

Sulfamethaxazole

The target of sulfamethoxazole is DHPS/*folP* (dihydropteroate synthase) in the folate pathway. However, DHPS was not present in the hypomorph library. Our analysis identifies *fas* (fatty acid synthase) as the most significantly depleted gene. However, it has been observed previously that *trpG* (anthranilate synthase) also interacts with folate synthesis by consuming chorismate as an intermediate (shared with the shikimate and tryptophan pathways) to make PABA, constituting a known chemical-genetic interaction (Johnson et. al, 2019). *trpG* comes up as the 8th ranked gene in terms of depletion (slope = -0.05; q-value = 0.97). Fig (20) shows the boxplot of *fas*, *trpG* and the distribution of the slopes of all the other genes when treated with sulfamethoxazole.

GSEA analysis based on Sanger categories indicate that the folic acid pathway comes up as the most significant pathway for the library of genes treated in sulfamethoxazole. This is a positive hit because sulfamethoxazole impacts the activity of folic acid pathway, including other genes such *as folB, folP1, and dfrA*, thus, validating our analysis. While individually, they did not have significantly negative slopes, they all exhibit a negative trend, which is statistically unlikely by chance. The pathway analysis results are summarized in Table (6).

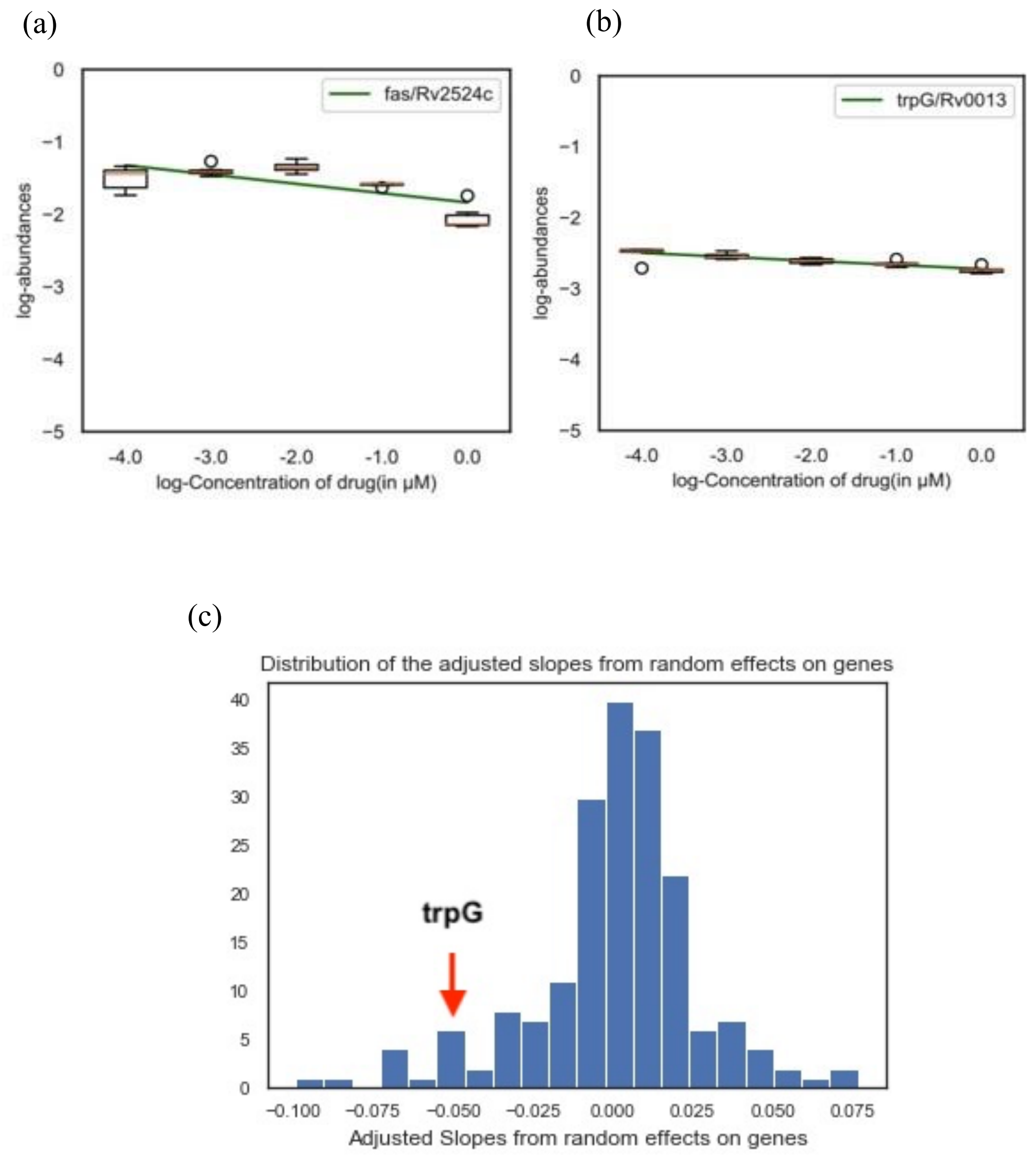

Figure 20 Boxplot of the slopes of (a) fas and (b) trpG and (c) Distribution of slopes of all genes when treated with sulfamethoxazole

Table 6 Summary of analysis of Sanger pathways for sulfamethoxazole

Pathway	Mean rank	ES	P-value	q-value	Description	genes
I.G.2	34.8	0.679	0.002	0.120	Folic acid	trpG/Rv0013(7) Rv0812/Rv0812(22) folB/Rv3607c(29) folP1/Rv3608c(52) dfrA/Rv2763c(64)

Fusidic Acid

The target of fusidic acid is *fusA*. *FusA* is elongation factor G, a component of the ribosome; hence fusidic acid is a translation inhibitor. But our current library does not include a hypomorphic strain for *fusA*. Our analysis for this drug indicates no significant hits. *ThyA* comes up as the topmost depleted gene, but it is likely to be false positive because it comes up for a lot of other drugs as well. Fig. (21) indicates the distribution of slopes for all the genes when treated with fusidic acid. As evident from the figure, there is no significant outlier when the library of genes is profiled for various concentrations of fusidic acid.

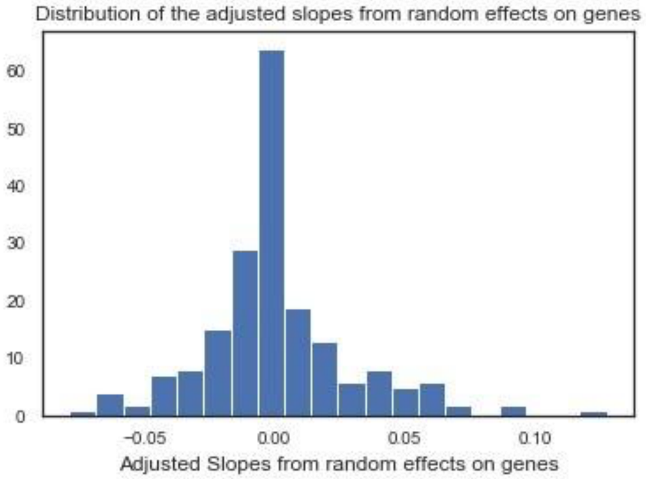

Figure 21 Distribution of slopes of all genes when treated with fusidic acid

Analysis of Published Hypomorph Data Created by Broad Institute

The chemical-genomics data created by the Broad institute (https://www.broadinstitute.org/chemical-biology/initiative-chemical-genetics) is used to validate our model. This data consists of the raw abundances of 155 mutants profiled at various concentrations of 4 drugs – rifampin, trimethoprim, methotrexate and BRD-4592.

Rifampin

The target of rifampin is *rpoB*, the RNA polymerase. Our analysis indicates that *rpoB* is indeed one amongst the top 10 depleted genes. Johnson et al., 2019 also has a similar observation in terms of the target of rifampin. Fig. (22) shows that *rpoB*, though not the most significantly depleted, has a steady decrease in the read abundances with increased concentration of the drug. The outliers in this case are *thyA* which is seen to have low slopes with other drugs as well. This is probably a non-specific artifact.

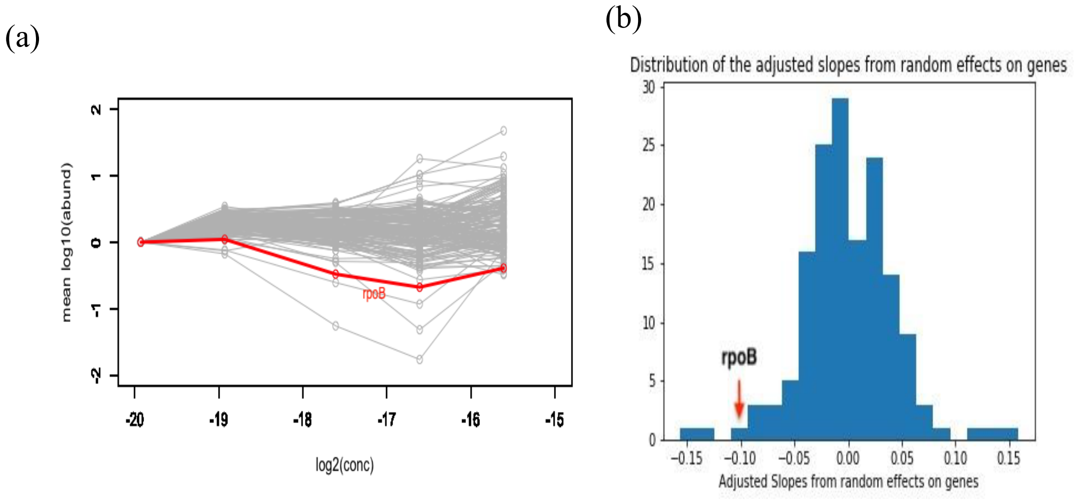

Figure 22 (a) Plot from our analysis, the red line indicates *rpoB.* Our observation is consistent with the one reported in the paper. (b) Histogram of the distribution of slopes of all the genes in this library.

Trimethoprim

Trimethoprim (TMP) is a widely used anti-tuberculosis drug which targets a gene dihydrofolate reductase, *dfrA* in the folate pathway. However, in this dataset, *dfrA* has a medium-to-positive slope, and hence does not appear as an interaction with TMP. In contrast, *trpG* has the 2nd most negative slope (slope = -0.005, adjusted p-value = 0.02), consistent with its role in the folate pathway. *TrpG* converts chorismate into PABA as the first step in the pathway. This interaction of *trpG* with TMP is also reported by Johnson et. al using ConCensusGLM. This is further evident from our plot in Fig (23).

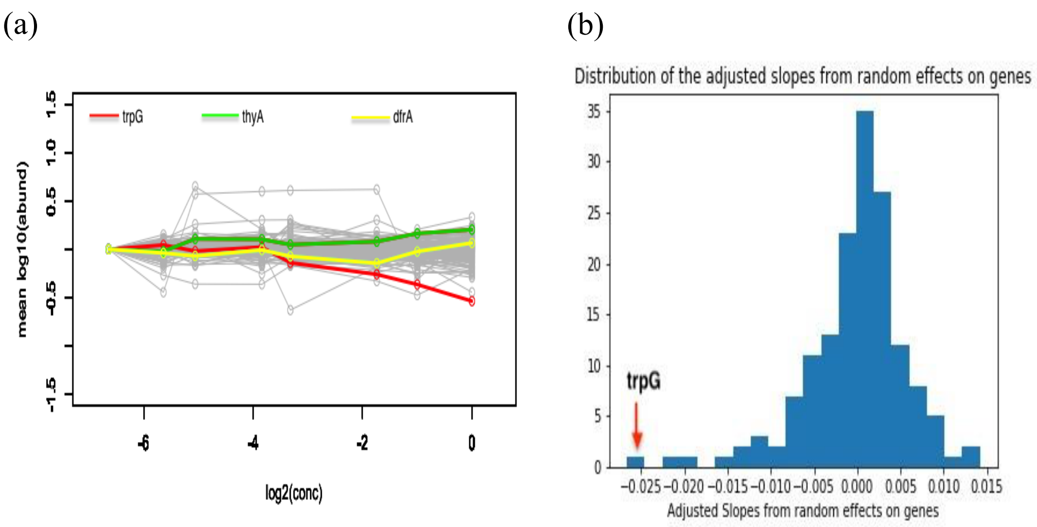

Figure 23 (a) This plot represents the trends of the log-abundances when the library is treated with trimethoprim. (b) Distribution of the slopes of all the genes treated with trimethoprim.

Methotrexate

Like trimethoprim, methotrexate also targets genes related to the folate pathway. Our analysis yields *trpG* as the one with the most negative slope. This is expected as *trpG* indeed belongs to the folate pathway. However, this slope is not significant as evident from Fig (24). There is no significant outlier in this case (Table 7).

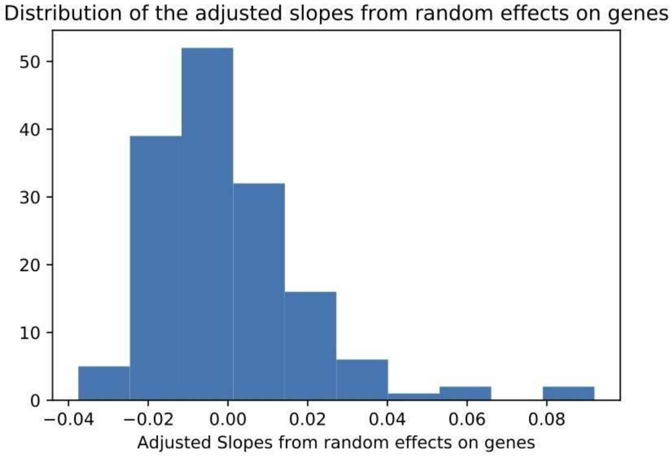

Figure 24 Distribution of slopes of all genes when treated with methotrexate.

Table 7 Summary of the slopes of genes when treated with methotrexate

gene	Slope(random)	Slope adj.	p-val(adj.)	q-value(adj.)
trpG	-0.0707	-0.037	0.0231	0.9736
Rv2190c	-0.0581	-0.0264	0.0787	1
pcnA	-0.0567	-0.026	0.0818	1
aceE	-0.05415	-0.0257	0.0843	1
pstP	-0.0621	-0.0252	0.0887	1

BRD-4592

The target of BRD-4592 is *trpA,* a gene in the pathway for synthesizing tryptophan (which is essential in these growth conditions). However, this gene is not present in the hypomorph library. On evaluating our analysis for this drug, we observe no significant hits. This is expected because the target gene is not a part of the library we are working with. Fig. (25) indicates that the distribution of slopes has no outliers. In this case, the results of our model are inconclusive.

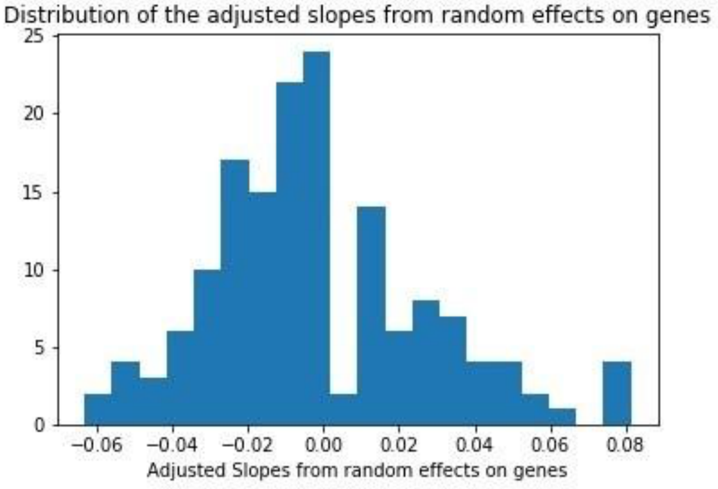

Figure 25 Distribution of slopes of all genes when treated with BRD-4592

Copper

Copper is known to be bactericidal at high concentrations, but at lower concentrations, it induces a tolerance mechanism involving genes such as *ricR* and *csoR* as sensor-regulators, cation transporter *ctpV*, metallothionein *mymT*, multi-copper oxidase *mmcO*, *socAB*, and *lpqS* (Darwin, 2015). None of these genes is essential in-vitro under regular growth conditions. We constructed a new hypomorph library with 465 essential genes and selected it for growth on 3 different carbon sources – glycerol, acetate, and cholesterol – in the presence of varying concentrations of copper sulfate (1 to 8 μM) with 3 different *sspB* expression strengths for different degrees of expression based on the Tet promoter (denoted *sspB*-1, *sspB*-2, and *sspB*-6), resulting in different degrees of proteolytic depletion of hypomorph targets. Analysis of the chemical-genomics data showed only two genes that had statistically significant depletion (*gyrA* in *sspB*-2/glycerol, and *thyA* in *sspB*-6/glycerol). While interactions with *thyA* were observed with several other drugs, it is unclear what the relevance of *gyrA* (the DNA gyrase that is the target of fluoroquinolones) is to copper exposure. However, GSEA analysis using the Sanger functional categories (Cole et al., 1998) yields an intriguing insight. Genes in the Murein Sacculus pathway (*murA, murD, murE, murF*) are enriched, in that they have negative slopes as a group (slopes ranging from -0.02 to -0.06). This effect appears to be independent of carbon source, as the murein pathway is ranked at the top in glycerol, acetate, and cholesterol, though occurring at different *sspB* strengths (different levels of proteolytic degradation) (Table 8). Fig. (26) shows the plot of the log-abundances of the library of genes when treated with copper in cholesterol (*sspB*-1). The genes in this pathway all have a weak but consistent negative trend when treated with copper. *mur* genes play an essential role in synthesizing the peptidoglycan (PG) layer in the cell wall (i.e., muramic acid as a component of lipid II, used to transport pentapeptides to the cell-wall for PG assembly and cross-linking), thereby maintain cell wall integrity. Our results are consistent with previous reports of a connection between peptidoglycan synthesis and copper sensitivity. Copper has been shown to specifically inhibit L, D-transpeptidases in *E. coli* (Peters et al., 2018). A similar mechanism in *M. tuberculosis* could explain the chemical-genetic interaction with the *mur* genes. Inhibition of LDTs by copper could be sensitized by depletion of *mur* genes, which are in the same PG pathway, producing excess growth impairment through reduction of cell-wall integrity. This is a novel observation for *M. tuberculosis*.

Table 8 Summary of ranks and adjusted p-value of the Murein Sacculus pathway (based on murADEF) when treated with copper, in 3 different carbon sources, with 3 different sspB strengths. The ranks are out of 122 total Sanger pathways.

Carbon source	Rank in sspB-1 (Adjusted p-val)	Rank in sspB-2 (Adjusted p-val)	Rank in sspB-6 (Adjusted p-val)
Cholesterol	**1 (0)***	**1 (0.04)***	1 (0.09)
Glycerol	8 (0.58)	4 (0.11)	**1 (0.04)***
Acetate	**1(0)***	3 (0.14)	7 (0.42)

Asterisks indicate significance (adjusted p-value < 0.05).

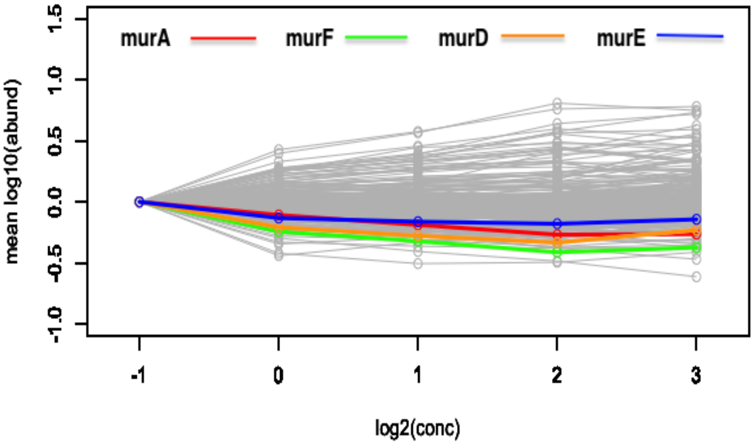

Figure 26 This figure represents the trend of the log-abundances relative to the abundance at 0xMIC with respect to log-concentrations of copper. The highlighted lines representing *murA, murF, murD, murE* shows a systematic depletion for increasing copper concentration.

CHAPTER V

DISCUSSION

The CGA-LMM approach uses linear mixed models to identify genes in a hypomorph library that interact with drugs (or growth inhibitors), thus potentially yielding insights into targets, pathways, or mechanisms for action. Conceptually, this approach to analyzing chemical-genomics data is based on the synergy between drug pressure and protein-depletion of drug targets in the library. In general, presence of an inhibitor (at sub-MIC concentrations) would be expected to partially inhibit the growth of all the members of a hypomorph library. Independently, when essential genes are depleted, such as by targeting them for degradation by the ClpXP protease, growth will be inhibited, though the degree of growth impairment between different mutants might vary. Chemical-genomics experiments are designed to look for synergies between these two effects. The depletion specifically of the target of a given drug should be hypersensitive to that mutant to the drug, resulting in more depletion than the rest of the population. Recall that, although the presence of an antibiotic might reduce the population density of the culture overall, this reduction is effectively factored out in the normalization process, where barcode counts from deep sequencing are converted into relative abundances of genes. The total number of reads for any given sample is arbitrary, depending on loading of sample on the instrument. So, based on relative abundances, each gene will potentially exhibit some degree of reduction in abundance due to the fitness deficit caused by protein degradation. But for the target of a drug, these effects combine, producing additional depletion, analogous to super-additive effects between two synergistic drugs (Chou, 2010). Furthermore, ideally, we are looking for genes where the hypersensitization is concentration-dependent, that is, depletion in abundance increases with increasing concentrations of drug. Genes which exhibit the same depletion across all concentrations would be ruled out as effects due to depletion of an essential protein; genes which exhibit depletion at just a single concentration, but not higher concentrations, would be ruled out as false positives.

In the approach we have described for analyzing chemical-genomic data, a key aspect is assessing the depletion of knock-down mutants across multiple concentrations, which is expressed through regression coefficients (representing "slopes"). This approach captures genes that exhibit a robust trend, or concentration-dependent effect, where increasing concentration of drug causes increasing (or equal) depletion. This contrasts with other approaches, which assess the depletion of a gene at each concentration independently, in comparison to a no-drug control (typically as a log-fold-change) (Johnson et al., 2019; Li et al., 2014) The advantage of a regression-based approach is that it takes more data into account, by integrating information across multiple concentrations, and thus is less susceptible to spurious fluctuations in observed counts that might drop to a low abundance at one concentration but not others, generating a false positive. Our approach filters out such false positives by requiring the depletion of a gene observed at one concentration is reinforced by similar depletion at higher concentrations. It is important to acknowledge that not every CG-interacting gene responds in a uniform way to increasing drug concentration. The concentration-dependence is not always linear (graded decrease), but sometimes decreases precipitously at a critical concentration (like a cliff, similar to classic enzyme inhibition curves, and the steepness of the slope can be influenced by cooperativity.) Nonetheless, the regression model will still detect such cases as a decreasing trend overall with a negative slope.

Importantly, we observed that sometimes, genes implicated in response to, or tolerance of, treatment with an antibiotic are only weakly depleted individually. However, if multiple genes in the target pathway are represented in the library, it might be possible to detect the interaction through pathway analysis. We observed this effect for both exposure to bedaquiline (ATP synthase genes) and copper (*mur* genes). Even though none of the pathway members might be statistically significant on their own, if there is a systematic effect, where each of the pathway genes exhibits partial depletion (negative slopes), it could be detected as statistically significant, indicating synergy between the drug and the pathway. This shows that sensitivity of detection drug targets that are members of a complex can be enhanced because other members of the complex (and hence pathway) can collectively show depletion effects. A similar phenomenon is observed when the hypomorph library is treated with various concentrations of isoniazid. This drug is known to increase sensitivity to knockdown the genes related to the fatty acid synthesis pathways. CGA-LMM followed by the GSEA analysis indicates that the genes belonging to the synthesis of fatty and mycolic acid and lipid biosynthesis show a systematic significant depletion as compared to the other genes in the library. Sulfamethoxazole depletes genes in the folic acid pathway such as *folP1, folB and dfrA*. Though none of these genes individually are significantly depleted, the pathway has a downward trend in the log-abundances (mean rank: 34.6, adjusted p-value: 0.007) when exposed to higher concentration of the drug. To study the influence of the drugs on the pathways, we have used the GSEA analysis on the ranks based on predicted slopes of the genes from the CGA-LMM approach on the Sanger and COG categories and GO terms. However, a shortcoming of the current analysis is that even if a few of the genes of a pathway are significantly depleted, then the entire pathway is identified as being impacted. Thus, pathway-based analysis based on the GSEA approach is sensitive to the top hits identified by the CGA-LMM model. The pathway analysis is meaningful only if multiple genes of the pathways are in the library under consideration. If the hypomorph library is such that it just comprises 1-2 genes of each pathway in the library, then the results on such a library can be inaccurate/misleading.

Computationally, our approach utilizes a linear mixed model (LMM) to assess these concentration-dependent depletion effects for each gene. While, theoretically, the observations for each gene could be used to fit a regression model for each gene independently, we chose the LMM framework because it enables the fitting of all the data simultaneously, with separate slopes and intercept coefficients for each gene. Importantly, the gene-specific parameters are treated as random effects in the LMM. This means the parameters for concentration-dependence are assumed to be unique for each gene, though they are assumed to be drawn from some multivariate normal background distribution. Although the variance of the parameters such as slope are unknown a priori, the variance is estimated empirically for the data, representing an inferred population over all the slopes. The population of slopes inferred by the LMM is exploited in determining the significance of drug-gene interactions. Our approach extracts the slope for each gene and compares it to the population using a p-value based on the Z-score, based on the mean and variance over the slopes for all the genes (extracted from the covariance matrix estimated in the LMM). This approach effectively identifies "outliers", or genes whose slopes are significantly more negative than the rest of the population. This approach is different from the conventional Wald test approach used to test if the coefficients are significant in a generalized linear model. A Wald test identifies any coefficients that are significantly different than 0. This can lead to an excessive number of hits, as observed with similar methods like ConCensusGLM (91 out of 152 reported as hits for trimethoprim), which often found many or all genes to have p-values < 10^{-10} and had to rely on other criteria such as an LFC threshold to prioritize likely candidate interactions. The perspective behind our more conservative approach is that the abundance of genes in a chemical-genomics experiment often exhibits variability due to unknown (or uncontrolled) factors, which are difficult to anticipate. This can produce multiple genes whose abundance slightly increases or decreases with concentration. It must be remembered that the hypomorph library is being subjected as a culture to two stresses simultaneously, possibly inducing a variety of intracellular adaptation mechanisms. Furthermore, there are potentially multiple sources of noise in the DNA sample preparation and sequencing steps. Acknowledging that we do not know all such factors affecting the experiment, we use the variability of the slopes over the whole library as a surrogate to estimate the net effect of these factors on the variability of slopes, and we focus on genes which exhibit a depletion beyond what is seen for the rest of the population. As seen in our experiments, this more conservative approach produces many fewer significant interactions, though hopefully enriching for true positives while filtering out false positives. In some cases, no genes might be detected as significant outliers, such as was the case for fusidic acid, the target of which, *fusA*, was not in our hypomorph library. We view this as an acceptable (though less informative) outcome; while one can always rank all the genes in a library by slope, the gene with the most negative slope does not necessarily mean it is a genuine interaction, especially if it is or near the range of the rest of the population.

The LMM methodology can also be applied, in principle, to analyzing data from CRISPRi libraries (Rock et. al, 2017). CRISPRi technology is rapidly supplanting methods such as ClpXP-mediated depletion as a way of generating hypomorph libraries. CRISPRi enables many more mutants to be profiled in parallel, through expression of sgRNAs, which knock-down transcription of target genes though binding of a catalytically dead Cas9 gene, blocking the RNA polymerase. The abundance of individual sgRNAs can again be assessed efficiently through counting nucleotide barcodes using next-generation sequencing. Nonetheless, the objective of the experiment is the same: to detect proteins whose depletion synergizes with exposure to an inhibitor. However, the LMM model would probably have to be adapted to take into account the relative strengths of different sgRNAs. There can be multiple candidate sgRNAs per gene (tens to hundreds), and the strength of binding (DNA: RNA hybridization) and hence dCas9 recruitment depends on a combination of similarity to the consensus PAM sequence, as well as GC-content of the complementary sequence. Since different sgRNAs are expected to confer different degrees of depletion, which affects the degree of synergy with the drug at each concentration, sgRNA strength would have to be incorporated in the model as a covariate, with the same goal of detecting genes that exhibit concentration-dependent depletion. Note that a regression-based approach that accounts for dependence on both drug concentration as well as sgRNA strength could be an improvement over other published methods for CRISPRi analysis, such as MAGeCK (Li et al., 2014), which only compared one concentration at a time to the no-drug control (i.e., log-fold-changes, instead of slopes), and which averaged the depletion effect over all sgRNAs for a given gene, regardless of sgRNA strength.

With regard to copper exposure, the most interesting result we observed was the interaction with *mur* genes (*murA, murD, murE,* and *murF*). While a low intracellular level of copper is required, e.g., as a co-factor for some metal-dependent enzymes like cytochrome C oxidase (Neyrolles, Wolschendorf, Mitra, & Niederweis, 2015), high concentrations of copper have long been known to have antibacterial properties. In mycobacteria, several genes have been identified to be involved in copper tolerance at moderate levels, including those in the csoR (ctpV) and ricR operons (mmcO, mymT, etc.) (Darwin, 2015), which are up regulated as concentrations of Cu2+ reach above 0.5 mM. However, these copper-tolerance genes are generally non-essential, and were not represented in the hypomorph library. Interestingly, pathway analysis revealed the several genes in the *mur* pathway display consistent depletion effects, including *murA, murD, murE, and murF*. These genes produce enzymes needed for assembly of precursors of peptidoglycan in the cell wall, specifically UDP-N-acetylmuramyl pentapeptide. It is possible that cell wall integrity affects the intracellular penetration by copper ions by acting as a passive barrier. This is supported by separate work showing that deletion of

L, D-transpeptidases in *E. coli*, which crosslink peptidoglycan, affect the sensitivity of cultures to copper (Peters et al., 2018). Additional experiments are needed to validate this chemical-genetic interaction, but it suggests that co-administration with copper could sensitize cells to cell-wall inhibitors or could be used to facilitate screening for novel inhibitors, e.g., of *murA*, potentially leading to novel combination therapeutics.

Future extension of this research involves exploring hierarchical Bayesian models to solve these problems. It would allow use to estimate posterior distributions over the regression parameters (slopes for each gene) thus resulting in a more integrated way of handling noise and testing significance than slope adjustment and *Z*-scores in our model. Furthermore, we could assert better control over the model by specifying reasonable hyperparameters for prior distributions over model parameters.

CHAPTER VI

CONCLUSION

This research is focused on developing statistical models to analyze chemical-genomic interactions. We have shown how mixed linear models can be used to quantify the behavior of genes in a library of hypomorph strains treated with various inhibitors to identify chemical-genomic interactions. These results are subjected to evaluating the statistical significance to identify the protein or the pathway exhibiting significant depletion when treated with increasing concentrations of the drug. The model was validated on a publicly available chemical-genomics dataset published by the Broad Institute. Additionally, the model was evaluated in multiple hypomorph libraries in *M. tuberculosis* developed by our collaborators that were treated with several anti-tuberculosis drugs such as levofloxacin, moxifloxacin, bedaquiline, fusidic acid, fidaxomicin, sulfamethoxazole. The proposed CGA-LMM method was used to identify the known targets of these drugs. The approach was then applied to evaluating genes implicated in tolerance of exposure to copper as an anti-bacterial.

REFERENCES

Andries, K., Verhasselt, P., Guillemont, J., Gohlmann, H. W., Neefs, J. M., Winkler, H., . . . Jarlier, V. (2005). A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science, 307(5707), 223-227. doi:10.1126/science.1106753

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25(1), 25-29. doi:10.1038/75556

Boyaci, H., Chen, J., Lilic, M., Palka, M., Mooney, R. A., Landick, R., . . . Campbell, E. A. (2018). Fidaxomicin jams Mycobacterium tuberculosis RNA polymerase motions needed for initiation via RbpA contacts. Elife, 7. doi:10.7554/eLife.34823

Chou, T. C. (2010). Drug Combination Studies and Their Synergy Quantification Using the Chou-Talalay Method. Cancer Research, 70(2), 440-446. doi:10.1158/0008-5472.Can-09-1947

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., . . . Barrell, B. G. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature, 393(6685), 537-544. doi:10.1038/31159

Darwin, K. H. (2015). Mycobacterium tuberculosis and Copper: A Newly Appreciated Defense against an Old Foe? J Biol Chem, 290(31), 18962-18966. doi:10.1074/jbc.R115.640193

Draper, N. R., & Smith, H. (1998). Applied regression analysis (3rd ed.). New York: Wiley.

Evans, J. C., & Mizrahi, V. (2015). The application of tetracyclineregulated gene expression systems in the validation of novel drug targets in Mycobacterium tuberculosis. Front Microbiol, 6, 812. doi:10.3389/fmicb.2015.00812

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res, 43(Database issue), D261-269. doi:10.1093/nar/gku1223

Johnson, E. O., LaVerriere, E., Office, E., Stanley, M., Meyer, E., Kawate, T., . . . Hung, D. T. (2019). Large-scale chemical-genetics yields new M. tuberculosis inhibitor classes. Nature, 571(7763), 72-78. doi:10.1038/s41586-019-1315-z

Kim, J. H., O'Brien, K. M., Sharma, R., Boshoff, H. I., Rehren, G., Chakraborty, S., . . . Schnappinger, D. (2013). A genetic strategy to identify targets for the development of drugs that prevent bacterial persistence. Proc Natl Acad Sci U S A, 110(47), 19095-19100. doi:10.1073/pnas.1315860110

Kim, J. H., Wei, J. R., Wallach, J. B., Robbins, R. S., Rubin, E. J., & Schnappinger, D. (2011). Protein inactivation in mycobacteria by controlled proteolysis and its application to deplete the beta subunit of RNA polymerase. Nucleic Acids Res, 39(6), 2210-2220. doi:10.1093/nar/gkq1149

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., . . . Liu, X. S. (2014). MAGeCK enables robust identification of essential genes from genome-scale

CRISPR/Cas9 knockout screens. Genome Biol, 15(12), 554. doi:10.1186/s13059-014-0554-4

Lin, K., O'Brien, K. M., Trujillo, C., Wang, R., Wallach, J. B., Schnappinger, D., & Ehrt, S. (2016). Mycobacterium tuberculosis Thioredoxin Reductase Is Essential for Thiol Redox Homeostasis but Plays a Minor Role in Antioxidant Defense. PLoS Pathog, 12(6), e1005675. doi:10.1371/journal.ppat.1005675

Lindstrom, M. L., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. Biometrics, 46(3), 673-687.

Neyrolles, O., Wolschendorf, F., Mitra, A., & Niederweis, M. (2015). Mycobacteria, metals, and the macrophage. Immunol Rev, 264(1), 249-263. doi:10.1111/imr.12265

Peters, K., Pazos, M., Edoo, Z., Hugonnet, J. E., Martorana, A. M., Polissi, A., . . . Vollmer, W. (2018). Copper inhibits peptidoglycan LD-transpeptidases suppressing beta-lactam resistance due to bypass of penicillin-binding proteins. Proc Natl Acad Sci U S A, 115(42), 10786-10791. doi:10.1073/pnas.1809285115

Rock, J. M., Hopkins, F. F., Chavez, A., Diallo, M., Chase, M. R., Gerrick, E. R., . . . Fortune, S. M. (2017). Programmable transcriptional repression in mycobacteria using an orthogonal CRISPR interference platform. Nat Microbiol, 2, 16274. doi:10.1038/nmicrobiol.2016.274

Schnappinger, D., & Ehrt, S. (2014). Regulated Expression Systems for Mycobacteria and Their Applications. Microbiol Spectr, 2(1). doi:10.1128/microbiolspec.MGM2-0018-2013

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102(43), 15545-15550. doi:10.1073/pnas.0506580102

Vilcheze, C., Av-Gay, Y., Barnes, S. W., Larsen, M. H., Walker, J. R., Glynne, R. J., & Jacobs, W. R., Jr. (2011). Coresistance to isoniazid and ethionamide maps to mycothiol biosynthetic genes in Mycobacterium bovis. Antimicrob Agents Chemother, 55(9), 4422-4423. doi:10.1128/AAC.00564-11

Vilcheze, C., & Jacobs, W. R., Jr. (2019). The Isoniazid Paradigm of Killing, Resistance, and Persistence in Mycobacterium tuberculosis. J Mol Biol, 431(18), 3450-3461. doi:10.1016/j.jmb.2019.02.016

Wei, J. R., Krishnamoorthy, V., Murphy, K., Kim, J. H., Schnappinger, D., Alber, T., . . . Rubin, E. J. (2011). Depletion of antibiotic targets has widely varying effects on growth. Proceedings of the National Academy of Sciences of the United States of America, 108(10), 4176-4181. doi:10.1073/pnas.1018301108

Wellington, S., Nag, P. P., Michalska, K., Johnston, S. E., Jedrzejczak, R. P., Kaushik, V. K., . . . Hung, D. T. (2017). A small-molecule allosteric inhibitor of Mycobacterium tuberculosis tryptophan synthase. Nat Chem Biol, 13(9), 943-950. doi:10.1038/nchembio.2420

Wright, D. B. (2017). Some Limits Using Random Slope Models to Measure Academic Growth. 2(58). doi:10.3389/feduc.2017.00058

Zewotir, T., & Galpin, J. S. (2007). A unified approach on residuals, leverages and outliers in the linear mixed model. Test, 16(1), 58-75. doi:10.1007/s11749-006-0001-2