BAYESIAN ANALYSIS OF TRANSPOSON MUTAGENESIS DATA

A Thesis

by

MICHAEL A. DEJESUS

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

MASTER OF SCIENCE

May 2012

Major Subject: Computer Science

BAYESIAN ANALYSIS OF TRANSPOSON MUTAGENESIS DATA

A Thesis

by

MICHAEL A. DEJESUS

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

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Thomas R. Ioerger
Committee Members,	James C. Sacchettini
	Tiffani L. Williams
Head of Department,	Duncan W. Walker

May 2012

Major Subject: Computer Science

ABSTRACT

Bayesian Analysis of Transposon Mutagenesis Data. (May 2012)

Michael A. DeJesus, B.S., University of Puerto Rico at Mayagüez

Chair of Advisory Committee: Dr. Thomas R. Ioerger

Determining which genes are essential for growth of a bacterial organism is an important question to answer as it is useful for the discovery of drugs that inhibit critical biological functions of a pathogen. To evaluate essentiality, biologists often use transposon mutagenesis to disrupt genomic regions within an organism, revealing which genes are able to withstand disruption and are therefore not required for growth. The development of next-generation sequencing technology augments transposon mutagenesis by providing high-resolution sequence data that identifies the exact location of transposon insertions in the genome. Although this high-resolution information has already been used to assess essentiality at a genome-wide scale, no formal statistical model has been developed capable of quantifying significance. This thesis presents a formal Bayesian framework for analyzing sequence information obtained from transposon mutagenesis experiments. Our method assesses the statistical significance of gaps in transposon coverage that are indicative of essential regions through a Gumbel distribution, and utilizes a Metropolis-Hastings sampling procedure to obtain posterior estimates of the probability of essentiality for each gene. We apply our method to libraries of *M. tuberculosis* transposon mutants, to identify genes essential for growth in vitro, and show concordance with previous essentiality results based on hybridization. Furthermore, we show how our method is capable of identifying essential domains within genes, by detecting significant sub-regions of open-reading frames unable to withstand disruption. We show that several genes involved in PG biosynthesis have essential domains.

To my family and friends.

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Thomas Ioerger, for his assistance and insight through out my studies. His impressive knowledge and infectious work ethic have been invaluable in this entire process. I would also like to thank my committee members, Dr. James C. Sacchettini and Dr. Tiffani L. Williams, for their guidance and input.

I would like to thank our collaborators Jason Zhang, Jennifer Griffin, and Christopher Sassetti for their help and cooperation in this research.

Finally, I would like to thank my family for their support, and &TOTSE for making me the kid I am today.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

A. Motivation

Determining what genes are essential for the survival of a given organism is of great interest to biologists and researchers. Knowledge of essentiality information for an organism enables the development of new drugs that inhibit essential genes, thus interfering with growth of an infectious bacteria. Furthermore, understanding which genes are essential allows scientists to have a better understanding of the evolutionary origins of life, and to better understand the function these genes play in an organism. In order to identify essential genes, libraries of mutant organisms that have had regions of their DNA disrupted by transposons have been created. New advances in sequencing have allowed for the rapid sequencing of large number of such mutants at the same time. By sequencing large libraries of these mutants, a new set of high-resolution sequence data is now available capable of revealing which areas of the genome are potentially disruptable and non-essential to the organism. Although this high-resolution sequence data has the potential of providing a wealth of new information about essentiality, this data also poses a new set of problems that make any quantitative analysis of this data challenging. By sequencing libraries of mutants that survived transposon insertion in their DNA, we can get an accurate picture of sites within the genome that can tolerate interruption. However genomic regions lacking insertions do not necessarily imply that the region is essential to the organism. These areas may represent sites that were simply missed by chance during mutagenesis but are otherwise non-essential to the organism. Furthermore, many essential genes are

---

The journal model is *IEEE Transactions on Automatic Control.*

able to withstand some insertions within their coding regions. While transposon insertions are supposed to disrupt the gene, in reality genes are often able to tolerate insertion in the N- and C-terminus, as the protein may still be translated and able to fulfill its biological function in spite of the insertion [1, 2]. Although at first one may be tempted to determine essentiality based on whether a gene shows evidence of insertions or not, these challenges (like the fact that some essential genes may withstand insertions) make this type of simplistic analysis impractical.

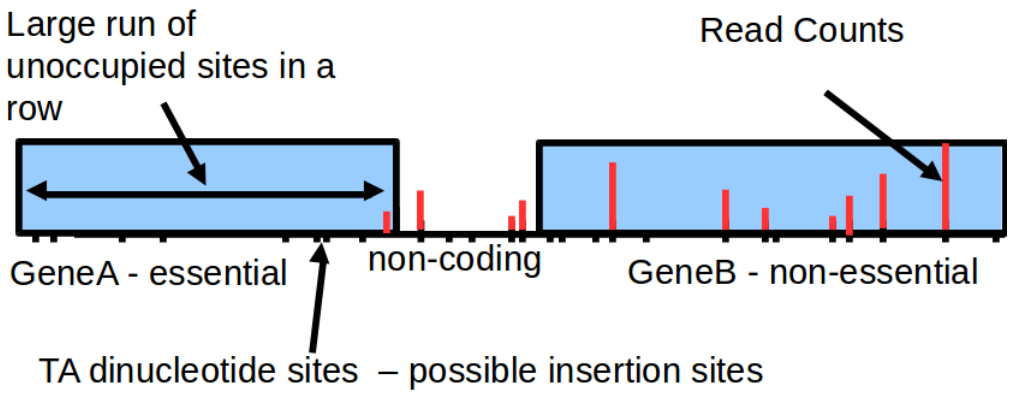


Fig. 1.: Diagram of Transposon Insertions in Essential and Non-Essential Genes

A more plausible analysis may be to use the proportion of insertions within a gene to assess essentiality. With this point of view, one may label those genes with a significantly lower number of insertions as essential and those with a significantly larger number of insertions as non-essential. However this approach is flawed as well. In reality, genes may code for multiple domains and some of these domains might

be essential for growth while others not. For example in *M. tuberculosis* Rv3198c (UvrD2) was shown to contain an essential N-terminal helicase domain (Pfam00580) and a non-essential C-terminal HRDC domain (Pfam00570) [3]. These domains play different roles within organism. While the role of the HRDC domain remains unknown, Williams et. al. showed that it does not significantly affect ATPase or helicase activity, where as the ATPase activity of the helicase domain was shown to be necessary for growth. Because this region sustains insertions, any attempt to model essentiality solely on the proportion of insertions will have trouble picking out these essential regions or essential domains.

Our approach to analyzing this sequence data is to examine at the maximum consecutive sequence of non-insertions in a row within any given gene. Because the Himar1 transposon used in these experiments is capable of inserting at any TA dinucleotide site within the genome, we can identify sites where it is missing. By using a Gumbel (Extreme Value) distribution, we can quantify the expected run length of non-insertions in a consecutive sequence of TA sites and determine whether an observed run length within a gene significantly deviates from our expectations. Those genes with longer-than-expected runs of non-insertions are less likely to be non-essential, since they imply the gaps of non-insertions are unlikely to be produced by chance. We use a Bayesian framework based on this Gumbel model to formally develop our analysis, estimating the parameters of our model by using Metropolis Hastings sampling algorithm.

The following section provides an overview of the background necessary to understand the basis of this statistical analysis. Section 1 contains a brief explanation of what is meant by an essential gene, and how this information is of use to biologists. Section 2 gives an overview of the transposon mutagenesis experiment and the relevant biology behind the sequence data, as well as a brief review of the related

literature surrounding these experiments and previous attempts to use this data to determine gene essentiality. Finally Section 3 explains the statistical framework that underlies our model.

B. Background

1. Essentiality

The purpose of transposon mutagenesis experiments is to identify which genes are essential to an organism. An essential gene is defined as one whose loss is lethal to the organism under a certain environmental condition. For example, genes that are involved in core metabolism function, protein translation, or DNA replication, are known to be essential in most organisms. The growth conditions of the organism are an important factor in determining whether a gene is essential or not. While many genes are essential to an organism in any given situation, some genes are only essential if a particular function is necessary for the organism in its current environment (e.g., presence or absence of a particular nutrient). Furthermore, transposon mutagenesis experiments can also be used to determine essential genes in vivo during infection in animal models [4]. By making inferences about the essentiality of genes in a particular growth condition, new insight is gained that shed lights on what roles and functions those essential genes might play within the organism. With such information, new drug candidates can be developed that are capable of inhibiting a certain protein or disrupting its function, and therefore targeting infectious bacteria. For example, the first-line anti-tuberculosis medication is isoniazid, which inhibits a key enzyme (enoyl-ACP reductase) necessary for biosynthesis of the mycolic acid required in the cell wall that is essential for *M. tuberculosis* [5]. Furthermore, essentiality information can make a more thorough understanding the evolutionary history of bacteria possible;

by analyzing this data we can get a picture of the minimum set of genes needed for a bacterial organism [6].

2. Transposon Mutagenesis

One of the most important techniques available to answer the question of essentiality is transposon mutagenesis. Transposons are small fragments of DNA (typically 1-2kb long) that can insert within the chromosomes of an organism [7]. Although transposons occur naturally in most bacteria, transposon insertions can also be mediated in-vitro, forcing new insertions to take place within the organism. The Mariner family of transposons are of particular interest as they have been shown to insert at random sites within the genome of bacterial organisms [8, 9]. The Himar1 transposon, for example, has shown specificity for arbitrary TA dinucleotides [10, 11]. This characteristic enables the construction of large libraries of mutants that have random regions of their DNA disrupted by the transposon insertions. It is these libraries of mutants that can help provide a better understanding of the the essential genes within an organism. Once a library of mutants is created, these mutants are then cultured and grown under an environmental condition of interest. Any transposon insertion within the coding region of a gene should interrupt the translation of its protein, usually destroying its function (see Figure 2). Therefore, those mutants capable of growing under specific conditions are those with insertions in genes that did not play any essential function for growth in this environment.

Once a library of mutants is created, it is necessary to identify the precise location where insertions took place to identify those regions that are not essential for growth. What follows is a brief review of the methods that have been developed to identify where these disruptions took place, and the previous attempts to determine essentiality from transposon mutagenesis experiments. This will hopefully put our

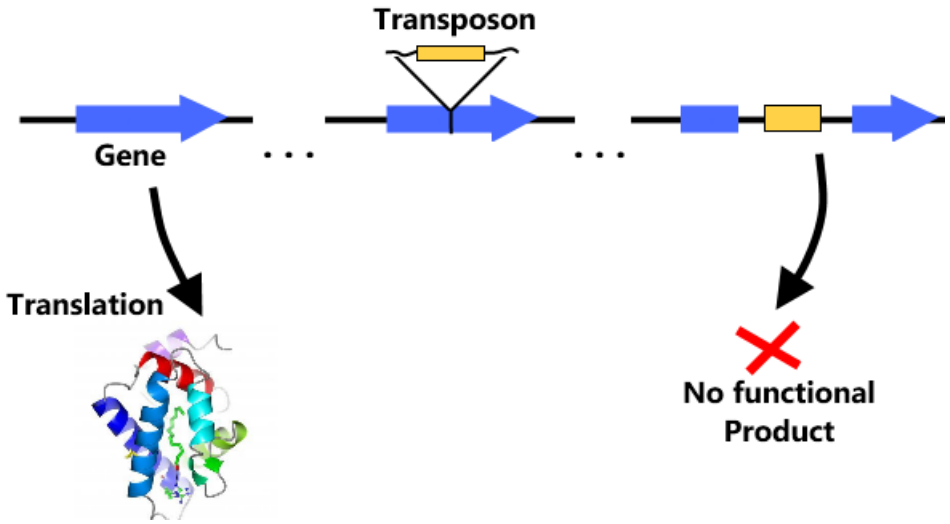


Fig. 2.: Visual Depiction of Transposon Mutagenesis. An essential gene codes for a protein which is translated by the organism. If a transposon disrupts the coding region of the essential gene, no functional product is created and the absence of this functioning protein prevents growth.

contributions into context, and show both the usefulness of new sequence technology as well as the importance of the statistical analysis we have developed.

a. TraSH

In 2001 Sassetti et. al [12] introduced a method called Transposon Site-Hybridization (TraSH) to identify essential genes within mycobacteria. Although there had been earlier ways of determining the survival of transposon mutants, these methods were far more labor intensive and generally unable to handle large libraries of mutants or a large number of genes at the same time. TraSH overcame these limitations by using micro-array hybridization to determine what genes in the mutant libraries where being expressed and which ones where not. Primer extension was used to amplify from the regions at the ends of the transposon out into the surrounding genomic regions, and these products where then identified by hybridization to gene-specific probes. After

hybridization takes place, micro-arrays detect the fluorescence signals from the probes and quantify the ratio of hybridization for the genes, which is subject to considerable noise in the read-out. Using this approach, genes necessary for optimal growth in a variety of organisms have been identified (e.g., *M. tuberculosis*, *H. influenza*, and *B. anthracis* [13, 14, 4]). In *M. tuberculosis* 614 out of 3,989 genes were initially identified as essential using TraSH [14]. However one substantial limitation of TraSH experiments was that it was incapable of identifying the exact coordinate where the transposon insertion took place. Although TraSH revealed what genes were being disrupted, it did not provide the high-resolution data (i.e., coordinates of insertions) necessary to interpret the effect at a molecular level.

b. High Density Mutagenesis and Deep-Sequencing

Traditional sequence methods can be used to overcome the problem of low-resolution information, and pin-point the exact coordinates in the genome where the insertions took place. However traditional sequencing was impractical for analyzing the large and complex libraries of transposon mutants that were available at the time. With the development of high-throughput sequencing and deep-sequencing, large libraries of mutants can be sequenced at the same time, providing high-resolution information about the location of the insertions.

High-density mutagenesis coupled with deep-sequencing (for example, using next-generation sequencers from Illumina or Roche) is the latest method used to determine the essentiality of genes, and has been used successfully to determine essentiality in a number of different organisms and growth conditions [15, 16, 17]. Although these sequencing techniques have been used for several years, no standard method for analyzing the output data exists. Previous methods for analyzing the data have relied on ad-hoc criteria. Gawronski et. al. [15] for example, required the exclusion

of insertions in the first and last 5-20% of the coding region of a gene so as to remove spurious insertions in essential genes.

In this thesis we introduce a novel approach to identifying essential genes by modeling the insertions at TA dinucleotides sites within each gene as a set of Bernoulli trials (implicitly assuming independence between sites), and then detecting statistically significant gaps of non-insertions within the genes [17]. This new way of analyzing high-resolution sequence data is developed into a Bayesian statistical analysis, allowing rigorous probabilistic estimates of essentiality from this data.

3. Statistical Framework for Analyzing Essentiality

In order to understand how sequence data from transposon mutagenesis experiments can be used to estimate the essentiality of genes, observations of insertions at TA sites are compared by analogy to coin-tossing. In a regular coin tossing scenario, we are confronted with a finite number of coin tosses resulting from a coin with a certain probability of heads and tails. In such a domain, we are often interested in knowing the probability of heads or tails, or the likelihood of observing a given pattern of insertions. Furthermore, by knowing the pattern of insertions, we can make inferences on the weight of the coin that is likely responsible for the observations.

With this analogy in mind, we can simplify the information contained within the sequence data that results from sequencing libraries of transposon mutants. Because the Mariner transposon inserts at random TA dinucleotides sites within the genome, there is a finite number of places where insertions can take place. Using the coin analogy, we can model the presence of insertions as independent coin tosses, with each gene containing a finite set of tosses depending on the number of TA dinucleotide sites that exist within it. If a TA site happens to have an insertion, we can say the outcome of that toss was that of "heads". If it does not have an insertion, we can

say that the outcome for that toss was that of "tails". By using this analogy, we can turn the sequence data into a set of Bernoulli trials from which we can gather important statistics that help us gain an understanding about essentiality. We can estimate the probability of insertion (i.e., probability of heads) and the probability of non-insertion (i.e., probability of tails) in essential and non-essential genes, and calculate the likelihood of observing the pattern of insertions.

In particular, by using this analogy we can determine the maximum run of non-insertions (i.e., biggest run of tails in a row) within a given gene, and use that to calculate how unexpected this observation was. Other types of information can also be obtained from these insertion patterns. For instance, we could characterize the proportion TA sites where insertions where observed by using a Binomial distribution, or we could identify those genes that are completely devoid of insertions. However, we believe that the run of non-insertions is more indicative of essentiality. For instance, because some genes have multiple domains with different functions, both of these alternative approaches would have trouble correctly evaluating their essentiality. A Binomial model may characterize a gene as non-essential based on the proportion of insertions in a non-essential domain, and completely miss a large run of non-insertions indicative of an essential domain, which the Gumbel model is able to detect.

CHAPTER II

METHODS

A. Overview

From the sequence data of a transposon mutant library, we obtain reads mapping to TA dinucleotide sites (TA sites) throughout the genome. Using this set of reads we create a list of all the TA sites within the the genome, and the number of reads that mapped to each individual site (read counts). Since our analysis depends on the transposon insertions that took place within a given gene, we adopt a binary representation of the data and represent those locations containing transposon insertions with "1", and those locations lacking insertions with "0".

By parsing the data in this manner, we can represent the TA sites within a given gene as a set of Bernoulli trials, with success and failure representing observations of insertion and non-insertion (i.e., 1 or 0) at any given TA site. Using Bernoulli trials allows us to model the insertions at different sites as independent from each other. Given a sequence of Bernoulli trials corresponding to each gene, we can then characterize the longest run of non-insertions in a row with a Gumbel (Extreme Value) distribution and determine if this run is significantly longer than expected.

Genes within this framework are represented as a mixture of two assignments: non-essential genes (assigned a value of 0) and essential genes (assigned a value of 1). Another possible category of genes could be those genes for which a disruption causes a growth-defect in the organism, however we do not make that distinction in our analysis. Section B describes this mixture model in a Bayesian framework. Section C presents the sampling methodology used to estimate the parameters of the Gumbel model, as well as the essentiality assignments.

B. Bayesian Mixture Model

1. Likelihood

Let $Y_i = \{r_i, n_i\}$ represent our observations for the i-th gene for $i = 1...G$, where r_i and n_i represent the total number of TA sites and the largest run of non-insertions observed in each gene. The essentiality assignments for all genes is represented by the unknown variable Z, with the individual assignment for i-th gene represented by the boolean vector Z_i which accepts binary values of 0 and 1 for non-essential and essential. These two classes of genes represent the two categories found in the mixture model. The mixture coefficient representing the prevalence of the category in the mixture is given by $\omega = \{\omega_1, \omega_0\}$. Finally, we assume a global non-insertion probability, ϕ_0, that governs probability of non-insertions across all non-essential genes. This is 1 minus the insertion density observed at non-essential genes.

We wish to estimate a complete joint probability density, $p(Z, Y, \phi_0)$, from which we can derive posterior estimates of essentiality of each gene, conditional on the data $p(Z|Y, \phi_0)$. To accomplish this we rewrite this joint probability in terms of the likelihood of the data and our prior expectations: $p(Y|Z, \phi_0) * p(Z) * p(\phi_0)$. We assume independence among genes, so our likelihood can be written as a product of our individual observations: $p(Y|Z, \phi_0 \propto \prod_i p(Y_i|Z, \phi_0)$. We use sampling methods to derive estimates of these posterior probabilities.

a. Non-Essential Genes

Our model depends on characterizing the expected length of the longest run of non-insertions within non-essential genes. To accomplish this, we use the Gumbel (Extreme Value) distribution. The Gumbel distribution models the distribution of extreme or maximum values obtained from a finite set of independent and identically

distributed samples. By maximizing over repeated samples of values, the shape of the Gumbel distribution is skewed to the right, producing a "fatter" tail in the right side of the distribution, allowing for extreme values to have a higher probability than being observed than they normally would with the underlying distribution. The Gumbel distribution has the following form:

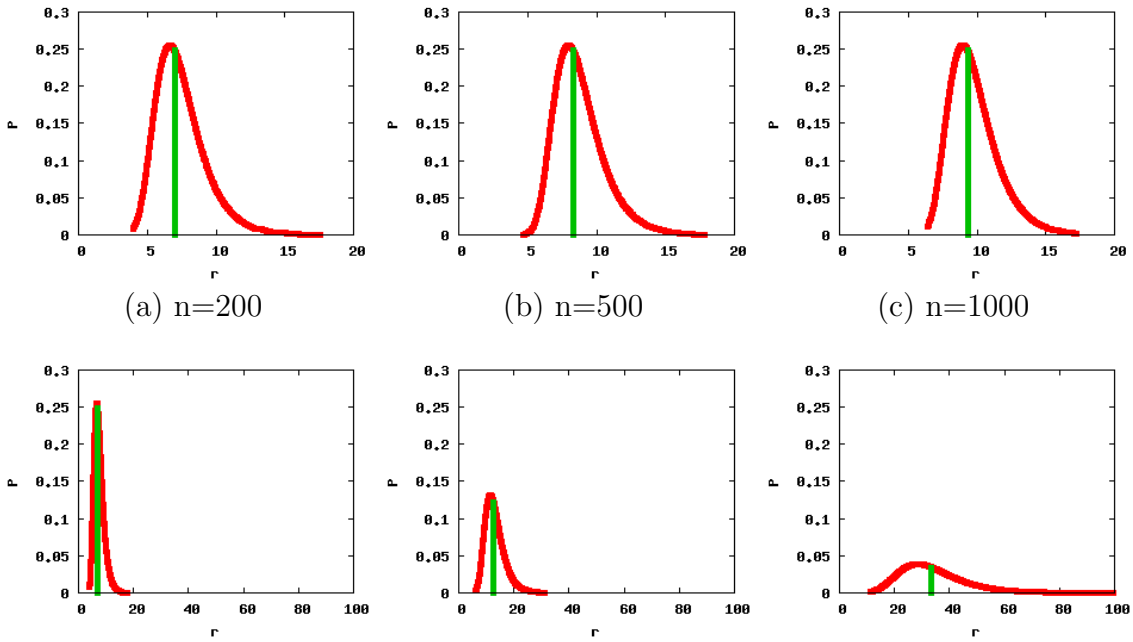
$$\begin{aligned} Gumbel(x;\mu,\sigma) &:= \frac{1}{\sigma}e^{-z-e^{-z}} \\ z &= \frac{x-\mu}{\sigma} \end{aligned} \tag{2.1}$$

where μ and σ are the parameters of the underlying distribution which govern the location and scale of the function, which can be any function belonging to the exponential family of distributions. In analogy to coin-tossing, these parameters are functions of the probability of non-insertion, ϕ_0, and of the total number of trials, n [18]:

$$\begin{aligned} \mu &= \log_{\frac{1}{\phi_0}}(n(1-\phi_0)) \\ \sigma &= \frac{1}{\log\frac{1}{\phi_0}} \end{aligned} \tag{2.2}$$

Figure 3 shows distributions of the longest runs of heads in a series of coin tosses, and the expected run, for different values of n and different values of ϕ_0. The expected maximum run scales up logarithmically in n and $1-\phi_0$ as n.

Because the Gumbel distribution depends on ϕ_0, we must estimate its value from our data. Previously [17] we estimated the ϕ_0 parameter in an ad-hoc manner by averaging the frequency of insertions within non-essential genes, removing those genes we pre-determined as essential (based on TraSH analysis done by Sassetti et al. [14]). In this formal Bayesian framework, we treat ϕ_0 as a Bayesian parameter and



(a) n=200 (b) n=500 (c) n=1000

(d) ϕ_0=0.5 (e) ϕ_0=0.7 (f) ϕ_0=0.9

Fig. 3.: Gumbel Distributions with Different Values of ϕ_0 and n. The vertical bar shows the expected maximum run according to the Gumbel distribution.

estimate it by sampling from its conditional density. Using this Gumbel model, the likelihood for a given non-essential gene based on the maximum number of consecutive TA sites without insertions, r, is:

$$p(Y|\phi_0) = \frac{1}{\sigma} e^{\left(-\frac{x-\mu}{\sigma} - e^{\left(-\frac{x-\mu}{\sigma}\right)}\right)} \tag{2.3}$$

with μ and σ parameters as defined in formula (2.2).

b. Essential Genes

We use a relatively simple uniform distribution for essential genes. The rational behind this choice of distribution is that our model is designed so that essential genes are defined in contrast to non-essential genes. Those genes which have an unusually

long run of non-insertions according to the Gumbel distribution will get classified as essential by contrast. By choosing an uniform distribution for essential genes, we can make use of the naturally small likelihood of large runs of non-insertions being explained by our model of non-essential genes.

$$p(Y|\phi_0, Z_i = 1) = U(r) = u \tag{2.4}$$

We use $u = 10^{-2}$ for all genes except for those with very small maximum runs of non-insertions (i.e. max run less than 5), where $u = 0$. The rationale for this is that those genes with a very small run of non-insertions would never be considered essential through an analysis of insertions.

c. Complete Data Likelihood

By making an independence assumption, the complete data likelihood of our model can be expressed as the product of independent likelihoods for all genes G. We can further decompose this likelihood into a product over all the non-essential genes times the product over all the essential genes:

$$\begin{aligned} p(Y_{obs}|\phi_0, Z) &= \prod_{i}^{G} p(Y_i|\phi_0, Z_i) \\ &= \prod_{Z_i=0} Gumbel(r_i, \mu_1, \sigma_1) \times \prod_{Z_i=0} U(r_i) \\ &= \prod_{Z_i=0} \left[\frac{1}{\sigma} e^{\left(-\frac{r_i-\mu_i}{\sigma_i} - e^{\left(-\frac{r_i-\mu_i}{\sigma_i}\right)} \right)} \right] \times \prod_{Z_i=1} [U(r_i)] \end{aligned} \tag{2.5}$$

2. Prior Probabilities

a. Prior Probability of ϕ_0

Our model depends on estimating the posterior probability of non-insertion at non-essential genes, ϕ_0, which is used in our Gumbel model. To quantify our prior expectations of this parameter, we use a Beta distribution as our prior:

$$\pi(\phi_0) = Beta(\phi_0; \alpha_0, \beta_0) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \phi_0^{\alpha_0 - 1}(1 - \phi_0)^{\beta_0 - 1} \tag{2.6}$$

where α_0 and β_0 are hyper-parameters that capture our expectations for ϕ_0. The Beta distribution is often used as a prior distribution for continuous variables like probabilities or percentages (i.e., variables bounded from 0 to 1), since the Beta distribution is conjugate with a number of different distributions (e.g., Binomial), which simplifies some calculations.

b. Prior Probability of Z

The prior probability of a complete essentiality assignment, Z, is given by a Binomial distribution:

$$\pi(Z) = Binomial(\omega_1; G, K_z) = \binom{G}{K_z} \omega_1^{K_z}(1 - \omega_1)^{G - K_z} \tag{2.7}$$

where ω_1 is the mixing coefficient for "essential" genes, G is the total number of genes, and K_z is the sum of the binary vector if essentiality assignments (i.e., $K_z = \Sigma Z_i$).

3. Full and Conditional Distributions

In order to estimate values of the missing data, Z, and parameter ϕ_0, we need to derive the conditional densities of these variables to sample from. Given our likelihood

formulas and our prior expectations, we can write the full joint probability formula for our model as:

$$\begin{aligned} p(Y, Z, \phi_0) &= p(Y|\phi_0, Z) \times \pi(\phi_0) \times \pi(Z) \\ &= \prod_{i}^{G} p(Y_i|\phi_0, Z_i) \times \pi(\phi_0) \times \pi(Z) \\ &= \left[\prod_{Z_i=0} Gumbel(r_i, \mu_1, \sigma_1) \times \prod_{Z_i=1} U(r_i) \right] \times \pi(\phi_0) \times \pi(Z) \end{aligned} \tag{2.8}$$

Having derived the complete joint distribution (2.5) we can then derive conditional distributions for the missing data, Z, and parameter ϕ_0 which we can then use to compute posterior estimates of these values.

a. Conditional Distribution for ϕ_0

In order to derive our posterior distribution for the ϕ_0 parameter, we make use of proportionality to cancel out any constants within the conditional distribution.

$$\begin{aligned} p(\phi_0|Y, Z) &\propto p(Y|\phi_0, Z) \times \pi(\phi_0) \times \pi(Z) \\ &\propto p(Y|\phi_0, Z) \times \pi(\phi_0) \\ &\propto \left[\prod_{Z_i=0} Gumbel(r_i, \mu_1, \sigma_1) \times \prod_{Z_i=1} U(r_i) \right] \times Beta(\phi_0; \alpha_0, \beta_0) \end{aligned} \tag{2.9}$$

b. Conditional Distribution for Z_i

Finally, in order to sample essentiality assignment for all genes, we must also derive the posterior distribution for each individual Z_i (i.e., essentiality assignment of each gene):

$$
\begin{aligned}
p(Z_i|Y, Z_{\{-i\}}, \phi_0) &\propto p(Y|\phi_0, Z_{\{-i\}}) \times \pi(Z_i) \times \pi(Z_{\{-i\}}) \\
&= p(Y_i|\phi_0, Z_{\{-i\}}) \times \pi(Z_i) \\
&= \left[Gumbel(r_i, \mu_1, \sigma_1)^{1-Z_i} \times U(r_i)^{Z_i}\right] \times \pi(Z_i)
\end{aligned} \tag{2.10}
$$

where $Z_{\{-i\}}$ is the vector of essentiality, Z, minus the i-th essentiality assignment, and $\pi(Z_i)$ is equal to the mixing coefficient for the category of gene specified by Z_i (i.e., ω_1 for $Z_i = 1$ and ω_0 for $Z_i = 0$).

C. Sampling

Once we have our conditional distributions for the missing data, Z, and our probability of non-insertion in non-essential genes, we wish to generate a sample that would represent the joint distribution. By obtaining a sample of values taken from this distribution, we can find estimates of the posterior probabilities for these parameters. Although ultimately we are interested in estimating the essentiality of all genes, the challenge is obtaining these estimates without knowing the probability of non-insertion, ϕ_0, beforehand. By sampling from the conditional density of parameter ϕ_0 at the same time as we sample Z, we can obtain estimates of the individual essentiality assignment without having to know or guess parameter ϕ_0. Since the posterior distributions or our model do not have known forms, we must utilize a sampling procedure that allows us to sample from arbitrary distributions. For our method, we use a Metropolis-Hastings algorithm (MH) to sample from the posterior distribution of ϕ_0 (2.9), and take a Gibbs Sampling step at each iteration to sample from the posterior distributions of Z_i (2.10).

1. Gibbs Sampling

Gibbs sampling is one of the most popular Markov-Chain Monte-Carlo (MCMC) sampling procedures used in Bayesian inference. The general idea behind Gibbs sampling is that, while a full joint density may be difficult or impossible to sample from, if the joint density can be reduced to conditionals with known forms, we can effectively sample from the series of conditional probabilities and generate a sequence of MCMC estimates that closely approximate the full joint density of interest [19]. The general Gibbs sampling procedure is explained in Algorithm 1. By splitting the joint density into conditional probabilities from which we can easily sample, we can arrive at a MCMC sample from the entire conditional probability density. We sample from these conditional probabilities in a iterative fashion, using the most recently sampled value of the previous parameter in the conditional probability of the parameter that is to be sampled next.

Result: MCMC Sample of Joint Density p(θ_1,θ_2,θ_3...θ_k)

Assign random starting values, S, to the vector of parameters $\Theta^{j=0}$, and set j=0;

while $j <$ *Desired Sample Size* **do**

- set j = j + 1;
- Sample p($\theta_1^j \mid \theta_2^{j-1}, \theta_3^{j-1} ... \ \theta_k^{j-1}$) ;
- Sample p($\theta_2^j \mid \theta_1^j, \theta_3^{j-1} ... \ \theta_k^{j-1}$) ;
- Sample p($\theta_3^j \mid \theta_1^j, \theta_2^j ... \ \theta_k^{j-1}$) ;
- ...;
- Sample p($\theta_k^j \mid \theta_1^j, \theta_2^j, \theta_3^j ... \ \theta_{k-1}^{j-1}$) ;

end

Algorithm 1: General Gibbs Sampling Algorithm

In order to easily sample values of the posterior distribution of Z_i, we calculate the posterior distribution for both values of Z_i, $p(Z_i = 0|Y, Z_{\{-i\}}, \phi_0)$ and $p(Z_i = 1|Y, Z_{\{-i\}}, \phi_0)$ and sample from them as Bernoulli trial with probability proportional to their posterior density:

$$z_i^{(j)} \sim Bernoulli(p_1) \tag{2.11}$$

$$p_1 = \frac{p(Z_i = 1|Y, Z_{\{-i\}}, \phi_0) \times \pi(Z_i = 1)}{p(Z_i = 1|Y, Z_{\{-i\}}, \phi_0) \times \pi(Z_i = 1) + p(Z_i = 0|Y, Z_{\{-i\}}, \phi_0) \times \pi(Z_i = 0)}$$

$$p_1 = \frac{Gumbel(r_i|Z_i = 1, Z_{\{-i\}}, \phi_0) \times \omega_1}{Gumbel(r_i|Z_i = 1, Z_{\{-i\}}, \phi_0) \times \omega_1 + U(r_i|Z_i = 1, Z_{\{-i\}}, \phi_0) \times \omega_0}$$

2. Metropolis Hastings

Although Gibbs sampling works well when conditional probabilities have a known distribution that is easy to sample from, it does not work when a form of the conditional probability we need to sample from is unknown. To sample from the posterior distribution for ϕ_0 we use the Metropolis-Hastings (MH) algorithm. While there are other methods capable of sampling from arbitrary distributions (e.g., rejection sampling, inversion sampling), these methods have drawbacks such as inefficiency or intractable analytical derivations. Rejection sampling, for example, requires one to find an function that envelopes the target distribution, and this often not an easy task if one does not know the shape of the target distribution, or the domain of the function is infinite. Furthermore, rejection sampling often requires rejecting many samples before accepting a variable as coming from the desired distribution making

the method impractical.

The MH algorithm circumvents these problems by using a proposal distribution that generates perturbed new candidate values to accept or reject. For example, a Gaussian distribution centered around the last accepted value, θ^{j-1}, and a small variance can be used. The values drawn from this proposal distribution are then accepted or rejected probabilistically; accepting if $f(\theta^t) > f(\theta^{t-1})$ or with probability proportional to $\frac{f(\theta^t}{f(\theta^{t-1})}$ if $f(\theta^t) < f(\theta^{t-1})$, where $f(\theta)$ is the conditional probability density of parameter θ [19]. While this may appear similar to rejection sampling, the proposal function does not have to envelope the target function. This effectively performs a random walk around the distribution of interest, with the parameter θ migrating around those regions within the distribution that are most likely given the data. By using the MH algorithm, we can sample from the likelihood function for non-essential genes and get an update of the ϕ_0 parameter and the essentiality assignment of all genes, Z. Algorithm 2 shows the random-walk MH algorithm, which uses a normal distribution with $\mu = \theta^{j-1}$ and $\sigma^2 = v$ to propose new candidates, as it applies to our domain.

Result: MCMC Samples of density $p(Z|Y,\phi_0)$ and $p(\phi_0|Y,Z)$

Assign starting values to $\phi_0^{j=0}$, and Z, and set j = 0;

while *$j <$ Desired Sample Size* **do**

set j = j + 1;

Draw candidate parameter ϕ_0^c from normal distribution Gaussian(ϕ_0^{j-1}, 0.001);

Compute ratio R = $\frac{p(\phi_0^c|Y,Z)}{p(\phi_0^{j-1}|Y,Z)}$;

Draw u $\sim$ U(0,1) ;

if $R > u$ **then**

Set $\phi_0^{(j)} = \phi_0^c$;

else

Set $\phi_0^{(j)} = \phi_0^{j-1}$;

end

$\omega_1^{(j)} = Beta(\alpha_w + K_z, \beta_w + G - K_z)$ $\omega_0^{(j)} = 1 - \omega_1^{(j)}$

for $i \leftarrow 1$ **to** G **do**

$p_1 = \frac{Gumbel(r_i|Z_i=1,Z_{\{-i\}},\phi_0)\times\omega_1}{Gumbel(r_i|Z_i=1,Z_{\{-i\}},\phi_0)\times\omega_1 + U(r_i|Z_i=1,Z_{\{-i\}},\phi_0)\times\omega_0}$ $Z_i^{(j)} \sim Bernoulli(p_1)$;

end

end

Algorithm 2: Random-Walk Metropolis-Hastings Algorithm for Sampling ϕ_0 and Z

CHAPTER III

RESULTS

In this chapter we evaluate our model by applying it to deep-sequence data from several transposon mutant libraries, focusing on growth of *M. tuberculosis* in vitro. We compare those genes identified as essential to previous essentiality results in TB, as well as show how our method can be used to identify essential domains within genes. In addition, we perform a differential analysis of genes essential for growth on cholesterol, which is needed by TB within a host during infection. Lastly, we examine the convergence of the sampling procedure used to estimate the parameters and estimates of essentiality for our model.

A. Essentiality Analysis of *M. tuberculosis*

We applied our Bayesian analysis on deep-sequencing data obtained libraries of *M. tuberculosis* (TB) Himar1 transposon mutants grown in minimal media and 0.1% glycerol (library constructed by J. Griffin) [17]. The TB genome is 4,411,654bp long and contains a total of 3,989 open reading frames (ORFs) [20]. TB contains a total of 74,605 TA sites within its genome, with 62,847 of them occurring in coding regions. Although the average number of TA sites within an ORF is 15.9 TA sites per gene, 41 ORFs do not contain any TA dinucleotides within them. We utilized reads from two independent libraries, which we summed together in order to get higher sampling of the TA sites. The libraries were sequenced with an Illumina GAII sequencer, and a read length of 36bp (6-8 million reads per library). Of the total TA sites in the genome, 44,350 had reads mapping to them showing evidence of a transposon insertion at those locations, 31,715 of which were at TA sites within the ORFs. We assume that sites with a small amount of reads (i.e., less than 5) represent spurious reads

possibly due to sequencing errors, and therefore those sites were treated as lacking any insertions (i.e. "0").

The sampling process was run for 50,000 iterations, providing essentiality estimates for all genes, as well as the parameter ϕ_0. Parameters were initialized as follows:

- ϕ_0: The probability of non-insertion for non-essential genes was initially set as $\phi_0 = 0.5$, meaning a 50% chance of non-insertion.
- α_w, β_w: The hyper-parameters for our mixing coefficient were set to $\alpha_w = 600$, $\beta_w = 3400$, to quantify our expectation that roughly 15% of the genome should be essential.
- Z: The vector of essentiality assignments, Z, was initialized according to the assignments found by Griffin et al. [17].
- v: The variance parameter for the proposal distribution of the MH sampling procedure is set to $v = 0.001$.

To ensure that the algorithm mixes well and the samples obtained are uncorrelated, the first 1,000 samples are treated as a "burn-in" period and discarded, and then only every 20th sample is kept there forward. Convergence of the Metropolis Hastings sampling procedure is examined in Section C.

Once the final trimmed sample is obtained, the estimate for the probability of non-insertions at non-essential genes, ϕ_0, and the posterior probabilities of essentiality, Z_i, are estimated by averaging the sampled obtained:

$$p(Z_i|Y) = \frac{1}{n}\Sigma_t Z_i^{(t)}$$

Figure 4 shows the trajectory of the ϕ_0 parameter and the percentage of genes labeled as essential after the first 1,000 iterations. obtained. The mean value of ϕ_0 across the sample was $\phi_0 = 0.344 \pm 0.005$. This parameter represents the probability of non-insertion at non-essential genes, hence 66% of the TA sites had insertions in non-essential genes (i.e., relatively high density). To verify that our result makes sense, we can calculate the frequency of non-insertions across those genes that our method ultimately infers to be non-essential as the proportion of sites without insertions divided by the total number of sites within those genes. This empirical estimate had a value of $\phi_{emp} = 0.358$, very similar to the value estimated by the model.

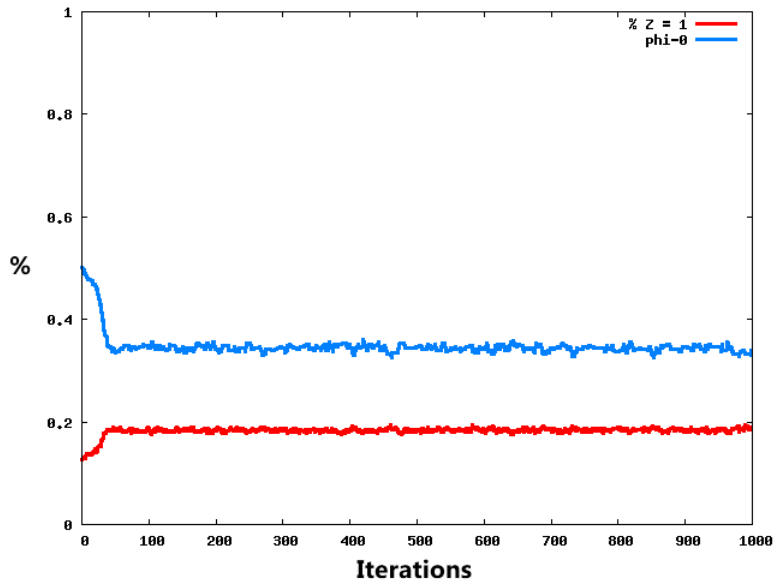


Fig. 4.: Trajectory of ϕ_0 and Percent of Essentials Genes During Sampling. The blue line shows values for the first 1,000 samples of the ϕ_0 parameter. The red line shows the proportion of genes labeled as essential for the first 1,000 iterations.

B. Essentiality Results and Comparisons

Because our method depends on identifying unusually long runs of non-insertions within the set of TA sites of a given gene, we expect our statistical analysis to predict those genes with larger runs of non-insertions relative to the entire set of TA sites to be essential with higher probability. Figure 5 shows a plot of TA count, n, and maximum run of non-insertions, r, for essential and non-essential genes. Essential genes generally lie along the diagonal, as these are genes where the maximum run of sites without insertions equals the total number of TAs within the gene.

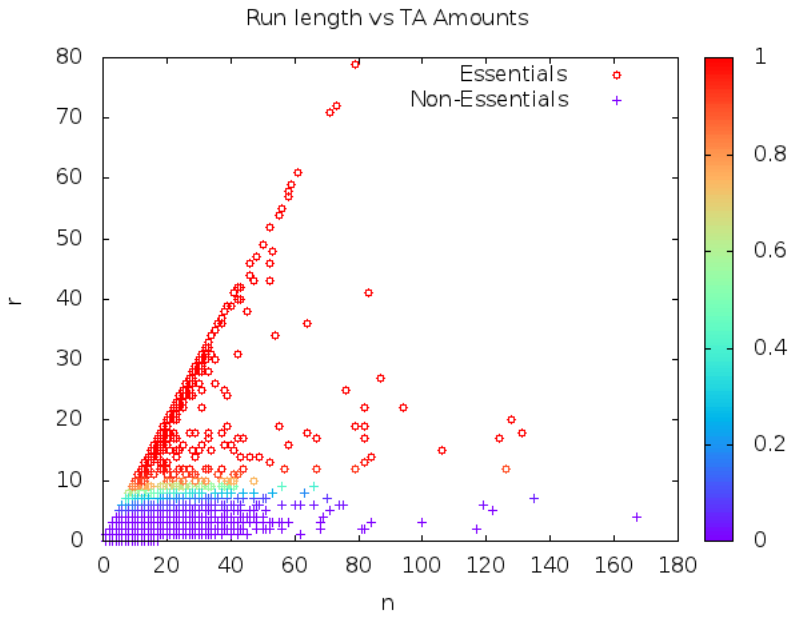


Fig. 5.: Plot of Maximum Run Length vs Number of TA Sites for Each Gene. The maximum run of non-insertions is plotted against the number of TA sites within the gene. The color gradient represents the posterior probability, Z_i, for the genes.

Figure 6 shows a cumulative plot of the posterior probabilities for all genes (i.e., Z_i). Of all the genes, 2933 have a posterior probability of essentiality less than

0.05 (i.e., $Z_i < 0.05$ "non-essential") and 531 which have a posterior probability of essentiality greater than 0.95 (i.e., $Z_i > 0.95$). These genes represent those which we are confident of the essentiality inferred by our method. This leaves a total of 482 which have a posterior probability between 0.05 and 0.95 (i.e., $0.05 < Z_i < 0.95$ "essential"), which are those genes for which the essentiality estimate may be less reliable. In general these genes are those for which the run of non-insertions is not strongly indicative of either category given the total number of TA sites within the gene.

Table I contains some statistics for these classes of genes. As expected, the average length of the maximum run for essential genes (19.68) was significantly higher than that of non-essential genes (1.80). Non-Essential genes contained a significantly higher amount of insertions (10.04) compared to other categories, however the average number of insertions within essential genes (2.16) was greater than zero, confirming that our method is not sensitive to a small amount of insertions within essential genes. Finally, essential genes were larger on average than non-essential genes (average of 499.32 amino acids and 24.58 TA sites, compared to 304.93 amino acids and 14.16 TA sites). This difference in average size may be due to the fact that shorter genes are unlikely to contain sufficient TA sites to produce significant run of non-insertions, generally lowering the average size of non-essential genes.

To determine whether our Bayesian analysis produces results compatible with what is known about the essentiality of individual genes within *M. tuberculosis*, we compared our predictions with a list of genes whose essentiality has been previously determined. Genes that are involved in core biological functions (e.g., DNA replication, metabolism) are well-known in the literature as essential to sustain bacterial life. Table II shows a list of some of these genes, along with some known to be non-essential, their biological function, and the essentiality assignments inferred from our

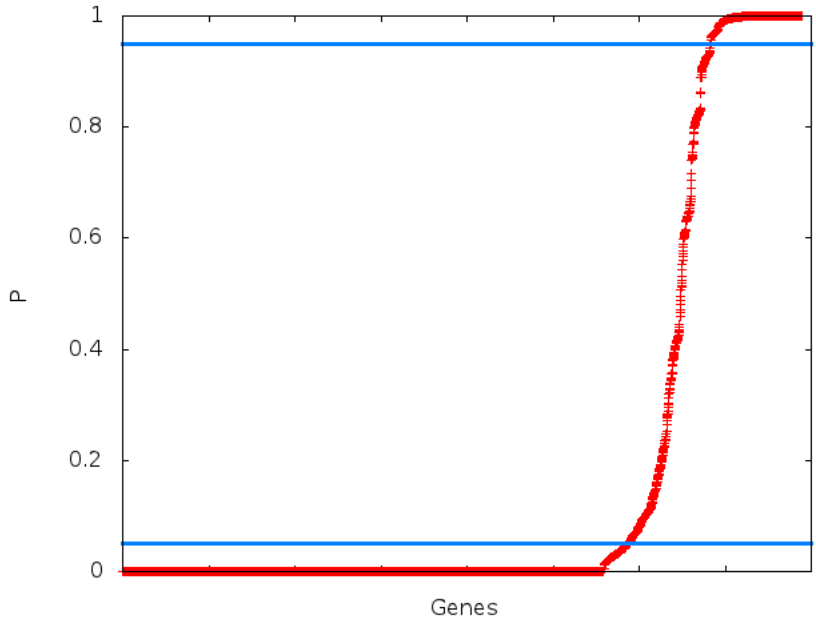


Fig. 6.: Cumulative Posterior Probability Estimates for All Genes. Genes with confident essentiality predictions are found on top (essential) and bottom (non-essential) of the curve, with those for which we are less confident in the middle of the curve. The blue lines represent the thresholds for essential ($Z_i > 0.95$) and non-essential ($Z_i < 0.05$) genes.

statistical analysis. For example, *inhA* and *gyrA* are both known to be essential for survival; the former is responsible for producing an enzyme necessary for biosynthesis of mycolic acids used in the cell wall [5], and the latter is needed to unwind DNA during replication [21]. InhA is the target of isoniazid a first line drug for treatment of tuberculosis, and GyrA is the target of fluoroquinolones, a family of broad-spectrum antibiotics used as second line treatments for tuberculosis. Examples of non-essential genes are the family of PGRS genes, and genes involved in PDIM biosynthesis. The PGRS family of genes have an unknown role in mycobacteria, however they are known to be mostly unnecessary for growth [11]; we find only 2 out of 67 (i.e., PE_PGRS57 and PE_PGRS54) to be essential in TB and this could be due to poor sequencing in GC-rich regions. PDIM (phenol phthiocerol dimycolate) is a surface polyketide

Table I.: Statistics for Essentials, Non-Essentials and Uncertain Genes.

	Total	**Average**			
	Genes	**Length (amino-acids)**	**# TA sites**	**# Insertions**	**Max run**
Non-Essentials (<0.0.5)	2933	304.93	14.16	10.04	1.80
Intermediate (0.05 – 0.95)	482	352.28	16.98	5.68	6.98
Essentials (>0.95)	531	499.32	24.58	2.16	19.68

necessary for infection in-vivo, but not for growth in in-vitro [22].

1. Comparison to Other Essentiality Results

a. Sassetti et. al. 2003

Sassetti et. al. had previously used the TraSH method to identify those genes necessary for optimal growth of TB in-vitro[14]. They identified 614 genes essential for growth in-vitro, by culturing a library of mutants on 0.2% glucose + 7H10 (rich media). In order to compare against this dataset, we classify our genes as essential so long as their posterior probability is greater than 0.5 (i.e., $p(Z_i|Y) > 0.5$), thus forcing the set of intermediate genes to be classified as their most probable category. In addition to determining which genes are essential or non-essential, Sassetti et.al. were also capable of quantifying the growth rate of mutants by determining the hybridization ratio of individual genes to the TraSH probes. This allowed them to characterize a third category of genes, those whose disruption causes a growth defect in the organism. However, they were unable to determine essentiality for 813 genes for which they could not obtain hybridization ratios. On the other hand, sequencing is able to provide data on all genes so long as they contain at least one TA site.

Table III outlines the agreement between our predictions and the TraSH results.

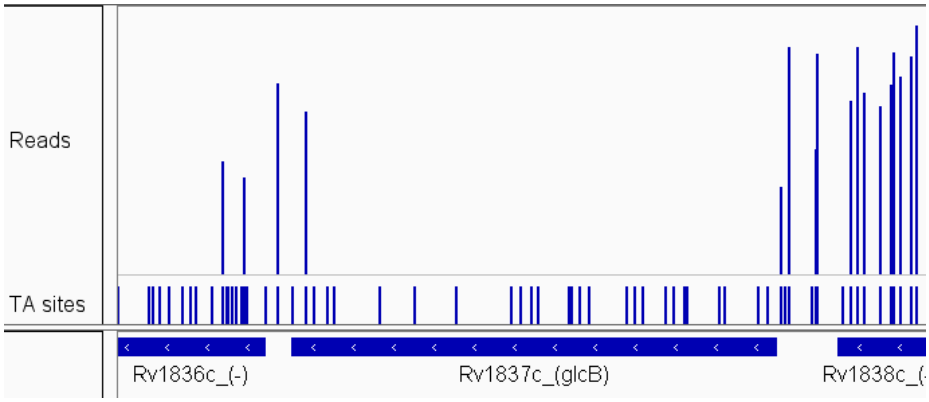
Table II.: Predictions on Genes with Experimentally Determined Essentiality. n is the number of TA sites, r is the length of the maximum run of non-insertions, and p is the posterior probability of essentiality calculated by the Bayesian method.

Orf	**Gene Name**	**n**	**r**	**p**	**Experimental Essentiality**	**Function**	**References**
Rv0001	dnaA	32	31	1.00	Essential	DNA replication	Greendyke et al. [23]
Rv0006	gyrA	46	44	1.00	Essential	DNA replication	Von Groll et al. [21]
Rv0014c	pknB	24	24	1.00	Essential	Signaling	Lougheed et al. [24]
Rv0046c	ino1	17	17	1.00	Essential	Inositol synthesis	Movahedzadeh et al. [25]
Rv0189c	ilvD	23	23	1.00	Essential	Amino acid biosynthesis	Singh et al. [26]
Rv0236c	aftD	40	39	1.00	Essential	Cell wall synthesis	Skovierova et al. [27]
Rv0334	rmlA	19	19	1.00	Essential	Cell wall synthesis	Qu et al. [28]
Rv0486	mshA	13	10	0.91	Essential	Mycothiol synthesis	Buchmeier and Fahey [29]
Rv0757	phoP	12	9	0.78	Essential	Signaling	Goyal et al. [30]
Rv0902c	prrB	15	13	0.99	Essential	Membrane transporters	Haydel et al. [31]
Rv0903c	prrA	10	9	0.79	Essential	Membrane transporters	Haydel et al. [31]
Rv1018c	glmU	24	24	1.00	Essential	Cell wall biosynthesis	Zhang et al. [32]
Rv1483	fabG1	13	13	0.99	Essential	Mycolic acid synthesis	Gurvitz [33]
Rv1484	inhA	10	10	0.92	Essential	Mycolic acid synthesis	Molle et al. [34]
Rv1485	hemZ	25	25	1.00	Essential	Heme biosynthesis	Parish et al. [35]
Rv2130c	mshC	25	24	1.00	Essential	Mycothiol biosynthesis	Buchmeier and Fahey [29]
Rv0242c	fabG	11	6	0.15	Non-Essential	Fatty acid synthesis	Gurvitz [33]
Rv0980c	PE_PGRS18	13	1	0.00	Non-Essential	Unknown	Banu et al. [36]
Rv1067c	PE_PGRS19	13	4	0.00	Non-Essential	Unknown	Banu et al. [36]
Rv1068c	PE_PGRS20	12	5	0.05	Non-Essential	Unknown	Banu et al. [36]
Rv2930	fadD26	40	8	0.28	Non-Essential	PDIM biosynthesis	Domenech and Reed [22]
Rv2931	ppsA	81	2	0.00	Non-Essential	PDIM biosynthesis	Domenech and Reed [22]
Rv2940c	mas	82	2	0.00	Non-Essential	PDIM biosynthesis	Domenech and Reed [22]
Rv2941	fadD28	47	4	0.00	Non-Essential	Fatty acid degradation	Cole et al. [20]
Rv2942	mmpL7	42	3	0.00	Non-Essential	Membrane transport	Domenech et al. [37]

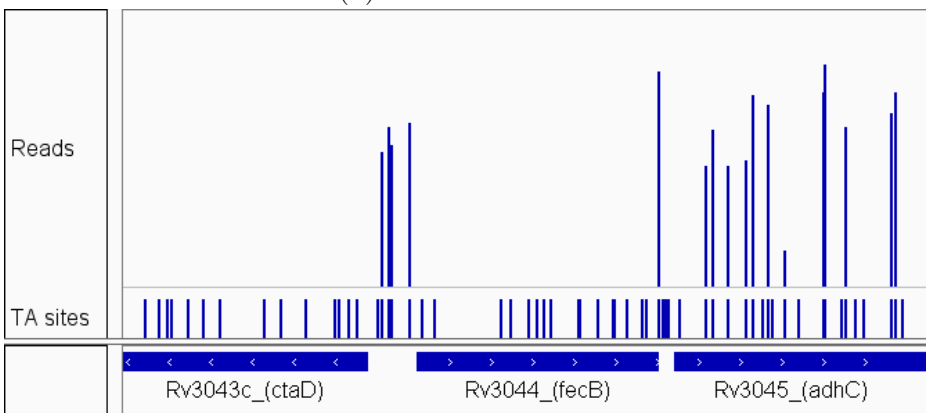
In general the results of the Bayesian method match those of Sassetti et. al., agreeing on 69% of essentials, and 98% of non-essentials. The results are also consistent with the non-essential genes identified in the DeADMAn experiments (with our results matching 1,750 of the 1,925 - 90.9% - of the non-essential genes reported) [38] . However there are a few disagreements. For example, Sassetti et. al. found that *PPE34*, *PPE35*, *mmpL2*, *mmpL12* were non-essential while our method predicts them to be essential due to significant gaps of transposon insertions within them. MmpL genes are membrane transporters, and only mmpL3 is thought to be essential [37].

On the other hand, Sassetti et. al. predict a number of genes to be essential which our method predicts to be non-essential, like *hycP* and *hycQ* (putative hydrogenases). These differences may be due to different growth conditions between our libraries, as the library from created by Sassetti et. al. were grown in rich-media in the presence of glucose, while the libraries utilized in this analysis were grown in minimal-media in the presence of glycerol.

Two genes that the Bayesian method predicts to be essential that were indicated as non-essential by Sassetti. et. al are *glcB* and *fecB*. Insertion patterns shown in Figure 7 clearly indicate that these genes are unable to withstand insertions. GlcB encodes for malate synthase in TB, which was originally thought to be necessary only for growth on fatty-acids as part of a glyoxylate shunt [39], but has recently been shown to be essential on other carbon sources like dextrose (Sacchettini lab, submitted). Our data confirms this by showing GlcB is necessary for growth on glycerol as well. FecB is involved in iron transport (ferric dicitrate) and is not expected to be essential in minimal media due to redundancy with other iron acquisition mechanisms like mycobactin [40], yet shows only one insertion at the C-terminus out of 20 total TA sites.



(a) Rv1837c - GlcB



(b) Rv3044 - FecB

Fig. 7.: Examples Classified as Non-Essential by Sassetti 2003

Table III.: Comparison of Essentiality Predictions with TraSH analysis. We compare the results obtained by Sassetti. et. al with those obtained with our Bayesian method for all 3989 genes in *M. tuberculosis*, with genes divided into the four categories of essentiality considered.

		Bayesian Method			
		Essential	**Non-Essential**	**No-Data**	**Total**
Sassetti-03	**Essentials**	427	186	1	614
	Non-Essential	114	2400	6	2520
	Growth-Defect	11	31	0	42
	No-Data	151	626	36	813
	Total	703	3243	43	3989

b. Binomial Model

In addition to previous essentiality assignments, we also compare our results with an alternative model based on the Binomial distribution. The Binomial model infers essentiality based on the proportion of insertions observed within genes regardless of their order, while our model determines essentiality based on significant consecutive TA sites lacking insertions. To make inferences about essentiality, we model the gene categories as a mixture of Binomial distributions, with different parameters θ_0 and θ_1 representing the probability of insertion at non-essential genes and essential genes respectively. These distributions express the probability of observing the amount of insertions within a gene. In specific, the probability of observing k_i out of n_i insertions within a given gene i, is given by the following likelihood:

$$p(Y_i|\theta, Z) = Binomial(\theta; k_i, n_i)$$
$$= \binom{n_i}{k_i}\theta^{k_i}(1-\theta)^{n_i-k_i}$$

Our prior expectations for parameter θ our described by a beta distribution:

$$\pi(\theta) = Beta(\theta; \alpha, \beta)$$

with hyper-parameters α and β. Because the Beta distribution is conjugate with the Binomial distribution, our conditional probability for the parameter θ becomes a new Beta distribution, with updated parameters:

$$p(\theta|Y, Z) = \prod_{i=1}^{G} Binomial(\theta; k_i, n_i) \times Beta(\theta; \alpha, \beta)$$
$$= Beta(\theta; \alpha + \Sigma k_i, \beta + \Sigma n_i - \Sigma k_i)$$

Using Gibbs sampling, we obtain samples of parameters θ_0 and θ_1 as well as the essentiality assignments Z_i, which are used to estimate posterior probabilities of essentiality as in the Bayesian method. After running a Gibbs sampling procedure for 50,000 iterations, estimates for the parameters were as follows: $\theta_0 = 0.660 \pm 0.002$ and $\theta_1 = 0.088 \pm 0.002$, implying 66% insertion density in non-essential genes (similar to the Gumbel estimate) and 8.8% in essential genes.

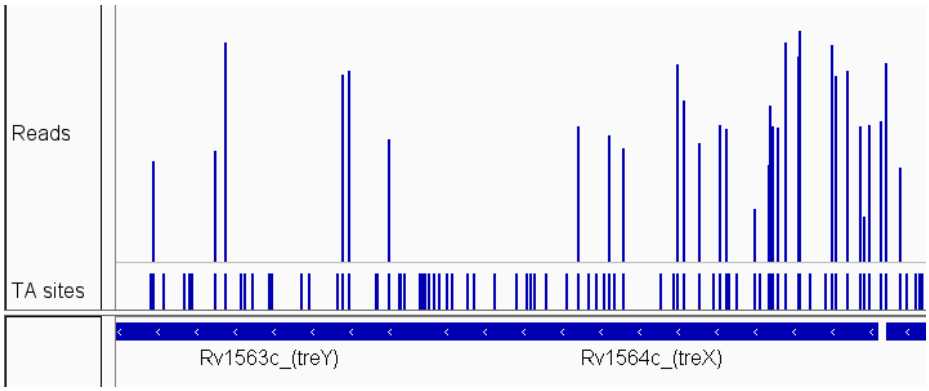
Table IV compares our results to the Binomial model. Although both methods seem to agree in general, the Binomial model predicts a significantly larger number of essential genes, inferring that 24.05% of genes in TB are essential. This discrepancy in the amount of essential genes predicted, may suggest that the proportion of insertions within genes is a not as good an indicator of essentiality as large gaps of insertions.

A Binomial model of essentiality may infer a gene is essential because it contains less insertions than expected, yet ignore that these insertions covered all areas of the gene exhibiting the small runs of insertions characteristic of non-essential genes (e.g., Rv2148c, Rv2382c, Rv1698, Rv0241c, Rv1548c).

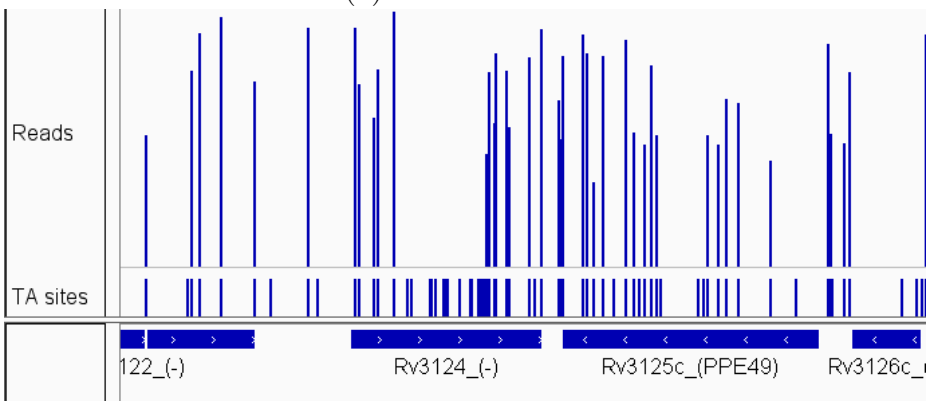
Table IV.: Comparison with the Binomial Model. Results obtained by a Binomial model of essentiality compared with those obtained by our Bayesian (Gumbel) method for all 3989 genes in *M. tuberculosis*.

		Bayesian Method			
		Essential	**Non-Essential**	**No-Data**	**Total**
Binomial Model	**Essentials**	668	291	0	959
	Non-Essential	36	2952	0	2520
	No-Data	0	0	43	43
	Total	704	3243	43	3989

Figure 8 shows the insertion patterns of some example genes to highlight the cases where the two methods disagree. TreX (Rv1564) is predicted by the Binomial model to be non-essential due to its large portion of insertions (i.e., 34%), however our Bayesian model predicts this gene to be essential due to the large stretch of non-insertions at the C-terminus of the gene (i.e., run of 14 TA sites in a row without insertions); this large gap may represent a significant essential region that the Binomial model is not capable of identifying. TreX is involved in glycogen degradation and trehalose synthesis and this pathway is thought to be essential [41, 42]. TreX has three domains, with a gap corresponding to the C-terminal domain [43] ($Z_i = 0.995$). Domain analysis is discussed in the next section. Another example of a gene predicted



(a) Rv1564 - TreX



(b) Rv3124

Fig. 8.: Examples of Disagreement With Binomial Model. Gene *treX* is predicted to be essential due to a run of 14 non-insertions in a row in the C-terminus. (8b) Gene Rv3124 is predicted to be essential due to its run of 16 non-insertions in a row in the middle of the gene.

to be non-essential by the Binomial model is gene Rv3124, a transcriptional regulator of molybdopterin biosynthesis [44]. Although Rv3124 shows a high a proportion of insertions near the N- and C- terminus of the gene, it also contains a significant gap (i.e. run of 16 non-insertions in a row, with $Z_i = 0.999$) in the middle of the gene suggestive of an essential region.

2. Essential Domains

To test our hypothesis that this Bayesian method is capable of capturing information about putative essential domains within genes, we obtained domain predictions from the Pfam database and compared them to the regions devoid of insertions within essential genes. Pfam predictions are based on Hidden Markov models for known protein families, and are used for predicting domains by analyzing amino-acid sequences and matching them to manually curated database of proteins [45]. Although protein structures of individual genes can be experimentally determined and are a much more reliable source of information, the majority of genes in *M. tuberculosis* have no known structure with only 8.5% of the ORFs in TB having their structure solved (deposited in PDB) [46]. Pfam predictions, however, allow us to obtain potential domain information on nearly all genes. After obtaining the domain predictions from Pfam, they were matched to our predicted essential genes. Table V contains some statistics for our results. The analysis was limited to those genes predicted to be essential by the Bayesian method, as these represent those genes which contain the significant stretches of non-insertions that are suggestive of essential regions. Of the 704 genes we predicted as essential, 687 of these had at least one domain prediction in the Pfam database. Of these 687 genes, 320 completely lacked insertions suggesting the entire gene is essential. This left 367 genes with a potential to contain both essential and non-essential domains. Since the domain predictions obtained from Pfam may not actually coincide with these gaps, the start and end of the domains within all 367 genes were matched with start and end of the runs of non-insertions in these genes. Only 276 genes contained domains that fell completely within the span of the largest run of non-insertions observed, hence showing no evidence of insertion for the region spanned by the domain. After obtaining this subset of genes we calculated the

span of TA sites contained within the domain, and focused on those which accounted for a significant gap of non-insertions given our model. We identified 117 genes with significant stretches of non-insertions that correlated with predicted domains.

Table V.: Statistics of Domains Within Essential Genes. The 704 essential genes obtained by our Bayesian Method are analyzed in order to identify those which contain Pfam domain predictions that coincide with meaningful gaps of non-insertions.

Essentials with Domains	Completely Essential	Contain Non-Essential Regions	Domain Matches Longest Run	Significant Run Within Domain
687	320	367	276	117

Figure 9 contains some examples of those genes with significant runs of non-insertions coinciding with the domain predictions from Pfam. Rv3190 encodes for two C-terminal protein domains (sugar-binding and extracellular domains) and a N-terminal, MviN-like, domain which regulates peptidoglycan biosynthesis and has been shown to be essential for growth in mycobacteria. This protein is actually a flippase of lipid-II and is regulated by interaction with FhaA (Rv0020c), which is phosporylated by PknB [47]. Insertions in Rv3910 are found only in the C-terminal domains, but not the N-terminal membrane domain, implying it alone is necessary for growth. Rv2051c (Ppm1) is involved in cell-wall glycolipid synthesis, an essential role within mycobacteria, and shows evidence of an essential domain (Pfam family: - PF0535.21) within its C-terminus which matches previous analyses of this gene [48]. Rv0018c (serine/threonine phosphatase) contains an essential catalytic domain within its N-terminus, and has been shown to dephosphorylate Rv0020 (FhaA) coun-

teracting phosphorylation by PknB [49]. Transposon insertions are only observed in the extracellular domain of unknown function.

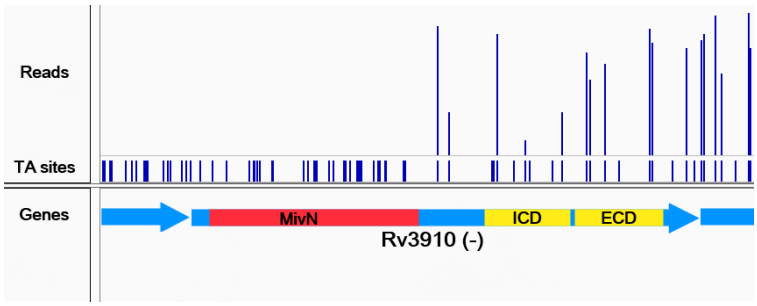
3. Low Density Dataset

To evaluate our method on other datasets with different insertion density (prepared by J. Zhang), we ran our analysis a different library of *M. tuberculosis* mutants grown on glycerol, but with a much lower proportion of insertions. This dataset contained significantly fewer transposon insertions in coding regions (i.e., 23,399 - 36.3% - compared to 31,715 - 50.4% in our first glycerol dataset), and therefore the set of TA sites in the genome were under-sampled, providing a more difficult challenge for estimating essentiality. Under-sampled datasets will likely contain longer stretches of non-insertions as TA sites in these libraries are much more likely to be missed by the sparse transposon insertions. Table VI contains a comparison of results for both datasets.

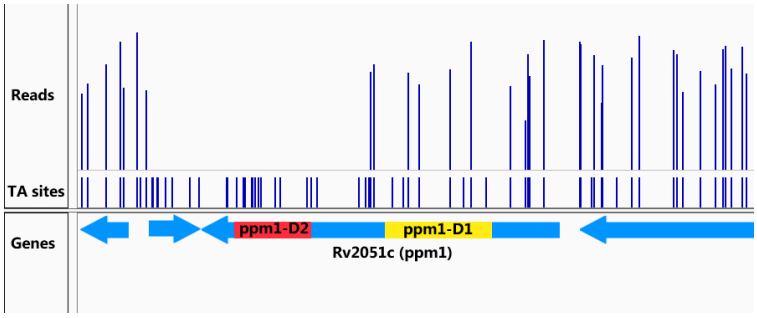
Table VI.: Comparison of Under-sampled Dataset and Regular Dataset.

	Undersampled	Normal
Essentials:	304	704
Non-Essentials:	3679	3243
Total:	3983	3947

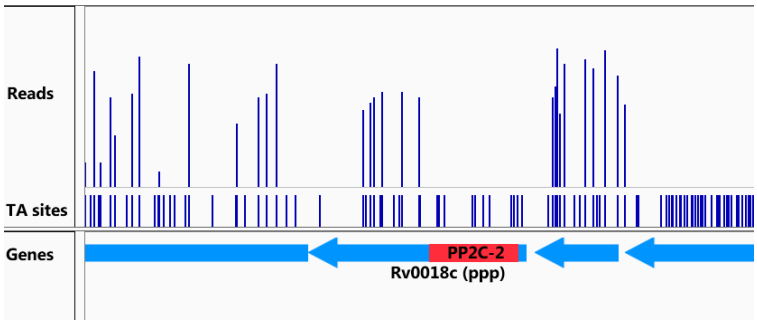
Our method finds a significantly smaller number of essentials in the under-sampled dataset. The probability of non-insertions estimated for the under-sampled



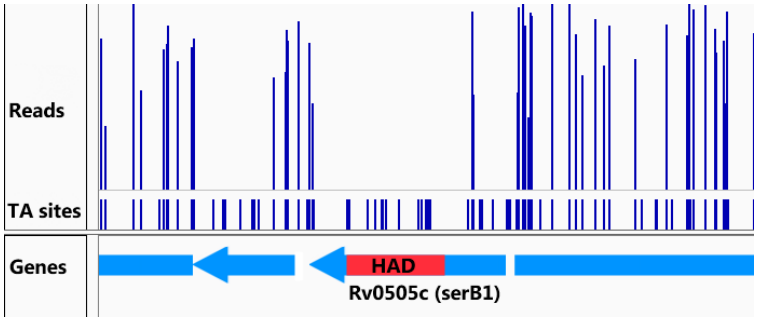
(a) Rv3910 (-)



(b) Rv2051c (ppm1)



(c) Rv0018c (ppp)



(d) Rv0505c (serB1)

Fig. 9.: Example Genes with Essential Domains. Essential domains are indicated in red, and non-essential domains are indicated in yellow.

dataset was $\phi_0 = 0.592$, which is significantly higher than the probability estimated in the original glycerol dataset, $\phi_0 = 0.344$. Because of this higher probability of non-insertion, all genes in the under-sampled dataset will be expected to have longer runs of non-insertions, even those which can withstand disruption. However, the Bayesian model is able to compensate for the lower insertion frequency, and does not predict an excess of essential genes. It is conservative in its predictions given the sparsity of the data. By increasing the expected maximum run, fewer genes will be predicted to be essential due to the fact that the total number of TA sites they contain is not large enough to produce significant runs according to the Gumbel model, given such a high probability of non-insertion.

4. Glycerol vs. Cholesterol

To test our analysis on mutants grown in different environmental conditions, we analyzed sequencing results for three independent libraries of TB mutants grown in minimal media and 0.01% cholesterol [17]. Cholesterol is thought to be a significant carbon source in macrophages, and thus mimics environmental conditions found during infection [50]. Like we did for the glycerol sequence data, we also summed the reads across all three independent libraries of cholesterol. This allowed us to have a denser dataset with a higher probability of all TA sites being sampled.

Table VII shows a list of the top genes our Bayesian method predicts to be essential for growth in cholesterol, and non-essential for growth in glycerol. All of these genes have previously been shown to be associated with cholesterol catabolism and/or fatty-acid degradation [17, 51]. For example, HsaD has been shown to catalyze the hydrolytic cleavage of a carbon - carbon bond in cholesterol ring degradation, and therefore is essential for growth in cholesterol media but not glycerol [52]. Interestingly, ChoD, a gene annotated as cholesterol oxidase, turns out to be non-essential as

Table VII.: Genes Differentially Essential for Growth on Cholesterol But Not Glycerol. 28 genes were selected that have a posterior probability of essentiality > 0.9 for cholesterol and < 0.1 for glycerol. This subset was enriched for genes known to be associated with cholesterol catabolism (8 out of 28, shown).

		Posterior Probability of Essentiality		
Gene	**Name**	**Glycerol**	**Cholesterol**	**Function**
Rv3556c	fadA6	0.091	0.996	acetyl-CoA acetyltransferase
Rv3543c	fadE29	0.029	0.999	acyl-CoA dehydrogenase
Rv3562	fadE31	0.000	0.952	acyl-CoA dehydrogenase
Rv3526	kshA	0.000	1.000	ketosteroid hydroxolase
Rv3540c	ltp2	0.000	0.999	ketoacyl-CoA thiolase
Rv3544c	fadE28	0.000	0.999	acyl-CoA dehydrogenase
Rv3568c	hsaC	0.000	0.998	dienoate hydrolase
Rv3569c	hsaD	0.000	0.991	dienoate hydrolase

shown by laboratory experiments [17]. Conversely, we also find genes that are non-essential for growth in cholesterol, yet necessary for growth in glycerol. For instance, GlpK (glycerol kinase) is essential for glycerol metabolism [53], and is predicted by our method to be essential for growth in glycerol but not for growth in cholesterol.

C. Convergence of Sampling Procedure

Our statistical analysis depends on obtaining an MCMC sample of the ϕ_0 parameter (i.e., probability of non-insertion in non-essential genes) to estimate posterior probabilities of essentiality. We obtain estimates of ϕ_0 by sampling its conditional probability given the data through the MH algorithm. Since the MH algorithm samples from the conditional distribution of a parameter given the rest, one after another, one potential concern is that these distributions might not mix well; that is, that they might not adequately explore the space of the distribution of interest. Parameters may get "stuck" sampling one area of the distribution, and influence the sampling of the other parameters. For these reasons, we eliminate the first 1,000 samples of the ϕ_0 parameter to ensure that the MH algorithm reaches a point where it is mixing well. This is referred to as the "burn-in" period [19]. In order to validate our final sample of the ϕ_0 parameter, Figure 10 presents the trajectory of the sample, which shows its values across the remaining iterations. Note that, while there is variation, a stable trend has been established.

A potential problem with MCMC samplers is that sampled values might be correlated with each other. By generating a Markov-Chain for sampling, any value at time t may actually be correlated with previous samples at time $t - k$. If the algorithm is producing results that are highly-correlated, then the sampler may not

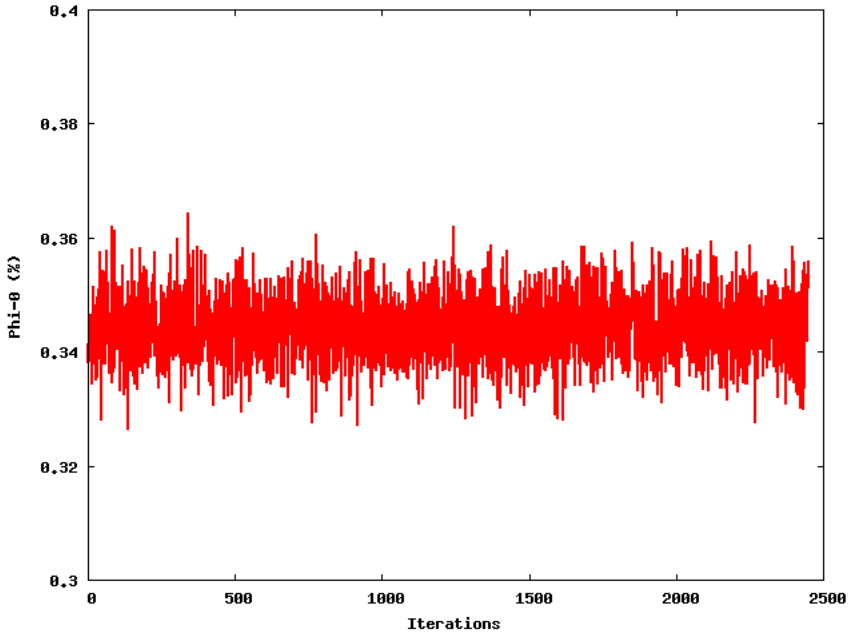


Fig. 10.: MCMC Sample of the ϕ_0 Parameter. The sampling procedure reaches "convergence", correctly sampling the ϕ_0 density.

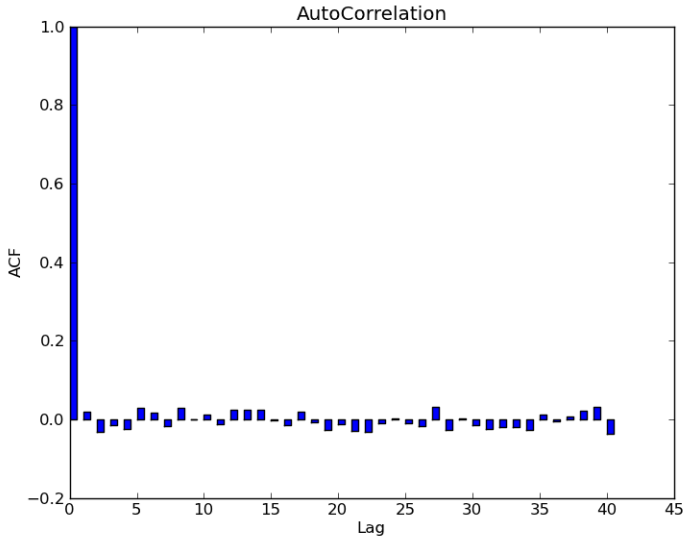


Fig. 11.: Auto-Correlation of MCMC Sample of the ϕ_0 Parameter. Low auto-correlation values with lag greater than zero show the samples are uncorrelated with each other at subsequent time steps, a potential problem for MCMC sampling procedures.

be truly exploring the distribution of interest in a random manner. To test whether our MCMC scheme was producing correlated values, we calculated the autocorrelation coefficient to a maximum lag of 50. Figure 11 shows a plot of the auto-correlation of the MCMC sample for the ϕ_0 parameter. The low values show that samples at $\Delta t \geq 1$ apart are effectively uncorrelated.

CHAPTER IV

DISCUSSION AND CONCLUSION

Which genes are essential to the survival of a bacterial organism is an important question scientists wish to answer as it allows scientists discover potential new drug targets, and learn more about an organism's evolution. By using transposon mutagenesis experiments, researchers can create libraries of mutant organisms that have had portions of their DNA interrupted. Using these libraries, scientists can extract information about which genes can sustain insertions without affecting an organisms survival, and therefore which genes are essential and non-essential to the organism. Using next-generation sequencing, scientists are able to determine precisely where these transposon insertions took place, providing with high-resolution information on non-essential regions in the genome.

The addition of this new sequence data necessitates a new method to analyze it that can exploit the high-resolution information to determine essentiality. We developed a Bayesian statistical analysis method to analyze this data and make rigorous predictions about the essentiality of individual genes. Using this method we have analyzed sequence data from a library of mutants of *M. tuberculosis* bacteria, and improved our understanding of essentiality within this organism.

The key insight in our model is the use of the Gumbel distribution to model the expected length of the maximum run of non-insertions within genes. This allows our analysis to determine whether the largest run of non-insertions in a particular gene is statistically significant, and therefore suggestive of a region that cannot withstand insertions. By modeling non-essential genes in this manner, our method is then able to pick out those genes that contain significantly longer runs of non-insertions than what we would normally expect, without being sensitive to a small number of insertions at

the N- or C- terminus of a gene. Furthermore, by using Metropolis-Hastings sampling we are able to obtain estimates of posterior probabilities of essentiality for all genes that quantify the confidence we have on our predictions.

Using this method, we get results that are 89% consistent with previous analysis of TraSH data. Many of these genes were expected to be essential given their indispensable role within bacterial organisms (e.g., GyrA, DnaA, InhA). However, our Bayesian analysis also identified some new essential genes required for growth in glycerol (e.g., GlcB). The main sources for disagreements between the findings of this Bayesian analysis and Sassetti et. al. are likely due to the differences in growth media used when creating both libraries, our method's ability to identify essential domains within genes that may otherwise be characterized as non-essential, as well as the fact that sequencing provides high-resolution coordinates of individual insertions which was not possible with hybridization. Utilizing our Bayesian method, we performed a differential analysis between transposon mutants grown on glycerol and those grown on cholesterol, where we obtained results which coincided with other analyses that have compared both of these growth media.

Because our model is based on an analysis of long stretches of the genome lacking any evidence of disruption, our method is capable of highlighting domains within genes that may be essential. Genes can code for multiple domains, and these domains may play different biological roles within the organism. If a gene contains an essential domain within its coding region, then that domain will be unable to withstand any insertions. By highlighting those areas that have unusually large gaps in insertions, our method is capable of picking out genes that contain evidence suggestive of essential regions. Although previous analyses have used data from deep-sequencing to determine essentiality, those methods used ad-hoc criteria or assumptions about parameters and do not produce rigorous statistical scores. Our method may be one

possible way of using transposon mutagenesis experiments to suggest potential new essential domains within genes whose protein structure is unknown, by estimating posterior probabilities. For example, using our method we found genes with essential domains (e.g. Rv0018c, Rv3910) that match Pfam predictions of domains, and whose essentiality is supported in the literature.

The Gumbel distribution depends on an estimate of the probability of non-insertion within non-essential genes as an internal (unobservable) parameter. However, by using a Bayesian statistical framework, we can estimate this parameter by sampling from its probability density function and thus effectively integrate over this parameter. Using this framework, we do not require an a priori estimate of this parameter to determine essentiality, but instead let our analysis find the distribution of this parameter that is suggested by the data. Previously, we used an approximation based on the frequency of insertions at TA sites within genes that are "probable-essentials" (i.e., containing insertions at 20% or more their TA sites). By not requiring assumptions or ad-hoc estimates of this parameter, we can apply our analysis to different datasets where this parameter may be significantly different or difficult to estimate without a formal framework. For example, we can use this method to determine essentiality within libraries of transposon mutants that have been under-sampled. Under-sampled libraries contain fewer transposon insertions, therefore the probability of non-insertions will be artificially high due to a lack of insertion coverage; however, by estimating this parameter based on the data, our Gumbel model is capable of adjusting and picking out stretches of non-insertion that are statistically significant, even for a high probability of non-insertion.

Another important feature of our method is its ability of estimating the confidence we have about our essentiality results. By generating samples from the posterior densities of essentiality, we get a measure of how likely it is that a gene be essential

while exploring the distribution of the parameters and missing data in our model. This allows us to assign high confidence to those predictions within genes that have consistently been inferred to be essential, and lower confidence to those genes for which our model can infer essentiality in both ways.

Finally, because our method depends on consecutive sequences of TA sites lacking insertions, and not on the simple presence or absence of insertions within a gene, our method is not sensitive to insertions at the N- or C- terminus of a gene, which essential genes have been shown to tolerate occasionally [2, 1].

Although our method has several strengths, it also has some potential limitations that would be useful to consider as future improvements. While our method can successfully determine areas in the sequence information that contains unusually long gaps lacking any reads, it is does so by taking a binary approach to the sequence information: if there are reads mapping to a TA site, we consider it as an insertion (1), if there are no reads we consider that site lacking any insertions (0). By doing so, however, we lose any potential information that the magnitude of reads or read counts mapping to that particular site would have given us. In reality, this information may contain useful information about essentiality. For instance, Sassetti et al. [14] were able to characterize those genes which may cause growth-defects in the organism once interrupted by quantifying their ratio of hybridization to the hybridization probes. Similarly, one may be able to identify this other category of genes by taking into consideration the counts of reads mapping to their given insertion sites, which may be significantly lower than expected from an average non-essential gene. On the other hand, read-counts might not accurately represent the prevalence of these insertions in the mutants sequenced. Read-counts can be subject to "PCR-bias" when amplification is not equally efficient across the templates used [54]. This may lead to artifacts that may render read-counts difficult to interpret.

Another limitation of our model is that it does not take into account the distance between TA sites. By treating TA sites as a sequence of independent Bernoulli trials, our model loses any meaningful information that might be contained in the distance between two TA sites. For example, if two TA sites are too far apart from each other, then observations at these TA sites may not accurately represent the essentiality of the genomic region between them. The model could be extended to take this information into account when assessing statistical significance.

Bibliography

[1] V. Smith, K. N. Chou, D. Lashkari, D. Botstein, and P. O. Brown, "Functional analysis of the genes of yeast chromosome V by genetic footprinting," *Science*, vol. 274, pp. 2069–2074, Dec 1996.

[2] B. J. Akerley, E. J. Rubin, A. Camilli, D. J. Lampe, H. M. Robertson, and J. J. Mekalanos, "Systematic identification of essential genes by in vitro mariner mutagenesis," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, pp. 8927–8932, Jul 1998.

[3] A. Williams, C. Gthlein, N. Beresford, E. C. Bttger, B. Springer, and E. O. Davis, "Uvrd2 is essential in mycobacterium tuberculosis, but its helicase activity is not required," *Journal of Bacteriology*, vol. 193, no. 17, pp. 4487–4494, 2011. [Online]. Available: http://jb.asm.org/content/193/17/4487.abstract

[4] C. M. Sassetti and E. J. Rubin, "Genetic requirements for mycobacterial survival during infection," *Proceedings of the National Academy of Sciences*, vol. 100, no. 22, pp. 12 989–12 994, 2003. [Online]. Available: http://www.pnas.org/content/100/22/12989.abstract

[5] D. A. Rozwarski, G. A. Grant, D. H. Barton, W. R. Jacobs, and J. C. Sacchettini, "Modification of the NADH of the isoniazid target (InhA) from Mycobacterium tuberculosis," *Science*, vol. 279, pp. 98–102, Jan 1998.

[6] J. I. Glass, N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, C. A. Hutchison, H. O. Smith, and J. C. Venter, "Essential genes of a minimal bacterium," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, pp. 425–430, Jan 2006.

[7] F. Hayes, "Transposon-based strategies for microbial functional genomics and proteomics," *Annu. Rev. Genet.*, vol. 37, pp. 3–29, 2003.

[8] D. J. Lampe, M. E. Churchill, and H. M. Robertson, "A purified mariner transposase is sufficient to mediate transposition in vitro." *the The European Molecular Biology Organization Journal*, vol. 15, no. 19, pp. 5470–5479, 1996.

[9] E. J. Rubin, B. J. Akerley, V. N. Novik, D. J. Lampe, R. N. Husson, and J. J. Mekalanos, "In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 4, pp. 1645–1650, 1999.

[10] J. Ashour and M. K. Hondalus, "Phenotypic mutants of the intracellular actinomycete rhodococcus equi created by in vivo himar1 transposon mutagenesis," *Journal of Bacteriology*, vol. 185, no. 8, pp. 2644–2652, 2003. [Online]. Available: http://jb.asm.org/content/185/8/2644.abstract

[11] G. Lamichhane, M. Zignol, N. J. Blades, D. E. Geiman, A. Dougherty, J. Grosset, K. W. Broman, and W. R. Bishai, "A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: Application to mycobacterium tuberculosis," *Proceedings of the National Academy of Sciences*, vol. 100, no. 12, pp. 7213–7218, 2003. [Online]. Available: http://www.pnas.org/content/100/12/7213.abstract

[12] C. M. Sassetti, D. H. Boyd, and E. J. Rubin, "Comprehensive identification of conditionally essential genes in mycobacteria," *Proceedings of the National Academy of Sciences*, vol. 98, no. 22, pp. 12 712–12 717, 2001. [Online]. Available: http://www.pnas.org/content/98/22/12712.abstract

[13] W. A. Day, S. L. Rasmussen, B. M. Carpenter, S. N. Peterson, and A. M. Friedlander, "Microarray analysis of transposon insertion mutations in bacillus anthracis: Global identification of genes required for sporulation and germination," *Journal of Bacteriology*, vol. 189, no. 8, pp. 3296–3301, April 15, 2007. [Online]. Available: http://jb.asm.org/content/189/8/3296.abstract

[14] C. M. Sassetti, D. H. Boyd, and E. J. Rubin, "Genes required for mycobacterial growth defined by high density mutagenesis," *Molecular Microbiology*, vol. 48, no. 1, pp. 77–84, 2003. [Online]. Available: http://dx.doi.org/10.1046/j.1365-2958.2003.03425.x

[15] J. D. Gawronski, S. M. S. Wong, G. Giannoukos, D. V. Ward, and B. J. Akerley, "Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for haemophilus genes required in the lung," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 38, pp. 16 422–16 427, 2009.

[16] G. C. Langridge, M.-D. Phan, D. J. Turner, T. T. Perkins, L. Parts, J. Haase, I. Charles, D. J. Maskell, S. E. Peters, G. Dougan, and et al., "Simultaneous assay of every salmonella typhi gene using one million transposon mutants." *Genome Research*, vol. 19, no. 12, pp. 2308–2316, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19826075

[17] J. E. Griffin, J. D. Gawronski, M. A. DeJesus, T. R. Ioerger, B. J. Akerley, and C. M. Sassetti, "High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism," *PLoS Pathog*, vol. 7, no. 9, p. e1002251, 09 2011.

[18] M. F. Schilling, "The longest run of heads," *College of Mathematics Journal*, vol. 21, pp. 196–207, 1990.

[19] S. M. Lynch, *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, 2007, iSBN 978-0-387-71264-2.

[20] S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell, "Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence," *Nature*, vol. 393, no. 6685, pp. 537–544, 1998. [Online]. Available: http://dx.doi.org/10.1038/31159

[21] A. Von Groll, A. Martin, P. Jureen, S. Hoffner, P. Vandamme, F. Portaels, J. C. Palomino, and P. A. da Silva, "Fluoroquinolone resistance in Mycobacterium tuberculosis and mutations in gyrA and gyrB," *Antimicrob. Agents Chemother.*, vol. 53, pp. 4498–4500, Oct 2009.

[22] P. Domenech and M. B. Reed, "Rapid and spontaneous loss of phthiocerol dimycocerosate (PDIM) from Mycobacterium tuberculosis grown in vitro: implications for virulence studies," *Microbiology (Reading, Engl.)*, vol. 155, pp. 3532–3543, Nov 2009.

[23] R. Greendyke, M. Rajagopalan, T. Parish, and M. V. Madiraju, "Conditional

expression of Mycobacterium smegmatis dnaA, an essential DNA replication gene," *Microbiology (Reading, Engl.)*, vol. 148, pp. 3887–3900, Dec 2002.

[24] K. E. Lougheed, S. A. Osborne, B. Saxty, D. Whalley, T. Chapman, N. Bouloc, J. Chugh, T. J. Nott, D. Patel, V. L. Spivey, C. A. Kettleborough, J. S. Bryans, D. L. Taylor, S. J. Smerdon, and R. S. Buxton, "Effective inhibitors of the essential kinase PknB and their potential as anti-mycobacterial agents," *Tuberculosis (Edinb)*, vol. 91, pp. 277–286, Jul 2011.

[25] F. Movahedzadeh, D. A. Smith, R. A. Norman, P. Dinadayala, J. Murray-Rust, D. G. Russell, S. L. Kendall, S. C. Rison, M. S. McAlister, G. J. Bancroft, N. Q. McDonald, M. Daffe, Y. Av-Gay, and N. G. Stoker, "The Mycobacterium tuberculosis ino1 gene is essential for growth and virulence," *Mol. Microbiol.*, vol. 51, pp. 1003–1014, Feb 2004.

[26] V. Singh, D. Chandra, B. S. Srivastava, and R. Srivastava, "Downregulation of Rv0189c, encoding a dihydroxyacid dehydratase, affects growth of Mycobacterium tuberculosis in vitro and in mice," *Microbiology (Reading, Engl.)*, vol. 157, pp. 38–46, Jan 2011.

[27] H. Skovierova, G. Larrouy-Maumus, J. Zhang, D. Kaur, N. Barilone, J. Korduláková, M. Gilleron, S. Guadagnini, M. Belanova, M. C. Prevost, B. Gicquel, G. Puzo, D. Chatterjee, P. J. Brennan, J. Nigou, and M. Jackson, "AftD, a novel essential arabinofuranosyltransferase from mycobacteria," *Glycobiology*, vol. 19, pp. 1235–1247, Nov 2009.

[28] H. Qu, Y. Xin, X. Dong, and Y. Ma, "An rmlA gene encoding d-glucose-1-phosphate thymidylyltransferase is essential for mycobacterial growth," *FEMS Microbiol. Lett.*, vol. 275, pp. 237–243, Oct 2007.

[29] N. Buchmeier and R. C. Fahey, "The mshA gene encoding the glycosyltransferase of mycothiol biosynthesis is essential in Mycobacterium tuberculosis Erdman," *FEMS Microbiol. Lett.*, vol. 264, pp. 74–79, Nov 2006.

[30] R. Goyal, A. K. Das, R. Singh, P. K. Singh, S. Korpole, and D. Sarkar, "Phosphorylation of PhoP protein plays direct regulatory role in lipid biosynthesis of Mycobacterium tuberculosis," *J. Biol. Chem.*, vol. 286, pp. 45 197–45 208, Dec 2011.

[31] S. E. Haydel, V. Malhotra, G. L. Cornelison, and J. E. Clark-Curtiss, "The prrAB two-component system is essential for Mycobacterium tuberculosis viability and is induced under nitrogen-limiting conditions," *J. Bacteriol.*, vol. 194, pp. 354–361, Jan 2012.

[32] W. Zhang, V. C. Jones, M. S. Scherman, S. Mahapatra, D. Crick, S. Bhamidi, Y. Xin, M. R. McNeil, and Y. Ma, "Expression, essentiality, and a microtiter plate assay for mycobacterial GlmU, the bifunctional glucosamine-1-phosphate acetyltransferase and N-acetylglucosamine-1-phosphate uridyltransferase," *Int. J. Biochem. Cell Biol.*, vol. 40, pp. 2560–2571, 2008.

[33] A. Gurvitz, "The essential mycobacterial genes, fabG1 and fabG4, encode 3-oxoacyl-thioester reductases that are functional in yeast mitochondrial fatty acid synthase type 2," *Mol. Genet. Genomics*, vol. 282, pp. 407–416, Oct 2009.

[34] V. Molle, G. Gulten, C. Vilcheze, R. Veyron-Churlet, I. Zanella-Cleon, J. C. Sacchettini, W. R. Jacobs, and L. Kremer, "Phosphorylation of InhA inhibits mycolic acid biosynthesis and growth of Mycobacterium tuberculosis," *Mol. Microbiol.*, vol. 78, pp. 1591–1605, Dec 2010.

[35] T. Parish, M. Schaeffer, G. Roberts, and K. Duncan, "HemZ is essential for heme biosynthesis in Mycobacterium tuberculosis," *Tuberculosis (Edinb)*, vol. 85, pp. 197–204, May 2005.

[36] S. Banu, N. Honore, B. Saint-Joanis, D. Philpott, M. C. Prevost, and S. T. Cole, "Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens?" *Mol. Microbiol.*, vol. 44, pp. 9–19, Apr 2002.

[37] P. Domenech, M. B. Reed, and C. E. Barry, "Contribution of the Mycobacterium tuberculosis MmpL protein family to virulence and drug resistance," *Infect. Immun.*, vol. 73, pp. 3492–3501, Jun 2005.

[38] G. Lamichhane, S. Tyagi, and W. R. Bishai, "Designer arrays for defined mutant analysis to detect genes essential for survival of Mycobacterium tuberculosis in mouse lungs," *Infect. Immun.*, vol. 73, pp. 2533–2540, Apr 2005.

[39] J. D. McKinney, K. Honer zu Bentrup, E. J. Munoz-Elias, A. Miczak, B. Chen, W. T. Chan, D. Swenson, J. C. Sacchettini, W. R. Jacobs, and D. G. Russell, "Persistence of Mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase," *Nature*, vol. 406, pp. 735–738, Aug 2000.

[40] D. Wagner, F. J. Sangari, A. Parker, and L. E. Bermudez, "fecB, a gene potentially involved in iron transport in Mycobacterium avium, is not induced within macrophages," *FEMS Microbiol. Lett.*, vol. 247, pp. 185–191, Jun 2005.

[41] G. M. Seibold and B. J. Eikmanns, "The glgx gene product of corynebacterium glutamicum is required for glycogen degradation and for fast adaptation to hyperosmotic stress," *Microbiology*, vol. 153, no. Pt 7, pp. 2212–20, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17600065

[42] R. Kalscheuer, K. Syson, U. Veeraraghavan, B. Weinrick, K. E. Biermann, Z. Liu, J. C. Sacchettini, G. Besra, S. Bornemann, and W. R. Jacobs, "Self-poisoning of Mycobacterium tuberculosis by targeting GlgE in an alpha-glucan pathway," *Nat. Chem. Biol.*, vol. 6, pp. 376–384, May 2010.

[43] H. N. Song, T. Y. Jung, J. T. Park, B. C. Park, P. K. Myung, W. Boos, E. J. Woo, and K. H. Park, "Structural rationale for the short branched substrate specificity of the glycogen debranching enzyme GlgX," *Proteins*, vol. 78, pp. 1847–1855, Jun 2010.

[44] P. Mendoza Lopez, P. Golby, E. Wooff, J. Nunez Garcia, M. C. Garcia Pelayo, K. Conlon, A. Gema Camacho, R. G. Hewinson, J. Polaina, A. Suarez Garcia, and S. V. Gordon, "Characterization of the transcriptional regulator Rv3124 of Mycobacterium tuberculosis identifies it as a positive regulator of molybdopterin biosynthesis and defines the functional consequences of a non-synonymous SNP in the Mycobacterium bovis BCG orthologue," *Microbiology (Reading, Engl.)*, vol. 156, pp. 2112–2123, Jul 2010.

[45] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Res.*, vol. 38, pp. D211–222, Jan 2010.

[46] M. T. Ehebauer and M. Wilmanns, "The progress made in determining the Mycobacterium tuberculosis structural proteome," *Proteomics*, vol. 11, pp. 3128–3133, Aug 2011.

[47] C. L. Gee, K. G. Papavinasasundaram, S. R. Blair, C. E. Baer, A. M. Falick, D. S. King, J. E. Griffin, H. Venghatakrishnan, A. Zukauskas, J. R. Wei, R. K. Dhiman,

D. C. Crick, E. J. Rubin, C. M. Sassetti, and T. Alber, "A phosphorylated pseudokinase complex controls cell wall synthesis in mycobacteria," *Sci Signal*, vol. 5, p. ra7, 2012.

[48] S. S. Gurcha, A. R. Baulard, L. Kremer, C. Locht, D. B. Moody, W. Muhlecker, C. E. Costello, D. C. Crick, P. J. Brennan, and G. S. Besra, "Ppm1, a novel polyprenol monophosphomannose synthase from Mycobacterium tuberculosis," *Biochem. J.*, vol. 365, pp. 441–450, Jul 2002.

[49] K. E. Pullen, H. L. Ng, P. Y. Sung, M. C. Good, S. M. Smith, and T. Alber, "An alternate conformation and a third metal in PstP/Ppp, the M. tuberculosis PP2C-Family Ser/Thr protein phosphatase," *Structure*, vol. 12, pp. 1947–1954, Nov 2004.

[50] A. K. Pandey and C. M. Sassetti, "Mycobacterial persistence requires the utilization of host cholesterol," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 4376–4380, Mar 2008.

[51] R. Van Der Geize, K. Yam, T. Heuser, M. H. Wilbrink, H. Hara, M. C. Anderton, E. Sim, L. Dijkhuizen, J. E. Davies, W. W. Mohn, and et al., "A gene cluster encoding cholesterol catabolism in a soil actinomycete provides insight into mycobacterium tuberculosis survival in macrophages," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 6, pp. 1947–1952, 2007. [Online]. Available: http://eprints.kingston.ac.uk/19454/

[52] N. Lack, E. D. Lowe, J. Liu, L. D. Eltis, M. E. Noble, E. Sim, and I. M. Westwood, "Structure of HsaD, a steroid-degrading hydrolase, from Mycobacterium tuberculosis," *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, vol. 64, pp. 2–7, Jan 2008.

[53] D. J. V. Beste, M. Espasa, B. Bonde, A. M. Kierzek, G. R. Stewart, and J. McFadden, "The genetic requirements for fast and slow growth in mycobacteria," *PLoS ONE*, vol. 4, no. 4, p. e5349, 04 2009.

[54] S. G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M. F. Polz, "PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample," *Appl. Environ. Microbiol.*, vol. 71, pp. 8966–8969, Dec 2005.

VITA

Michael A. DeJesus attended the University of Puerto Rico at Mayagüez, receiving a B.S. degree in Computer Science in 2003. He went on to attend Texas A&M University, working under Dr. Thomas Ioerger in the area of bioinformatics. His interests include bioinformatics, artificial intelligence, and evolutionary computation.

Mr. DeJesus may be reached at Department of Computer Science, Texas A&M University, 301 Harvey R. Bright Building, College Station, TX 77843-3112.

His e-mail is michael.dejesus@tamu.edu

The typist for this thesis was the author.