STATISTICAL ANALYSIS OF TRANSPOSON SEQUENCING DATA TO DETERMINE ESSENTIAL GENES

A Dissertation

by

MICHAEL A. DE JESUS ANEIRO

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Thomas R. Ioerger
Committee Members,	Tiffani L. Williams
	James C. Sacchettini
	Ricardo Gutierrez-Osuna
Head of Department,	Dilma Da Silva

December 2016

Major Subject: Computer Science

ABSTRACT

Transposon Sequencing (TnSeq) has become a popular biological tool for assessing the phenotypes of large libraries of bacterial mutants at the same time. This allows for high-throughput identification of genes which are essential for growth, thus providing valuable information about the function of those genes and the discovery of potential drug targets that could lead to treatments.

However, analysis of data obtained from TnSeq is challenging as it requires estimating unknown parameters from data that is often noisy and likely coming from a mixture of different phenotypes. In addition, the classification of essentiality is not known a priori, requiring unsupervised methods capable of identifying key features in the data to confidently determine essentiality.

We present several models capable of identifying essential genes while overcoming the difficulties that are present in analyzing TnSeq data. Together, these methods provide ways to analyze TnSeq data in one or multiple conditions, confined within gene boundaries or across the entire genome, and while reducing the impact of noise and outliers that are often present in this type of data.

To my mother, father, brother, 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, Dr. Ricardo Gutierrez-Osuna, and Dr. Tiffani L. Williams, for their guidance and input.

I would like to thank our collaborators Jason Zhang, Jennifer Griffin, Subhalaxmi Nambi, Richard Baker, Clare Smith, Andrew Olive, 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 weird kid I am today.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Motivation

Identifying genes that are essential for growth of bacterial organisms is important for the development of drugs that inhibit the function of a crucial protein, thus possibly becoming a target for treatment of infectious bacteria. For instance isoniazid and rifampicin, which are first-line drugs used to treat tuberculosis, both bind to proteins that play essential roles (InhA and DNA Polymerase respectively) thus preventing the growth of the pathogen [1, 2]. Furthermore, if the essentiality of a gene is shown to depend on a particular condition, this can provide valuable insight on the function of unknown proteins. In order to identify which genes are essential, individual bacilli are mutated in such a way that the function of one (or more) of its genes is interrupted. This can be done in a high-throughput fashion by using a small fragment of DNA (called a transposon) to disrupt random locations in the genome, thus allowing for the creation of large libraries of mutants. However, bottlenecks in previous sequencing methods did not allow for the sequencing of large libraries of mutants at the same time. New advances in sequencing have made it possible to rapidly sequence an entire library of thousands of such mutants simultaneously. By sequencing large libraries of transposon mutants, researchers have access to high-resolution sequence data that reveals which specific areas of the genome can be disrupted by a transposon, providing valuable information about their need for an organism's viability.

Although this high-resolution sequence data has the potential of providing a wealth of new information about essentiality, there are a large number of issues that make any analysis of this data challenging. As mutant libraries represent those bacteria that survived a transposon insertion in their DNA, the resulting data reflects only those regions which are capable of tolerating disruption. However, genomic regions lacking any insertions do not necessarily imply that the area is essential to the organism. These areas may represent sites that were simply missed by chance (as not all potential insertion sites are saturated when construction a transposon library) but are otherwise non-essential to the organism (and thus would have tolerated insertion of a transposon if one had occurred there). In this sense, TnSeq data can be thought of as a missing data problem where the disruptability of empty sites is unknown and most be determined statistically. Furthermore, as the essentiality of genes is not known beforehand, unsupervised methods are required to confidently determine which genes are necessary. While transposon insertions are supposed to disrupt the function of the genomic regions where they insert, in reality essential genes are often able to tolerate some insertions. For instance, it is common to observe insertions in the N- and C-terminus of a gene as the protein may still able to be translated and fulfill its biological function despite the insertion [3, 4, 5]. Thus, although initially one may be tempted to determine essentiality based on whether a gene has evidence of insertions or not, more sophisticated statistical models are required.

1.2 Background

1.2.1 Transposon Mutagenesis

In order to determine essentiality it useful to observe how an organism copes with the loss of a gene's function. A popular technique used to disrupt the function of a genomic region is transposon mutagenesis. In transposon mutagenesis, a mutation is mediated through the insertion of a transposon at a random position in the genome. Transposons are small fragments of DNA (typically 1-2kb long) that can insert within the chromosomes of an organism [6]. Different transposons exist with varying characteristics like local-sequence preference [7, 8, 9]. The Himar1 transposon, for example, has shown specificity for arbitrary TA dinucleotides [8]. As the entire DNA sequence of a bacterial organism can be known beforehand through sequencing, the preference for TA dinucleotides can be used to know which sites may be targeted by the transposon.

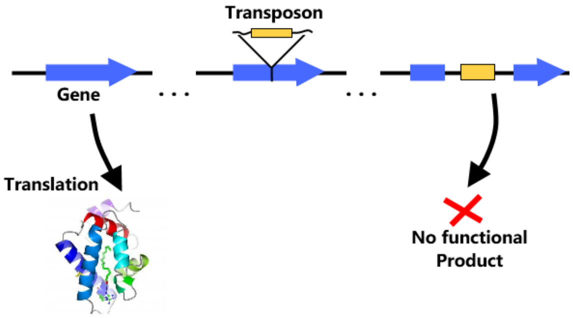


Figure 1.1: Example diagram of transposon mutagenesis

Using transposon mutagenesis, large libraries of mutants are constructed with disruptions at random locations of the genome. The resulting libraries are grown under environmental conditions of interest, thus potentially revealing phenotypes for the mutants.

Determining what disruption resulted in the phenotype observed, however, requires identifying the location of the transposon insertion.

1.2.2 Transposon Site-Hybridization (TraSH)

Transposon Site-Hybridization (TraSH) was a previous attempt to determine essentiality using transposon mutant libraries. In order to identify the location of insertions, TraSH used micro-array hybridization to figure out what genes in the mutant libraries where being expressed and which ones where not [10]. 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 (Figure 1.2). Thus, genes which are non-essential for growth are those which hybridize to the probes in rates similar to the background rate. Those genes which are essential for growth, and thus do not have surviving mutants in the library, would represent those which hybridize as significantly lower rates. Using this approach, genes essential for growth in variety of organisms like *M. tuberculosis* and *B. anthracis* were identified [11, 12, 13].

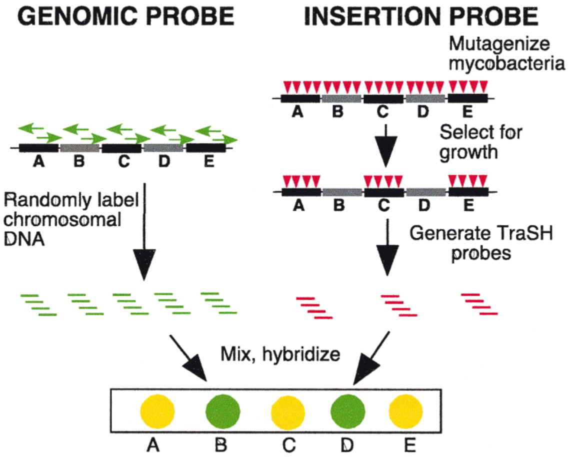


Figure 1.2: Diagram of the TraSH method. Source: Sassetti (2003) [11]

Although one can obtain a general a idea of what genes were being expressed through the use of TraSH, it does not provide high-resolution information about where the insertions took place. In addition, a serious statistical treatment of the data produced by TraSH is made difficult by the fact that the assessment of essentiality is limited to measurements of hybridization at a few probes (i.e. 4) for each gene. With the advent of new sequencing technology, it become possible to determine the exact location of insertions, thus replacing the need for TraSH.

1.2.3 Transposon Sequencing (TnSeq)

With the development of high-throughput sequencing, large libraries of mutants can be sequenced at the same time, providing the exact location of the transposon insertions. This technique, called transposon sequencing (TnSeq), helps overcome many of the limitations of previous methods. Once libraries of transposon mutants are created and exposed to the desired conditions, the surviving mutants are sequenced using deep-sequencing. The resulting sequence reads are mapped to the genome to determine the precise coordinate of the transposon insertions. While the number of reads mapping to any given location should be proportional to the number of mutants in the library, the number of reads can be affected by artifacts like PCR amplification. More modern protocols utilize barcodes to ensure that counts at a position represent unique insertions [14].

TnSeq has been successfully used to analyze essentiality in a number of different organisms and growth conditions [15, 16, 17]. However, despite being used since 2009, there was a lack of rigorous statistical methods capable of analyzing the large amount of data produced by TnSeq. Initial attempts to determine essentiality through TnSeq tended to rely on arbitrary thresholds of fitness, or ad-hoc criteria to overcome the difficulties of analyzing this data. For instance, since essential genes are often capable of sustaining insertions at the N- and C- termini, these methods often required the exclusion of the first and last 5-20% of a gene's coding region so as to not label these genes as non-essential [15].

1.3 Related Work

One of the earliest approaches to analyzing TnSeq data was work done by Natalie Blades and Karl W. Broman [18]. Their method utilizes a multinomial function to characterize the number of mutants with transposon insertions unique to a gene, as well as the number of mutants with transposon insertions that occur in coding regions that are shared (overlap) between two adjacent genes. This last feature is meant to capture the uncertainty that exists when the coding regions of adjacent genes overlap, and thus a transposon insertion cannot be definitively assigned to one particular gene. One important limitation of the approach used by Blades and Broman, is that it assumes that any gene with a transposon insertion is considered to be non-essential. Thus, they focused on estimating the probability of a gene being essential given that it was devoid of insertions, the number of sites within the gene, and the overall saturation of the library. This assumption is useful in the case of libraries with very low saturation (i.e. libraries with few mutants and little sequence data available) as was common before the advancement in sequencing technology. However,

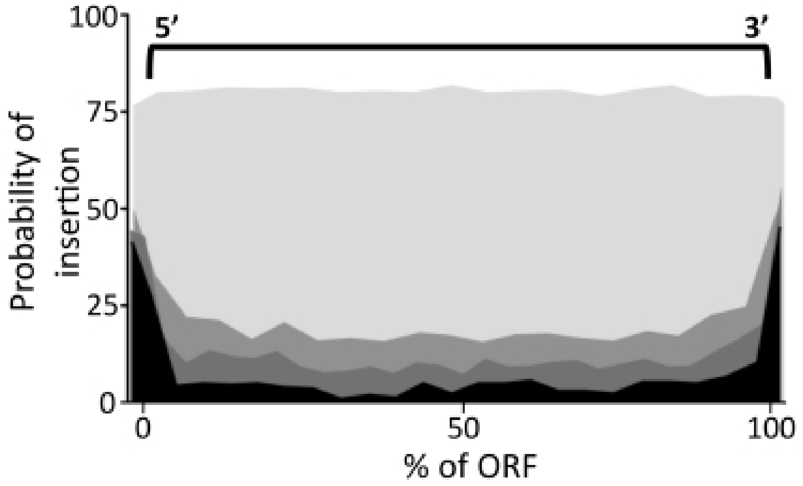


Figure 1.3: Frequency of insertions as percentage of the ORF. Darker shades represent genes with higher probability of being essential. Essential genes have a high likelihood of observing insertions near the 5' and 3' ends, and a non-zero probability of containing insertions across the entire coding-region. Source: Griffin et al. (2011) [17]

it is not true that a gene with a transposon insertion must be non-essential. Most genes, including essential genes, have a high probability of tolerating insertions within the 5' and 3' terminal ends (N- and C- termini) of the gene, but may tolerate insertions in other parts of the gene as well (Figure 1.3).

To address the insertability observed at the terminal ends of genes, methods would often ignore insertions that occurred within a predetermined distance of the termini. For instance Gawronski et al. [15] excluded insertions that occurred in the first 5% and last 20% of a gene's coding region. Criteria like this require assuming a certain fixed cut-off distance at which the transposon insertions would disrupt a genes functions when they occur in the ends. However, this is unlikely to be equal for all proteins. More importantly, essential genes can allow insertions at other areas of the gene, if they contain non-essential protein domains or at linkers within genes [19].

Other methods for analyzing TnSeq data have borrowed from the RNA-Seq literature. Like TnSeq, RNA-Seq is based on counts of reads obtained from sequencing [20]; thus much of the same methodology can be used for both types of data. A popular choice for analyzing RNA-Seq data is edgeR [21], which utilizes a Negative Binomial distribution to assess the likelihood of the read-counts observed within a gene. One of the advantages of the Negative Binomial model is that it can represent over-dispersion in the data that is typically observed in biological datasets. It can also be extended to account for the abundance of sites with no insertions (i.e. counts of zero), by using a zero-inflated Negative Binomial model. Given the success of the Negative Binomial model in analyzing RNA-Seq data, Zomer et al. [22] developed software called ESSENTIALS which utilizes edgeR as the underlying analysis method to distinguish between essential and non-essential genes.

Unfortunately, ESSENTIALS is also hindered by the presence of insertions at the terminal ends or non-essential domains of essential genes. As edgeR takes the total number of read-counts that occur within a gene, any essential gene which tolerates some insertions will tend to have a higher number of read-counts compared to essential genes that are completely devoid of insertions. Thus, ESSENTIALS tends to underestimate the number of essential genes, classifying genes that can tolerate some insertions, as non-essential.

Most of the methods mentioned so far limited to determining essentiality within predefined genetic boundaries like the coding-region of a gene. However, the essentiality of regions outside of genes is also important, as these may include genomic features that play important roles (e.g. like binding sites for transcription factors or DNA methylation sites). Methods meant to analyze the essentiality of an entire genome often depend on a sliding-window approach, where the read-counts that occur within an specified window of insertions sites (or nucleotides) are compared to some null distribution [23, 24].

Aside from not being limited to coding-regions, another advantage of these approaches is that they are not susceptible to insertions at the terminal ends of ORFs or within domains. However, sliding-window approaches require predetermined choice for length of the sliding-window, which can significantly impact how much influence is exerted by neighboring insertions sites. A more rigorous methodology would be preferable in these cases, as it is often not obvious which window-size would be best (or more justified) for determining essentiality.

1.4 Scope and Contribution

In this dissertation, several statistical models for determining essentiality from TnSeq data are presented. Each method discussed here represents a novel approach to overcoming the difficulties of analyzing TnSeq data, and led to a publication that outlined the methodology:

- Section 2 presents a Bayesian model of essentiality that identifies the unusually long stretches of sites without insertions that are typical of essential genes (or domains) [17, 25]. In doing so, it overcomes the difficulties that exist in identifying genes that can tolerate insertions at their terminal ends or contain non-essential domains.

- Section 3 presents a hierarchical Bayesian model of essentiality that estimates individual insertion probabilities for each gene, thus relaxing assumptions made by most other models (which typically assume genes share a single parameter representing insertion probability or mean read-count) [26, 27].
- Section 4 presents a Hidden Markov Model (HMM) to determine the essentiality of an entire genome [28]. This HMM set itself apart from most other methods since it has four states representing four different categories of essentiality (or levels of fitness).
- Section 5 describes how to identify conditionally essential genes (genes that are essential in one condition but not another), and presents two different methods that normalize datasets with different saturation levels, number of reads, and skew [29, 30].
- Section 6 presents a novel method capable of identifying genetic interactions by comparing TnSeq datasets of different strains grown under two experimental conditions.

These (and other) methods are provided in a graphical software package called TRANSIT [31].

2 DETERMINING ESSENTIAL GENES BY DETECTING UNUSUALLY LONG GAPS*

2.1 Introduction

The primary goal of analyzing transposon mutagenesis datasets is to determine which genes are essential for growth under a specific condition. As transposon insertions disrupt the function of the genomic regions they insert in, those that occur in essential genes will render the organism unable to carry out necessary functions for survival.

The underlying idea behind the Gumbel Model of essentiality is that genes that are essential will have unusually long stretches of the genome without any observable transposon insertions [17, 25]. These regions would appear as “gaps” in the transposon insertion pattern (Figure 2.1), and would indicative of the essentiality of a region as they are unlikely to occur by chance. On the other hand, because the organism is capable of tolerating insertions in non-essential regions, those areas should exhibit gaps that are as long as would be expected given the saturation of the library.

Because the Himar1 transposon inserts specifically at TA dinicleotides, the number of TA sites in a given gene (and which of them had an insertion) can be easily determined from the data. In an analogy to a sequence of coin tosses, each TA site can be viewed as an independent Bernoulli trial with a parameter representing the probability of success.

*DeJesus, M.A., Zhang, Y.J., Sassettti, C.M., Rubin, E.J., Sacchettini, J.C., and Ioerger, T.R., “Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries”, Bioinformatics, 2013, 29(6), 695-703, by permission of Oxford University Press.

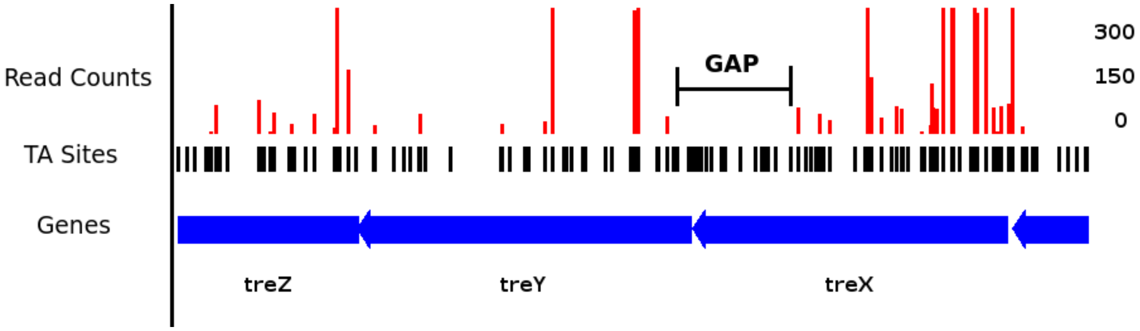


Figure 2.1: Insertion pattern for TreX, a gene involved in trehalose biosynthesis. The presence of a long gap - a region of sites without any insertions - suggests the gene may code for a protein domain that plays an essential role.

A sequence of TA sites in a row without any transposon insertions (or "run") is therefore analogous to a sequence of heads in a row. The expected value and the variance of the largest run of heads in a row are given by the following equations [32]:

$$ER_n = \log_{1/\phi}(n(1-\phi)) + \gamma/ln(1/\phi) - 1/2 + r_1(n) + \varepsilon_1(n) \tag{2.1}$$

$$VarR_n = \pi^2/6ln^2(1/\phi) + 1/12 + r_2(n) + \varepsilon_2(n) \tag{2.2}$$

where n represents the number of coin tosses (or sites), ϕ represents the probability of heads (or probability of non-insertion), and r_1, r_2, ε_1, and ε_2 are small correction terms.

To model the distribution of the largest run in a series of trials, the Extreme Value distribution (or Gumbel distribution) is utilized. The Gumbel distribution is part of the exponential family of distributions, and has the following form:

$$\begin{aligned} Gumbel(x;\mu,\sigma) &:= \frac{1}{\sigma}e^{-z-e^{-z}} \\ z &= \frac{x-\mu}{\sigma} \end{aligned} \tag{2.3}$$

with location and scale parameters μ and σ respectively. 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.

Combining Equations 2.1 and 2.2 with the formulas for the mean, and variance of the Gumbel distribution (ignoring the negligible correction terms for simplicity) results in the following parameters:

$$\mu = \log_{\frac{1}{\phi}}(n(1-\phi)) \tag{2.4}$$

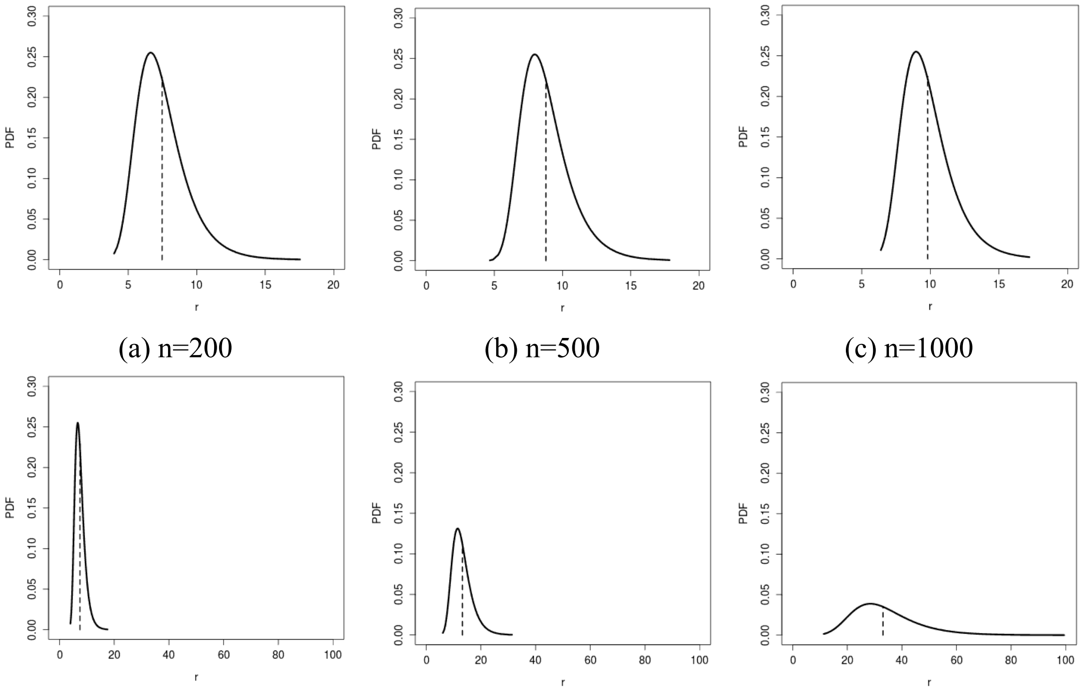
$$\sigma = \frac{1}{\ln \frac{1}{\phi}} \tag{2.5}$$

where n is the number of trials (or TA sites) and ϕ is the probability of observing a head (or empty TA site).

Figure 2.2 shows distributions of the longest runs of heads in a series of coin tosses for different values of n and different values of ϕ. The expected maximum run scales up logarithmically in n and $1-\phi$ as n.

followingcreditlineappearswhereverthematerialisused:author, title,journal,year,volume,issuenumber,pagination,bypermissionofOxford University-Pressorthesponsoringsocietyifthejournalisasocietyjour

Although the number of sites in each gene is known beforehand, the probability of



(a) n=200

(b) n=500

(c) n=1000

(d) $\phi_0 = 0.5$

(e) $\phi_0 = 0.7$

(f) $\phi_0 = 0.9$

Figure 2.2: Gumbel distributions with different values of ϕ and n.
The first row of figures shows how the distribution behaves while varying the number of trials, n, and using a fixed probability of success $\phi = 0.5$. The second row shows how the distribution behaves while varying the probability of success, ϕ,in a fixed number of trials, $n = 200$. The vertical dashed lines shows the expected maximum run according to the Gumbel distribution.

non-insertion, ϕ, is unknown. This parameter is crucial to estimate in order to determine the probability of a gene being essential. Section 2.2 formally describes the Bayesian model of the data while Section 2.3 describes the posterior distributions and how the variables are estimated.

2.2 Data Model

Let $Y_i = \{r_i, s_i, n_i\}$ represent our observations for the i-th gene for $i = 1...G$, where n_i represents the total number of TA sites, r_i represents the number of TA sites spanned by the largest run of non-insertions, and s_i represents the number of nucleotides spanned by the largest run. 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 genes respectively. 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, $p(Z,Y,\phi_0)$, which combines the observed data as well as the unobserved parameters of the model. Using Bayes theorem we can use the joint distribution to derive conditional distributions from which we can obtain estimates of the essentiality for the genes 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:

$$(Y|Z,\phi_0) * p(Z) * p(\phi_0) \tag{2.6}$$

Since we assume independence among genes the likelihood can be written as a product of the individual observations:

$$p(Y|Z,\phi_0 \propto \prod_i p(Y_i|Z,\phi_0) \tag{2.7}$$

$$= \prod_i p(s_i,r_i,n_i|Z_i,\phi_0) \tag{2.8}$$

Due to the definition of the joint probability, the joint likelihood of the data (i.e. $p(s_i,r_i,n_i)$) may be specified different ways (i.e. $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$:

$$\begin{aligned} p(Y_i|Z,\phi_0) &= p(r_i,n_i,s_i|Z=0,\phi_0) \\ &= p(r_i,n_i|Z=0,\phi_0) \times p(s_i|r_i,Z=0,\phi_0) \\ &= p(s_i|Z=1,\phi_0) \times p(r_i,n_i|s_i,Z=1,\phi_0) \end{aligned}$$

This fact is useful different distributions can be used to represent the mixture of essential and non-essential genes. Sections 2.2.1 and 2.2.2 show the specification of the likelihoods for non-essential and essential genes respectively.

2.2.1 Likelihood for Non-Essential Genes

As non-essential genes are not required by the organism, they are expected to withstand disruption at levels that correspond to the probability of insertion in the library (i.e. the saturation of the library). The length of the maximum run of insertions in a non-essential gene should therefore conform to the Gumbel distribution, given the non-insertion probability ϕ_0:

$$p(r_i|Z_i = 0, \phi_0, \omega_1) = Gumbel(x; m, \tau) = \frac{1}{\tau} e^{-z-e^{-z}} \tag{2.9}$$

where m and τ represent location and scale parameters.

Since r_i and s_i are highly correlated, we model their dependence as linear-Gaussian distribution, with covariance matrix $\Sigma = \left[\left[\sigma_r^2, \sigma_{r,s}\right], \left[\sigma_{r,s}, \sigma_s^2\right]\right]$ estimated *a priori* from empirical data:

$$p(s_i|r_i, Z = 0, \phi_0, \omega_1) \sim N(s_i - \lambda_r r_i, \sigma_r^2) \tag{2.10}$$

were λ_r and σ_r are the parameters of the Normal distribution, derived from the Linear Gaussian relationship (i.e. $\lambda_r = \frac{\sigma_{r,s}}{\sigma_r}$) observed in the data.

The likelihood for a non-essential gene is therefore:

$$\begin{aligned} p(Y_i|Z, \phi_0) &= p(r_i, n_i, s_i|Z = 0, \phi_0) \\ &= p(r_i, n_i|Z = 0, \phi_0) \times p(s_i|r_i, Z = 0, \phi_0) \\ &= Gumbel(x; m, \tau) = \frac{1}{\tau} e^{-z-e^{-z}} \times \sim N(s_i - \lambda_r r_i, \sigma_r^2) \end{aligned}$$

2.2.2 Likelihood for Essential Genes

Unlike non-essential genes, those genes which are necessary for the growth of the organism will have stretches of TA sites lacking insertions that are longer than would we expected by chance. This requires using different distributions to describe the likelihood.

$$p(r_i, s_i | Z_i = 1, \phi_0, \omega_1) =$$
$$p(s_i | Z = 1, \phi_0, \omega_1) \times p(r_i | s_i, Z = 1, \phi_0, \omega_1)$$

We model the likelihood of observing a span of nucleotides (s_i) with a normalized sigmoid (logistic) function that is relatively uniform as long as the gene contains a gap that is as large as a typical protein domain. Using this likelihood allows our method to disambiguate those cases where the run of non-insertions actually represents a smaller or larger segment of the genome than suggested by the number of consecutive TA sites without insertions:

$$p(s_i | Z_i = 1) = \Omega(s_i; \delta) = \frac{C}{1 + e^{0.01*(\delta - x)}} \tag{2.11}$$

where δ is the mean number of nucleotides spanned by an average protein domain, and C is a normalization constant. Previous studies of the length of domains within proteins have found the average size to be roughly 100 amino-acids or 300bp [33]. Using this threshold for δ, the likelihood of observing a given span s_i is more or less uniform, except it is near 0 if the the longest run of non-insertions spans less than about 300bp.

As with non-essential genes, the likelihood of observing a span of nucleotides r_i given s_i is modeled through a linear-Gaussian dependence similar to Equation (2.10), but with an inverse relationship (i.e. $N(r_i - \lambda_s s_i, \sigma_s^2)$). The joint likelihood of the observations at essential genes is therefore:

$$p(r_i, s_i | Z_i = 1, \phi_0, \omega_1) = \Omega(s_i) \times N(r_i - \lambda_s s_i, \sigma_s^2) \tag{2.12}$$

2.3 Parameter Estimation

To estimate the parameters of interest, including the probability of essentiality, the posterior distributions of these unknown parameters must first be derived. A prior distribution for these parameters is also required.

2.3.1 Posterior Distribution of ω

The prior distribution of the mixture coefficient, ω, can be taken to be a Beta distribution. The Beta distribution is a common choice for a prior on a probability parameter as its support is defined in the interval [0,1] and it is conjugate with other common distributions. The posterior distribution is derived as follows:

$$\begin{aligned} p(\omega_1 \mid Y, Z, \phi_0) &\propto \pi(Z \mid \omega_1) \times \pi(\omega_1) \\ &\propto Binomial(K_{z1}; \omega_1, G) \times Beta(\omega_1; \alpha_w, \beta_w) \\ &\propto Beta(\omega_1; \alpha_w + K_{z1}, \beta_w + G - K_{z1}) \end{aligned} \tag{2.13}$$

2.3.2 Posterior Distribution of Z

The probability of essentiality is estimated through the indicator variable, Z_i, which indicates which mixture (or essentiality class) the gene belongs to. As there are two possible essentiality classes, the posterior is given for both possible values (i.e., $Z_i = 1$ and $Z_i = 0$):

$$\begin{aligned} & p(Z_i = 1 \mid Y, Z_{\{-i\}}, \phi_0, \omega_1) \\ & \propto p(s_i \mid Z_i = 1) \times p(r_i \mid s_i, Z_i = 1) \times \pi(Z_i = 1 \mid \omega_1) \\ & \propto \Omega(s_i) \times N(r_i - \lambda_s s_i, \sigma_s^2) \times \omega_1^{Z_i=1}(1-\omega_1)^{1-Z_i=1} \end{aligned} \tag{2.14}$$

$$\begin{aligned} & p(Z_i = 0 \mid Y, Z_{\{-i\}}, \phi_0) \\ & \propto p(r_i \mid Z_i = 0, \phi_0) \times p(s_i \mid r_i, Z_i = 0) \times \pi(Z_i = 0 \ \mid \omega_1) \\ & \propto Gumbel(r_i \mid m, \tau) \times N(s_i - \lambda_r r_i) \times \omega_1^{Z_i=0}(1-\omega_1)^{1-Z_i=0} \end{aligned} \tag{2.15}$$

As there are only two possible values, the posterior probability of an individual Z_i therefore a Bernoulli distribution:

$$Z_i = Bernoulli(\frac{p_1}{p_1 + p_0})$$

$$p_1 = p(r_i, s_i | Z_{\{-i\}}, \phi_0) \times \omega_1$$

$$p_0 = p(r_i, s_i | Z_{\{-i\}}, \phi_0) \times (1 - \omega_1)$$

2.3.3 Posterior Distribution of ϕ_0

As a prior for the non-insertion probability, ϕ_0, the Beta distribution is chosen. As the likelihood of non-essential genes is the only one which depends on ϕ_0, others can be discarded as constants with respect to ϕ_0. Unfortunately, the remaining likelihood is a product of Gumbel distributions (for individual non-essential genes). This likelihood is not conjugate with any known distribution, thus the resulting posterior distribution does not have standard form that is easy to sample from:

$$\begin{aligned} p(\phi_0 \mid Y, Z, \omega_1) &\propto p(Y \mid Z, \phi_0, \omega_1) \times \pi(\phi_0) \times \pi(Z \mid \omega_1) \times \pi(\omega_1) \\ &\propto \prod_{i}^{G} p(r_i, s_i \mid Z_i, \phi_0, \omega_1) \times \pi(\phi_0) \times \pi(Z \mid \omega_1) \times \pi(\omega_1) \\ &\propto \prod_{i=1}^{non} Gumbel(r_i \mid m, \tau) \times \pi(\phi_0) \end{aligned} \tag{2.16}$$

Because this posterior distribution does not have an standard form, another approach to approximating it must be used.

2.3.4 Metropolis-Hastings

In order to sample from the posterior density of the ϕ_0 parameter, we utilize a random-walk Metropolis Hastings (MH) algorithm. The MH algorithm is capable of sampling from arbitrary distributions of interest by proposing new candidate values from a proposal distribution that depends on the last accepted value, $\phi_0^{(j-1)}$. A common choice for this proposal distribution is a Normal distribution with mean equal to the last acceppted value and with small variance, v. Candidate values are accepted or rejected probabilistically, depending on their relative likelihood.

Algorithm 1 presents the sampling scheme used to sample the posterior densities of ϕ_0 and Z_i, and ω. A MH step is taken to sample ϕ_0, individual values of Z_i are sampled for each gene, and finally the mixture coefficient, ω, is sampled given the current indicator vector.

```
Algorithm: Random-Walk Metropolis-Hastings
Result: MCMC Samples of density $p(Z|Y,\phi_0)$ and $p(\phi_0|Y,Z)$
Assign starting value to $\phi_0$, and initialize $Z$ based on proportion of insertions
  within individual genes (i.e. If $\frac{|TA|_i}{n_i} < 0.1$ then $Z_i = 1$ else $Z_i = 0$);
for j=1 to desired sample size do
    Draw candidate parameter $\phi_0^c$ from Normal distribution, N($\phi_0^{j-1}$, $v$);
    Compute ratio R = $\frac{p(\phi_0^c|Y,Z)}{p(\phi_0^{j-1}|Y,Z)}$ ;
    Draw $u$ from uniform distribution on [0,1] ;
    if $u < R$ then
        Set $\phi_0^{(j)} = \phi_0^c$;
    else
        Set $\phi_0^{(j)} = \phi_0^{j-1}$ ;
    end
    Let $K_z$ equal the number of genes with $Z_i^j = 1$;
    Let $G$ be the total number of genes;
    Sample $\omega_1^{(j)} \sim Beta(\alpha_w + K_z, \beta_w + G - K_z)$;
    for $i \leftarrow 1$ to $G$ do
        $p_1 = p(r_i, s_i|Z_i = 1, Z_{\{-i\}}, \phi_0) \times \omega_1$ ;
        $p_0 = p(r_i, s_i|z_i = 0, Z_{\{-i\}}, \phi_0) \times (1 - \omega_1)$ ;
        Sample $Z_i^{(j)} \sim Bernoulli(\frac{p_1}{p_1 + p_0})$ ;
    end
end
```

Algorithm 1: Random-Walk Metropolis-Hastings Algorithm for Sampling ϕ_0 and Z

Since the MH algorithm samples from the conditional distributions of the parameters 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, it is common to eliminate an initial number of samples to to ensure that the MH algorithm reaches a point where it is mixing well before the samples are used to achieve estimates. This is referred to as the “burn-in” period [34].

Another 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 can exhibit some correlation with previous samples at time $t - k$. If the algorithm is producing results that are highly-correlated, then the sampler may not be truly exploring the distribution of interest in a random manner. This form of auto-correlation can be “trimmed” by discarding every s-th sample, thus effectively making the remaining samples uncorrelated. Once an adequate sample is obtained from the MH procedure, the sample can be used to estimate the parameters of interest.

2.4 Results

The Gumbel Model was applied to 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) [35]. 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.

2.4.1 Essentiality Results

After obtaining the sample from the MH procedure, the posterior probability of essentiality for all genes is estimated by averaging over the sample of essentiality values, $\bar{Z}_i$. Genes with $\bar{Z}_i < 0.05$ are classified as non-essential (i.e. $Z_i = 1$ in less than 5% of the final sample), and genes with $\bar{Z}_i > 0.95$ are classified as essential. A total of 757 genes are categorized as essential by this criterion. The remaining genes represent those genes for which the method is unable to reach an essentiality assignment with confidence. Figure 2.3 shows a plot of the sorted $\bar{Z}_i$ values for all the genes, with the blue lines representing the thresholds of essentiality and non-essentiality. Notice that the majority of genes have a small probability of being essential (i.e. low Z_i) which expected in most bacterial genomes.

Table 2.1: Statistics for essentials, non-essentials and uncertain genes. Non-essential genes are those with $Z_i < 0.05$, Essential genes are those with $Z_i > 0.95$. Average span is in nucleotides.

	Total	**Average**			
	Genes	**TA Sites**	**Insertions**	**Max Run**	**Span**
Essentials	757	20.50	1.87	16.35	969.55
Uncertain	242	17.43	7.50	5.27	400.74
Non-Essentials	2703	15.69	10.77	2.05	55.47

Table 2.1 reports statistics for the different categories of genes. On average essential genes contained significantly longer maximum runs of non-insertion (16.35) than non-essential genes and these runs spanned a larger number of nucleotides (969.5), which is consistent with our expectations for essentiality. Non-essential genes contained a larger number of insertions on average (15.69). Although essential genes contained only a small

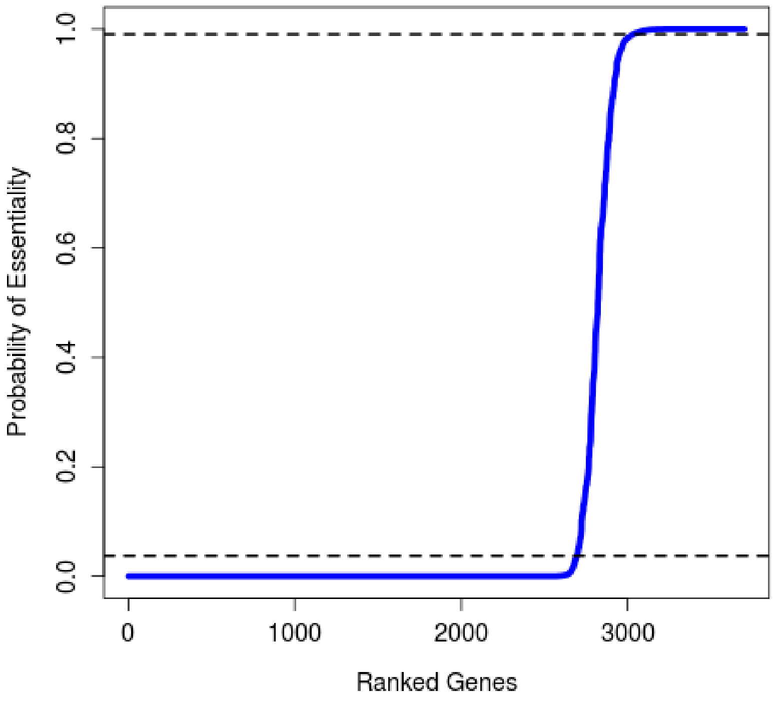
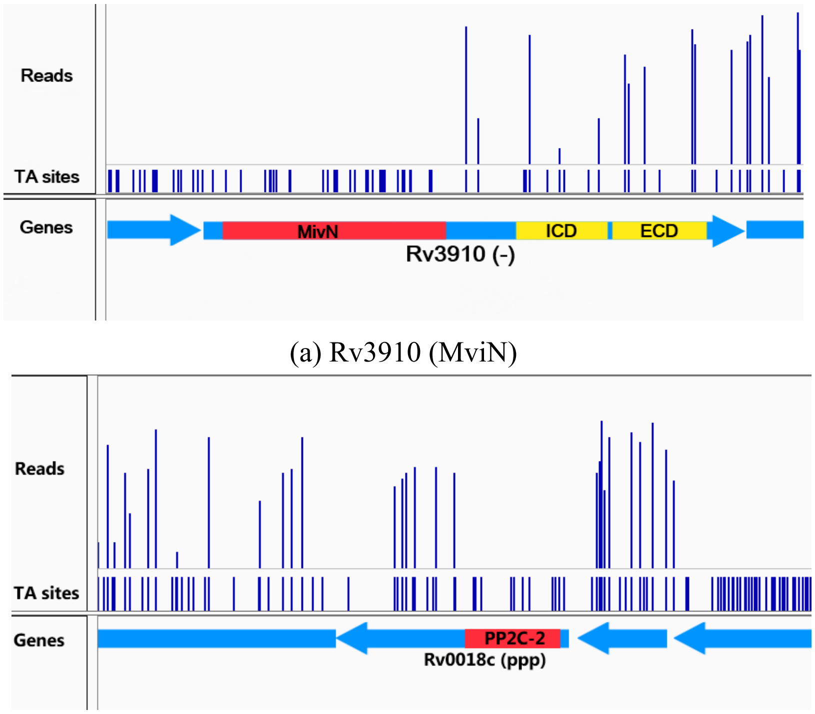


Figure 2.3: Plot of sorted $\bar{Z}_i$ values for all genes. The average Z_i of the final sample for all genes was estimated, and plotted in ascending order. The dashed lines represent the respective thresholds for the two categories of essentiality: $\bar{Z}_i > 0.9902$ and $\bar{Z}_i < 0.0371$

number of insertions (1.87) this number was greater than zero, indicating that the method is capable of detecting essential genes with a small number of insertions, provided they contain a long enough run of non-insertions suggestive of an essential region.

Figure 2.4 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 [36]. Insertions in Rv3910 are found only in the C-terminal domains, but not the N-terminal mem-



(a) Rv3910 (MviN)

(b) Rv0018c (ppp)

Figure 2.4: Example genes with essential domains. Essential domains are indicated in red, and non-essential domains (as predicted by PFAM) are indicated in yellow.

brane 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 [37]. Rv0018c (serine/threonine phosphatase) contains an essential catalytic domain within its N-terminus, and has been shown to dephosphorylate Rv0020 (FhaA) counteracting phosphorylation by PknB [38]. Transposon insertions are only observed in the extracellular domain of unknown function.

2.4.2 Concordance with Previous Results

The essentiality of the entire *M. tuberculosis* genome has been characterized previously using transposon-site hybridization [11, 12]. We compare our essentiality inferences to previous results to verify that our method achieves results that are consistent with expectations of the essentiality in *M. tuberculosis*. Sassetti et al. utilized Transposon Site Hybridization (TraSH) to characterize the genes necessary for optimal growth *in vitro*, for a library of transposon mutants grown on 0.02% glucose and rich-media (7H10). While our method analyzes deep sequencing of transposon libraries, TraSH utilizes hybridization of gene-specific probes to quantify the level of fluorescence being emitted by hybridization probes to determine which genes are being interrupted in the library of mutants. Table 2.2 contains a comparison between the two methods.

Table 2.2: Comparison of essentiality predictions with TraSH analysis. The results obtained by Sassetti. et. al are compare with those obtained with our Gumbel method for genes in *M. tuberculosis*. Genes for which the TraSH method did not produce data, are labeled "no-data". Genes with less than four TA sites were labeled "Short" as they could not be analyzed by the Gumbel method.

		Gumbel Method				
		Essential	**Uncertain**	**Non-Essential**	**Short**	**Total**
TraSH (Sassetti)	**Essentials**	**457**	46	82	29	614
	Growth-Defect	11	2	28	1	42
	Non-Essential	123	116	**2137**	144	2520
	No-Data	166	78	456	113	813
	Total	757	242	2703	287	3989

Sassetti et al. included an additional category of genes representing those whose interruption causes growth-defects (i.e. slower growth); our method does not make this distinction. Excluding these, the two methods show agreement in 74.4% of essentials, and 84.8% of non-essentials for a total of 82.8% across both categories. There were only 82 genes predicted to be essential by TraSH but not by our method, and 123 genes predicted to be non-essential by TraSH but found to be essential by our method.

Some of these differences could be due to the different growth conditions of the libraries. For example, because our library was grown on glycerol we find genes necessary for glycerol metabolism as essential, such as GlpK (glycerol kinase). Other differences may be due to incomplete sequence coverage (e.g. gaps in PE_PGRS genes, which are highly GC-rich and hard to sequence). Two out of 62 PE_PGRS genes in the H37Rv genome were classified as essential by our model because of large regions without insertions, though genes in this family are generally believed to be non-essential [39]. Over-representation of PE_PGRS gene among essentials was also noted in other transposon library analyses using sequencing [40].

One notable difference is that Sassetti et al. found *glcB* to be non-essential, however insertion pattern clearly indicate that this gene was unable to tolerate insertions in the libraries of mutants analyzed. GlcB encodes for malate synthase in *M. tuberculosis*, which was originally thought to be necessary only for growth on fatty-acids as part of a glyoxylate shunt [41], but has recently been shown to be essential on other carbon sources like dextrose [42]. A complete absence of transposon insertions in Rv1837c was also observed in the DeADMAn studies [40]. Our data suggests that GlcB is also essential for growth on glycerol (in liquid culture with minimal media), showing a significant run of non-insertions

(25 out of 27 - spanning 2078 nucleotides, p($Z_i = 1$)=1.0). It should be be noted that in the original TraSH data, GlcB had a hybridization ratio of 0.41, which is near the threshold for essentiality (< 0.20).

3 MODELING INDIVIDUAL INSERTION FREQUENCIES[*]

3.1 Introduction

One limitation of the Gumbel method introduced in Section 2 is that it assumes a global insertion (or non-insertion) frequency that is shared by all non-essential genes. While this assumption makes the equations manageable, it is unlikely to be true. In reality, losing the function of a gene (by disrupting its function with a transposon) is likely to lead to different fitness costs to the organism depending on the function being disrupted and the biological (metabolic) costs to the organism.

This variability in insertion probability is evident in libraries of transposon mutants. Figure 3.1 shows a histogram of the observed number of insertions within windows of 20 TA sites (gray bars), for a transposon mutant library of *M. tuberculosis* [17]. The resulting distribution of the number of insertions is more dispersed than what would be expected with a fitted binomial distribution (black line). This suggests that the insertion frequency is not constant, but instead varies depending on the genomic region being considered. Assuming an insertion probability that is globally constant will ignore this variability, and lead to less reliable predictions.

In this Section a new hierarchical model of essentiality is introduced which overcomes this limitation [26, 27]. This method utilizes a binomial likelihood to model the

[*] © 2014 IEEE. Reprinted, with permission, from DeJesus, M.A. and Ioerger, T.R., "Capturing uncertainty by modeling local transposon insertion frequencies improves discrimination of essential genes", IEEE Transactions on Computational Biology and Bioinformatics, May 2014.

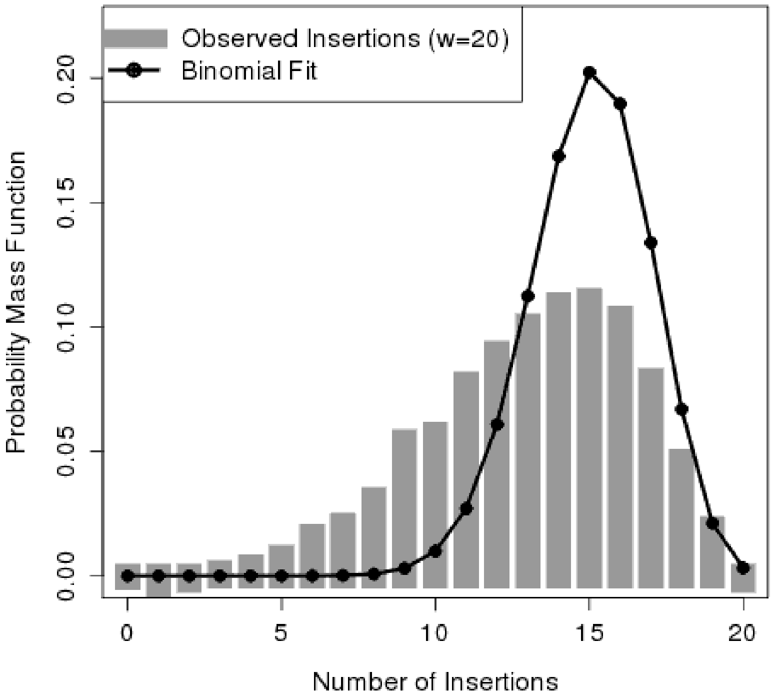


Figure 3.1: Histogram of the number of insertions within windows of 20 TA sites (gray bars). A beta-binomial model with a variable insertion frequency is capable of fitting the observed data (black line).

insertions within the genes. As with the Gumbel model, insertions are treated as Bernoulli events with a probability of success representing the insertion probability. This insertion probability, however, is allowed to be different for each gene.

The Bayesian framework on which these models is based on allows for a hierarchical extension by applying a prior distributions on the parameters of interest. This hierarchical approached improves the prediction of essential genes by taking into consideration the natural variability of insertion probabilities observed in the data as well as the length of the genes into account. Section 3.2 describes the model in detail, while Sections 3.3 and Section 3.4 briefly goes over the parameter estimation and results.

3.2 Hierarchical Model

For the all genes $i \in \{1...G\}$, let $Y_i = \{k_i, n_i\}$ represent the data for the i-th gene, consisting of the number of insertions, k_i, and the total number of TA sites, n_i. Each gene i contains a latent variable θ_i, which represents the insertion probability for this gene. The genes are modeled as a mixture of non-essential and essential genes, with an indicator variable, $Z_i = \{0, 1\}$, indicating whether the i-th gene belongs to the class of non-essential (0) or essential (1) genes. The mixture coefficient, ω_1, represents the probability of a gene belonging to the essential class (with the probability of belonging to the non-essential class $\omega_0 = 1 - \omega_1$).

3.2.1 Complete Data Likelihood

For each gene i, the likelihood of observing k_i insertions out of n_i TA sites is given by a binomial distribution with success probability θ_i. Assuming genes are independent of each other, the complete data likelihood is given by the product of binomial distributions over all the genes:

$$\prod_{i}^{G} \text{Binomial}(k_i|n_i, \theta_i) \tag{3.1}$$

3.2.2 Prior Probabilities

The distribution of individual insertion probabilities, θ_i is modeled by a mixture of two Beta distributions: one modeling the probability of insertion for "essential" genes, and another modeling the insertion probability at non-essential genes:

$$\begin{aligned} \theta_i|Z_i = 0 &\sim \text{Beta}(\kappa_0\rho_0,\ \kappa_0(1-\rho_0)) \\ \theta_i|Z_i = 1 &\sim \text{Beta}(\kappa_1\rho_1,\ \kappa_1(1-\rho_1)) \end{aligned} \tag{3.2}$$

Under this parameterization (i.e. $\alpha = \kappa\rho$ and $\beta = \kappa(1-\rho)$), the ρ parameter represents the mean insertion probability (i.e. mean of the distribution). On the other hand, the κ parameter can be thought of as the number of observations. This is because in the common parameterization the sum $\alpha+\beta$ can represent the number of Bernoulli trials depending on the application. Under this parameterization $\alpha+\beta = \kappa\rho+\kappa(1-\rho) = \kappa$. Thus, with larger values of κ the distribution becomes tighter around the mean (i.e. ρ).

Because the ρ parameters represent probabilities, requiring support for values in the range $[0,1]$, Beta distributions are chosen as priors:

$$\begin{aligned}\rho_0 &\sim \text{Beta}(\alpha_0, \beta_0) \\ \rho_1 &\sim \text{Beta}(\alpha_1, \beta_1)\end{aligned} \tag{3.3}$$

where α_0, β_0, α_1, and β_1 are hyper-parameters for the beta distribution.

As the κ parameters require support for values in the range $[0, \inf)$, gamma distributions are chosen as priors:

$$\begin{aligned}\kappa_0 &\sim \text{Gamma}(a_0, b_0) \\ \kappa_1 &\sim \text{Gamma}(a_1, b_1)\end{aligned} \tag{3.4}$$

where a_0, b_0, a_1, and b_1 are hyper-parameters describing the shape and and scale of the respective distributions.

The prior distribution for the indicator variable, Z_i, is given by the Bernoulli distribution, with probability of success ω_1, which represents the probability of a gene belonging to the class of essential genes:

$$Z_i \sim \text{Bernoulli}(\omega_1) \tag{3.5}$$

Finally, the prior distribution for ω_1 is given by a Beta distribution:

$$\omega_1 \sim \text{Beta}(\alpha_\omega, \beta_\omega) \tag{3.6}$$

3.3 Parameter Estimation

3.3.1 Conditional Distributions

Below, the conditional distributions for the parameters of the essential genes are given (the corresponding distributions for the non-essential parameters are defined in a similar manner). For an individual insertion probability, the conditional distribution is a beta distribution with updated parameters:

$$p(\theta_i|k_i,\kappa,\rho,Z_i=1) \propto \text{Beta}(\theta_i|\kappa_1\rho_1+k_i,\ \kappa_1(1-\rho_1)+n_i-k_i)$$

The beta distributions depend on parameters ρ_1 and κ_1 which are distributed as follows:

$$p(\kappa_1|k_i,\theta_i,\rho_1,Z_i=1) \tag{3.7}$$

$$\propto \text{Beta}(\theta_i|\kappa_1\rho_1,\ \kappa_1(1-\rho_1)) \times \text{Gamma}(\kappa|a_1,b_1) \tag{3.8}$$

$$p(\rho_1|k_i,\theta_i,\kappa_1,Z_i=1) \tag{3.9}$$

$$\propto \text{Beta}(\theta_i|\kappa_1\rho_1,\ \kappa_1(1-\rho_1)) \times \text{Beta}(\rho_1|\alpha_1,\beta_1) \tag{3.10}$$

Finally, the individual indicator variable, Z_i, is given by a Bernoulli distribution:

$$p(Z_i=1|k_i,\theta_i,\kappa_1,\rho_1,\omega_1) = \text{Bernoulli}\left(\frac{p_1}{p_1+p_0}\right)$$

where,

$$p_1 = \mathrm{Beta}(\theta_i | \kappa_1 \rho_1 + k_i,\ \kappa_1 (1-\rho_1) + n_i - k_i) \times \omega_1$$

$$p_0 = \mathrm{Beta}(\theta_i | \kappa_0 \rho_0 + k_i,\ \kappa_0 (1-\rho_0) + n_i - k_i) \times (1-\omega_1)$$

3.3.2 Metropolis-Hastings

As with the Gumbel model, parameters are estimated using MCMC samples obtained through the Metropolis-Hastings algorithm. Because the binomial likelihood (3.1) and the beta priors (3.2) are conjugate, the resulting conditional distribution can be sampled from directly. This is a special case of the Metropolis-Hastings algorithm called the Gibbs Sampler, where the proposal density is always accepted, and thus the MH ratio will never be rejected.

However, this is not the case for the conditional distributions of the ρ and κ parameters (Equations (3.10) and (3.8)), therefore the Metropolis Hastings algorithm is necessary to sample from these (non-standard) distributions. A combination of Gibbs Steps and MH steps can be used obtain samples for all the parameters (See Algorithm 2).

3.4 Results

Our method was applied to deep-sequencing data from mutant libraries of the H37Rv strain of *M. tuberculosis* [17, 25]. The library was grown in minimal media and 0.1% glycerol. The surviving mutants were sequenced with an Illumina GAII sequencer, with a read length of 36 bp, producing 6 to 8 million reads. These reads were mapped to the

```
Algorithm: Random-Walk Metropolis-Hastings
Result: MCMC Samples of the densities $p(Z_i|Y,\Theta,\rho,\kappa)$ and $p(\theta_i|Y,\rho,\kappa)$ for
        $i \in \{1...G\}$
Assign starting values to $\theta_i, \rho_0, \kappa_0, \rho_1, \kappa_1$ and initialize $Z_i$ based on proportion of
  insertions within individual genes.
for j=1 to desired sample size do
    //Gibbs Steps - $\theta_i$
    for $i \leftarrow 1$ to $G$ do
        Sample $\theta_i \sim$ Beta($\rho\kappa + k_i, \kappa(1-\rho) + n_i - k_i$)
    end

    //MH Step - $\rho_0$
    Draw candidate parameter $\rho_0^c$ from Normal distribution, N($\rho_0^{j-1}$, $v$) and
      accept according to MH ratio $\frac{f(\rho_0^c)}{f(\rho_0^{i-1})}$

    //MH Step - $\kappa_0$
    Draw candidate parameter $\kappa_0^c$ from Normal distribution, N($\kappa_0^{j-1}$, $v$) and
      accept according to MH ratio $\frac{f(\kappa_0^c)}{f(\kappa_0^{i-1})}$

    //MH Step - $\rho_1$
    Draw candidate parameter $\rho_1^c$ from Normal distribution, N($\rho_1^{j-1}$, $v$) and
      accept according to MH ratio $\frac{f(\rho_1^c)}{f(\rho_1^{i-1})}$

    //MH Step - $\kappa_1$
    Draw candidate parameter $\kappa_1^c$ from Normal distribution, N($\kappa_1^{j-1}$, $v$) and
      accept according to MH ratio $\frac{f(\kappa_1^c)}{f(\kappa_1^{i-1})}$

    Let $K_z$ equal the number of genes with $Z_i^j = 1$
    Let $G$ be the total number of genes
    Sample $\omega_1^{(j)} \sim Beta(\alpha_w + K_z, \beta_w + G - K_z)$
    //Gibbs Steps - $Z_i$
    for $i \leftarrow 1$ to $G$ do
        $p_1 = p(k_i|Z_i = 0, \rho_1, \kappa_1) \times \omega_1$
        $p_0 = p(k_i|Z_i = 0, \rho_0, \kappa_0) \times (1 - \omega_1)$
        Sample $Z_i^{(j)} \sim Bernoulli(\frac{p_1}{p_1 + p_0})$
    end
end
```

Algorithm 2: Random-Walk Metropolis-Hastings Algorithm for Sampling values of θ_i and Z_i for all genes i

H37Rv genome, producing read counts at each TA site in the genome.

The H37Rv genome is 4.41 million bp long and contains 3,989 open-reading frames (ORFs) [35]. Of these ORFs, 3947 contain at least 1 TA site, with an average of 15.9 TA sites per ORF. The remaining 42 ORFs, which do not contain a TA site, were not considered in this analysis as their essentiality cannot be determined with libraries built with the Himar1 transposon.

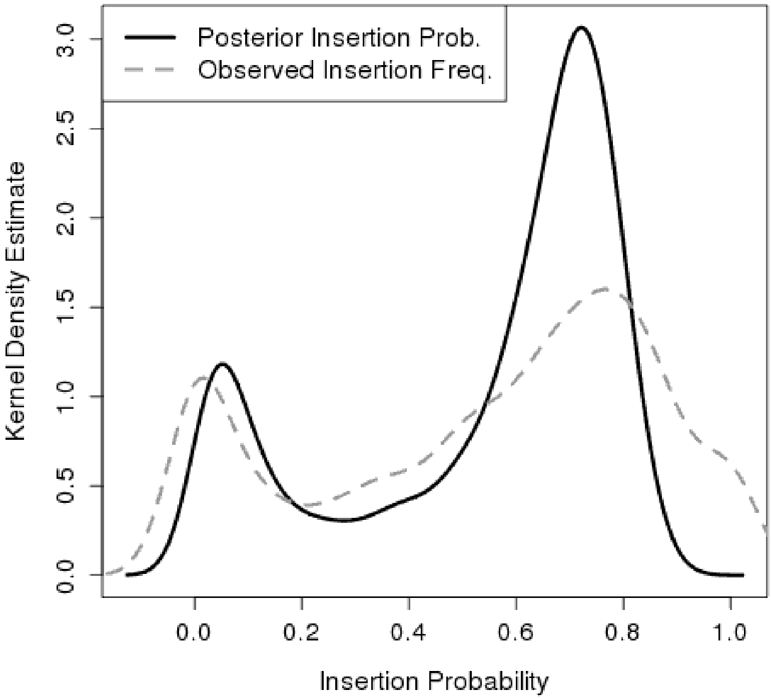


Figure 3.2: Kernel density estimates for the mean posterior insertion probability (black-solid) and observed insertion frequency (gray-dashed) for all the genes.

A sample of 52,000 values was obtained with the independent Metropolis Hastings algorithm. In order to make sure that the MCMC chain converged before parameters were estimated, the first 2,000 samples were discarded as part of the burn-in period. The remaining 50,000 samples were used to estimate the posterior mean of the parameters of the model.

3.4.1 Insertion Frequencies

Samples of the individual probabilities were obtained for all genes. The mean insertion frequency, $\bar{\theta}_i$, was estimated from these samples. Figure 3.2 contains a density plot of the mean insertion probability (black-line). The plot shows two peaks ($\theta = 0.052$ and $\theta = 0.721$) corresponding to the mixture of essential and non-essential genes. For comparison, the insertion frequency observed in the data (i.e. $\frac{k_i}{n_i}$) is plotted as well (gray dashed line). The mean insertion probability resembles the observed frequency, with sharper peaks at the posterior modes.

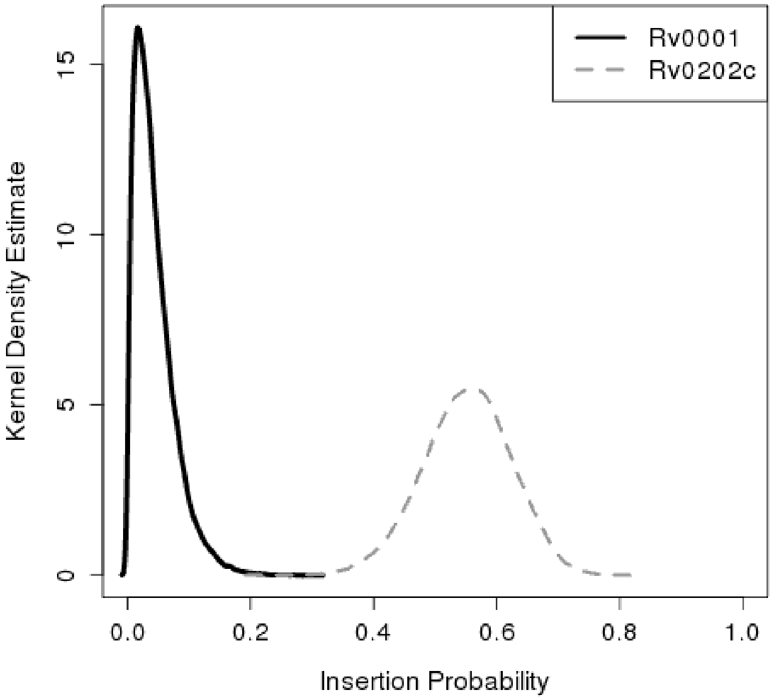


Figure 3.3: Kernel density estimates for the posterior insertion probability of DnaA (Rv0001), a known essential gene involved in DNA repair, and MmpL11 (Rv0202c), a known non-essential gene believed to function as a transmembrane protein.

The samples of insertion probability for the genes reflect our expectations for essential and non-essential genes. Figure 3.3 shows density plots of the samples for DnaA (Rv0001) and MmpL11 (Rv0202c). DnaA is a known essential gene involved in DNA repair. It contains a total of 32 TA sites with a single insertion in the C-terminus. Its mean insertion probability is $\bar{\theta}_i = 0.044$, corresponding to the small probability of observing an insertion in this essential gene. On the other hand, MmpL11 is a transmembrane transport protein determined to be non-essential in knock-out experiments [43]. It contains insertions in 20 out of 39 TA sites, with a mean insertion probability of $\bar{\theta}_i = 0.551$, consistent with expectations of non-essential genes.

3.4.2 Essentiality Results

To estimate the probability of a gene being essential, the sample of individual essentiality values, Z_i, was averaged for all genes ($\bar{Z}_i = \frac{1}{n}\sum Z_i$). A method analogous to the Benjamini-Hochberg procedure for posterior probabilities was used to obtain the thresholds of essentiality [44]. Setting the False Discovery Rate at 5%, genes with $\bar{Z}_i > 0.99304$ are classified as essential, and genes with $\bar{Z}_i < 0.0391$ are classified as non-essential. Those genes that do not meet these thresholds are classified as Uncertain.

Comparison to the TraSH Method

The essentiality of the M. tuberculosis genome has been assessed before, through the Transposon Site Hybridization method [11, 12]. This method quantifies the amount of luminescence that is observed in probes that hybridize to each of the genes in the genome [10]. Hybridization ratios were obtained from libraries of *M. tuberculosis* grown in rich media and glucose, and these where used to characterize genes as essential, non-essential or growth-defect (representing those genes which lead to reduced growth rate). Genes for which the hybridization ratio could not be obtained were classified as “No-Data”.

Table 3.1: Essentiality comparison between the TraSH method and the Local-Frequency Model.

		Local-Frequency Model			
		Essential	**Uncertain**	**Non-Essential**	**Total**
TraSH (Sassetti)	**Essentials**	**329**	257	28	614
	Growth-Defect	5	20	17	42
	Non-Essential	36	682	**1796**	2514
	No-Data	80	412	285	777
	Total	450	1371	2126	3947

Table 3.1 shows a breakdown of the results from the TraSH method and the local-frequency model. Of the 614 genes predicted to be essential by TraSH, 28 are predicted to be non-essential by the local-frequency model. Although these genes are predicted to be essential by the TraSH experiments, they contained a large number of insertions in the library analyzed (average $\theta_i = 0.72$). This high insertion frequency suggests the discrepancy could be due to differences in the growth media between the two libraries.

In addition to these 28 genes, the methods disagree on 36 other genes which are classified as essential by the local frequency model and Non-Essential by TraSH. Similarly, these genes contain a small number of insertions (average $\theta = 0.03$) in the library, which suggests that these genes are essential in the library analyzed, and the discrepancy may be due to the difference in the construction of the libraries.

Comparison to the Global Frequency Model

To determine the effect of relaxing the assumption of a constant insertion frequency, we compare our results to a binomial model with global insertion frequencies. Two “global” insertion frequencies, θ_0 and θ_1, are shared across the genes belonging to a given class of essentiality (i.e. essential and non-essential genes). Using Gibbs sampling, samples for the parameters θ_0 and θ_1 are obtained, as well as the essentiality assignments Zi. Estimates of the probability of essentiality are calculated by averaging the samples, as in the individual-frequency model. After running the Gibbs sampling procedure for 52,000 iterations, estimates for the parameters were as follows: $\bar{\theta}_0 = 0.684 \pm 0.002$ and $\bar{\theta}_1 = 0.102 \pm 0.002$, implying a 68.4% insertion density in non-essential genes and 10.2% in essential genes.

Table 3.2: Essentiality comparison between the Global Frequency Model (GFM) and the Local Frequency Model

		Local-Frequency Model			
		Essential	**Uncertain**	**Non-Essential**	**Total**
GFM	**Essentials**	**450**	259	0	709
	Uncertain	0	**603**	0	603
	Non-Essential	0	509	**2126**	2635
	Total	450	1371	2126	3947

Table 3.2 compares the results from the individual-frequency and global-frequency models. Overall, the local-frequency model is more conservative than the global-frequency model, predicting more uncertain genes (1,371 vs 603). In fact, all 709 genes classified as essential by the global-frequency model are classified as either essential (450) or uncer-

tain (259) in the local frequency model. The same is true for non-essential genes, where the global-frequency model predicts 2,635 non-essential genes, while the local-frequency model predicts 2,126 of these to be essential and classifies the rest (509) as uncertain.

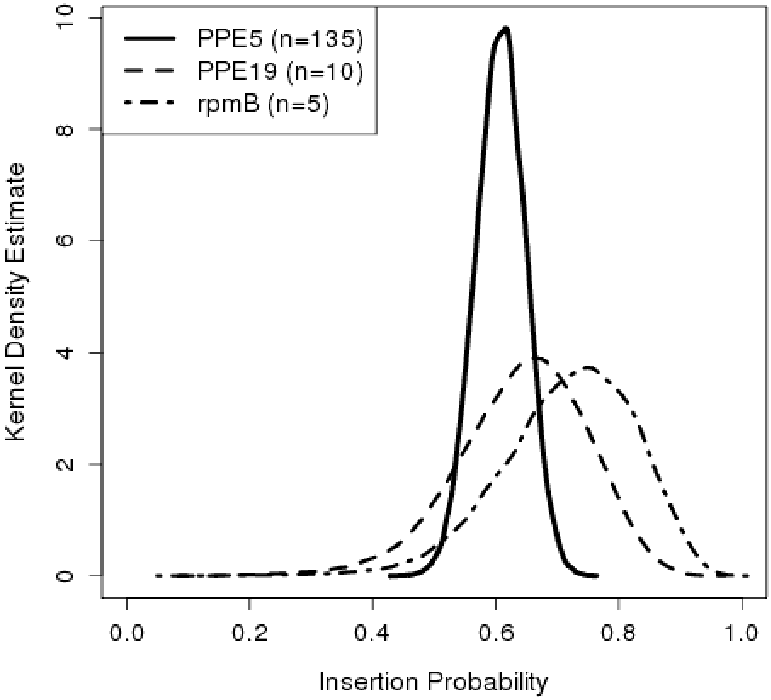


Figure 3.4: Insertion density for PPE5 (solid), PPE19 (dashed) and RpmB (dot-dash). All three genes contained an observed insertion frequency of 0.7, although they had different sizes (# TA sites). The insertion density of the genes reflects the uncertainty that exists in smaller genes as these contain a smaller number of TA sites (Bernoulli trials).

This tendency to be more conservative in its predictions is due to the fact that the local-frequency model is able to capture the uncertainty there is with smaller genes. By sampling from a beta-binomial model, the lower number of TA sites (i.e. Bernoulli trials) leads to an increased variance. Figure 3.4 shows a density plot of the sampled insertion density for PPE5, PPE19, and RpmB. All these genes have an observed insertion density of 0.7 (i.e. $\frac{k_i}{n_i} = 0.7$), however they have different number of TA sites (PPE5=135, PPE19=10, and RpmB=5). While the global-frequency model classifies all these genes as non-essential, the local-frequency model classifies RpmB as Uncertain because it takes into account the increased uncertainty due to the smaller number of TA sites. The "shifting" of the mode of these distributions is due to the fact that smaller genes will regress towards the mean of the distribution of non-essential insertion frequencies (i.e. $\bar{\rho}_0 = 0.69$) as there are more strongly affected by this parameter.

4 ANALYZING SEQUENTIAL READ-COUNTS THROUGHOUT THE GENOME[*]

4.1 Introduction

While the analysis methods in Section 2 and 3 focused on determining the essentiality of specific genes, these methods may miss other important genomic elements that occur outside of gene boundaries (e.g. like promoters which occur upstream of the gene start sites). In addition, these methods were based on a Bernoulli interpretation of the data which ignores the magnitude of the read-counts. Valuable information about essentiality may be present in the magnitude of read-counts, as those mutants which suffered not fitness cost (or perhaps saw fitness improvement) will grow faster than those which suffer a cost to their fitness because an important function was disrupted by the transposon insertion.

This Section describes a novel method for analyzing Tn-Seq data using Hidden Markov Models (HMMs) [28]. HMMs are useful for analyzing sequential datasets, in which a sequence of observed values is explained by an underlying state sequence (i.e. "essentiality" of each site, which is not directly observed). For example, the genome of an organism can be viewed as an alternating sequence of essential and non-essential regions. An HMM can be designed to incorporate information from read counts at individual TA sites to infer the probability distribution over states, and then use the Viterbi algorithm to

[*] Reprinted with permission from "A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data" by DeJesus, M.A. and Ioerger, T.R., 2013. BMC Bioinformatics, 14:303, Copyright [2013] by BioMed Central.

infer the most likely state sequence. The sequential-dependence of the model (conditional probability of a state conditioned on the previous neighboring site) helps disambiguate the interpretation of each site, thereby coupling neighboring sites together. The resulting state transition model affords a ‘smoothing’ of the read-count data, where, for example, TA sites with no insertions in non-essential regions (e.g. because they are absent from the library) are tolerated because neighboring sites have insertions. However, if a consecutive sequence of TA sites with no insertions is long enough, the most probable state sequence, as determined by the Viterbi algorithm, switches locally to essential, providing a different labeling of that region.

The incorporation of read-counts in this HMM requires defining appropriate likelihood functions. This model utilizes the the geometric distribution to represent the distribution probability of read-counts in non-essential regions, reflecting the fact that sites with high read counts (far above average) are observed with much lower frequency than those with lower read counts. Furthermore, the transition probabilities of the HMM must be carefully defined so that the minimum length of essential regions matches our expectations. In ad A major contribution of this paper is to show how to calibrate these parameters so that the performance of the HMM will be reasonable and robust across a range of datasets, including those with high or low insertion density (a function of the diversity of insertion library), and those with high or low mean read counts (a function of how much sequencing data is collected).

In addition, we extend the HMM with two extra states, one representing regions with particularly "low" read counts, and one representing regions with higher than average read-counts. Genes belonging to the former class of genes have been characterized before in *M. tuberculosis* and referred to as "growth-defect" genes [11], as these are genes whose disruption leads to impaired growth of the organism. We continue this convention here, labeling those genes with depressed read-counts as "growth-defect" (despite the fact that these genes code for proteins whose normal function contribute to growth) to be consistent with the prior literature. Growth-defect regions are not completely devoid of insertions (as essential regions would be), but have a lower number of insertions than non-essential regions (on average), suggesting that these clones did not grow as well and had lower abundance due to competition with other clones in the library.

Similarly, the latter class of genes (i.e. those with higher than average read-counts) are labeled "growth-advantage" genes. These could represent genes that have a metabolic cost (e.g. biosynthesis of a secreted toxin) and are not necessary for growth in vitro. The addition of these two states to our HMM allows it to distinguish regions in Tn-Seq data with suppressed or unusually high read counts in a statistically rigorous way.

The HMM in this application is defined in a straightforward way (see Rabiner for details [45]). We are given a sequence of observations, $c_1..c_n$, which represent read counts at each TA site throughout the genome. We assume a generative model in which the read count at each site is determined by the local state of each site, which is hidden (i.e. not directly observable). Each TA site is assumed to be in one of four states: q_{ES} (essential), q_{GD} (growth-defect), q_{NE} (non-essential), q_{GA} (growth-advantage) (See Figure 4.1).

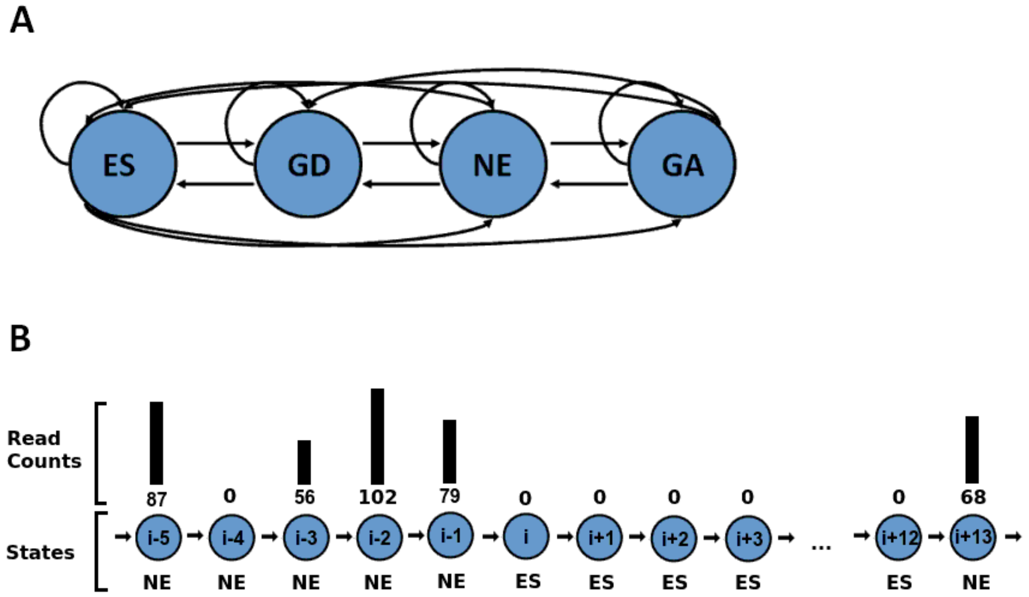


Figure 4.1: (A) Diagram of the fully-connected HMM structure. From left to right, the states represent read counts of increasing magnitude (essential, growth-defects, non-essential, and growth-advantage). (B) Diagram of the states for a local sequence of ∼20 TA sites, with state labels (underneath), transitions (from q_{i-5} to q_{i+13}) and their corresponding emissions (i.e. read counts). A transition is shown from the non-essential state to the essential state at time i, as the essential state is most likely to explain the consecutive observations of no insertions (from q_i to q_{i+13})

4.2 Model

From a given sequence of observations (read counts), we want to infer the most probable state sequence $q_1..q_n$ that could have generated it, based on the joint probability of counts and states:

$$\arg\max_{qi..qn} p(q_1...q_n, c_1...c_n) \tag{4.1}$$

HMMs are based on the Markov property, i.e. that observations and successor states only depend on the current state and are conditionally independent of previous history:

$$p(c_i|q_1,...,q_i) = p(c_i|q_i) \tag{4.2}$$

$$p(q_{i+1}|q_1,...,q_n,c_1,...,c_n) = p(q_{i+1}|q_i,c_{i+1}) \tag{4.3}$$

Thus, because of this conditional independence, the total joint probability can be written as:

$$p(q_1,...,q_n,c_1,...,c_n) = p(q_1)\prod p(q_{i+1}|q_i,c_i)p(c_i|q_i) \tag{4.4}$$

The model we propose depends critically on specifying an appropriate likelihood function for read counts. In Tn-Seq experiments, the distribution of read counts can be approximated through a geometric distribution, in that sites with lower counts are more common, and sites with high counts (far above average) are much more rare. An example histogram in shown in Figure 4.2 (taken from an *M. tuberculosis* H37Rv dataset [17]).

Thus we model the likelihood function (i.e. emission probability) for q_{NE} as geometric:

$$p(c_i|q_{NE};\theta) \sim (1-\theta)^{c_i}\theta \tag{4.5}$$

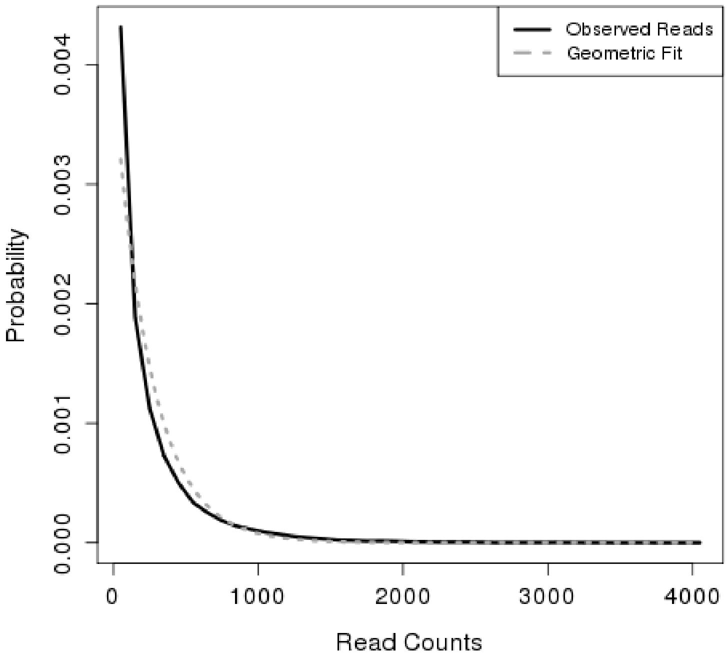


Figure 4.2: Histogram of read-counts for a library of *M. tuberculosis* transposon mutants (black, solid vertical lines), fitted with a geometric distribution with parameter $\theta = 1/\bar{c}$ (dashed line).

The function is parameterized by θ, which represents the Bernoulli probability of insertion for the geometric distribution. The maximum-likelihood estimate for this parameter is $\theta = 1/\bar{c}$, where $\bar{c}$ is the mean read count at non empty TA sites.

We also use geometric distributions as likelihood functions for the other states. For q_{ES}, we set θ very near to 1 (e.g. 0.99), making sites with 0 counts highly probable, but also allowing sites with 1-2 reads (which could be spurious reads due to base call errors). For q_{GD} we set θ to be $\theta_{GD} = 1/(0.01 \times \bar{c} + 2)$ (where $\bar{c}$ represents the mean), reflecting the fact that the growth-defect state must represent approximately $\sim 100\times$ lower read counts than q_{NE} but cannot be less than 1 (converges to 2, in the limit, for very low coverage datasets). For the growth-advantage state, q_{GA}, we set θ using five times the mean read count (i.e. $\theta_{GA} = \frac{1}{5\bar{c}}$), to capture sites with significantly more insertions ($> 5\times$) locally than what is observed on average in the genome. The net effect is that the overlapping densities of the four likelihood functions produce four distinct regions where each one dominates individually, as shown in Figure 4.3.

Another critical aspect of our model is the definition of the state transition probabilities, as these determine the degree of smoothing of the HMM. Let the transition matrix be defined as $T_{ab} = p(q_{i+1} = b | q_i = a)$. The basic assumption is that the probability of self-transition, T_{aa}, should be nearly 1 for all states, while T_{ab} should be nearly 0 for $a \neq b$ (off-diagonal elements in the T matrix). This assumption controls the rate at which the HMM transitions from state to state, requiring a significant change in read-counts to justify a transition and smoothing over spurious reads. For simplicity, we use a fully symmetric

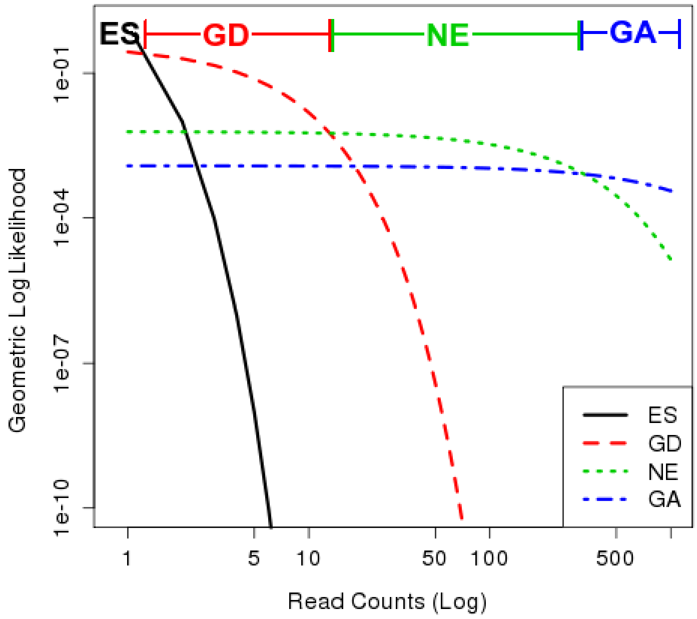


Figure 4.3: Log-log plot of geometric likelihood functions for the essential, growth-defect, non-essential and growth-advantage states.

matrix, and we allow any state to transition to any other state (i.e. we do not force sites to progress in a sequence, such as $q_{ES} \rightarrow q_{GD} \rightarrow q_{NE}$). The magnitude of T_{aa} determines the tendency of the model to stay in one state for a certain number of steps before being forced into another state that better fits the data. This depends on several factors, including: a) the expected minimum length of essential regions (number of TA sites), and b) the relative magnitudes of the likelihood functions, which are competing to explain the read counts.

To estimate the expected minimum length of essential regions, we utilize the geometric distribution. The geometric distribution describes the probability of observing a run of successes in a row, which can be used to model the distribution of run lengths. This depends on the insertion probability in non-essential regions. Because the insertion density of the library will include essential regions with insertion probabilities which are not representative of non-essential regions. To alleviate this bias, we estimate the insertion probability, p_{ins}, empirically by discarding regions with 10 or more TA sites in a row lacking insertions, and calculating the insertion density in the remaining areas. Once the insertion probability is estimated, the minimum length of essential regions, r^*, is taken to be the smallest run such that the geometric probability is less than 0.01 (i.e. $r^* = \text{argmin}\, P(r|1-p_{ins}) < 0.01$). Typically r^* is in the range of 5-10 TA sites, depending on the dataset. The self-transition probability is then set as follows:

$$T_{aa} = 1 - (\lambda_{NE}(0))^{r*}$$

where $\lambda_{NE}(0)$ represents the likelihood of observing a read-count of zero in a non-essential region. The rationale for this formula is that the cost of staying in a state such as q_{NE} through a region devoid of insertions, must balance the penalty incurred for observing sites with 0 read counts ($\lambda_{NE}(0)$) and the number of such TA sites in a row which are likely to be observed in non-essential regions (r^*).

We will show empirically in the Results section that this adaptive method for setting the transition probabilities leads to an appropriate assignment of state labels for a variety of types of datasets, and we will examine the resulting length distribution of states produced.

Finally, given this definition of the HMM, we use the Viterbi algorithm to calculate the most probable state sequence for a given set of read counts [45]. Briefly, the Viterbi algorithm is a dynamic programming algorithm in which the probability of each state at step i is calculated based on the state-probability distribution from the previous step:

$$p(q_i = a) = \max p(q_{i-1}) \times p(q_i|q_{i-1}) \times p(c_i|q_i) \tag{4.6}$$

After computing this incrementally for $i = 1..N$, a back-trace is made from the most probable terminal state q_n^* to extract the sequence of states based on which states were used for updates at each step. Because the Viterbi algorithm requires the multiplication of small probabilities, and the state sequence for analyzing transposon insertions is large, an HMM may incur underflow problems. To overcome this issue, the probabilities are normalized at each iteration, as described by Rabiner et al [45].

4.3 Results

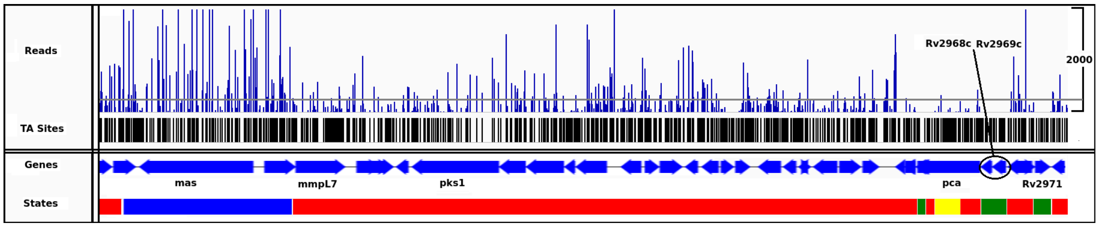


Figure 4.4: Read counts and state classifications for a ∼57 kb region of the H37Rv genome is shown. Essential regions are shown in green, growth-defect regions in yellow, non-essential regions in red, and growth-advantage regions in blue. Read counts are truncated at 2,000 (with a max of ∼3,000 in this region), and the mean read count in the library is represented by a gray horizontal line.

The HMM method was applied to a transposon mutant library of *M. tuberculosis*, constructed by Griffin et al [17]. This library was grown on minimal media and 0.1% glycerol, and was sequenced on an Illumina GAII sequencer with a 36 bp read length, resulting in approximately 6 million reads. The reads were mapped to the H37Rv genome, and the read counts at each location in the genome were quantified (i.e. $c_1..c_N$). The H37Rv genome is 4,411,532 bp in length, with a GC-content of 65.6%. It contains a total of 74,605 TA sites, spaced on average 59 bp apart. The overall insertion density, defined as TA sites with at least one insertion ($c_i \geq 1$), is 54.18% (39,762) of all possible insertion sites. The average read-count at these locations is $\bar{c} = 195$ (discarding the top 5% for robustness).

The mean read count was used to calculate the θ parameter for the emission probabilities of the four states as described above. Using these parameters, the most likely sequence of states responsible for the observations was obtained through the Viterbi algorithm. This sequential ordering of states provides an assessment of the essentiality of the entire H37Rv genome, regardless of gene boundaries.

A total of 16.6% of the genome is labeled by the essential state (q_{ES}). This is close to the expectations for bacterial organisms, where roughly 10%-15% of the genome is considered to be essential [46]. The majority of sites are labeled non-essential (78%), with a small percentage of sites labeled as growth-defect and growth-advantage (4.1% and 1.3%). Essential states averaged a very small number of insertions and read counts (0.006 and 0.2 respectively), demonstrating that the HMM is associating the essential state with stretches devoid of insertions, though these locations can occasionally contain insertions with a very small number of reads so long as as the observations at neighboring sites are consistent with essentiality. In contrast non-essential regions have a mean insertion density of 70%, and mean read counts of 220 in this dataset. Growth-defect regions have some insertions but these are dramatically reduced (20% density and a 10-fold reduction in mean read counts). Insertion density in growth-advantage regions is almost saturated (90%), and mean read counts are on average $> 3\times$ larger.

Figure 4.4 shows the read counts and state labels observed in a representative $\sim$57 kb region of the genome. Genes are shown as blue arrows, and the corresponding state classifications are shown at the bottom of the figure. As evident from this figure, the HMM takes into consideration the fluctuation in read counts observed. Regions devoid of insertions are classified as essential (green), those with read-counts close to the average in the library are classified as non-essential (red), while those regions with lower and higher read counts than average are classified as growth-defects (yellow) and growth-advantage (blue) respectively. Notice that *mas* (mycocerosic acid synthase, which is involved in PDIM biosynthesis) has much higher read counts than the average, and is therefore identified as a growth-advantage region. A long region of the genome is identified as non-essential as it contains read-counts that are closer to the average, despite occasional large spikes in the read-counts. This region includes *mmpL7*, which matches the expectations that most genes in the MmpL family are non-essential in vitro [43].

4.3.1 Analysis of Essentiality of Individual Genes

While the Viterbi algorithm does not take into consideration gene boundaries when determining the labeling of states, it is often necessary to determine the essentiality of individual genes in the genome. To determine individual calls of essentiality, each gene is assigned the essentiality class belonging to the most frequent state found within its boundaries. However, because genes may contain a mixture of essential and non-essential domains, genes are also classified as essential if they contain sub-sequences of sites belonging to the q_{ES} state, which are statistically longer than expected. Thus a gene is also classified as essential if it has at least n sites labeled as q_{ES}, where n is 3σ above the expected maximum run length for the gene, based on the Extreme Value Distribution [17].

The essentiality assignments obtained through the HMM method can be validated by comparing to those obtained by Sassetti et al with the Transposon Site Hybridization (TraSH) method [10], which used a completely different experimental methodology for read-out (hybridization versus sequencing). This method has been used to assess the essentiality of *M. tuberculosis* in vivo and in vitro [11, 12], by quantifying hybridization to DNA microarrays imprinted with representative oligos for each gene. Due to the significantly different methodologies, a true comparison between these methods is difficult. For instance, Sassetti et al. recognized that TraSH probes for essential genes may actually hybridize to adjacent non-essential regions, particularly if the genes are small. While the HMM does not depend on hybridization, it may have a difficulty transitioning from one state to another depending on the size of the gene. In addition, libraries used by these methods were grown on different media and therefore are likely to identify genes that are involved in pathways that correspond to the specific growth media used.

Despite these limitations, there is significant agreement in their assessment of essentiality, with 89.9% of essential and non-essential genes in concordance with the previous results (70% concordance between essential genes, and 95% among non-essential genes). Approximately half of the genes labeled as ‘growth-defect’ by the HMM were previously determined to be essentials, and half as non-essentials, reflecting the borderline nature of these genes and the utility of having an intermediate category. These are discussed further below. 27 genes were called ‘growth-advantaged’ due to an excess of transposon insertions, and all of these were previously categorized as non-essentials.

Sassetti et al [11] also defined a set of 42 ‘growth-defect’ genes. Importantly, these were not characterized by experimentally determining growth rates in individual transposon-insertion mutants. Rather, they were identified as genes that matched the criterion for ‘non-essential’ on the first plating of the library (hybridization ratio > 0.4, range: 0.41-2.04), but which had much lower ratios upon re-plating (hybridization ratio < 0.2, thus matching the criterion for ‘essential’). The interpretation of these genes is that transposon insertions were not lethal, but that the mutants had a slower growth rate, resulting in gradual depletion in the library due to competition during culturing. In the experiment from which the dataset we use was derived [17], the DNA for sequencing was extracted from the library immediately after selection, thus corresponding to the ‘first plating’. Consistent with this, most of these genes (29/42) exhibited transposon insertions in our dataset and were categorized by the HMM as non-essential. We speculate that, if the library had been expanded after selection, clones with insertions in these genes would have gradually decreased in abundance.

Although the methods disagree on essentiality of some genes, some of these disagreements may be due to differences in the growth media, as well as the different interpretations of essentiality. For example *glpK*, a glycerol kinase, is necessary for glycerol metabolism (and therefore essential when grown on glycerol), but it is not necessary when the library is grown on glucose (as in the original TraSH experiment). In addition, these differences can also be due to the fact that we identify genes containing essential domains as ”essential”, while this distinction was not made in the original TraSH experiments. In fact, all of the genes classified as essential by the HMM and as non-essential by the TraSH method are devoid of insertions in the majority of their TA sites or contain stretches that are significantly longer than expected, suggesting these genes are essential in this library on glycerol. Among these genes are *ppm1* (Rv2051c) and *ppp* (Rv0018c), which independent experiments have shown contain essential domains [37, 36].

In addition to the TraSH method, we compare our results to those obtained with the reads-based method developed by Zhang et al [23]. This method is capable of assessing the essentiality of the entire genome by looking at the read counts that fall within windows of 400-600 bp, and estimating a p-value for each of these windows in the genome to quantify how these regions deviate from expectations. Our results correlate well with the results obtained by window-based method, with a 93.72% match in the classification of genes (i.e. essential and growth-defects genes, as determined by our HMM, matching essential and domain-essential genes determined by the window-based method, and non-essential and growth-advantage genes matching non-essential genes). In addition, the essential and growth-defect states had TA sites with an average p-value of 0.049, and non-essential and growth-advantage states an average p-value of 0.538 (as determined by the window-based method).

4.3.2 Performance on Other Datasets

To demonstrate that the HMM works on other datasets, we ran it on a Tn-Seq dataset from *H. influenza* (in vitro dataset SD2, [15]). The *H. influenza* KW20 genome is less than half the size of *M. tuberculosis* (1,830,138 bp, 1724 genes) but significantly more AT-rich (GC content = 38%), so there are more TA sites (131,960) but they are spaced more closely ($\sim$14 bp apart). The Tn-Seq dataset contains 736,631 reads, hitting only 37.9% of the TA sites, with a mean read count of 11.2 (per non-zero site). Running the HMM on this lower-density dataset results in 372 genes being labeled as essential, 1150 as non-essential, 211 as growth-defect, and 6 as growth-advantage. This distribution is very close to the assignments determined by Gawronski et al. [15], who found 363 essentials (with insertions in $<$5% of TA sites in the 5-80% region of the ORF), and 211 growth-defect genes (with insertion frequencies of 5-40%). The overlap (intersection) between the essential genes detected by both their method and ours was 94% (341 genes), and the intersection between their list of growth-defect genes and ours was 60% (127).

The overlap between essential genes found by the HMM method and those found by Gawronski et al. significantly larger than the overlap between the TraSH method described above (i.e. 94% vs. 70%). This high level of agreement between the two comparisons suggests that the quality of the data used in the analysis (i.e. high-resolution sequencing data vs. hybridization ratios) contributes significantly to the quality of the analysis.

In addition, we applied the HMM method to three modified datasets, constructed to represent libraries of different sizes and different volumes of sequencing data. These datasets were constructed by modifying the original H37Rv library analyzed before, to emulate cases where transposon mutant libraries may be sparse or where the amount of sequencing performed on the library is lower (i.e. less reads).

The first dataset was constructed by setting the read counts at random TA sites to zero (i.e. $c_i = 0$), thus lowering the mean insertion density of the dataset while keeping the magnitude of the remaining read-counts the same. This dataset emulates libraries with significantly less diversity of insertions. The second dataset was constructed by randomly perturbing approximately one-half of the reads, lowering the magnitude of these reads while keeping the total number of insertions equal. This dataset represents libraries for which the amount of sequencing performed is significantly less, producing read counts with lower magnitudes. The final dataset was a combination of these two operations, resulting in a dataset with both lower insertion density and lower mean read count.

The HMM is robust, and capable of adapting to libraries with very different insertion densities and mean read counts, providing results which are generally consistent with each other. The fraction of the genome labeled as essential is approximately the same in all four datasets (approximately 15%). Although the decreased density will result in longer stretches of the genome without a transposon insertion, the HMM is capable of adapting its parameters to become more conservative in designating regions without insertions as essential.

4.3.3 Growth-Defect and Growth-Advantage Genes

One of the principle advantages of our 4-state HMM is that it can distinguish local regions of the genome with significantly depressed or elevated read counts (transposon insertions). The former could represent genes whose disruption is not lethal but could lead to a growth-defect, resulting in a lower representation of clones in the library, and thus a lower abundance of sequencing reads [47]. By analogy, regions with significantly greater than average reads could represent genes whose disruption leads to a growth advantage. In the H37Rv dataset, there were 140 genes labeled as q_{GD} (growth-defect), and 27 genes labeled as q_{GA} (growth-advantage). These are discussed in turn below.

Among the genes labeled as growth-defect, there are several notable ones for which a biological explanation can be made. One of these is *pbpA*, a penicillin-binding protein in Mtb. Mutants have shown decreased growth rates and defective cell septation when *pbpA* is knocked out *M. smegmatis* [48]. In addition, the wild-type phenotype was restored by complementing in *pbpA* from *M. tuberculosis*, suggesting that *pbpA* plays an important role in cell-division and disruption of this gene might lead to impaired growth in *M. tuberculosis*. In fact, this region contains an average insertion density of 0.21, and an average read-count of 32, significantly below the global insertion density (0.52) and read-counts (257).

Recent structural and enzymatic studies have shown that *bfrB* and its ortholog, *bfrA*, are not completely interchangeable. Although they are both ferritin proteins used for iron storage, *bfrB* has a 20-aa C-terminal extension that enhances its iron oxidation activity

[49]. Thus growth of *bfrB* mutants might be hindered because *bfrA* cannot perform this function as efficiently. In fact, data from the original TraSH experiments shows that *bfrB* had a much lower hybridization ratio (0.73) compared to *bfrA* (2.63), suggesting clones with insertions in *bfrB* were less competitive.

Many genes in the mycobactin biosynthesis cluster (*mbtA-J*) are also labeled as growth-defect genes, suggesting that transposon mutants are viable but grow more slowly than wild-type. Because Mtb has only one (non-heme) iron acquisition system, which is mycobactin-dependent, these biosynthetic genes are essential in iron-depleted environments and non-essential in those environments that are rich in iron. Indeed, it has been shown that mycobactin-deficient mutants of Mtb, the growth rate is dependent on the iron concentration [50]. In the original TraSH experiments (plated on 7H10 medium, $\sim 150\,\mu$M Fe), *mbtB* was specifically shown to be cause a slow-growth phenotype when disrupted, with insertion mutants gradually decreasing in abundance in the library with successive platings [11].

Another interesting growth-defect gene is *glpX*. *glpK* (glycerol kinase), which is the first step in glycerol incorporation, is essential as expected (recall that this H37Rv dataset came from selection of the library on glycerol as a carbon source). *glpX* is a fructose-1,6-bisphosphatase, which also should be required when grown on gluconeogenic substrates by circumventing a non-reversible step in glycolysis pathway to generate glucose [51]. In Mtb the unexpected non-essentiality of *glpX* for growth on glycerol has been previously noted [52]. One possible explanation is that Rv2131c (*cysQ*), an inositol monophosphatase, might also have partial fructose-1,6-bisphosphatase activity [53].

icl (isocitrate lyase) is also identified as a growth-defect gene in this dataset. This is one of the two enzymes on the glyoxylate shunt, which has been shown to be critical for infection, based on attenuation of knockouts in mice [41]. As anticipated, *icl* is essential for growth on fatty-acid substrates like acetate [41]. However, recent evidence suggests that the glyoxylate shunt might play a role even in growth on other carbon sources such as carbohydrates. For instance, *icl* knockouts have displayed a growth-defect (2-4 day lag compared to wild-type) on glucose [54]. More recently, it has been shown that inhibitors of malate synthase (GlcB, the other enzyme of the glyxolate shunt) are active against cultures whether grown on acetate or glucose [42]. Thus, the fact that the HMM labels *icl* as a growth-defect region in this dataset obtained from growth on glycerol is consistent with these findings and suggests that *icl* plays an unexpected metabolic role in Mtb even when growing on carbon sources other than fatty acids.

Another gene identified as belonging to the growth-defect category is *treS*, which is involved in the trehalose pathway. Trehalose is one of the principle carbohydrates synthesized in mycobacteria. It is used in producing cell-wall glycolipid components (e.g. TMM and TDM, trehalose mono- and di-mycolates), and is inter-converted with other sugars like glucose and maltose. The latter are polymerized into intracellular glycogen (for energy storage) and capsular glucan. Several genes in this network have been shown to be essential in vitro, including *galU*, *glgA*, *glgB*, *pep2*, and *glgE* (all essential in our dataset). However, *treS* is labeled as a growth-defect gene. *treS* is responsible for interconverting trehalose and maltose [55, 56]. It is possible that the organism is sensitive to perturbations of this network (given the essentiality of nearby genes like *glgA*, and toxicity of intermediate metabolites like maltose-1-phosphate [57]). In fact, it was previously shown that transposon-insertion mutants of *treS*/Rv0126 display a slow-growth phenotype [11].

As noted before, our 4-state HMM is also capable of detecting regions with unexpectedly high read-counts that might confer growth-advantages to the organism when disrupted. One region of the genome that stands out is the PDIM locus, Rv2930-Rv2939. This locus contains genes involved in the biosynthesis of phthiocerol dimycocerosate (PDIM), including *fadD26* and *ppsABCDE*. In addition, other genes outside this locus believed to be involved in PDIM biosynthesis, like *papA5* and *mas*, are identified as well. These genes contain read counts well above the global average (~ 250). *fadD26* itself has a mean read count of 818, more than three times the average throughout the genome. *ppsDE* had a mean read count of 732, and *ppsABC* a mean read count of 463. PDIM is a cell-wall associated glycolipid that modulates the immune response in the host [58, 59]. Although it is required for virulence (as strains with disruptions of these genes are attenuated in animal models [60]), it is not required for survival in vitro [11, 17, 15]. In fact, biosynthesis of PDIM requires resources and imparts a metabolic cost, hence disruption of this pathway is advantageous to cells. Due to the increased metabolic cost, it is widely observed that *M. tuberculosis* stocks maintained in the lab frequently lose the ability to synthesize PDIM via acquisitions of mutations in these genes, often leading larger colony sizes [61]. This growth advantage and consequent selection effect likely explains why clones with transposon insertions in the PDIM locus are over-represented in the library.

5 IDENTIFYING CONDITIONALLY ESSENTIAL GENES: THE IMPORTANCE OF NORMALIZING READ COUNTS[*]

Because TnSeq datasets can come from different libraries, sample preparation protocols, and sequencing methodology, they often differ significantly in the amount of genomic material obtained or reads present. As such, proper normalization of read-counts is necessary when attempting to compare different datasets to avoid confusing differences in the libraries for different phenotypes. Several considerations need to be considered when properly normalizing read-counts. Sections 5.1 and 5.2 describe some of the most common issues as well as the normalization methods developed to address them [29, 30, 31].

5.1 Normalizing Insertion Density and Read-Counts

An obvious way of normalizing datasets is to make datasets share the same mean across non-zero sites (called here NZmean). The normalization is achieved by dividing by the total number of reads in the dataset by the total number of sites with at least one insertion, and using this as a scaling factor. While this normalization method works well when datasets differ primarily in the mean read-counts, it is susceptible when datasets have a significantly different insertion density or when there are outliers present.

Differences in the saturation of a genomic region is particularly problematic because a significantly lower number of insertions is often used as an indicator of essentiality, thus

[*] Electronic version of an article published as Journal of Bioinformatics and Computational Biology, 14(3), 2016. doi: 10.1142/S021972001642004X © World Scientific Publishing Company. http://www.worldscientific.com/worldscinet/jbcb

this could easily be mistaken for real biological differences. Libraries that are prepared with different protocols may lead to significantly different levels of saturation, but saturation may be different even among replicates from the same library and growth conditions. Figure 5.1 shows the insertion pattern of a PvdS in two replicates from a *M. tuberculosis* library. Although these are two replicates were exposed to the same condition (and thus should not exhibit any differences in essentiality) there is a significant difference in the frequency of insertions.

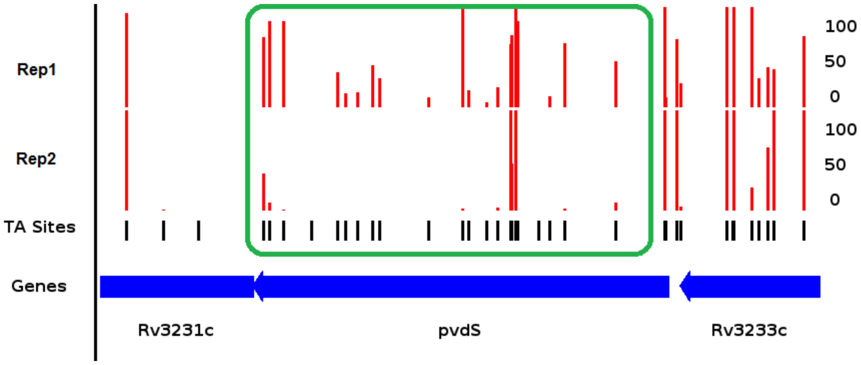


Figure 5.1: Top 100 reads from a *M. tuberculosis* TnSeq dataset. A large read-count with a magnitude $> 200,000$ is present in this dataset. This single site has a large impact on the mean read-count.

In addition to differences in density, it is also important to take outliers into consideration. As with many other biological datasets, TnSeq is noisy and can contain unusually large read-counts (Figure 5.2). Outliers such as those are likely due to problem in sequencing (like PCR amplification, which can lead to the abundance of some DNA fragments being amplified too severely). Nevertheless, normalization of TnSeq dataset must account for these artifacts so as to prevent estimates of the mean (or other relevant statistic used in normalization) from being affected by the outliers.

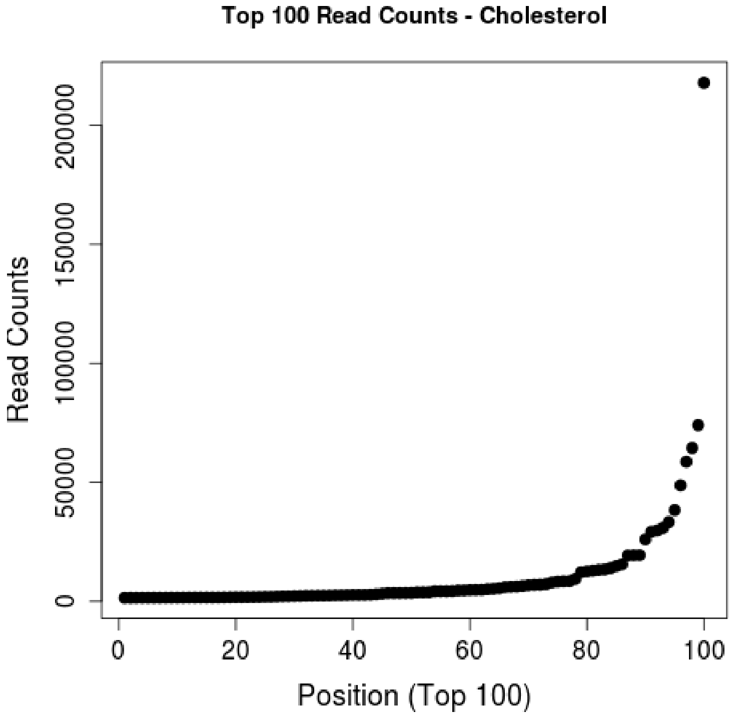


Figure 5.2: Top 100 reads from a *M. tuberculosis* TnSeq dataset. A large read-count with a magnitude $> 200,000$ is present in this dataset. This single site has a large impact on the mean read-count.

5.1.1 Trimmed Total Reads (TTR)

A simple normalization method that can address both of these issues is called Trimmed Total Reads (TTR). Like NZMean, TTR normalizes datasets so that the mean-read counts are equivalent. However, instead normalizing over non-zero counts, TTR normalizes datasets so that the average among all counts (including empty sites) is equal. This, it turns out, addresses differences in density as well as magnitude of reads. In order to minimize the influence of outliers, the bottom and top 5% of read-counts are trimmed thus resulting in a more robust estimator.

To see how TTR can address both differences in density and average read-counts, assume read-counts come from a mixture of a Normal distribution with parameters μ and σ, and a Bernoulli distribution:

$$f(x) = \begin{cases} \theta \times \mathrm{N}(x|\mu, \sigma) & x > 0 \\ \\ (1-\theta) \times \mathrm{Bern}(x|p=0) & x = 0 \end{cases} \tag{5.1}$$

with θ mixture coefficient and Bernoulli probability of observing a zero equal to 1 (i.e. probability of success, equal to zero). The expected value of this mixture is:

$$\mathrm{E}[f(x)] = \theta \times \mu \tag{5.2}$$

In this model, the Normal distribution is responsible for the non-zero read-counts (and thus the parameter μ is equivalent to the mean at non-zero sites i.e. "NZMean"), and the Bernoulli distribution is responsible for the counts of zero. The θ mixture coefficient dictates which distribution is responsible for the observation (i.e. insertion from a Normal distribution or a non-insertion), and is equivalent to the saturation of the library.

Given a set of TnSeq datasets $j = 1 \dots K$, their read-counts can be normalized to have the same expected value as a reference dataset, r, as follows:

$$\mathrm{E}[f_r(x)] = w_j \times \mathrm{E}\left[f_j(x)\right] \tag{5.3}$$

$$\theta_r \times \mu_r = w_j \times \left(\theta_j \times \mu_j\right) \tag{5.4}$$

$$w_j = \frac{\theta_r \times \mu_r}{\theta_j \times \mu_j} \tag{5.5}$$

where w_j is a multiplicative factor that scales the read-counts in dataset j so that its expected read-counts match the reference dataset:

$$f_r(x) = w_j \times f_j(x) \tag{5.6}$$

For simplicity, the reference dataset can be taken to be the first replicate of the control datasets (e.g. j = 1). In this case, the multiplicative factor for dataset 1 would be $w_1 = 1.0$

5.1.2 Example

Consider two “datasets” coming from mixtures with the following parameters:

Name	μ	σ	θ	$\mathrm{E}[f(x)]$
“dataset1”	500	50	0.5	250.0
“dataset2”	500	50	0.3	150.0

Despite having an equal mean over non-zero counts (i.e. μ aka “NZMean”), “dataset2” has an expected value that is lower due to its significantly lower insertion density.

We took a sample of 1000 counts from each mixture and observed a difference between the means of -101.62; close to the difference of -100 that is expected (i.e. $150 - 250$). We used resampling to get a distribution of the differences in means that would be expected if the datasets were equal under the null-hypothesis:

When the datasets are unnormalized, the observed difference would be extremely unexpected under the assumption that the datasets are equal thus leading us to (incorrectly) reject the null-hypothesis. Proper normalization of the datasets should avoid the rejection of the null-hypothesis, by taking the difference in saturation between the datasets into

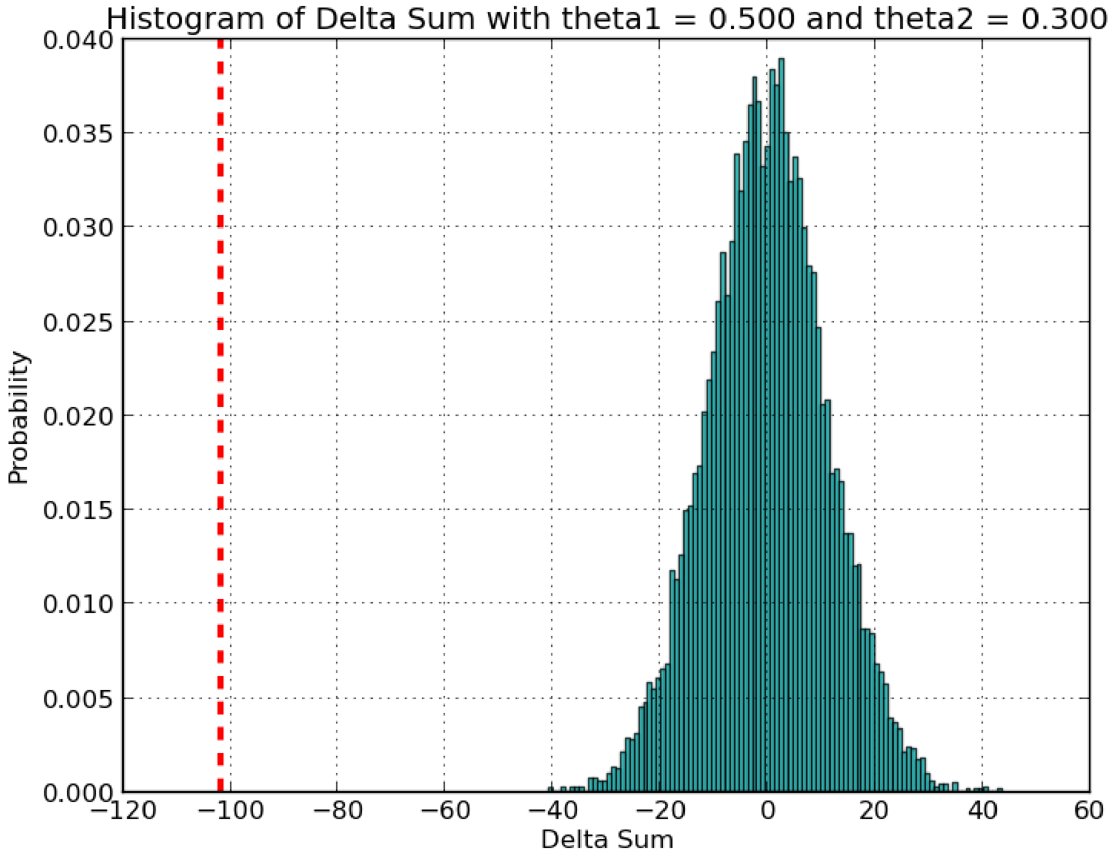


Figure 5.3: Histogram of the difference in means generated by permuting counts (including zeros) **before normalization**. The red-line represents the observed difference in means before normalization (-101.62)

consideration.

Applying the normalization procedure described above scales the datasets so that their expected values are equal, despite the differences in saturation. Although for this example we know the exact theoretical values of μ and θ, this would not be true in practice. To calculate the normalization factor, we estimate the parameters μ and θ for each dataset from the data using the sample means and sample densities. Note that in real TnSeq datasets this would include read-counts and insertions at both essential and non-essential regions. However, because non-essential genes greatly outnumber essential genes, the estimates should still be close to the true values. Using the estimated values for the parameters, we calculate the normalization factor as described above (using “dataset 1” as the reference):

$$w_2 = \frac{0.501 \times 500.34}{0.329 \times 498.99} = 1.527 \tag{5.7}$$

Multiplying “dataset2” by this factor would produce an expected value of $\mathrm{E}[w \times f_2(x)] = 229.05$, much closer to the expected value of “dataset1”. Note that the theoretical factor that would make the expected values truly equal would be $\frac{5}{3} = 1.\overline{666}$. The estimated factor, 1.527, is a relatively close approximation. Using the same sample used before, but normalizing “dataset2” by the estimated factor, yields the following resampling histogram (Figure 5.4):

The observed difference in means after normalization was (19.12), well within the range expected if the distributions were equal and thus the differences in insertion density do not cause the null-hypothesis to be rejected.

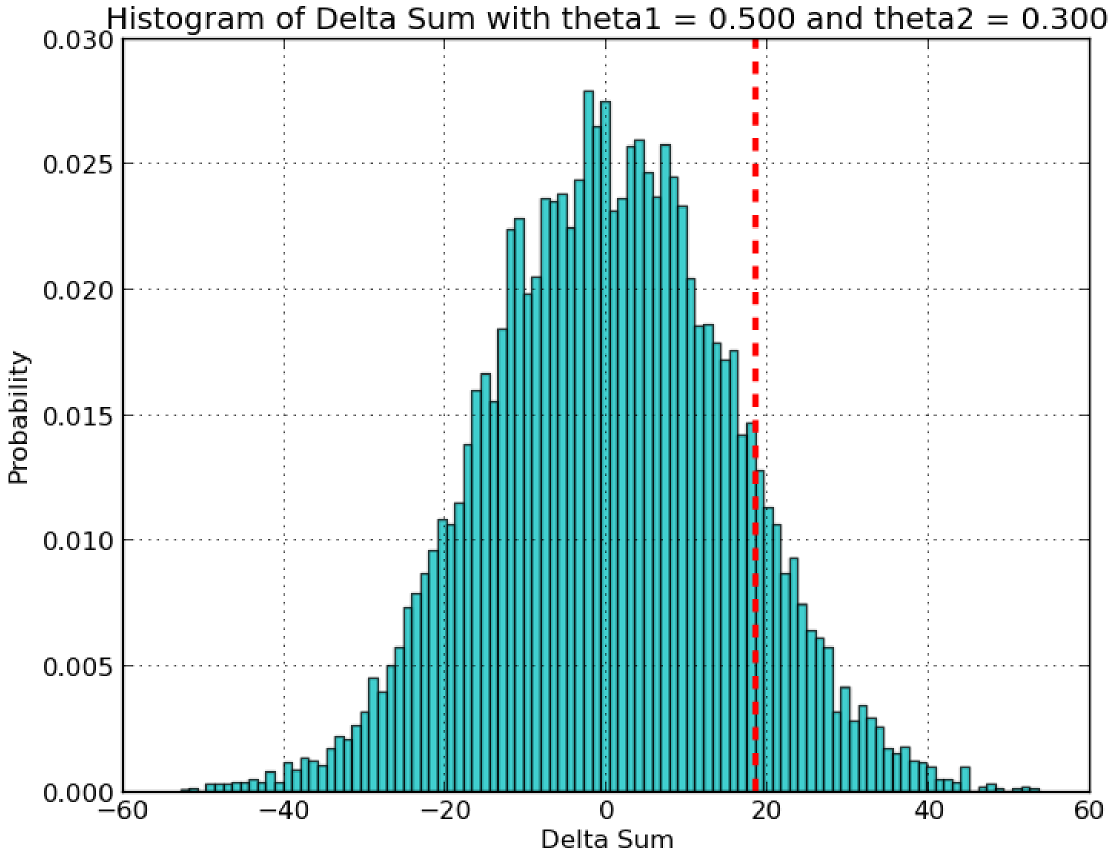


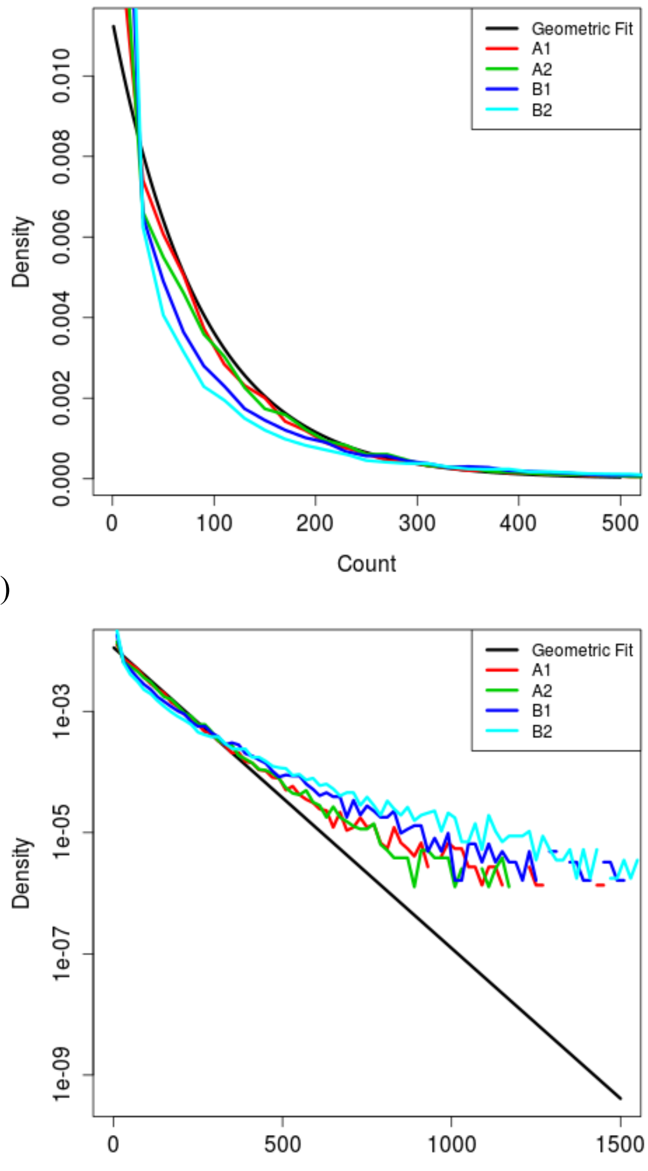
Figure 5.4: Histogram of the difference in means generated by permuting counts (including zeros) **after normalization** by TTR. The red-line represents the observed difference in means before normalization (-101.62)

5.2 Correcting for Skew in TnSeq Datasets

In practice, some datasets appear ”well behaved”, where the distribution of read counts tends to fit a simple geometric or negative-binomial distribution, while other datasets are skewed, with a few highly over-represented sites dominating the read-count distribution. While there is not a rigorous argument for why the distribution of read counts must be geometric, it is clear that in most datasets, TA sites with only a few reads (1-10) are most abundant, while sites with high counts (> 1000) are much less abundant. This can be observed in four representative datasets shown in Figure 5.5. The skew, especially at high counts, can be seen better on a log scale (Figure 5.5(b)).

These datasets are from a Himar1 Tn-mutant library in *M. tuberculosis*, where A1 and A2 are two replicates grown in vitro, and B1 and B2 representing in vivo datasets, where the library has been passaged through a mouse. Each dataset has 2 to 5 million reads distributed over 74,602 TA sites. Datasets A1 and A2 appear to fit a geometric distribution more closely than B1 and B2, which show greater skew. This can also be seen on a QQ-plot (quantile-quantile), where B1 and B2 veer farther away from the 1:1 diagonal than the in vitro datasets. Indeed, B1 and B2 have extremely high counts at a few individual sites (with max read counts of 6,009 and 16,146 respectively), compared to max counts of 1,693 and 1175 in the A1 and A2 datasets.

The effect of the skew observed in datasets like B1 and B2 (which is a common phenomenon in Tn-Seq) is that it can bias the statistical analysis of essential regions, especially for methods that depend on the read counts. Certainly, for genes containing the



(a)

(b)

Figure 5.5: (a) Histogram of non-zero read counts obtained from *M. tuberculosis* tn-mutant libraries. A1, A2 are replicates grown in vitro, and B1 and B2 are replicates grown in vivo. The black line represents a Geometric fit. (b) Histogram of read counts on a log scale.

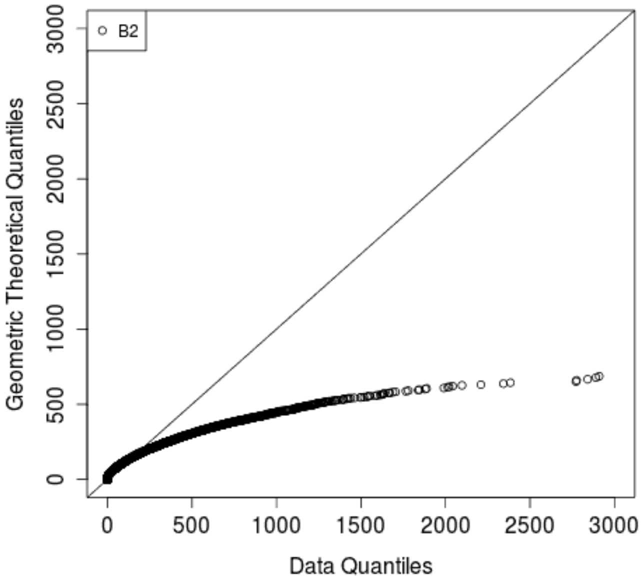


Figure 5.6: QQ-plot of the raw read counts for dataset B2, and the theoretical Geometric quantiles.

TA sites with high spikes in read counts, they will appear excessively non-essential, and it could make the gene appear differentially essential in other conditions. Conversely, the spikes in read counts at some TA sites will suppress the apparent level of reads at other sites, potentially making them appear relatively more essential.

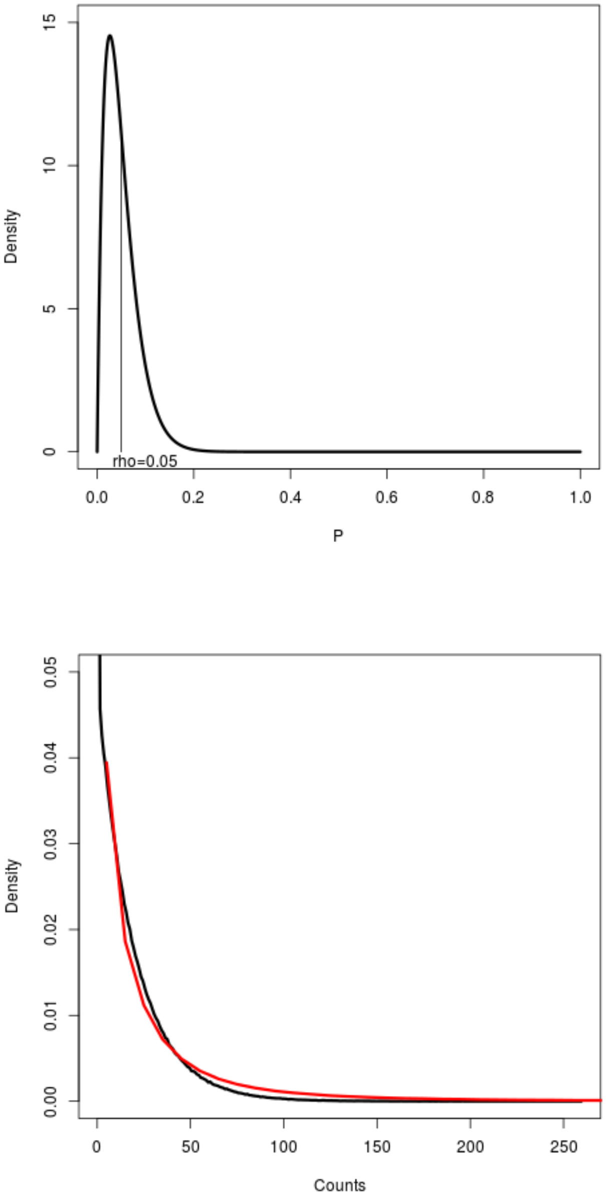
5.2.1 Beta-Geometric Correction

We propose a novel method for correcting for this skew in read-count distributions by fitting each dataset to a modified distribution called a **Beta-Geometric** distribution (Equation 1), and using this to adjust the observed read counts so they more closely fit a geometric. This approach is based on the observation that the skewed Tn-Seq datasets actually appear to fit not a single geometric with a single Bernoulli parameter, p, but the weighted sum (integral) of multiple geometric distributions with different values of p. As weights on p, we choose the Beta distribution, with parameters ρ and κ set so that the peak is around p. The Beta distribution has an extra degree of freedom representing dispersion around p (See Figure 5.7). This reflects a generative model in which individual cells in the Tn-mutant library have different growth rates, some growing slightly faster and some slightly slower than wild-type cells, depending on the location of the transposon insertion in their genome. This variability in growth rates will smear out the apparent abundance of read counts after selection (i.e. several rounds of doubling in selective conditions). In this model, the spikes in read counts would come from clones that had higher-than-average growth rates, for whatever reason (biological or random).

$$pdf(c;\rho,\kappa) = \int_0^1 Beta(p \mid \rho,\kappa) \times Geomtric(c \mid p)\, dp \tag{5.8}$$

5.3 Empirical Comparison of Normalization Methods

To assess the performance of different normalization methods, including the TTR and BGC methods described in Sections 5.1 and 5.2, replicate datasets are compared against each other. As these replicate datasets are grown under the same conditions, no statistically significant differences are expected if the datasets are properly normalized. To explore the limits of the normalization methods, some of the replicates are artificially modified to have lower density or include outliers, this posing a more stringent test of the different methods.



(a)

(b)

Figure 5.7: (a) Example of a Beta distribution with $\rho = 0.05$, and $\kappa = 40$. (b) Histogram of counts from a regular Geometric distribution ($p = 0.05$, black curve), and a Beta-Geometric distribution ($\rho = 0.05$, $\kappa = 40$, red).

5.3.1 Resampling

The permutation test can be used to detect significant differences in mean-read counts between genes in different conditions. For each gene, the read counts at all the TA sites and all replicates in each condition are summed, treating replicates within a condition as independent and identically distributed. The difference between the sum of read-counts at each condition is then calculated. The significance of this difference is evaluated by comparing to a resampling distribution generated from randomly reshuffling the observed counts at TA sites in the region among all the datasets. This creates a distribution of read count differences that might be observed by chance, assuming a null hypothesis that the two conditions are not in fact different. A p-value is then derived from the proportion of reshuffled samples that have a difference more extreme than that observed in the actual experimental data.

Due to the stochastic nature of read counts, there will be some variability in the respective some of read counts. If the difference between the sums of read counts falls within the bounds of the resampling distribution, this is interpreted as being due to chance. On the other hand, true conditionally essential genes will show a highly significant difference, as insertions in the locus will be observed in one condition but not the other resulting in a difference which is typically much larger than any of the differences observed by randomly re-shuffling read counts. Furthermore, this method can detect genes whose disruption leads to a reduction in fitness; that is, genes which are not absolutely essential in one of the conditions, but instead have lower read-counts in one of the conditions compared to the other.

The permutation test distinguishes which of these differences is statistically significant. p-values are derived from the fraction of samples that exceed the observed difference (See Figure 5.8), and this is adjusted for multiple comparisons by the Benjamini-Hochberg procedure.

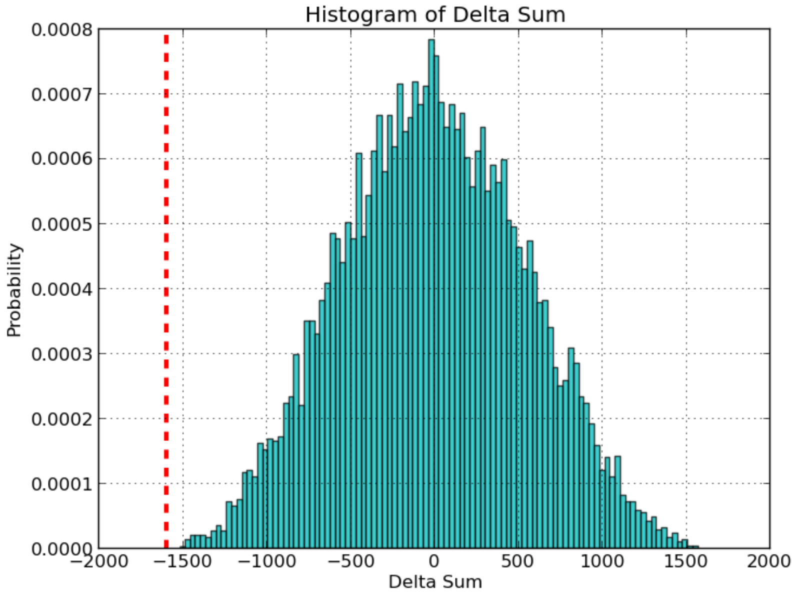


Figure 5.8: Resampling histogram for the *M. tuberculosis* gene Rv0017c, grown in vitro and in vivo. Rv0017c has 23 TA sites, and the sum of the observed counts at the TA sites in this genes *in vitro* was 1,318 and *in vivo* was 399, therefore the observed difference in counts is -918. To determine the significance of this difference, 10,000 permutations of the counts at the TA sites among the datasets was generated and the observed differences plotted as a histogram showing that a difference as extreme as -918 almost never occurs by chance. The p-value is determined by the tail of this distribution to be 0.003 (30 out of 10,000).

5.3.2 Comparison of Normalization Methods

In order to further explore how these methods behave, we compared normalization methods on four replicate datasets grown under the same condition (where no true-positives are expected). Two replicates were taken as "control" samples, and two other replicates as “experimental” samples. Read-counts in the two “experimental” datasets were artificially modified to add more noise or outliers, therefore posing a challenge for the normalization methods:

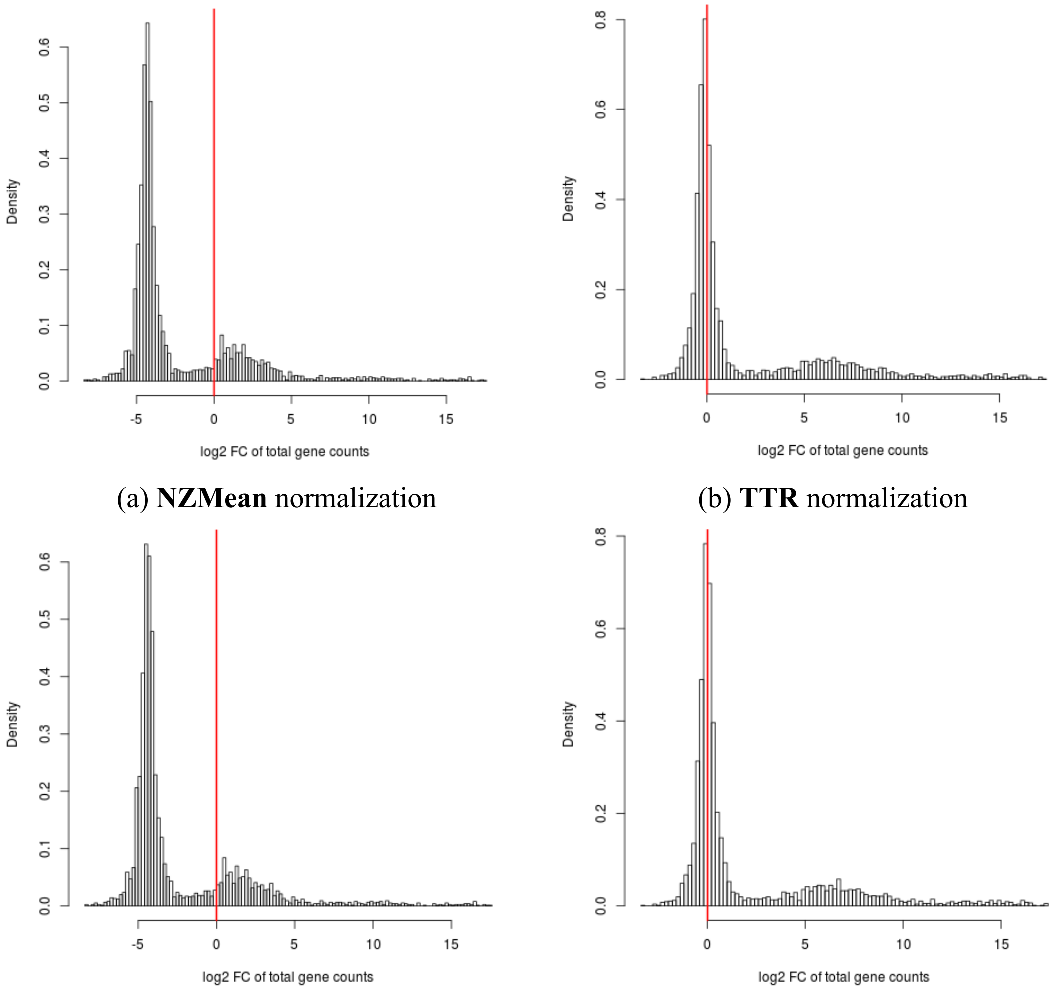
- Same Density & Same Counts - The raw data with no modification.
- Same Density & Double Counts - Density was kept the same but the counts at experimental samples were doubled.
- Half Density & Same Counts - Density for experiments samples was cut in half (artificially setting random counts to 0), and remaining counts were kept the same.
- Half Density & Double Counts - Density for experiments samples was cut in half (artificially setting random counts to 0), and remaining counts were doubled.
- Outliers 1 - A random number of outliers (between 30-40) were added to each experimental replicate. Outliers ranged from 10,000 to 200,000 .
- Outliers 2 - A random number of outliers (between 70-80) were added to each experimental replicate. Outliers ranged from 10,000 to 200,000 .
- Outliers 3 - A random number of outliers (between 110-130) were added to each experimental replicate. Outliers ranged from 10,000 to 200,000 .
- Outliers 4 - A random number of outliers (between 730-780) were added to each experimental replicate. Outliers ranged from 10,000 to 200,000 .

Table 5.1 shows the number of false-positives obtained by the different methods tested. The performance of several different normalization methods were compared relative to no normalization (“nonorm”). As expected, NZMean (which normalizes non-zero

Table 5.1: False positives ($p_{\text{adj.}} < 0.05$) obtained by each normalization method, after running on replicates of the libraries.

Experiment	**nonorm**	**nzmean**	**totreads**	**TTR**	**BGC**
Same Density & Same Counts	157	2	2	2	2
Same Density & Double Counts	1814	1	2	2	1
Half Density & Same Counts	5	405	3	2	2
Half Density & Double Counts	10	369	2	1	3
Outliers 1	161	823	913	2	2
Outliers 2	171	1805	1840	3	1
Outliers 3	162	2158	2193	1	2
Outliers 4	232	2176	2181	3	2

read-counts to be equal) is capable of normalizing datasets correctly when there are differences in the overall magnitude of the read-counts. However, when there are differences in saturation (and thus the frequency of empty sites is significantly different), it does a poor job. Total-reads normalization does a good job of handling saturation, however it has problems in the presence of outliers. TTR, and BGC performed nearly identically, able to remove the influence of outliers, in addition to handling differences in saturation and the magnitude of read-counts. The histograms of the datasets in the presence of outliers (700+) shows how robust TTR and BGC are to these outliers (Figure 5.9).



(a) **NZMean** normalization

(b) **TTR** normalization

(c) **Total Reads** normalization

(d) **BGC Normalization**

Figure 5.9: Histogram of log-fold change in mean read-count per gene after normalizing read-counts in the presence of outliers. NZMean (a) and Total Reads (b) are susceptible in the presence of outliers. On the other hand, TTR (b) and BGC (d) are robust to outliers, as the peak of the distribution is centered around zero as expected in replicate datasets.

6 DETERMINING INTERACTIONS BETWEEN GENES

6.1 Introduction

6.1.1 Genetic Interactions

One way to determine if two genes interact with each other (e.g. are involved in the same pathway) is to examine the fitness of a mutant that has mutations in both of the genes. If the two genes interact with each other, then the fitness of the double mutant (where the function of both genes has been impaired) should deviate from what would be expected given the individual mutations [62, 63].

Different models of a quantitative genetic interaction exist. These differ in how they define the expected fitness of the double mutant [62]. A common way of defining the expected fitness model of the double mutants is the multiplicative model, where the fitness cost incurred by the double mutant is the product of the individual fitness costs of the individual mutants. Figure 6.1 shows a visual representation of the expected fitness under the traditional multiplicative model (blue bars). When two genes (X and Y) do not interact with each other, then the fitness of the double mutant ($\Delta X \times \Delta Y$) is expected to be the product of the individual fitness costs.

If the fitness of the double mutant is even worse than expected, then this suggests a negative interaction where, for example, both genes might play redundant functions in an important pathway. The organism may be able to withstand the disruption of one of the genes due to the redundancy, but will incur a significant cost when both genes are disrupted

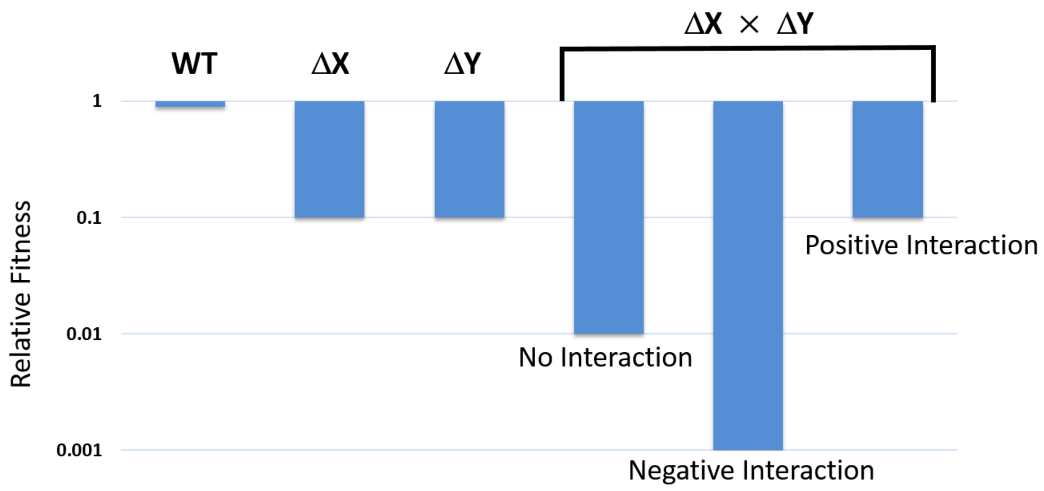


Figure 6.1: Visual representation of the multiplicative model of genetic interactions. If the double mutant ($\Delta X \times \Delta Y$) incurs a greater reduction in fitness than expected, then this suggests a negative interaction between gene X and gene Y. If the double mutant exhibits better fitness, then this suggests there is a positive interaction between them.

at the same time.

If the fitness of the double mutant is higher than expected, then this is said to be a "positive interaction". This could occur when, for example, one gene is produces a toxic intermediate which the other gene is responsible for eliminating. While there may be a large fitness cost when the gene responsible for eliminating the toxic product is disrupted (as the organism cannot remove the toxic product), the double mutant would exhibit an improvement in fitness (as it does not produce or have to eliminate the toxic intermediate any longer).

TnSeq can be a valuable tool for determining genetic interactions [47]. As transposons disrupt the function of the genes where they insert at, tn-mutant libraries can be used to obtain "knockouts" of all the genes in the genome. In the context of TnSeq, double mutants are obtained by creating tn-mutant library from a knockout strain (KO) where a gene has been experimentally knocked-out. Thus, a tn-mutant of this KO strain would have two genes disrupted: the original gene KO from which the library was created, and the individual tn-mutant which was interrupted by the insertion of a transposon. The phenotypes of the individual genes in the KO tn-mutant library can be compared to phenotypes from a wildtype strain (where single genes will be disrupted via transposons), allowing the determination of genetic interactions.

For TnSeq datasets, read-counts can be used as a measure of fitness. Mutant bacilli that have fitness advantage are expected to grow faster than those with significant fitness costs, and thus should result in more genetic material available when sequencing. Those bacteria that have trouble growing, on the other hand, will leave little genetic material to sequence and map to the genome as transposon insertions. However, because sequencing shows only a “snapshot” of the growth of the bacteria at a given moment, (at least) two different time-points are necessary to measure the change in fitness. Thus how read-counts change across the time-points should be indicative of the tn-mutant’s fitness.

Since the overall sequencing material obtained will vary between datasets and time-points, normalizing the read-counts across the datasets is crucial (See Section 5). This poses a unique problem for determining fitness using read-counts, as the effects of the genetic interactions will be different than expected under traditional models of genetic interactions.

Figure 6.2 illustrates how genetic interactions would look like when analyzing read-counts. While the fitness of a double mutant should be lower than single mutants when

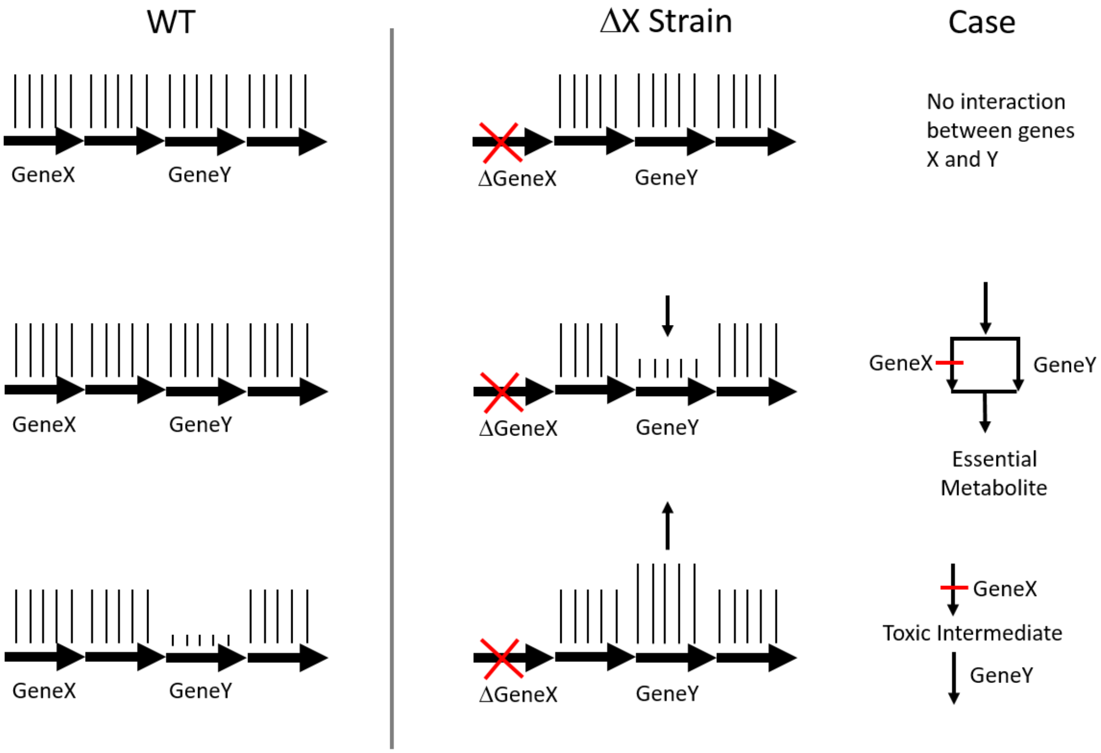


Figure 6.2: Depiction of genetic interactions in TnSeq data

there is no interaction, in the context of TnSeq, the read-counts will be normalized so that their (expected) read-counts are the same. A positive or negative interaction, therefore, is implied when read-counts are higher or lower than then average read-count in the dataset. Because fitness is measured across (at least) two different time-points, a method capable of analyzing how read-counts change across the different strains and time-points is necessary.

6.1.2 Analyzing Log Fold-Change

The goal of this method is to identify genes which show a significant change in the read-counts, given a tn-library created from a WT strain and tn-library created from a KO strain, each grown at two different time points. This allows for many different possible comparisons, as shown in 6.3A. While one could use the resampling method discussed in

Section 5 to compare pairs of datasets, it is not clear how this can be used to determine genetic interactions.

For instance, while resampling could be used to detect that a gene exhibited a 2-fold increase in read-counts across time-points (e.g. KO at Time 2 vs KO at Time 1), if a similar effect is seen in WT (where only one gene is disrupted), then this is not indicative of a genetic interaction but instead indicative of disrupting the function of that specific gene. Similarly, one could compare both the WT and KO strains at time 2, if a similar effect is seen at time 1 then one could not rule out that the observed enrichment in due to differences in the library as opposed to actual biological differences.

To identify genetic interactions, the change in fitness must be determined across both time-points and strain simultaneously. For example, if read-counts of a gene in the WT strain are relatively low at both time-points (Blue line in Figure 6.3, showing a small decrease), yet there is a significant increase in the same gene in the KO strain (yellow line) then this is suggestive of a positive genetic interaction.

The method discussed here estimates the $\log_2$ fold-change (logFC) in mean-read counts between the time points (i.e. Time-2 vs Time-1) for both strains separately. It then compares the results between the strains (i.e. KO vs WT), identifying genes which have significantly different logFC between the strains. This can be thought of as finding significant changes in enrichment, which is like comparing the slopes of two different lines. Figure 6.4 shows a visual representation of the general approach.

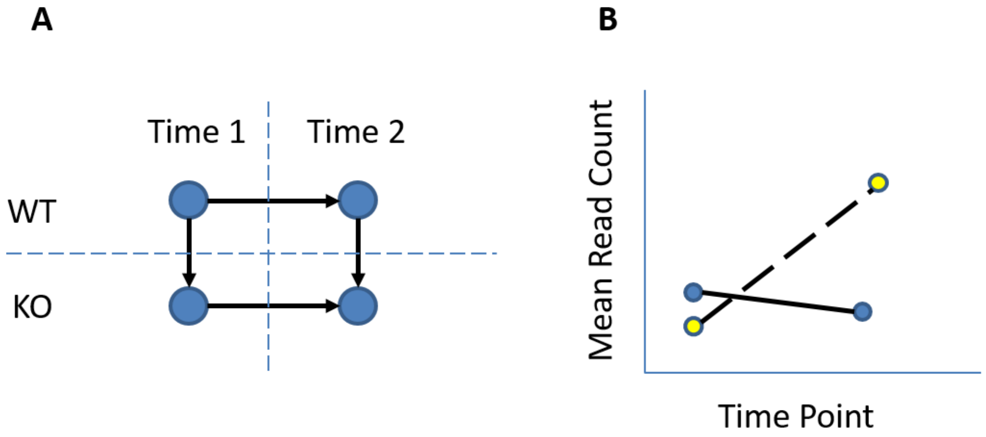


Figure 6.3: (A) Possible comparisons of different datasets available in this experimental setup. (B) Illustration of change in mean-read count across time-points between the strains.

6.2 Method

For a given gene and a given condition (e.g. WT-1, KO-1, WT-2, KO-2), read-counts are pooled across the replicates (e.g. combine replicates of WT-1) and model this set of observations as coming from a Normal distribution. Since read count data is traditionally modeled as Negative Binomial, then the mean, which is the quantity of interest, is approximately Normal due to the Central Limit Theoremr (see Section 5.1).

To get the distribution of the mean, priors are set on the parameters μ and σ, and then posterior distribution is derived for the parameter μ given the data:

Using the Normal likelihood:

$$X \sim \text{Normal}(\mu, \sigma)$$

Coupled with these priors:

$$p(\mu) = \text{Normal}$$

$$p(\sigma) = \text{Inverse-Gamma}$$

Results in the following posterior distribution for the mean:

$$p(\mu \mid X, \sigma) = \text{Normal}(\mu_n, \sigma_n)$$

where μ_n and σ_n are updated parameters based on the data and the prior information. Since the distribution of the means has a known form (Normal), which can be sampled easily. See Figure 6.4 for a visual representation of what follows.

Using this fact, samples are obtained for the distribution of means for each of the conditions. Thus, for each gene, there are vectors of representative samples for the mean read-counts of each condition: WT-1, KO-1, WT-2, and KO-2. These samples can be used to get Monte Carlo estimates of other distributions or values of interest by carrying out operations on the values on the sample. In order to get a representative sample of the distribution of logFC between time-points (e.g. WT-2 vs WT-1, and KO-2 vs KO-1), the $\log_2$ of the ratio between the samples of means is taken like follows:

$$\text{LFC-WT}_i = \log_2\left(\frac{\text{WT-32}_i}{\text{WT-0}_i}\right)$$

$$\text{LFC-KO}_i = \log_2\left(\frac{\text{KO-32}_i}{\text{KO-0}_i}\right)$$

for all samples $i \in \{1 \ldots S\}$. This results in two vectors, LFC-WT and LFC-KO, containing representative (Monte Carlo) samples of the distributions of logFC of mean read-counts for WT and KO respectively. In other words, these distributions represent how the variation in mean read-counts for a given strain after being in-vivo for the given time span (i.e. span between time point 1 and time point 2).

The difference of the two samples of logFC is used to get a single sample representing the distribution of the difference in logFC:

$$\begin{aligned}\Delta\text{LFC}_i &= \text{LFC-KO}_i - \text{LFC-WT}_i \\ &= \log_2\left(\frac{\text{KO-32}_i}{\text{KO-0}_i}\right) - \log_2\left(\frac{\text{WT-32}_i}{\text{WT-0}_i}\right)\end{aligned}$$

for all samples $i \in \{1 \ldots S\}$. This new vector contains a representative sample of the distribution of the differences between the logFC. It is this final vector of samples that is used to determine those genes with significant differences. To classify genes, the 95% Highest Density Interval (HDI) is calculated, representing the interval for the distribution where the true value of the difference in logFC will be in with 95% probability. A gene is considered to show a significant difference if the HDI region does not overlap with a [-0.5, 0.5] region around 0. This region is meant to represent those values of the Δ logFC that are practically equivalent to 0.0 (i.e. the Null Hypothesis of no difference between logFC). This region is called a “Region of Practical Equivalence” or ROPE [64, 65]

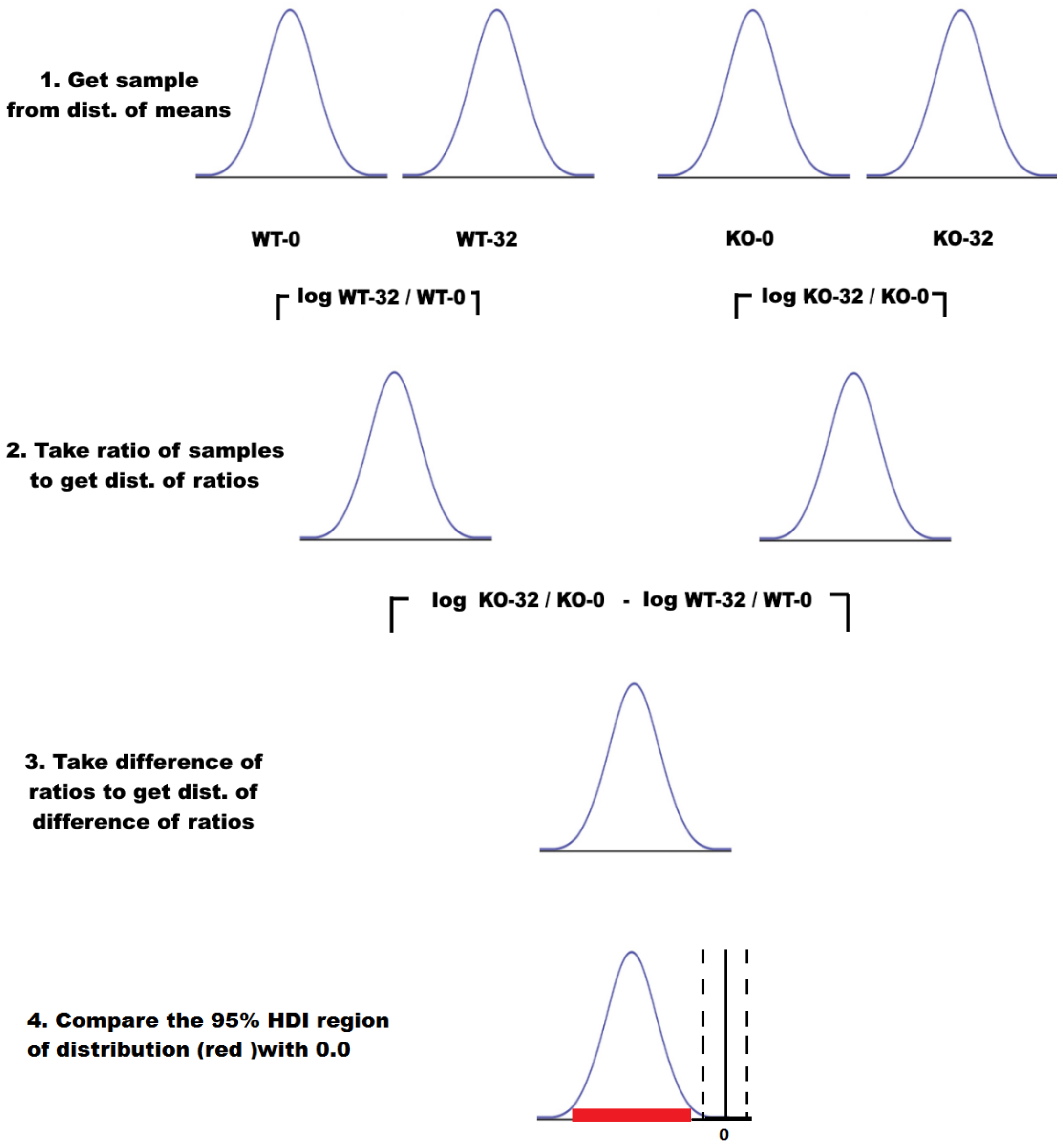

1. Get sample from dist. of means
WT-0
WT-32
KO-0
KO-32
log WT-32 / WT-0
log KO-32 / KO-0
2. Take ratio of samples to get dist. of ratios
log KO-32 / KO-0 - log WT-32 / WT-0
3. Take difference of ratios to get dist. of difference of ratios
4. Compare the 95% HDI region of distribution (red)with 0.0
0

Figure 6.4: Visual description of how the method works. Read picture from bottom up. (1) Distributions of the mean read-counts are generated for the 4 conditions: WT-0, WT-32, KO-0, and KO-32. (2) We calculate the logFC between the samples for each strain, to get two distribution of logFC. (3) We take the difference of the two logFC distributions to get a single distribution of the difference between logFC of the strains. (4) We compare the overlap of the distribution of the differences with the null hypothesis of no difference to assess significance.

In addition, the probability that the Δ logFC falls within the [-0.5, 0.5] window around 0 can also be calculated. This is useful for "ranking" the genes: the lower this probability, the farther away from 0.0 the distribution is and therefore the more significant the observation.

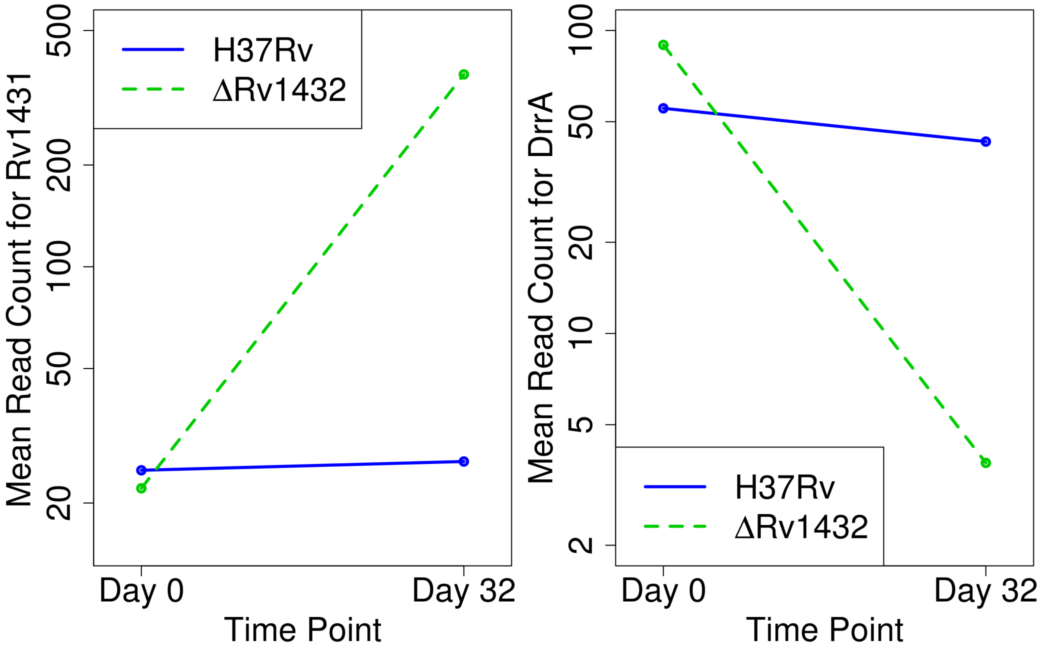
6.3 Results

To test the ability of this method to identify genetic interactions, we applied it to a wildtype strain (WT) and three separate knockout strains (KO) of M. tuberculosis. The KO strains each consisted of a single gene deletion: Rv1432, Rv2680, and Rv1565c. These three genes were chosen because they showed a growth-impairment in vivo and did not have a known function. Each of the four transposon mutant libraries (one WT and three KO strains) were inoculated into five C57BL/6 mice (for a total of 20 mice). After 24 hours (referred to here as "day 0" or"d0"), two mice in each group were sacrificed, and bacteria were recovered from the spleen and plated. These represented libraries *before* selection, to control for potential biases in the inoculation of the mice. The remaining three mice in each group were sacrificed after 32 days of infection (referred to here as "day 32" or "d32"), representing conditions *after* selection in vivo. This period of infection encompasses the full spectrum of immune responses, including adaptive immunity which is initiated approximately 10 days post-infection in this model. The libraries were sequenced, processed, and mapped to the H37Rv genome [31]. The replicates datasets had saturation ranging between 27% and 42%, with the mean template count at non-zero sites was in the range of 40-219 templates per site. The statistical analysis of genetic interactions described above was applied to the knockout libraries of Rv1432, Rv2680, and Rv1565c compared to wild-type (H37Rv). This analysis identified 135, 80, and 144 genes, respectively, that had an observed $\Delta \log \mathrm{FC}$ that was significantly different than zero, indicating potential genetic interactions with the knocked-out genes. Table 6.1 shows a breakdown of the results. The analytical framework presented was sufficient to resolve different classes of genetic interactions such as aggravating interactions, which result in a significantly lower fitness than expected. Positive interactions, which improve fitness compared to what would be expected in the double mutant, can be split in to alleviating interactions (which reduce the impact of deleting a gene) and suppressive interactions which completely suppress any negative effects. For example, a strong suppressive interaction was found between *rv1432* and the adjacent gene, *rv1431* (Figure 6.5a), which is in the same operon. This interaction was evident as a large increase in the number of read counts at sites in *rv1431* at day 32 in the KO strain. A plausible explanation for this observation may be that Rv1431 and

Table 6.1: Types of genetic interactions identified for the three knockout (KO) strains analyzed. Negative interactions result in reduced fitness for the double mutant (Aggravating). Positive interactions improve fitness relative to the expected fitness deficit of the double mutant (Alleviating), or completely suppress any negative effects of the double mutation (Suppressing).

Knockout	**Negative**	**Positive**	
Strains	**Aggravating**	**Alleviating**	**Suppressing**
ΔRv1432	11	58	11
ΔRv1565c	25	115	4
ΔRv2680	32	87	16

Rv1432 could be members of the following biochemical pathway: Rv1431 → toxic intermediate → Rv1432 → product. Thus deletion of *rv1432* would result in an accumulation of a toxic intermediate that is deleterious to the organism, but this effect would be suppressed when *rv1431* is deleted (and thus the toxic intermediate would not be produced at all). Mutants lacking all three members of the ABC efflux pump, DrrABC, decreased



(a) Rv1431

(b) DrrA

Figure 6.5: Plot of the mean read-counts (log-scale) for Rv1431 (panel **A**) and DrrA (panel **B**) between H37Rv (WT) and the knockout strain of Rv1432 (KO). Rv1431 illustrates a suppressive interaction with Rv1432, while DrrA shows an aggravating interaction.

specifically in the Δ*rv1432* library, defining an aggravating interaction (Figure 6.5b). The increased requirement for DrrABC in the *rv1432* knockout (an aggravating interaction) could suggest that the toxic intermediate hypothesized to be produced by *rv1432* may be exported by the the ABC efflux pump. Our analysis also identified 69 genes that exhibit a significant change in enrichment between WT libraries and the library lacking *rv2680*. Notably, eight of these genes are in the biosynthetic cluster for phthiocerol dimycocerosate (PDIM), a lipid that constitutes a significant fraction of the outer cell envelope of Mtb [66]. Read counts for each of these genes, *rv2930*-*rv2941* increase slightly ($\log\mathrm{FC}$ of ~ 0.5) in the WT library over the course of infection, but decrease in the KO library (approximately 10-fold). These effects result in $\Delta\log\mathrm{FC}$ scores of around -3.5 This effect was not observed in the other knockouts (Δ*rv1432*, Δ*rv1565c*), and hence is specific to Δ*rv2680*. This implies that the requirement for the PDIM locus is more stringent in the absence of *rv2680* (aggravating interaction). In addition, a large number of genes involved in the synthesis/-modification/transport of fatty acids or the very long-chain mycolic acid components of the cell envelope show differential enrichment, including *fadE7*, *cmaA2*, *mmaA3*, *fabG3*, *lpqQ*, *lpqD*, *lppJ*, and *lipG* (see network diagram in Figure 6.6). The anabolism of long-chain lipids, such as PDIM, plays an important but complex role in Mtb. Not only do these lipids serve important roles individually, but their synthesis is also linked to each other and to the overall metabolic state of the cell [67, 68]. Thus, decreasing the synthesis of one abundant lipid has been found to increase the synthesis of others, and to alter the balance of acyl-CoA metabolites that are central to carbon metabolism. As a result, it is not surprising that PDIM synthesis is a member of a genetic interaction network that contains a number of other genes involved in lipid metabolism, and our studies add *rv2680* to this functional network. These data strongly suggest that our analytical framework is sensitive enough to identify most members of an interacting biochemical pathway. The different relationships these and other genes demonstrate why it is necessary to accurately classify genetic interactions to generate testable hypotheses.

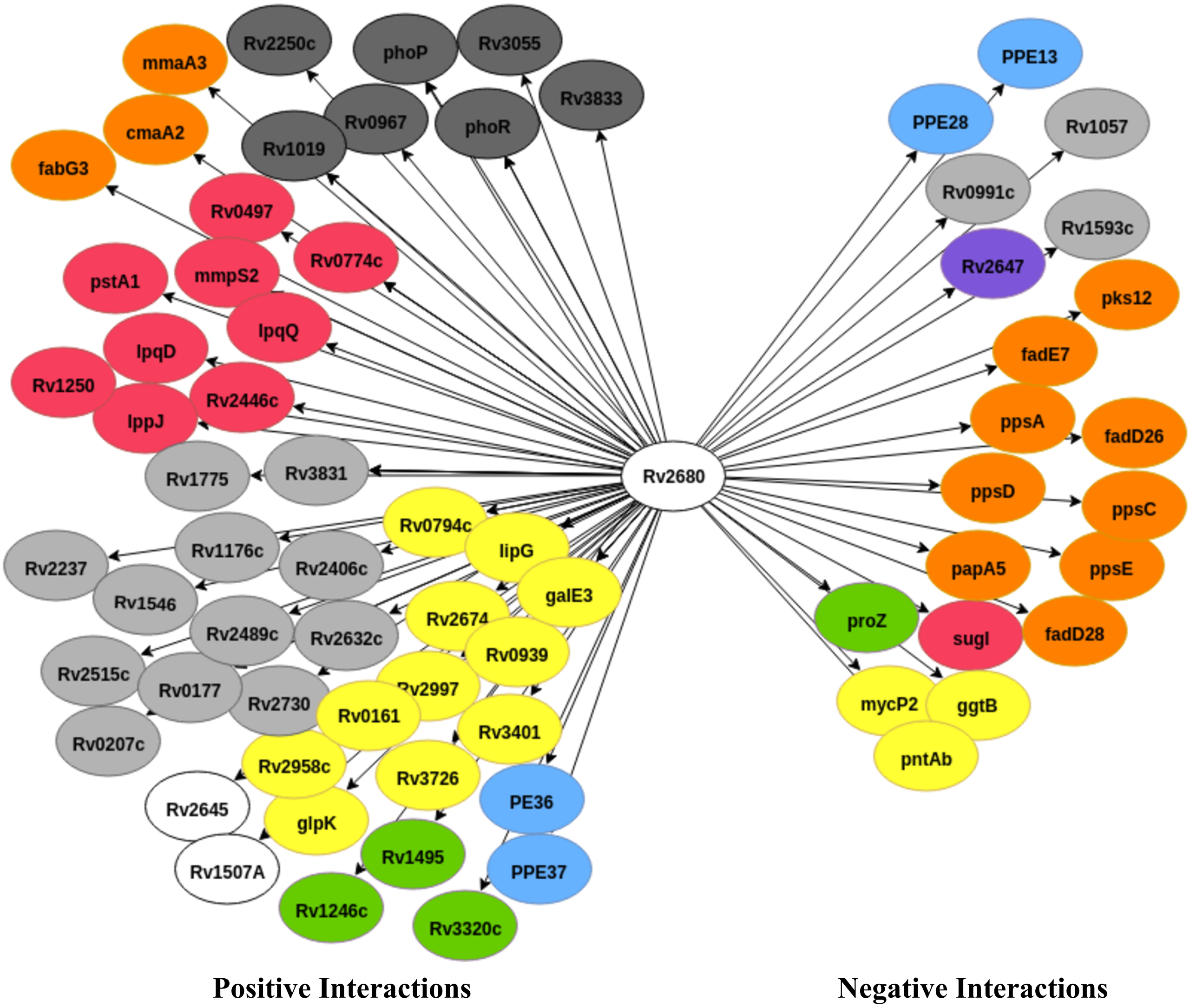


Figure 6.6: Genetic interactions with Rv2680. Genes on the left showed positive interactions, while genes on the right showed negative interactions. The genes are colored by functional category: Yellow: intermediary metabolism and respiration, Orange: lipid metabolism, Red: cell wall and cell processes, Blue: PE/PPE, Purple: insertion seqs and phages, Green: virulence, detoxification, adaptation, Light Grey: conserved hypotheticals, Dark Grey: regulatory proteins, White: Unknown.

7 CONCLUSIONS

7.1 Discussion

In this dissertation I have presented several statistical methods capable of analyzing data obtained from TnSeq experiments. TnSeq has become a valuable tool for assessing the phenotypes for large libraries of mutants at the same time, largely due to the high-throughput nature of transposon sequencing as well as the high-resolution data obtained [69]. Determining the precisely location in the genome that are tolerant to disruption provides useful information about the essentiality of bacterial genomes, and the phenotypes these mutants exhibit under selection. Regions responsible for coding proteins that play fundamental roles (like DNA repair) are almost certainly essential in all conditions, while other regions (like those responsible for the metabolism of a specific substrate) may only become essential under certain conditions. With this information, it is possible to understand the function of coding regions, possibly leading to the identification of novel drug targets whose disruption is lethal to the pathogen.

As the genes of interest are those that do not tolerate insertions, TnSeq can be thought of as a negative experiment. Thus, there is inherent ambiguity about whether a region is truly essential or is lacking insertions for some other reason. For instance, the transposon that is utilized may have some previously unknown disinclination for inserting in the presence of a local sequence of DNA. The curvature of the chromosome may also prevent insertions at certain areas, leading to chromosomal specific effects. In addition, it may be difficult to properly sequence some are as of the genome (for example, due to high GC content) making them appear as if they were unable to tolerate insertions. Some of these may be addressed by normalization or examining the local sequence pattern surrounding empty regions, but others, such as chromosomal position effects, may be limitations of this methodology.

In addition to these potential limitations, several other challenges make proper statistical analysis difficult. For instance, many genes are often observed to tolerate insertions in the beginning and ends of their coding regions, even when they code for proteins that play crucial roles for the survival of the organism. This biological difficulty is expected to be present in all organisms, representing an inherent challenge that makes trivial analysis of TnSeq data impossible. The protocol used to create and process the transposon libraries can also affect the quality of the resulting TnSeq data in significant ways. Chief among them is the choice of transposon used for the construction of the mutant libraries. Transposons may exhibit different local sequence biases for insertion, which would have important consequences for the analysis methods utilized. The methods presented in this dissertation assume that the protocol utilizes the Himar1 mariner-based transposon, which has shown a strong insertion preference for insertions at TA dinucleotide sites [8]. This choice allows the method to take advantage of the preference bias shown by the transposon to model TA sites in the genome as a series of Bernoulli trials where a success represents the presence of a transposon insertion. Other transposons, like the Tn5 transposon which inserts at any location, may require different choices for the statistical model to properly determine essentiality.

In addition, the experimental methodology can have a large effect on the variability and noise of the resulting read-counts. For instance, problems in the polymerase chain reaction (PCR) used to amplify the fragments of DNA obtained from the library may lead to errors and outliers [70]. The strong selection pressures the libraries are subjected to may also lead to skewed read-counts as the difference in fitness between the mutants is magnified in the selection process. Modern protocols try to overcome some of these obstacles by using barcodes to identify unique insertions, and thus reduce the impact of spikes due to problems with PCR amplification [14]. However, not all TnSeq datasets take advantage of these new techniques, and there exist other challenges inherent in this type of data. Thus, statistical methods must be able to overcome these problems, as well as the potential differences in the libraries being compared (like different levels of saturation, or amount of sequence material obtained). All these problems make analysis of TnSeq data difficult.

The methods presented in this dissertation, address these and other challenges present in the analysis of TnSeq. Sections 2, 3 and 4, focused on the general problem of determining essentiality of genes within a given condition. The Gumbel and Beta-Binomial methods (discussed in Sections 2 and 3 respectively) are proper Bayesian models that take advantage of the properties of the Himar1 transposon to quantify the likelihood of observed patterns of insertions. These models are limited to gene boundaries, as they require a finite set of trials for each gene. On the other hand, the HMM, (discussed in Section 4), tackles the problem of determining essentiality through out the entire genome, without having predefined gene boundaries; allowing it to identify entire regions that span several genes, and even non-coding regions, that are essential. In addition it takes into consideration the magnitude of read-counts (as opposed to simply the presence of an insertion), and thus can detect different levels of essentiality, like regions whose disruption leads to a growth-defect or a growth-advantage.

Sections 5 and 6 address the problem of comparing changes in fitness or essentiality across different conditions. As discussed in Section 5, it is crucial to normalize the read-counts in the datasets being analyzed in order to make them comparable. This helps ensure that the differences detected are not false positives (due to outliers from problems in PCR amplification, or skew), and instead represent real differences observed between the datasets. Section 6 introduces a method that detects significant changes in enrichment, which allows one to detect genetic interactions in a high-throughput manner, identifying interaction networks. Together, these represent a comprehensive set of methods capable of extracting valuable information about essentiality from TnSeq data. They are made available in a software package called TRANSIT [31], which simplifies their use and hopefully provides access to these computationally complex methods to a larger audience.

7.2 Future Work

7.2.1 Extend Models to Work with Other Transposons

Most of the methods outlined in this dissertation assume TnSeq data was obtained from libraries created using the Himar1 transposon, which is a popular choice for transposon mutagenesis. It, like other mariner transposons, has shown preference for inserting at TA dinucleotides, which has important implications for the choice of essentiality of models. As a consequence, the assumptions made in the models presented in this dissertation may not apply to other choice of transposon, like the Tn5 transposon which can insert at any position in the genome. Thus, there is a need to modify and extend these models of essentiality to work on other transposons, specifically addressing their properties.

7.2.2 Take Spacing of TA Sites Into Consideration

As the distribution of TA dinucleotides throughout the genome is stochastic, the spacing between any pair of sites is variable. Hence, some TA sites may be very close together while others may be several hundred nucleotides apart. In the limit, TA sites that are side-by-side may be effectively redundant (as they provide the same information about essentiality), while those which are far apart may represent entirely different regions (as the space between them may code for a different genomic feature or protein). To date, models have generally avoided this issue given the difficulty of tackling such a problem. How to extend these (and other models of essentiality) to take the distance between TA sites into consideration is still an open question.

7.2.3 Differential Comparison of the Entire Genome

The methods outlined in Sections 5 and 6 are capable of comparing datasets, thus identifying conditionally essential genes (i.e. genes which are essential under one condition but not another). However, these methods are limited to pre-defined genetic boundaries like the coding region of a gene. Other areas of the genome, such as non-coding regions upstream of a gene, may code for important features (like promoters or or σ-factors) which may play crucial roles. Conditional essentiality in these regions would be missed by methods which are limited by gene boundaries. Thus, research ample need for a method capable of analyzing the change in essentiality between two (or more) conditions, across the entire genome.

REFERENCES

[1] 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.

[2] C. Calvori, L. Frontali, L. Leoni, and G. Tecce, “Effect of rifamycin on protein synthesis,” *Nature*, vol. 207, pp. 417–418, Jul 1965.

[3] 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.

[4] 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.

[5] B. Christen, E. Abeliuk, J. M. Collier, V. S. Kalogeraki, B. Passarelli, J. A. Coller, M. J. Fero, H. H. McAdams, and L. Shapiro, “The essential genome of a bacterium,” *Mol. Syst. Biol.*, vol. 7, p. 528, 2011.

[6] B. McCLINTOCK, “The origin and behavior of mutable loci in maize,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 36, pp. 344–355, Jun 1950.

[7] W. S. Reznikoff, “The Tn5 transposon,” *Annu. Rev. Microbiol.*, vol. 47, pp. 945–963, 1993.

[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] F. C. Luft, “Sleeping Beauty jumps to new heights,” *J. Mol. Med.*, vol. 88, pp. 641–643, Jul 2010.

[10] C. M. Sassetti, D. H. Boyd, and E. J. Rubin, “Comprehensive identification of conditionally essential genes in mycobacteria,” *PNAS*, vol. 98, no. 22, pp. 12712–12717, 2001.

[11] 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.

[12] C. M. Sassetti and E. J. Rubin, “Genetic requirements for mycobacterial survival during infection,” *PNAS*, vol. 100, no. 22, pp. 12989–12994, 2003.

[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.

[14] J. Long, M. DeJesus, D. Ward, R. Baker, T. Ioerger, and C. Sassetti, "Identifying essential genes in *Mycobacterium tuberculosis* by global phenotypic profiling.," in *Methods in Molecular Biology: Gene Essentiality* (L. J. Lu, ed.), vol. 1279, Springer, 2015.

[15] J. D. Gawronski, S. M. 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," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, pp. 16422–16427, Sep 2009.

[16] G. C. Langridge, M. Phan, D. Turner, T. Perkins, L. Parts, J. Haase, I. Charles, D. Maskell, S. 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.

[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, p. e1002251, 09 2011.

[18] N. J. Blades and K. W. Broman, "Estimating the number of essential genes in a genome by random transposon mutagenesis," Tech. Rep. MSU-CSE-00-2, Dept. of Biostatistics Working Papers, Johns Hopkins University, July 2002.

[19] 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.

[20] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat. Methods*, vol. 5, pp. 621–628, Jul 2008.

[21] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139–140, Jan 2010.

[22] A. Zomer, P. Burghout, H. J. Bootsma, P. W. Hermans, and S. A. van Hijum, "ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data," *PLoS ONE*, vol. 7, no. 8, p. e43012, 2012.

[23] Y. J. Zhang, T. R. Ioerger, C. Huttenhower, J. E. Long, C. M. Sassetti, J. C. Sacchettini, and E. J. Rubin, "Global assessment of genomic regions required for growth in Mycobacterium tuberculosis," *PLoS Pathog.*, vol. 8, p. e1002946, Sep 2012.

[24] S. Solaimanpour, F. Sarmiento, and J. Mrazek, "Tn-seq explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries," *PLoS ONE*, vol. 10, no. 5, p. e0126070, 2015.

[25] M. A. DeJesus, Y. J. Zhang, C. M. Sassetti, E. J. Rubin, J. C. Sacchettini, and T. R. Ioerger, "Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries," *Bioinformatics*, vol. 29, pp. 695–703, Mar 2013.

[26] M. A. DeJesus and T. R. Ioerger, “Improving discrimination of essential genes by modeling local insertion frequencies in transposon mutagenesis data,” in *BCB* (J. Gao, ed.), p. 144, ACM, 2013.

[27] M. A. DeJesus and T. R. Ioerger, “Capturing Uncertainty by Modeling Local Transposon Insertion Frequencies Improves Discrimination of Essential Genes,” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 12, no. 1, pp. 92–102, 2015.

[28] M. A. DeJesus and T. R. Ioerger, “A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data,” *BMC Bioinformatics*, vol. 14, p. 303, 2013.

[29] M. DeJesus and T. Ioerger, “Reducing type i errors in tn-seq experiments by correcting the skew in read count distributions,” in *7th International Conference on Bioinformatics and Computational Biology (BICoB 2015)*, 2015.

[30] M. A. DeJesus and T. R. Ioerger, “Normalization of transposon-mutant library sequencing datasets to improve identification of conditionally essential genes,” *J Bioinform Comput Biol*, p. 1642004, Jan 2016.

[31] M. A. DeJesus, C. Ambadipudi, R. Baker, C. Sassetti, and T. R. Ioerger, “TRANSIT–A Software Tool for Himar1 TnSeq Analysis,” *PLoS Comput. Biol.*, vol. 11, p. e1004401, Oct 2015.

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

[33] S. J. Wheelan, A. Marchler-Bauer, and S. H. Bryant, “Domain size distributions can predict domain boundaries,” *Bioinformatics*, vol. 16, pp. 613–618, Jul 2000.

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

[35] S. T. Cole, R. Brosch, and J. Parkhill, “Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence,” *Nature*, vol. 393, no. 6685, pp. 537–544, 1998.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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," *PNAS*, vol. 100, no. 12, pp. 7213–7218, 2003.

[41] 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.

[42] I. Krieger, J. Freundlich, V. Gawandi, J. Roberts, V. Gawandi, Q. Sun, J. Owen, M. Fraile, S. Huss, K. Duncan, J.-L. Lavandera, T. Ioerger, and J. Sacchettini, "Structure-guided discovery of phenyl diketo-acids as potent inhibitors of M. tuberculosis malate synthase," *Chemistry & Biology*, 2012.

[43] 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.

[44] P. Muller, G. Parmigiani, and K. Rice, "Fdr and bayesian multiple comparisons rules," in *Proceedings of the ISBA 8th World Meeting on Bayesian Statistics*, (Benidorm,

Spain), Juner 2006.

[45] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” in *Proceedings of the IEEE*, pp. 257–286, 1989.

[46] S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D’Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A. L. Barabasi, Z. N. Oltvai, and A. L. Osterman, “Experimental determination and system level analysis of essential genes in Escherichia coli MG1655,” *J. Bacteriol.*, vol. 185, pp. 5673–5684, Oct 2003.

[47] T. van Opijnen, K. L. Bodi, and A. Camilli, “Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms,” *Nat. Methods*, vol. 6, pp. 767–772, Oct 2009.

[48] A. Dasgupta, P. Datta, M. Kundu, and J. Basu, “The serine/threonine kinase PknB of Mycobacterium tuberculosis phosphorylates PBPA, a penicillin-binding protein required for cell division,” *Microbiology (Reading, Engl.)*, vol. 152, pp. 493–504, Feb 2006.

[49] G. Khare, V. Gupta, P. Nangpal, R. K. Gupta, N. K. Sauter, and A. K. Tyagi, “Ferritin structure from Mycobacterium tuberculosis: comparative study with homologues identifies extended C-terminus involved in ferroxidase activity,” *PLoS ONE*, vol. 6, no. 4, p. e18570, 2011.

[50] J. J. De Voss, K. Rutter, B. G. Schroeder, H. Su, Y. Zhu, and C. E. Barry, "The salicylate-derived mycobactin siderophores of Mycobacterium tuberculosis are essential for growth in macrophages," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 97, pp. 1252–1257, Feb 2000.

[51] F. Movahedzadeh, S. C. Rison, P. R. Wheeler, S. L. Kendall, T. J. Larson, and N. G. Stoker, "The Mycobacterium tuberculosis Rv1099c gene encodes a GlpX-like class II fructose 1,6-bisphosphatase," *Microbiology (Reading, Engl.)*, vol. 150, pp. 3499–3505, Oct 2004.

[52] D. J. 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, 2009.

[53] X. Gu, M. Chen, H. Shen, X. Jiang, Y. Huang, and H. Wang, "Rv2131c gene product: an unconventional enzyme that is both inositol monophosphatase and fructose-1,6-bisphosphatase," *Biochem. Biophys. Res. Commun.*, vol. 339, pp. 897–904, Jan 2006.

[54] E. J. Munoz-Elias and J. D. McKinney, "Mycobacterium tuberculosis isocitrate lyases 1 and 2 are jointly required for in vivo growth and virulence," *Nat. Med.*, vol. 11, pp. 638–644, Jun 2005.

[55] G. Chandra, K. F. Chater, and S. Bornemann, "Unexpected and widespread connections between bacterial glycogen and trehalose metabolism," *Microbiology (Reading, Engl.)*, vol. 157, pp. 1565–1572, Jun 2011.

[56] T. Sambou, P. Dinadayala, G. Stadthagen, N. Barilone, Y. Bordat, P. Constant, F. Levillain, O. Neyrolles, B. Gicquel, A. Lemassu, M. Daffe, and M. Jackson, "Capsular glucan and intracellular glycogen of Mycobacterium tuberculosis: biosynthesis and impact on the persistence in mice," *Mol. Microbiol.*, vol. 70, pp. 762–774, Nov 2008.

[57] 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.

[58] C. Astarie-Dequeker, L. Le Guyader, W. Malaga, F. K. Seaphanh, C. Chalut, A. Lopez, and C. Guilhot, "Phthiocerol dimycocerosates of M. tuberculosis participate in macrophage invasion by inducing changes in the organization of plasma membrane lipids," *PLoS Pathog.*, vol. 5, p. e1000289, Feb 2009.

[59] M. A. Kirksey, A. D. Tischler, R. Simeone, K. B. Hisert, S. Uplekar, C. Guilhot, and J. D. McKinney, "Spontaneous phthiocerol dimycocerosate-deficient variants of Mycobacterium tuberculosis are susceptible to gamma interferon-mediated immunity," *Infect. Immun.*, vol. 79, pp. 2829–2838, Jul 2011.

[60] J. S. Cox, B. Chen, M. McNeil, and W. R. Jacobs, "Complex lipid determines tissue-specific replication of Mycobacterium tuberculosis in mice," *Nature*, vol. 402, pp. 79–83, Nov 1999.

[61] 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.

[62] R. Mani, R. P. St Onge, J. L. Hartman, G. Giaever, and F. P. Roth, “Defining genetic interaction,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 3461–3466, Mar 2008.

[63] P. Beltrao, G. Cagney, and N. J. Krogan, “Quantitative genetic interactions reveal biological modularity,” *Cell*, vol. 141, pp. 739–745, May 2010.

[64] J. K. Kruschke, “Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison,” *Perspect Psychol Sci*, vol. 6, pp. 299–312, May 2011.

[65] J. K. Kruschke, “Bayesian estimation supersedes the t test,” *J Exp Psychol Gen*, vol. 142, pp. 573–603, May 2013.

[66] D. E. Minnikin, L. Kremer, L. G. Dover, and G. S. Besra, “The methyl-branched fortifications of Mycobacterium tuberculosis,” *Chem. Biol.*, vol. 9, pp. 545–553, May 2002.

[67] M. Jain, C. J. Petzold, M. W. Schelle, M. D. Leavell, J. D. Mougous, C. R. Bertozzi, J. A. Leary, and J. S. Cox, “Lipidomics reveals control of Mycobacterium tuberculosis virulence lipids via metabolic coupling,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, pp. 5133–5138, Mar 2007.

[68] W. Lee, B. C. VanderVen, R. J. Fahey, and D. G. Russell, “Intracellular Mycobacterium tuberculosis exploits host-derived fatty acids to limit metabolic stress,” *J. Biol. Chem.*, vol. 288, pp. 6788–6800, Mar 2013.

[69] T. van Opijnen and A. Camilli, “Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms,” *Nat. Rev. Microbiol.*, vol. 11, pp. 435–442, Jul 2013.

[70] L. Garibyan and N. Avashia, “Polymerase chain reaction,” *J. Invest. Dermatol.*, vol. 133, p. e6, Mar 2013.