# Improving discrimination of essential genes by modeling local insertion frequencies in transposon mutagenesis data*

Michael A. DeJesus, Thomas R. Ioerger
Department of Computer Science
Texas A&M University
College Station, TX 77843
{mad, ioerger}@cs.tamu.edu

## ABSTRACT

Transposon mutagenesis experiments enable the identification of essential genes in bacteria. Deep-sequencing of mutant libraries provides a large amount of high-resolution data on essentiality. Statistical methods developed to analyze this data have traditionally assumed that the probability of observing a transposon insertion is the same across the genome. This assumption, however, is inconsistent with the observed insertion frequencies from transposon mutant libraries of *M. tuberculosis*.

We propose a modified binomial model of essentiality that can characterize the insertion probability of individual genes in which we allow local variation in the background insertion frequency in different non-essential regions of the genome. Using the Metropolis-Hastings algorithm, samples of the posterior insertion probabilities are obtained for each gene, and the probability of each gene being essential is estimated. We compare our predictions to those of previous methods and show that, by taking into consideration local insertion frequencies, our method is capable of making more conservative predictions that better match what is experimentally known about essential and non-essential genes.


BCB'13, September 22 - 25 2013, Wshington, DC, USA

ACM 978-1-4503-2434-2/13/09$15.00.
http://dx.doi.org/10.1145/2506583.2506610

## Categories and Subject Descriptors

I.6.3 [**Computing Methodologies**]: Simulation and Modeling — *Applications*; J.3 [**Computer Applications**]: Life and Medical Sciences — *Biology and Genetics*

## General Terms

Sequence Analysis, Essentiality, Hierarchical Models

## 1. INTRODUCTION

Knowledge of which genes are essential for the growth of an organism enables the development of new drugs that target these genes, thus preventing its growth [9]. A common way to determine which genes are essential in bacterial organisms is through transposon mutagenesis experiments. In these experiments, large libraries of mutants are created by subjecting individual bacilli to transposon mutations. Transposons are small fragments of DNA that are capable of inserting within the genome, thereby disrupting the genomic regions where they insert. The Himar1 transposon is frequently used in transposon mutagenesis experiments, as it is known to insert at random TA dinucleotide sites ("TA sites") within the genome. This specificity to TA sites can be exploited through sequencing, as the possible insertion locations can be known beforehand.

Early attempts to use transposon mutant libraries to assess essentiality utilized micro-array hybridization to determine which genes where being expressed and which ones where not [13, 14, 15]. Although these methods were capable of assessing which genes where disrupted, they did not provide detailed information about where the insertions took place. With the development of next-generation sequencing, large libraries of transposon mutants can be sequenced at the same time, providing high-resolution information about which areas in the genome can be disrupted.

Various statistical methods have been developed to analyze the data obtained with deep-sequencing, and asses the essentiality of bacterial organisms. Some of these methods have examined the relative number of transposon insertions that map to specific TA sites ("read counts"). For example, Zhang et al [16] developed a non-parametric test of mean read counts to assess the essentiality requirements for windows of TA sites throughout the genome. In addition to read-counts, other methods have focused on the relative frequency of the insertions (i.e. fraction of TA sites dis-

rupted). Blades and Broman [3] developed a Multinomial model to characterize the essentiality of libraries that had a small number of transposon insertions. This method was used to assess the genes necessary for growth of *M. tuberculosis* in vitro and in lung [11, 10]. Recently, we developed a Bayesian model of essentiality that used the extreme value distribution to determine the probability of observing large gaps devoid of any insertions that are characteristic of essential regions [8, 5]. Using the Metropolis Hastings algorithm, this method was able to avoid using a priori estimates of parameter values, instead estimating them by sampling from the corresponding posterior distributions.

One key assumption of this method is that the insertion probability is the same for all non-essential genes. While this assumption serves to simplify the statistical model, it is unlikely that all genes (or all genes within a given class of essentiality) share the same insertion probability. For instance, GC-rich regions can be difficult to sequence successfully, which may lead to incomplete sequence coverage in certain genomic regions leading to depressed read-counts and insertions. This could explain why PE_PGRS genes in the *M. tuberculosis* genome have been previously observed to contain large gaps devoid of insertions, and indeed have been characterized as essential by some statistical models, even though this family of genes is generally believed to be non-essential [2, 11]. Furthermore, because Himar1-based transposons have specificity to TA sites, and the distribution of TA sites within genes is variable, genes can contain different amount of TA sites within regions that can be disrupted (for example, in the N- and C- termini or within non-essential domains), which can lead to differences in the number of insertions observed.

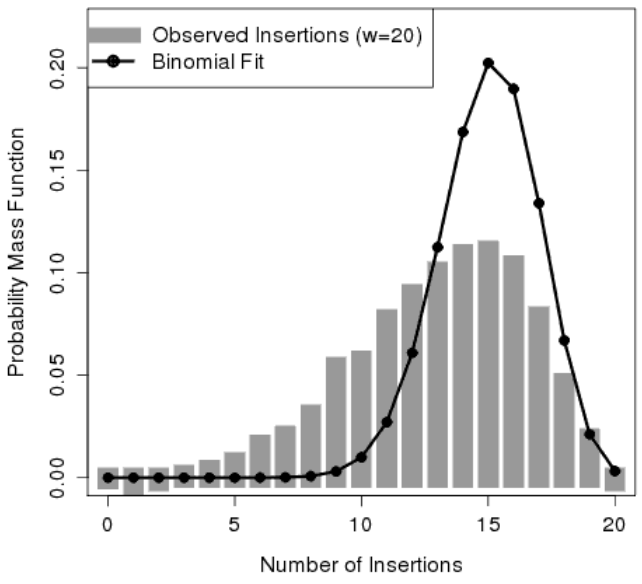


**Figure 1: Histogram of the number of insertions observed within windows of 20 TA sites (gray bars). The binomial distribution (black line) is incapable of fitting the over-dispersion of observed in the number of insertions.**

This variability in insertion probability is evident in libraries of transposon mutants. Figure 1 shows a histogram of the observed number of insertions (gray bars) within windows of 20 TA sites, for a transposon mutant library of M. tuberculosis [8]. A sliding window of 20 TA sites was shifted by one across the genome, and the number of insertions observed within the window was determined. The observed distribution of insertions (gray bar) is more dispersed than what would be expected with a binomial distribution (black line), suggesting that the insertion frequency is not constant but instead varies within genomic regions. Assuming an insertion probability that is globally constant will ignore this variability, and lead to less reliable predictions.

In this paper, we present a Bayesian method for analyzing deep-sequencing data from transposon mutagenesis experiments. Our method utilizes a binomial likelihood to model the insertions within the genes and a beta distribution to model the local insertion probability for each gene. The Metropolis-Hastings algorithm is used to estimate the parameters of the model and obtain the posterior probability of essentiality for each gene. The predictions of the model are then compared to previous results, and the effect of taking into consideration individual insertion probabilities is examined.

Thus the main contribution of this paper is to show how to extend Bayesian models of essentiality by relaxing the assumption of a global insertion frequency to a local-frequency model, where each gene can have its own local variation. This extension improves the prediction of essential genes by taking into consideration the variability of insertion probabilities observed in the data, and the length of the genes into account.

## 2. METHODS

The data obtained from sequencing the transposon mutant libraries is mapped to the genome, and the amount of reads matching individual TA sites ("read counts") is determined. The read counts are censored to a maximum value of 1, representing whether an insertion was observed at a particular TA site or not (i.e. a value of 1 indicates at least one insertion was observed, and a value of 0 indicates no insertions were observed). This model assumes that the insertion frequency is sufficient to determine the essentiality of genes. Although potentially relevant information about essentiality might be lost by censoring the read counts, read counts can also be unreliable if the sequencing was subject to PCR bias or amplification [1].

Under this representation TA sites are treated as Bernoulli events, with the presence or absence of an insertion indicating a success or a failure. For each gene, the number of TA sites and insertions it contains is determined, and these are treated as a series of independent trials. In addition, genes are assumed to belong to a mixture of two classes of essentiality: essential and non-essential genes. The insertion frequency for each of these classes of genes is modeled through a mixture of beta distributions. Finally, the Metropolis-Hastings algorithm is used to sample from the conditional distributions of the parameters, and the posterior probability of a gene belonging to a class essential class of genes is estimated.

## 2.1 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 $\theta_i$, 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, $\omega_1$, represents the probability of a gene belonging to the essential class (with the probability of belonging to the non-essential class $\omega_0 = 1 - \omega_1$).

### 2.1.1 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 $\theta_i$. Assuming genes are independent of each other, the complete data likelihood is given by the product of binomial distributions over all the genes:

$$\prod_i^G \text{Binomial}(k_i|n_i, \theta_i) \tag{1}$$

### 2.1.2 Prior Probabilities

The distribution of individual insertion probabilities, $\theta_i$ 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:

$$\begin{aligned} \theta_i|Z_i = 0 &\sim \text{Beta}(\kappa_0\rho_0,\ \kappa_0(1-\rho_0)) \\ \theta_i|Z_i = 1 &\sim \text{Beta}(\kappa_1\rho_1,\ \kappa_1(1-\rho_1)) \end{aligned} \tag{2}$$

Under this parametrization (i.e. $\alpha = \kappa\rho$ and $\beta = \kappa(1-\rho)$), the $\rho$ parameter represents the mean insertion probability (i.e. mean of the distribution). On the other hand, the $\kappa$ 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 $\kappa$ the distribution becomes tighter around the mean (i.e. $\rho$).

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

$$\begin{aligned} \rho_0 &\sim \text{Beta}(\alpha_0, \beta_0) \\ \rho_1 &\sim \text{Beta}(\alpha_1, \beta_1) \end{aligned} \tag{3}$$

where $\alpha_0$, $\beta_0$, $\alpha_1$, and $\beta_1$ are hyper-parameters for the beta distribution.

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

$$\begin{aligned} \kappa_0 &\sim \text{Gamma}(a_0, b_0) \\ \kappa_1 &\sim \text{Gamma}(a_1, b_1) \end{aligned} \tag{4}$$

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 $\omega_1$, which represents the probability of a gene belonging to the class of essential genes:

$$Z_i \sim \text{Bernoulli}(\omega_1) \tag{5}$$

Finally, the prior distribution for $\omega_1$ is given by a Beta distribution:

$$\omega_1 \sim \text{Beta}(\alpha_\omega, \beta_\omega) \tag{6}$$

### 2.1.3 Full Joint Distribution

Using the likelihood function (1) and the prior distributions (2, 4, 3, 5, 6) described above, the full joint distribution has the following form:

$$\begin{aligned} p(\mathbf{K},\Theta,\kappa_1,\rho_1,\kappa_0,\rho_0,\mathbf{Z},\omega_1) &= \prod_i^G p(k_i \mid n_i,\theta_i) \times p(\theta_i \mid \kappa_{Z_i},\rho_{Z_i}) \\ &\quad \times p(\kappa_1) \times p(\rho_1) \times p(\kappa_0) \times p(\rho_0) \times p(Z_i \mid \omega_1) \times p(\omega_1) \end{aligned} \tag{7}$$

$$\begin{aligned} &= \prod_i^G \text{Binomial}(k_i|n_i,\theta_i) \times [\text{Beta}(\theta_i|\kappa_1\rho_1,\ \kappa_1(1-\rho_1))]^{Z_i} \\ &\times [\text{Beta}(\theta_i|\kappa_0\rho_0,\ \kappa_0(1-\rho_0))]^{1-Z_i} \\ &\times \text{Gamma}(\kappa_0|a_0,b_0) \times \text{Beta}(\rho_0|\alpha_0,\beta_0) \\ &\times \text{Gamma}(\kappa_1|a_1,b_1) \times \text{Beta}(\rho_1|\alpha_1,\beta_1) \\ &\times \text{Bernoulli}(Z_i|\omega_1) \times \text{Beta}(\omega_1|\alpha_w,\beta_w) \end{aligned} \tag{8}$$

where $\mathbf{K} = \{k_1, k_2, ..., k_G\}$, $\Theta = \{\theta_1, \theta_2, ..., \theta_G\}$, and $\mathbf{Z} = \{Z_1, Z_2, ..., Z_G\}$.

### 2.1.4 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 $\rho_1$ and $\kappa_1$ which are distributed as follows:

$$p(\kappa_1|k_i,\theta_i,\rho_1,Z_i=1) \propto \text{Beta}(\theta_i|\kappa_1\rho_1,\ \kappa_1(1-\rho_1)) \times \text{Gamma}(\kappa|a_1,b_1)$$
$$p(\rho_1|k_i,\theta_i,\kappa_1,Z_i=1) \propto \text{Beta}(\theta_i|\kappa_1\rho_1,\ \kappa_1(1-\rho_1)) \times \text{Beta}(\rho_1|\alpha_1,\beta_1)$$

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

$$p(Z_i=1|k_i,\theta_i,\kappa_1,\rho_1,\omega_1) = \text{Bernoulli}\left(\frac{p1}{p1+p0}\right)$$

where,

$$p1 = \text{Beta}(\theta_i|\kappa_1\rho_1+k_i,\ \kappa_1(1-\rho_1)+n_i-k_i)\times\omega_1$$
$$p0 = \text{Beta}(\theta_i|\kappa_0\rho_0+k_i,\ \kappa_0(1-\rho_0)+n_i-k_i)\times(1-\omega_1)$$

## 2.2 Parameter Estimation

In order to estimate parameters of the model and the probability of the genes being essential, samples are obtained through the Metropolis-Hastings algorithm. The Metropolis-Hastings (MH) algorithm is a Markov Chain Monte Carlo (MCMC) method that can be used to sample from arbitrary functions which may be too difficult to sample from otherwise. Briefly, candidate values are generated from a proposal distribution and then accepted or rejected according to a ratio of the target function evaluated at the candidate value ($x_c$) and the last value ($x_{i-1}$) in the Markov chain: $MHRatio = \frac{f(x^c)}{f(x^{i-1})}$.

Because the binomial likelihood (1) and the beta priors (2) are conjugate, the resulting conditional distribution can be sampled from easily. However, this is not the case for the conditional distributions of the $\rho$ and $\kappa$ parameters. We use a combination of Gibbs Steps and MH steps to obtain samples for all the parameters (See Algorithm 1).

# 3. RESULTS

Our method was applied to deep-sequencing data from mutant libraries of the H37Rv strain of *M. tuberculosis* [8, 5]. 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 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) [4]. 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.

A sample of 52,000 values was obtained with the 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. The acceptance rate for the $\rho_0$ and $\rho_1$ parameters was 60% and 62%, and the acceptance rate for the $\kappa_0$ and $\kappa_1$ parameters was 67% and 72% respectively. Multiple chains of the MH sampler were run in an attempt to verify that the the sampler was not trapped in local minima, and was converging to the same area in parameter space.

**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 $\theta_i \sim \text{Beta}(\rho\kappa + k_i, \kappa(1-\rho)+n_i-k_i)$

  **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 1:** Random-Walk Metropolis-Hastings Algorithm for Sampling values of $\theta_i$ and $Z_i$ for all genes $i$

## 3.1 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 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.

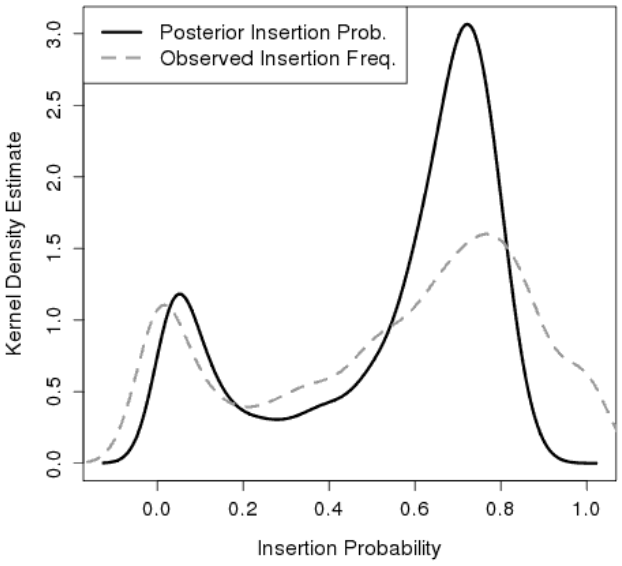


**Figure 2: Kernel Density Estimates for the mean posterior insertion probability (black-solid) and observed insertion frequency (gray-dashed) for all the genes.**

The samples of insertion probability for the genes reflect our expectations for essential and non-essential genes. Figure 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 [6]. 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.

## 3.2 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 [12]. 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.

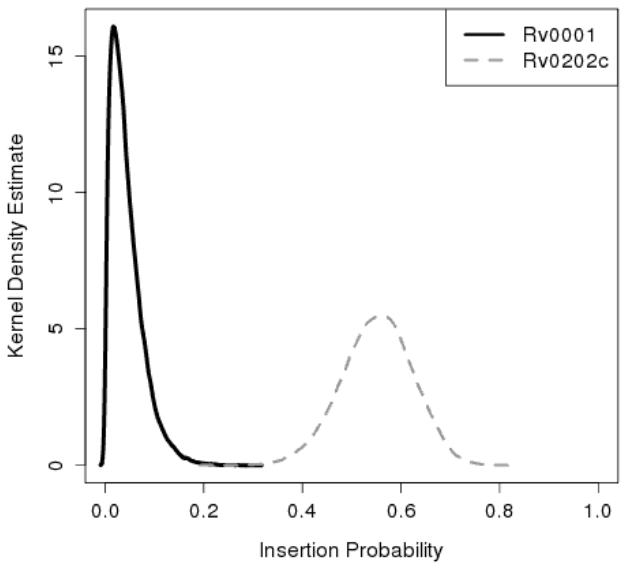


**Figure 3: Kernel Density Estimates for the posterior insertion probability of DnaA (Rv0001), a known essential gene involved in DNA repair, and MmpL11 (Rv0202c), a known non-essential gene believed to function as a transmembrane protein.**

### *3.2.1 Comparison to the TraSH Method*

The essentiality of the M. tuberculosis genome has been assessed before, through the Transposon Site Hybridization method [14, 15]. This method quantifies the amount of luminescence that is observed in probes that hybridize to each of the genes in the genome [13]. 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 1: Essentiality Comparison Between the TraSH method and the Local-Frequency Model.**

| | | Local Frequency Model | | | |
|---|---|---|---|---|---|
| | | Ess. | Unc. | Non. | Total |
| TraSH | Essential | 329 | 257 | 28 | 614 |
| | Growth-Def. | 5 | 20 | 17 | 42 |
| | Non-Ess. | 36 | 682 | 1796 | 2514 |
| | No-Data | 80 | 412 | 285 | 777 |
| | Total | 450 | 1371 | 2126 | 3947 |

Table 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.

### 3.2.2 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, $\theta_0$ and $\theta_1$, 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 $\theta_0$ and $\theta_1$ are obtained, as well as the essentiality assignments $Zi$. Estimates of the probability of essentiality are calculated by averaging the samples, as in the local-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 2: Essentiality Comparison Between the Global-Frequency Model (GFM) and the Local-Frequency Model**

| | | Local Frequency Model | | | |
|---|---|---|---|---|---|
| | | Ess. | Unc. | Non. | Total |
| GFM | Ess. | 450 | 259 | 0 | 709 |
| | Unc. | 0 | 603 | 0 | 603 |
| | Non. | 0 | 509 | 2126 | 2635 |
| | Total | 450 | 1371 | 2126 | 3947 |

Table 2 compares the results from the local-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 uncertain (259) in the local-frequency model.In addition, all 450 genes classified as essential by the local-frequency model are also classified as essential by the global-frequency model. The local-frequency model's tendency to be conservative is also 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.

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 that exists 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 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

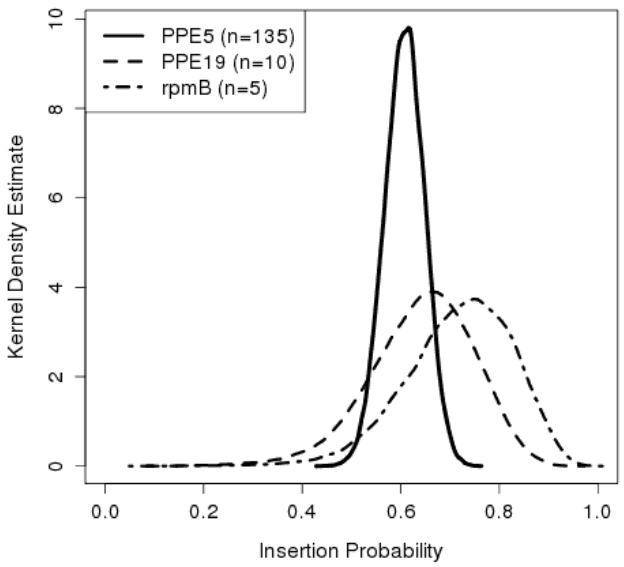


**Figure 4: Insertion Density for PPE5 (solid), PPE19 (dashed) and RpmB (dot-dash). All three genes contained an observed insertion frequency of 0.7, although they were different sizes (# of TA sites). The larger variance in the insertion density for PPE19 and rmpB reflects the greater uncertainty that exists in smaller genes.**

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.

### 3.2.3 Comparison to the Extreme Value Model

Previously we developed a Bayesian model for gene essentiality that utilized the extreme value distribution to determine the likelihood of observing a run of TA sites lacking insertions in a row. By taking the order of insertions into account, this method enabled the identification of domains within genes that contained both essential and non-essential regions. This is in contrast to the binomial model which does not take into consideration the order of TA sites. These two models of essentiality are compared in Table 3.

**Table 3: Essentiality Comparison Between the Extreme Value Model (EVM), and the Local-Frequency Model**

| | | Local Frequency Model | | | |
|---|---|---|---|---|---|
| | | Ess. | Unc. | Non. | Total |
| EVM | Ess. | 446 | 222 | 0 | 668 |
| | Unc. | 2 | 300 | 40 | 342 |
| | Non. | 2 | 685 | 2006 | 2693 |
| | Short | 0 | 164 | 80 | 244 |
| | Total | 450 | 1371 | 2126 | 3947 |

As with the global-frequency model, the local-frequency model is more conservative than the extreme value model, classifying 1,371 of the genes as Uncertain in contrast to the 342 classified by the extreme value distribution. Among

these genes are MmpL4, MmpL8 and MmpL9, which belong to the MmpL protein family, thought to be involved in polyketide biosynthesis, and lipid metabolism. Although only MmpL3 has been shown to be essential in knockout experiments [6], the extreme value model classifies mmpL4, mmpL8 and mmpL9 as essential because they contain gaps in insertion density that are longer than expected, despite also containing a relatively high insertion frequency elsewhere in the gene.

Of the 446 genes classified as essential by the local-frequency model, only 4 of these are classified as non-essential or uncertain by the extreme value model. All four genes have a small number of insertions (observed insertion density between 0.09 - 0.14), suggesting these genes are truly essential. Indeed, although the total number of TA sites in these genes ranges from 20-37, the length of the maximum run of non-insertions ranges from 8 to 12 TA sites. This suggests that the few insertions observed were capable of interrupting the run of non-insertions (e.g. one or two insertions occur in the middle of an otherwise empty gene), making them appear to be Non-Essential or Uncertain to the extreme value model (as the run of non-insertions was not sufficiently long).

Because the local-frequency model makes more conservative predictions depending on the size of the gene, it can make predictions even for those genes which contain only a very small number of TA sites within their boundaries. In contrast, the extreme value model ignores genes that are deemed too short (labeled "Short") by taking a threshold on length (i.e. $< 3$ TA sites or a span of nucleotides $< 150$bp ) and therefore excluding them from the analysis. Out of 244 genes classified as "Short" by the extreme value model, the local-frequency model classifies 164 genes as "Uncertain", without the need of an ad-hoc threshold on gene length.

As mentioned before, a potential downside of the binomial model is that it does not take into consideration the order of insertions and therefore can miss essential domains within genes. For example, genes Rv3910 and Rv0018c have been shown to code for essential protein domains involved in cell wall synthesis [7]. While the extreme value model is capable of identifying these genes as essential, the local-frequency model classifies them as Uncertain.

### 3.2.4 *Effect of the Essentiality Threshold*

Although the thresholds on the posterior probabilities of essentiality were determined through the same method for all Bayesian models (analogous to the Benjamini-Hochberg procedure for posterior probabilities [12]), this method leads to different thresholds depending on the posterior probabilities of the genes (originally, 0.9930, 0.9900, and 0.9902 for the local-frequency model, global-frequency model and extreme value model, respectively). This difference in the thresholds of essentiality may affect the number of essential (and non-essential) genes predicted by the models, as well as the agreement between them. To assess the effect of the threshold on the predictions of essential genes, we reduced the threshold on the posterior probability essentiality (from $> 0.99$ to $> 0.80$) and determined the number of essential genes predicted by the models (Figure 5).

As can be seen, the number of essential genes predicted by the models increases as the essentiality threshold is relaxed. However with the local-frequency model predicts less essential genes than the other models, making more conservative predictions despite the relaxation of the threshold. The overlap between the models also increases as the essentiality threshold is relaxed. These observations are also true in non-essential genes, where the local-frequency model is also more conservative in its predictions.

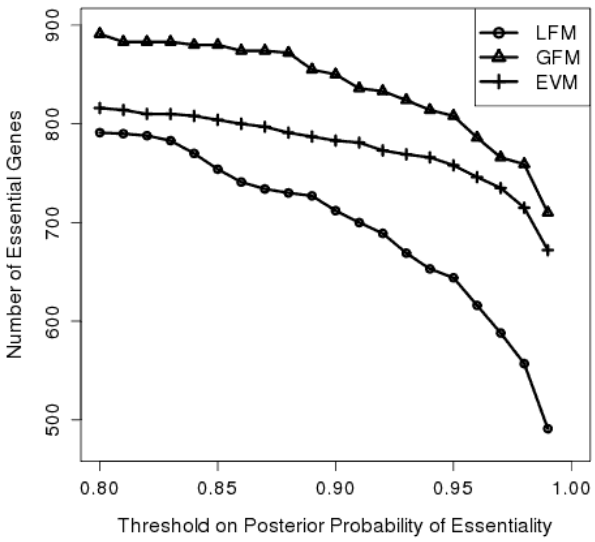


**Figure 5: Number of Essential genes predicted by the the Local-Frequency Model (LFM, circle), Global-Frequency Model (GFM, triangle), and the Extreme Value Model (EVM, cross).**

## 4. CONCLUSIONS

The intricacies of next-generation sequencing data necessitate the development of methods that can analyze this data in a robust way. Although assuming a global insertion frequency can simplify the statistical analysis of transposon mutagenesis data, it does not accurately represent expectations about the insertion probability of genes. We developed a Bayesian model that estimates the probability of essentiality for all the genes, taking into consideration the individual insertion probabilities. We applied this model to a library of *M. tuberculosis* transposon mutants, and found several cases which highlight the benefit of assuming an individual insertion frequency.

The insertion frequency of genes is not expected to be globally constant across the genes. Differences in sequencing coverage or errors in mapping reads to the genome can lead to different insertion frequencies between genes, even among those with the same class of essentiality (i.e. essential or non-essential genes). In addition, although our method characterizes all genes which lead to viable mutants as "non-essential", mutations can in fact lead to growth-defects that affect the growth rate of mutants. The severity of the growth-impairment resulting from mutation at these genes will affect the number of viable mutants observed in the library, and therefore the relative frequency of insertions observed for a given gene. By modeling the insertion frequency for each

gene, these effects can be taken into consideration where as they would otherwise be missed by assuming a global insertion frequency.

Previous methods which have assumed a global insertion frequency have been susceptible to these problems. For example, although the family of PE_GRS genes have been shown to be non-essential through knock-out experiments [2], some of these genes have been characterized as essential by previous methods. A possible reason for this might be that these genes contain regions with high GC content that are difficult to sequence, leading to stretches within the gene that are devoid of insertions.

While our binomial model is capable of modeling the insertion frequencies among the genes, it does so by considering only the presence or absence of insertions, and not the number of reads. Although the number of reads might be susceptible to problems in sequencing (e.g. PCR amplification), it has been successfully used to assess essentiality before [16, 17]. In addition, our method does not take into consideration the order of insertions (or non-insertions). A method we have previously developed, assessed the probability of observing a series of TA sites lacking insertions in a row. By using the extreme value distribution, this method was capable of identifying genes which contained essential and non-essential regions (like an essential domain). However, this method assumed a global insertion frequency, which meant it suffered from the limitations already outlined. By using a model that can take into consideration the local variation in insertion frequencies, as the model we describe here, these limitations can be overcome and more accurate predictions of essentiality can be made.

## 5. ACKNOWLEDGMENTS

*Funding:* **This work was supported by NIH grant U19 AI107774 (TRI)**.

## References

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

[2] 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.*, 44:9–19, Apr 2002.

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

[4] S. T. Cole, R. Brosch, and J. Parkhill. Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544, 1998.

[5] 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*, 29(6):695–703, Mar 2013.

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

[7] 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*, 5:ra7, 2012.

[8] 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*, 7(9):e1002251, 09 2011.

[9] S. Hasan, S. Daugelat, P. S. Rao, and M. Schreiber. Prioritizing genomic drug targets in pathogens: application to Mycobacterium tuberculosis. *PLoS Comput. Biol.*, 2(6):e61, Jun 2006.

[10] 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.*, 73(4):2533–2540, Apr 2005.

[11] G. Lamichhane, M. Zignol, N. J. Blades, D. E. Geiman, A. Dougherty, J. Grosset, K. W. Broman, and W. R. Bishai. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: Application to mycobacterium tuberculosis. *PNAS*, 100(12):7213–7218, 2003.

[12] 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.

[13] C. M. Sassetti, D. H. Boyd, and E. J. Rubin. Comprehensive identification of conditionally essential genes in mycobacteria. *PNAS*, 98(22):12712–12717, 2001.

[14] C. M. Sassetti, D. H. Boyd, and E. J. Rubin. Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*, 48(1):77–84, 2003.

[15] C. M. Sassetti and E. J. Rubin. Genetic requirements for mycobacterial survival during infection. *PNAS*, 100(22):12989–12994, 2003.

[16] 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.*, 8(9):e1002946, Sep 2012.

[17] 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*, 7(8):e43012, 2012.