Capturing Uncertainty by Modeling Local Transposon Insertion Frequencies Improves Discrimination of Essential Genes

Michael A. DeJesus and Thomas R. Ioerger

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 were obtained for each gene, and the probability of each gene being essential is estimated. We compared 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.

Index 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 [1]. 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 [2], [3], [4], [5], [6]. 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 were being expressed and which ones were not [7], [8], [9]. Although these methods were capable of assessing which genes were 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 assess 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. [10] 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 disrupted). Blades and Broman [11] 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 [12], [13]. 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 [14], [15]. Using the Metropolis Hastings (MH) 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 [12], [16]. Furthermore, because Himar1-based transposons have specificity to TA sites, and the distribution

• *The authors are with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843. E-mail: {michael.dejesus, ioerger}@cs.tamu.edu.*

Manuscript received 21 Nov. 2013; revised 12 May 2014; accepted 20 May 2014. Date of publication 28 May 2014; date of current version 30 Jan. 2015. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2014.2326857

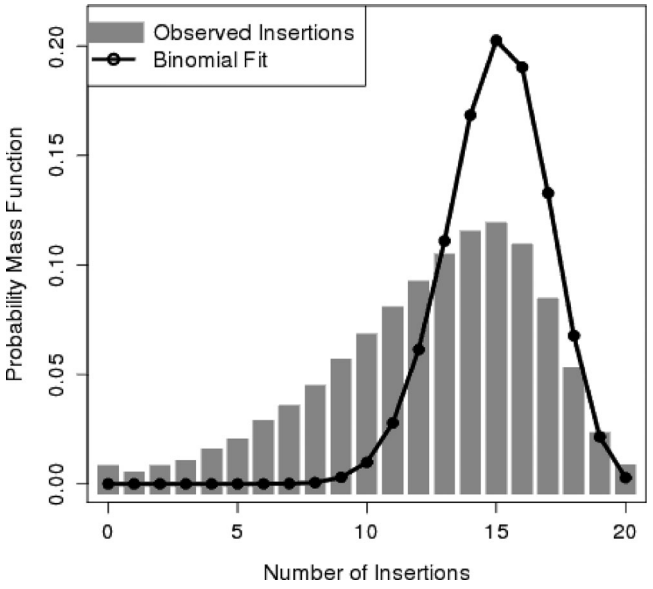

Fig. 1. Histogram of the number of insertions observed within windows of 20 TA sites (gray bars). The Binomial distribution (black line) is incapable of fitting the over-dispersion of observed in the number of insertions.

of TA sites within genes is variable, genes can contain different amount of TA sites within disruptable regions (for example, in the N- and C- termini or within non-essential domains), which can lead to differences in the number of insertions observed.

This variability in insertion probability is evident in libraries of *M. tuberculosis* transposon mutants [14]. Fig. 1 shows a histogram of the observed number of insertions (gray bars) in regions previously predicted to be non-essential [15]. Non-overlapping windows of 20 TA sites were taken across non-essential regions, 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 if the insertion probability were constant. In the case of a constant insertion probability, the number of insertions would be distributed as Binomial$(k \mid n, p)$ (black line). The over-dispersion observed in the data suggests the variability must be due to other factors aside from the stochastic nature of transposon insertions.

The variability in insertion frequency throughout the genome produces an inherent level of uncertainty when modeling the data. We look at the role of uncertainty when determining essentiality of genes. By modeling the local insertion probabilities, the model allows for cases where a gene may contain a lower insertion than average, yet still be considered non-essential. To quantify the uncertainty in the model we measure the entropy in the posterior probability distribution of essentiality for libraries of different levels of saturation.

In this paper, we present a hierarchical 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 were 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 (LFM), 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. We show that modeling individual insertion probabilities results in more conservative predictions which are consistent with expectations for libraries of transposon mutants.

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 were 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 [17].

Under this representation TA sites were 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 were treated as a series of independent trials. In addition, genes were 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 $i \in \{1 \ldots G\}$, let $Y_i = \{k_i, n_i\}$ represent the data for the ith 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 were modeled as a mixture of non-essential and essential genes, with an indicator variable, $Z_i = \{0, 1\}$, indicating whether the ith 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 $\omega_0 = 1 - \omega_1$).

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 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:

$$\prod_{i}^{G} \text{Binomial}(k_i \mid n_i, \theta_i). \tag{1}$$

2.1.2 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:

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

Under this parametrization (i.e., $\alpha = \kappa\rho$ and $\beta = \kappa(1-\rho)$), the 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 were 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 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 were 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 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:

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

Finally, the prior distribution for ω_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:

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

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

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

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 were defined in a similar manner). For an individual insertion probability, the conditional distribution is a Beta distribution with updated parameters:

$$\begin{aligned} &p(\theta_i \mid k_i, \kappa, \rho, Z_i = 1) \\ &\qquad \propto \text{Beta}(\theta_i \mid \kappa_1\rho_1 + k_i,\ \kappa_1(1-\rho_1) + n_i - k_i). \end{aligned}$$

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

$$\begin{aligned} &p(\kappa_1 \mid k_i, \theta_i, \rho_1, Z_i = 1) \\ &\qquad \propto \text{Beta}(\theta_i \mid \kappa_1\rho_1,\ \kappa_1(1-\rho_1)) \times \text{Gamma}(\kappa \mid a_1, b_1), \\ &p(\rho_1 \mid k_i, \theta_i, \kappa_1, Z_i = 1) \\ &\qquad \propto \text{Beta}(\theta_i \mid \kappa_1\rho_1,\ \kappa_1(1-\rho_1)) \times \text{Beta}(\rho_1 \mid \alpha_1, \beta_1). \end{aligned}$$

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

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

where,

$$\begin{aligned} p_1 &= \text{Beta}(\theta_i \mid \kappa_1\rho_1 + k_i,\ \kappa_1(1-\rho_1) + n_i - k_i) \times \omega_1, \\ p_0 &= \text{Beta}(\theta_i \mid \kappa_0\rho_0 + k_i,\ \kappa_0(1-\rho_0) + n_i - k_i) \times (1 - \omega_1). \end{aligned}$$

2.2 Parameter Estimation

In order to estimate parameters of the model and the probability of the genes being essential, samples were obtained through the Metropolis-Hastings algorithm. The Metropolis-Hastings 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 were 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 easily sampled. However, this is not the case for the conditional distributions of the ρ and κ parameters. We use a combination of Gibbs steps and MH steps to obtain samples for all the parameters (see Algorithm 1).

Algorithm 1: 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
$i \in \{1 \dots G\}$
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 step - θ_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 - ρ_0 ;
Draw candidate parameter ρ_0^c from Normal distribution, N(ρ_0^{j-1}, v) and accept according to MH ratio $\frac{f(\rho_0^c)}{f(\rho_0^{i-1})}$;

//MH step - κ_0
Draw candidate parameter κ_0^c from Normal distribution, N(κ_0^{j-1}, v) and accept according to MH ratio $\frac{f(\kappa_0^c)}{f(\kappa_0^{i-1})}$

//MH step - ρ_1
Draw candidate parameter ρ_1^c from Normal distribution, N(ρ_1^{j-1}, v) and accept according to MH ratio $\frac{f(\rho_1^c)}{f(\rho_1^{i-1})}$

//MH step - κ_1
Draw candidate parameter κ_1^c from Normal distribution, N(κ_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 step - 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 θ_i and Z_i for all genes i

3 Results

Our method was applied to deep-sequencing data from mutant libraries of the H37Rv strain of *M. tuberculosis* [14], [15]. The library was grown in minimal media and 0.1 percent 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.

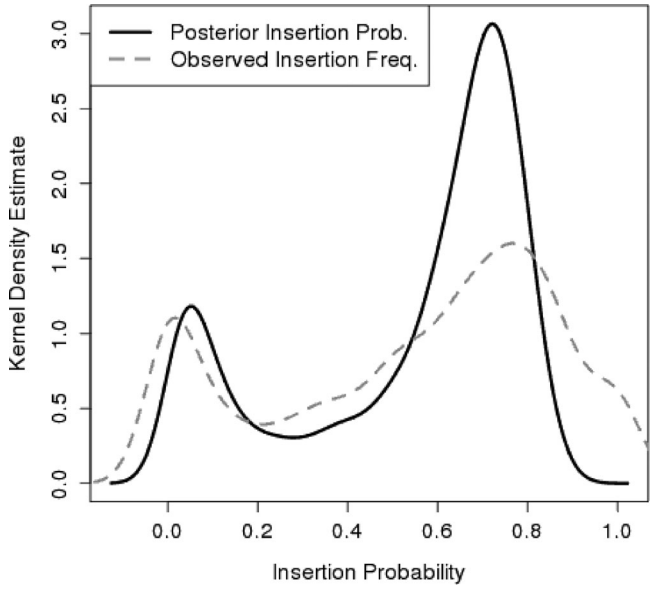

Fig. 2. Kernel density estimates for the mean posterior insertion probability (black-solid) and observed insertion frequency (gray-dashed) for all the genes.

The H37Rv genome is 4.41 million bp long and contains 3,989 open-reading frames (ORFs) [18]. Of these ORFs, 3,947 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 ρ_0 and ρ_1 parameters was 60 and 62 percent, and the acceptance rate for the κ_0 and κ_1 parameters was 67 and 72 percent respectively. Multiple chains of the MH sampler were run in an attempt to verify that the sampler was not trapped in local minima, and was converging to the same area in parameter space.

3.1 Insertion Frequencies

Samples of the individual probabilities were obtained for all genes. The mean insertion frequency, $\overline{\theta}_i$, was estimated from these samples. Fig. 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

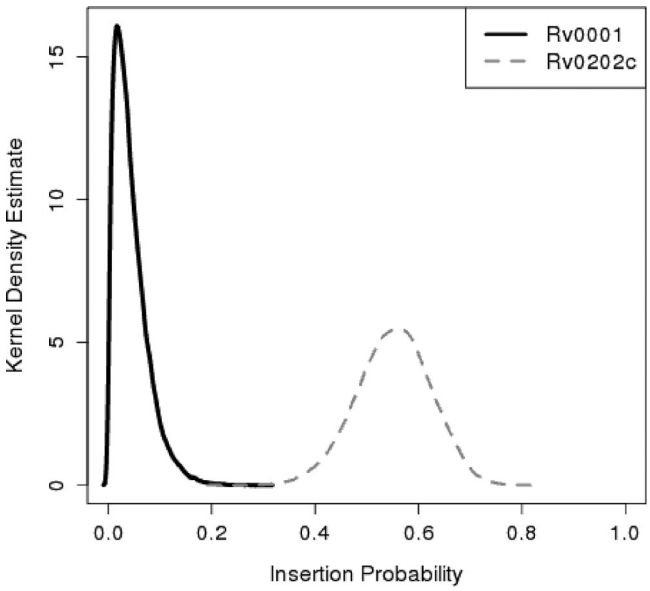

Fig. 3. Kernel density estimates for the posterior insertion probability of DnaA (Rv0001), a known essential gene involved in DNA repair, and MmpL11 (Rv0202c), a known non-essential gene believed to function as a transmembrane protein.

probability resembles the observed frequency, with sharper peaks at the posterior modes.

The samples of insertion probability for the genes reflect our expectations for essential and non-essential genes. Fig. 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 $\overline{\theta}_i = 0.044$, 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 [19]. It contains insertions in 20 out of 39 TA sites, with a mean insertion probability of $\overline{\theta}_i = 0.551$, 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 ($\overline{Z}_i = \frac{1}{n}\sum Z_i$). A method analogous to the Benjamini-Hochberg procedure for posterior probabilities was used to obtain the thresholds of essentiality. While the Benjamini-Hochberg procedure is designed for p-values, it can be modified to control the false discovery rate (FDR) for posterior probabilities [20]. In similar fashion to the Benjamini-Hochberg procedure, posterior probabilities are ordered in ascending order. Starting from the smallest posterior probability, the ratio $\frac{\alpha i}{n}$ is compared to the difference in posterior probability at position i and the mean posterior probability of the previous positions, $1 \ldots i-1$. Setting the FDR at 5 percent, this method produces the following thresholds: genes with $\overline{Z}_i > 0.99304$ were classified as essential, and genes with $\overline{Z}_i < 0.0391$ were classified as non-essential. Those genes that do not meet these thresholds were classified as uncertain.

TABLE 1
Essentiality Comparison between the TraSH Method and the Local Frequency Model

		Local Frequency Model			
		Ess.	Unc.	Non. Ess.	Total
TraSH	Ess.	329	257	28	614
	Growth-Def.	5	20	17	42
	Non. Ess.	36	682	1,796	2,514
	No-Data	80	412	285	777
	Total	450	1,371	2,126	3,947

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 (TraSH) [8], [9]. This method quantifies the amount of fluorescence that is observed in probes that hybridize to each of the genes in the genome [7]. Hybridization ratios were obtained from libraries of *M. tuberculosis* grown in rich media and glucose, and these were used to characterize genes as essential, non-essential or growth-defect (representing those genes for which transposon insertion leads to reduced growth rate). Genes for which the hybridization ratio could not be obtained were classified as "No-Data".

Table 1 shows a comparison of the results from the TraSH method and the Local Frequency Model, presented as a confusion matrix. Of the 614 genes predicted to be essential by TraSH, 28 were predicted to be non-essential by the Local Frequency Model. Although these genes were 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 were 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 were essential in the library analyzed, and the discrepancy may be due to the difference in the construction of the libraries.

A significant difference between the methods is the presence of the "Uncertain" class of genes for the Local Frequency Model. The LFM classifies around a quarter (1,371) of the genes to be uncertain. This is because there is inherent ambiguity in interpreting the insertion patterns in certain regions, necessitating a new category for representing those genes whose essentiality cannot be confidently determined through the data.

Tables 2 and 3 contain comparisons between the TraSH method and the Global Frequency Model (GFM) and 0Extreme Value Model (EVM) described in Sections 3.2.2 and 3.2.3.

Both of these models classify more genes as essential than the TraSH method (709 for the GFM and 668 for the EVM respectively, compared to 614 for TraSH). Approximately 15 percent of those genes are classified as non-essential by TraSH, a higher percentage than the LFM (8

TABLE 2
Essentiality Comparison between the TraSH Method and the Global Frequency Model

		Global Frequency Model			
		Ess.	Unc.	Non. Ess.	Total
TraSH	Ess.	443	110	61	614
	Growth-Def.	12	5	25	42
	Non. Ess.	105	295	2,114	2,514
	No-Data	149	193	435	777
	Total	709	603	2,635	3,947

TABLE 4
Essentiality Comparison between the Global Frequency Model and the Local Frequency Model

		Local Frequency Model			
		Ess.	Unc.	Non. Ess.	Total
GFM	Ess.	450	259	0	709
	Unc.	0	603	0	603
	Non. Ess.	0	509	2,126	2,635
	Total	450	1,371	2,126	3,947

percent). This phenomena is also true for non-essential genes where both the GFM and the EVM classify more genes as non-essential (2,635 and 2,693, respectively), and contain a larger disagreement.

These methods also predict less uncertain genes (603 for the GFM and 342 for the EVM) compared to the Local Frequency Model (1,371). This is partly due to the fact that the Local Frequency Model also estimates individual insertion probabilities, instead of assuming these parameters to be globally shared. However the Global Frequency and Extreme Value Models may also not be adequately capturing the uncertainty that is present in the data. We discuss these results in more depth in the corresponding sections (Sections 3.2.2 and 3.2.3), and discuss the issue of uncertainty in Section 3.3.

3.2.2 Comparison to the Global Frequency Model

To determine the effect of relaxing the assumption of a constant insertion frequency, we compared 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 were obtained, as well as the essentiality assignments Zi. Estimates of the probability of essentiality were 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: $\overline{\theta}_0 = 0.684 \pm 0.002$ and $\overline{\theta}_1 = 0.102 \pm 0.002$, implying a 68.4 percent insertion density in non-essential genes and 10.2 percent in essential genes.

Table 4 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 versus 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. Fig. 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

TABLE 3
Essentiality Comparison between the TraSH Method and the Extreme Value Model

		Extreme Value Model				
		Ess.	Unc.	Non. Ess.	Short	Total
TraSH	Ess.	430	75	81	28	614
	Growth-Def.	9	4	28	1	42
	Non. Ess.	94	151	2,131	138	2,514
	No-Data	135	112	453	77	777
	Total	668	342	2,693	244	3,947

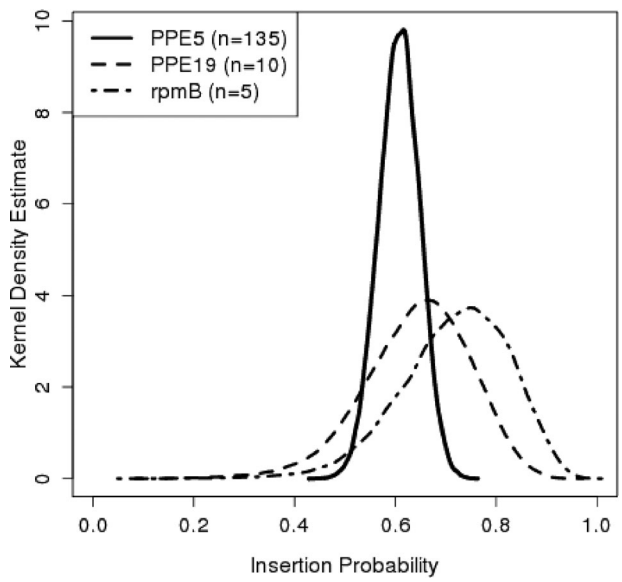

Fig. 4. Insertion density for PPE5 (solid), PPE19 (dashed) and RpmB (dot-dash). All three genes contained an observed insertion frequency of 0.7, although they were different sizes (# of TA sites). The larger variance in the insertion density for PPE19 and rmpB reflects the greater uncertainty that exists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

TABLE 5
Essentiality Comparison between the Extreme Value Model and the Local Frequency Model

		Local Frequency Model			
		Ess.	Unc.	Non. Ess.	Total
EVM	Ess.	446	222	0	668
	Unc.	2	300	40	342
	Non. Ess.	2	685	2,006	2,693
	Short	0	164	80	244
Total	450	1,371	2,126	3,947	

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., $\overline{p}_0 = 0.69$) as they 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 consecutive TA sites lacking insertions. 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 5.

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 Model. Among those genes are members of the MmpL protein family (e.g., MmpL4, MmpL8 and MmpL9), which are believed to be involved transport of lipids and glycolipids. Within this family of proteins, only MmpL3 has been shown to be essential in knockout experiments [19]. The Local Frequency Model classifies MmpL3 as essential and the remaining members of this family of proteins as either uncertain or non-essential. In contrast, the Extreme Value Model classifies MmpL4, MmpL8 and MmpL9 as essential because they contain gaps in insertion pattern that are longer than expected, despite also containing a relatively high insertion frequency elsewhere in the gene. By being more conservative in its predictions, the Local Frequency Model is able to more accurately predict the non-essentiality of genes like most of those in the MmpL family of proteins.

Of the 450 genes classified as essential by the Local Frequency Model, only four of these were 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 eight 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 [21]. 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 [20]), 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 (Fig. 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.

3.3 Capturing Uncertainty in the Data

As mentioned before, the predictions made by the Local Frequency Model are more conservative than the other models studied. This is primarily due to the fact that it estimates the local insertion frequencies for each of the genes. For instance, consider a gene with an insertion count just below the expected insertion frequency for non-essential genes; the Global Frequency Model would classify this gene as essential with relatively high confidence. However, the

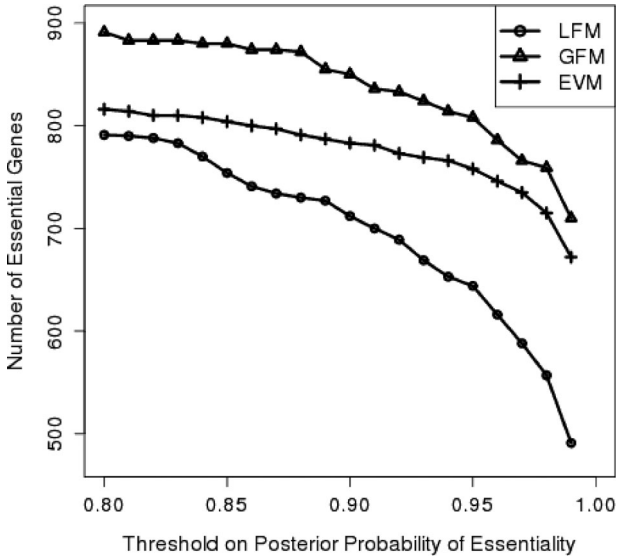

Fig. 5. Number of essential genes predicted by the Local Frequency Model (circle), Global Frequency Model (triangle), and the Extreme Value Model (cross) as a function of the threshold on the probability of essentiality. The Local Frequency Model predicts less essential genes, even as the threshold is relaxed.

extra degree of freedom in the Local Frequency Model (θ_i) offers another explanation for the data. The observed number of insertions could be low for this gene in general, despite being non-essential. Therefore, the Local Frequency Model allows for more variability in the classification of genes, hence the posterior distribution is more diffuse. To visualize this effect on the posterior probability of essentiality, the $\overline{Z}_i$ values for all genes were plotted in ascending order for the different models (Fig. 6). The probability of essentiality for the genes increases much more gradually in the Local Frequency Model compared to the other models. By estimating the local probability of insertion for all genes, instead of depending on a global parameter shared by genes, the model is capable of capturing the uncertainty in

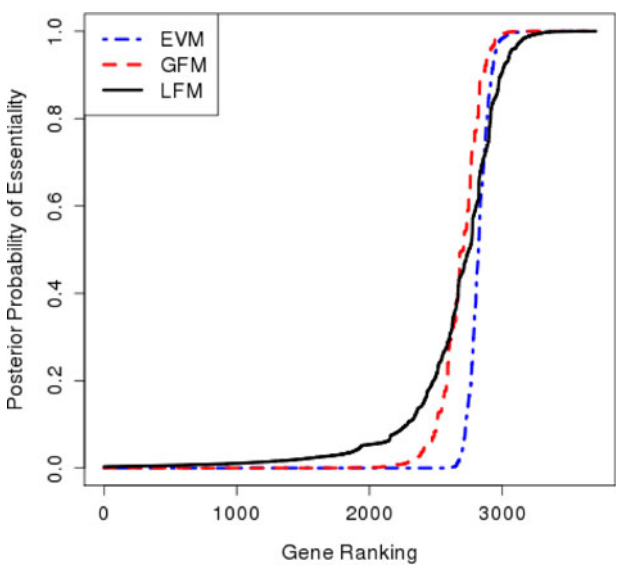

Fig. 6. Posterior probability of essentiality ($\overline{Z}_i$) in ascending order for the Extreme Value Model , Global Frequency Model and Local Frequency Model.

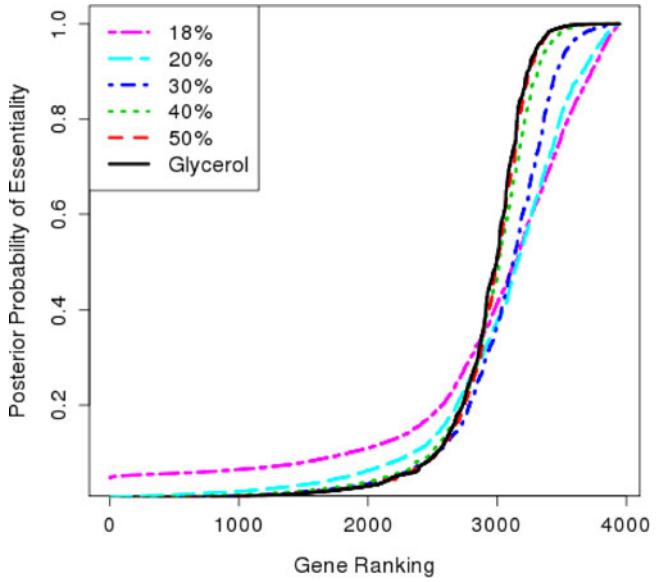

Fig. 7. Posterior probability of essentiality ($\overline{Z}_i$) in ascending order for transposon libraries of different insertion density. Mean insertion density for the original glycerol data set is 52 percent.

the data which would otherwise be missed by models which assume a global insertion probability. In contrast, the Global Frequency Model calls most genes either essential or non-essential with too high of confidence, and very few genes in between (603, $0.0391 < p < 0.99304$; see Table 4) are labeled uncertain.

While the uncertainty captured by the Local Frequency Model is due to the individual insertion probabilities (and the varying sizes of the genes), the uncertainty can also be affected by how many insertions are represented in the transposon library. The lower the saturation of the library (i.e., the less mutants containing insertions at different locations), the more difficult it is to distinguish between essential and non non-essential regions.

To assess the effect of the level of saturation of the library on the estimates of essentiality, the Local Frequency Model was used to analyze libraries spanning different levels of saturation. Libraries were artificially created from the original glycerol library (analyzed above), by setting read-counts of random TA sites to zero throughout the genome. This process was used to create libraries having a saturation (insertion density) ranging from 18 to 50 percent. Fig. 7 shows the resulting posterior probability distributions for these libraries compared to the original glycerol library.

As the saturation of the library decreases, the slope of the resulting distribution of essentiality decreases as well, reflecting the increased uncertainty that occurs with sparser data sets. The increase in uncertainty leads to a decrease in the number of essential genes predicted for the respective libraries. The increased uncertainty is also reflected in the distribution of insertion probabilities for the genes. Fig. 8 shows the distribution of the insertion probabilities estimated by the Local Frequency Model for the different libraries.

The two peaks of the distribution (representing the two classes of genes), slowly shift closer to the global insertion frequency in the library, increasing the uncertainty of classifying genes in between the two peaks. Once the saturation

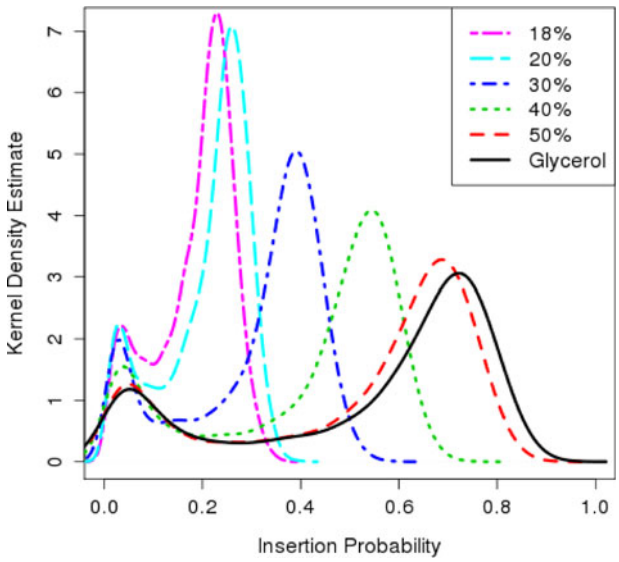

Fig. 8. Kernel density estimate of mean insertion probabilities ($\bar{\theta}_i$) for libraries different levels of saturation. Mean insertion density for the original glycerol data set is 52 percent.

of the library reaches a critical point ($\sim$16 percent) the two peaks collapse into one, making the two category of genes effectively indistinguishable from each other.

To measure the changes in uncertainty, the mean entropy of the posterior probability of essentiality was estimated ($\overline{H} = -\frac{1}{G}\sum_i^G p(Z_i)\log p(Z_i)$) for each of the different data sets. As expected, the entropy increases as the saturation of the data set diminishes (Fig. 9), reflecting the increased uncertainty that is captured by the Local Frequency Model by taking into consideration local insertion frequencies. Table 6 shows a breakdown of the predictions by the Local Frequency Model on the different libraries. The number of uncertain genes increases as the saturation of the library decreases, coinciding with the reduction of the space between peaks in Fig. 8.

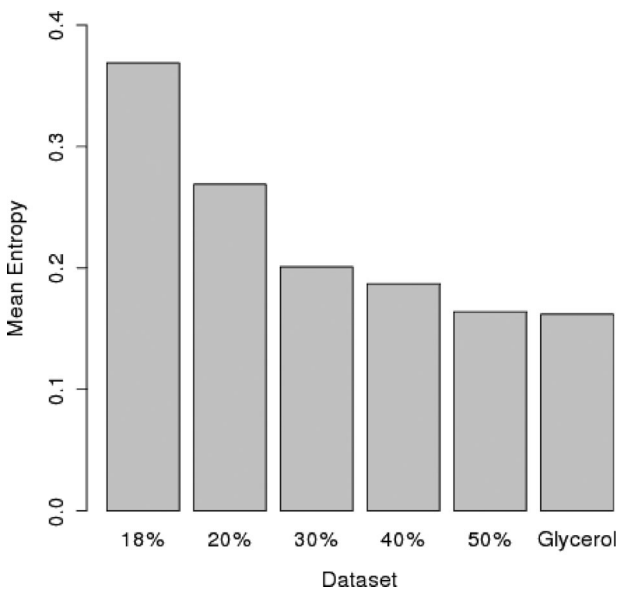

Fig. 9. Shannon entropy of the posterior probability of essentiality for the LFM on data sets with different levels of saturation. Mean insertion density for the original glycerol data set is 52 percent.

TABLE 6
Essentiality Results for Libraries with Different Levels of Saturation

Saturation	Essential	Uncertain	Non-Essential
18%	2	2,409	1,536
20%	2	2,296	1,649
30%	66	1,822	2,059
40%	334	1,592	2,021
50%	438	1,438	2,071
Glycerol	450	1,371	2,126

Mean insertion density for the original glycerol data set is 52 percent.

The higher entropy is also observed amongst the different models, where the Local Frequency Model shows the highest mean entropy (0.162), followed by the Global Frequency Model (0.051) and the Extreme Value Model (0.037).

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 patterns observed in real data, or realistic expectations about the variability of insertions in 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, the observed insertion frequency can be affected by the growth rate of the tranposon mutants. The severity of the growth-impairment resulting from the disruption of these genes will affect the number of viable mutants available for sequencing, and therefore the relative frequency of insertions observed for a given gene.

These effects, in addition to the stochastic nature of transposon insertions, are a source of uncertainty in the insertion data. By modeling the insertion frequency for each gene, these effects can be taken into consideration while they would otherwise be missed by assuming a global insertion frequency. Although increasing the degrees of freedom of a model may lead to over-fitting, we show that the predictions for the Local Frequency Model exhibit higher entropy than alternative models while also matching expectations of essentiality for *M. tuberculosis*. The entropy increases as the saturation of the library decreases, resulting in a increased number of uncertain genes being predicted due to the natural loss of statistical power.

By taking into consideration the uncertainty in the data, more accurate predictions can be made. Previous methods which assumed a global insertion frequency have been

susceptible to these problems. For example, although PE_GRS genes and several MmpL genes have been shown to be non-essential through knock-out experiments [16], some of these genes have been characterized as essential by previous statistical 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 [10], [22]. 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 "gaps", or a series of TA sites lacking insertions in a row. By using the Extreme Value distribution to quantify the statistical significance of these regions lacking insertions, this method was capable of identifying genes which contained essential and non-essential regions (like an essential domain). However, that method assumed a global insertion frequency, which meant it suffers from the limitations outlined before. A promising extension of our work may be to augment the Extreme Value Model to allow for local insertion probabilities at individual genes, therefore capturing the uncertainty that exists in transposon mutagenesis data and improving the assignments of essentiality.

Acknowledgments

This work was supported by the National Institutes of Health [U19 AI107774].

References

[1] S. Hasan, S. Daugelat, P. S. Rao, and M. Schreiber, "Prioritizing genomic drug targets in pathogens: Application to Mycobacterium tuberculosis," *PLoS Comput. Biol.*, vol. 2, no. 6, p. e61, Jun. 2006.

[2] D. J. Lampe, M. E. Churchill, and H. M. Robertson, "A purified mariner transposase is sufficient to mediate transposition in vitro," *Eur. Mol. Biol. Org. J.*, vol. 15, no. 19, pp. 5470–5479, 1996.

[3] E. J. Rubin, B. J. Akerley, V. N. Novik, D. J. Lampe, R. N. Husson, and J. J. Mekalanos, "In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 4, pp. 1645–1650, 1999.

[4] V. Pelicic, S. Morelle, D. Lampe, and X. Nassif, "Mutagenesis of Neisseria meningitidis by in vitro transposition of Himar1 mariner," *J. Bacteriol.*, vol. 182, no. 19, pp. 5391–5398, Oct. 2000.

[5] T. M. Maier, R. Pechous, M. Casey, T. C. Zahrt, and D. W. Frank, "In vivo Himar1-based transposon mutagenesis of Francisella tularensis," *Appl. Environ. Microbiol.*, vol. 72, no. 3, pp. 1878–1885, Mar. 2006.

[6] D. A. Rholl, L. A. Trunck, and H. P. Schweizer, "In vivo Himar1 transposon mutagenesis of Burkholderia pseudomallei," *Appl. Environ. Microbiol.*, vol. 74, no. 24, pp. 7529–7535, Dec. 2008.

[7] C. M. Sassetti, D. H. Boyd, and E. J. Rubin. (2001). Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Nat. Acad. Sci. USA* [Online]. *98(22)*, pp. 12712–12717. Available: http://www.pnas.org/content/98/22/12712.abstract

[8] C. M. Sassetti, D. H. Boyd, and E. J. Rubin, "Genes required for mycobacterial growth defined by high density mutagenesis," vol. 48, no. 1 pp. 77–84, 2003.

[9] C. M. Sassetti and E. J. Rubin. (2003). Genetic requirements for mycobacterial survival during infection. in *Proc. Nat. Acad. USA* [Online]. *100(22)*, pp. 12989–12994. Available: http://www.pnas.org/content/100/22/12989.abstract

[10] 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, no. 9, p. e1002946, Sep. 2012.

[11] N. J. Blades and K. W. Broman. (2002, Jul.). Estimating the number of essential genes in a genome by random transposon mutagenesis. Dept. Biostatist. Working Papers, Johns Hopkins Univ., Tech. Rep. MSU-CSE-00-2, Jul. 2002, [Online]. Available: http://biostats.bepress.com/jhubiostat/paper15

[12] G. Lamichhane, M. Zignol, N. J. Blades, D. E. Geiman, A. Dougherty, J. Grosset, K. W. Broman, and W. R. Bishai. (2003). A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: Application to Mycobacterium tuberculosis. *Proc. Nat. Acad. Sci. USA* [Online]. *100(12)*, pp. 7213–7218. [Online]. Available: http://www.pnas.org/content/100/12/7213.abstract

[13] 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, no. 4, pp. 2533–2540, Apr. 2005.

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

[15] 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, no. 6, pp. 695–703, Mar. 2013.

[16] 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.

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

[18] 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.

[19] 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.

[20] P. Muller, G. Parmigiani, and K. Rice, "FDR and bayesian multiple comparisons rules," in *Proc. ISBA 8th World Meeting Bayesian Stat.*, Benidorm, Spain, Jun. 2006, pp. 1–18.

[21] 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.

[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.

Michael A. DeJesus received the bachelor's and master's degrees in computer science from the University of Puerto Rico, Mayaguez, and Texas A&M University, in 2008 and 2012, respectively. He is currently working toward the PhD degree in computer science at Texas A&M. His research interests include machine learning and statistical pattern recognition techniques to analyze sequence data.

Thomas R. Ioerger received the BS degree (honors) in molecular and cell biology from the Pennsylvania State University in 1989, and the MS and PhD degrees in computer science from the University of Illinois in Urbana-Champaign, the latter in 1996. He is an associate professor in the Department of Computer Science and Engineering at Texas A&M University. His primary research interests include bioinformatics and machine learning.