Supplementary Material

Michael A. DeJesus
Thomas R. Ioerger

December 3, 2012

Contents

1 Derivation of Probability Densities

Let $Y_i = \{r_i, n_i\}$ represent our observations for the i-th gene for $i = 1...G$, where r_i and n_i represent the total number of TA sites and the largest run of non-insertions observed in each gene. The essentiality assignments for all genes is represented by the latent vector $Z = \langle Z_1, Z_2, Z_3...Z_G \rangle$, with the individual assignment for i-th gene represented by the boolean variable Z_i which accepts binary values of 0 and 1 for non-essential and essential. These two classes of genes represent the two categories found in the mixture model. The mixture coefficient representing the prevalence of the category in the mixture is given by $\omega = \langle \omega_1, \omega_0 \rangle$. Finally, we assume a global non-insertion probability, ϕ_0, that governs probability of non-insertions across all non-essential genes.

1.1 Likelihood

We assume independence among genes, so our likelihood can be written as a product of our individual observations:

$$\begin{aligned} p(Y \mid Z, \phi_0, \omega_1) &= \prod_{i=1}^{non} p(Y_i \mid Z_i, \phi_0, \omega_1) \times \prod_{i=1}^{ess} p(Y_i \mid Z_i, \phi_0, \omega_1) \\ &= \prod_{i=1}^{non} p(r_i, s_i \mid Z_i, \phi_0, \omega_1) \times \prod_{i=1}^{ess} p(r_i, s_i \mid Z_i, \phi_0, \omega_1) \\ &= \prod_{i=1}^{non} p(r_i \mid Z_i, \phi_0) \times p(s_i \mid r_i, Z_i) \times \prod_{i=1}^{ess} p(s_i \mid Z_i) \times p(r_i \mid s_i, Z_i) \\ &= \prod_{i=1}^{non} Gumbel(r_i \mid \mu, \tau) \times N(s_i - \lambda_r r_i, \sigma_r^2) \times \prod_{i=1}^{ess} Sigmoid(s_i) \times N(r_i - \lambda_s s_i, \sigma_s^2) \end{aligned} \tag{1}$$

1.2 Prior Probabilities

To quantify our prior expectations of this parameter, we use a Beta distribution as our prior:

$$\pi(\phi_0) = Beta(\phi_0; \alpha_0, \beta_0) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \phi_0^{\alpha_0 - 1} (1 - \phi_0)^{\beta_0 - 1} \tag{2}$$

where α_0 and β_0 are hyper-parameters that capture our expectations for ϕ_0. The prior probability of a individual essentiality assignment, Z_i, depends on the mixing coefficient ω_1, and is given by a Bernoulli distribution with probability ω_1:

$$\pi(Z_i \mid \omega_1) = Bernoulli(Z_i; \omega_1) = \omega_1^{Z_i} (1 - \omega_1)^{1 - Z_i}$$

The prior probability of an essentiality assignment for all genes, Z, is therefore given by a product of Bernoulli trials with probability ω_1:

$$\pi(Z \mid \omega_1) = \prod_{i}^{G} Bernoulli(\omega_1) = \omega_1^{K_z} (1 - \omega_1)^{G - K_z}$$

where G is the total number of genes, and K_z is the sum of the binary vector of essentiality assignments (i.e., $K_z = \Sigma Z_i$). Finally, our prior expectations for the mixing coefficient ω_1 are given by a Beta distribution:

$$\pi(\omega_1) = Beta(\omega_1; \alpha_w, \beta_w) = \frac{\Gamma(\alpha_w + \beta_w)}{\Gamma(\alpha_w)\Gamma(\beta_w)} \omega_1^{\alpha_w - 1} (1 - \omega_1)^{\beta_w - 1}$$

1.3 Full Joint Distribution

Using the data-likelihood and prior distribution, we derive a full joint distribution as follows:

$$
\begin{aligned}
p(Y, Z, \phi_0, \omega_1) &= \prod_{i=1}^{non} p(Y_i \mid Z_i, \phi_0, \omega_1) \times \prod_{i=1}^{ess} p(Y_i \mid Z_i, \phi_0, \omega_1) \times \pi(\phi_0) \times \pi(Z \mid \omega_1) \times \pi(\omega_1) \\
&= \prod_{i=1}^{non} p(r_i, s_i \mid Z_i, \phi_0, \omega_1) \times \prod_{i=1}^{ess} p(r_i, s_i \mid Z_i, \phi_0, \omega_1) \times \pi(\phi_0) \times \pi(Z \mid \omega_1) \times \pi(\omega_1) \\
&= \prod_{i=1}^{non} p(r_i \mid Z_i, \phi_0) \times p(s_i \mid r_i, Z_i) \times \prod_{i=1}^{ess} p(s_i \mid Z_i) \times p(r_i \mid s_i, Z_i) \times \pi(\phi_0) \times \pi(Z \mid \omega_1) \times \pi(\omega_1) \\
&= \prod_{i=1}^{non} Gumbel(r_i \mid \mu, \tau) \times N(s_i - \lambda_r r_i, \sigma_r^2) \times \prod_{i=1}^{ess} Sigmoid(s_i) \times N(r_i - \lambda_s s_i, \sigma_s^2) \\
&\times Beta(\phi_0; \alpha_0, \beta_0) \times \prod Bernoulli(\omega_1) \times Beta(\omega_1; \alpha_w, \beta_w)
\end{aligned} \tag{3}
$$

1.4 Conditional Distributions

In order to obtain posterior estimates of essentiality, we derive conditional probability densities from which to sample from, using proportionality, we remove terms that do not depend on the variable of interest. First we derive the conditional distribution for the ϕ_0 parameter, which our method depends on:

$$
\begin{aligned}
p(\phi_0 \mid Y, Z, \omega_1) &\propto \prod_{i=1}^{non} p(Y_i \mid Z_i, \phi_0, \omega_1) \times \prod_{i=1}^{ess} p(Y_i \mid Z_i, \phi_0, \omega_1) \times \pi(\phi_0) \times \pi(Z \mid \omega_1) \times \pi(\omega_1) \\
&\propto \prod_{i=1}^{non} p(r_i, s_i \mid Z_i, \phi_0, \omega_1) \times \pi(\phi_0) \\
&\propto \prod_{i=1}^{non} p(r_i \mid Z_i, \phi_0) \times p(s_i \mid r_i, Z_i) \times \pi(\phi_0) \\
&\propto \prod_{i=1}^{non} p(r_i \mid Z_i, \phi_0) \times \pi(\phi_0) \\
&\propto \prod_{i=1}^{non} Gumbel(r_i \mid \mu, \tau) \times Beta(\phi_0; \alpha_0, \beta_0)
\end{aligned} \tag{4}
$$

Next, we derive conditional distributions for the essentiality of an individual gene i, starting with the case where gene i is non-essential (i.e, $Z_i = 0$):

$$
\begin{aligned}
&p(Z_i = 0 \mid Y, Z_{\{-i\}}, \phi_0, \omega_1) \propto p(Y_i \mid Z, \phi_0, \omega_1) \times \pi(\phi_0) \times \pi(Z_{\{-i\}}, \omega_1) \times \pi(Z_i = 0 \mid \omega_1) \times \pi(\omega_1) \\
&\propto p(r_i, s_i \mid Z_i, \phi_0, \omega_1) \times \pi(Z_i = 0 \mid \omega_1) \\
&\propto p(r_i \mid Z_i, \phi_0) \times p(s_i \mid r_i, Z_i) \times \pi(Z_i = 0 \mid \omega_1) \\
&\propto Gumbel(r_i; \mu, \tau) \times N(s_i - \lambda_r r_i, \sigma_r^2) \times Bernoulli(Z_i = 0; \omega_1) \\
&\propto Gumbel(r_i; \mu, \tau) \times N(s_i - \lambda_r r_i, \sigma_r^2) \times (1 - \omega_1)
\end{aligned} \tag{5}
$$

followed by the case where gene i is essential (i.e, $Z_i = 1$):

$$
\begin{aligned}
&p(Z_i = 1 \mid Y, Z_{\{-i\}}, \phi_0, \omega_1) \propto p(Y_i \mid Z, \phi_0, \omega_1) \times \pi(\phi_0) \times \pi(Z_{\{-i\}}, \omega_1) \times \pi(Z_i = 1 \mid \omega_1) \times \pi(\omega_1) \\
&\propto p(r_i, s_i \mid Z_i, \phi_0, \omega_1) \times \pi(Z_i = 1 \mid \omega_1) \\
&\propto p(s_i \mid Z_i) \times p(r_i \mid s_i, Z_i) \times \pi(Z_i = 1 \mid \omega_1) \\
&\propto Sigmoid(s_i) \times N(r_i - \lambda_s s_i, \sigma_s^2) \times Bernoulli(Z_i = 1; \omega_1) \\
&\propto Sigmoid(s_i) \times N(r_i - \lambda_s s_i, \sigma_s^2) \times \omega_1
\end{aligned} \tag{6}
$$

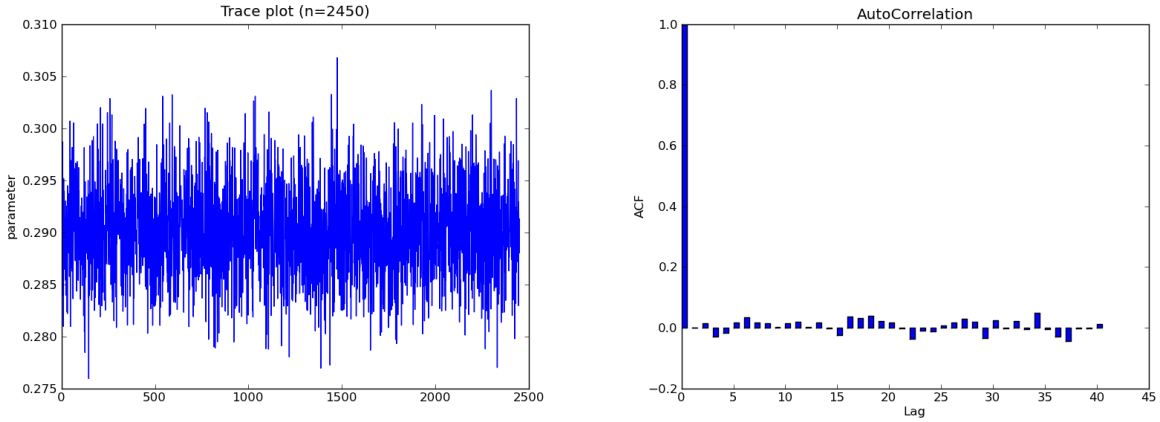
Finally, we derive a conditional distribution to sample the mixing coefficient ω_1 for the essential component:

$$\begin{aligned} p(\omega_1 \mid Y, Z, \phi_0) &\propto \pi(Z \mid \omega_1) \times \pi(\omega_1) \\ &\propto \prod Bernoulli(\omega_1) \times Beta(\omega_1; \alpha_w, \beta_w) \\ &\propto Beta(\omega_1; \alpha_w + K_{z1}, \beta_w + G - K_{z1}) \end{aligned} \tag{7}$$

2 Model Diagnostics

2.1 Convergence

Our statistical analysis depends on obtaining an MCMC sample of the ϕ_0 parameter (i.e., probability of non-insertion in non-essential genes) to estimate posterior probabilities of essentiality. We obtain estimates of ϕ_0 by sampling its conditional probability given the data through a MH sampling procedure. To ensure that values being sampled were uncorrelated and that the MH sampling procedure was mixing well, we discard the first 1,000 samples as a "burn-in" period and keep only every 20-th sample there after. Figure S1 presents a trajectory plot of the ϕ_0 sample, showing that our MH sampling explored the sample space without getting "stuck" in a particular region.



(a) Trace Plot of Sample for the ϕ_0 Parameter.

(b) Auto-Correlation of Sample for the ϕ_0 Parameter.

Figure S1: Convergence Diagnostics

To test whether or not the final sample contained uncorrelated values, we calculated the autocorrelation coefficient to a maximum lag of 50 for the sample of ϕ_0 values (See Figure S1 (a). The low values show that samples at $\Delta t \geq 1$ apart are effectively uncorrelated.

2.2 Sensitivity analysis

To assess the sensitivity of this result to the fixed parameters in the likelihood function for essential genes, we obtained results for different values of of d and k parameters of the sigmoid function. Figure 1 shows a cumulative plot of Z_i values for different combinations of these parameters, with horizontal black lines representing the final thresholds of essentiality and non-essentiality. As can be seen by this figure, the k parameter has little effect on the final result. On the other hand, a two-fold increase and decrease of the δ parameter significantly changes the slope of the graph as well as then number of non-essential genes estimated. This is consistent with the fact that the d parameter represents the expected span of nucleotides for essential domains. This has been empirically determined to be approximately 300 nucleotides.

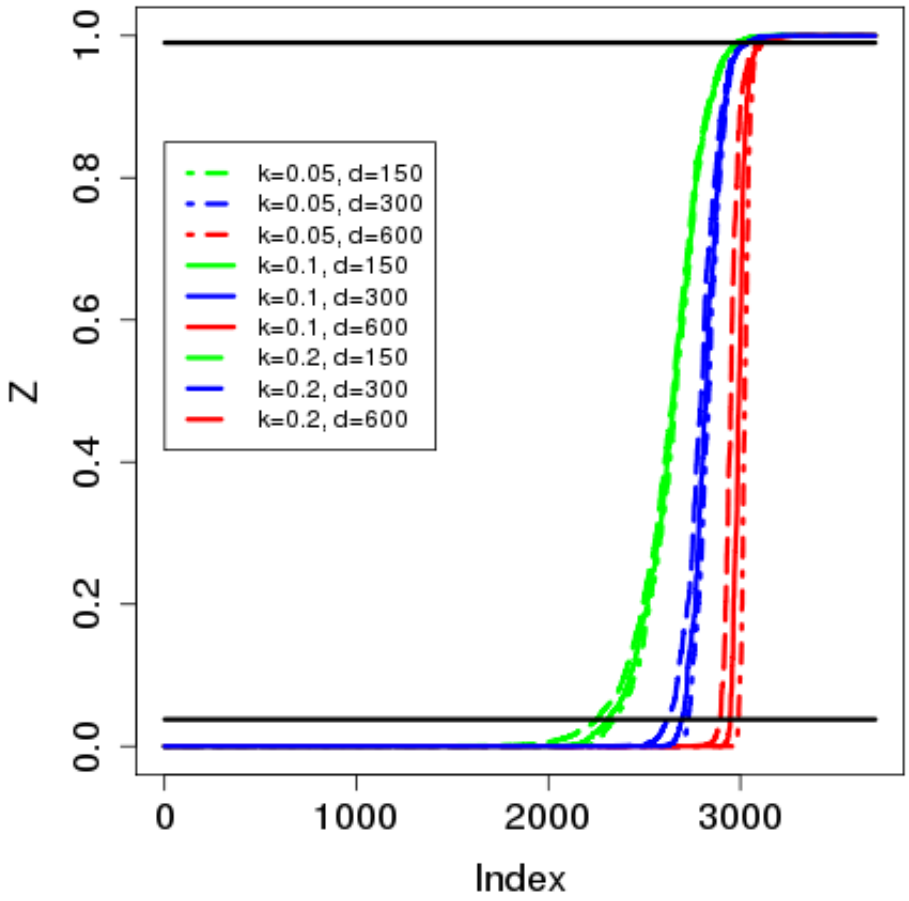


Figure S2: Sensitivy analysis. The MH sampling procedure was run for different values of κ and δ, to determine the sensitivity of the results to these fixed hyper-parameters. Green, Blue and Red lines correspond to $\delta = \{150, 300, 600\}$. Dot-Dash lines, Full lines, and Dashed lines correspond to $\kappa = \{0.05, 0.1, 0.2\}$.

3 Low Density Dataset

To evaluate our method on other datasets with different insertion density, we ran our analysis a different library of *M. tuberculosis* mutants grown on glycerol (prepared by Y. Zhang), but with a significantly smaller frequency of insertions. This dataset contained fewer transposon insertions in coding regions (i.e., 23,399 - 36.3% - compared to 31,715 - 50.4% in our first glycerol dataset), and therefore the set of TA sites in the genome were undersampled. Table S4 contains a comparison of the results of our method on both datasets.

Table S4: Comparison Between Low Density Dataset and Regular Dataset.

		Low Density Dataset				
		Essential	**Intermediate**	**Non-Essential**	**No-Data**	**Total**
Regular	Essentials	351	348	58	0	757
	Uncertain	16	163	63	0	242
	Non-Essential	49	794	1860	0	2703
	No-Data	1	8	79	199	287
	Total	417	1313	2060	199	3989

Our method finds a significantly smaller number of essentials in the low density dataset. The probability of non-insertion estimated for the low density dataset was $\phi_0 = 0.592$, which is significantly higher than the probability estimated in the original glycerol dataset, $\phi_0 = 0.290$. Because of this higher probability of non-insertion, all genes in the low density dataset are expected to have longer runs of non-insertions, even those which can withstand disruption. However, the Bayesian model is capable of compensating for the lower insertion frequency observed in low density datasets by increasing the expected length of the maximum run of non-insertions, and requiring significantly larger runs of non-insertion than would be required in a regular or high density dataset.

4 Comparison to Other Statistical Methods

4.1 H. influenza - Gawronski et al.

In order to compare to other methods, we analyzed data from *H. influenza* and compared our results to those obtained by Gawronski *et al.* (2009). Gawronski et al. characterized 135 genes in *H. influenza* specifically required for infection by comparing mutants grown in vitro to mutants grown in vivo (mouse lung). Their analysis compared the fraction of TA sites with insertions in vitro and the survival index, which was the ratio of the number of reads in vivo versus the number of reads in vitro. They required genes to sustain insertions in $> 40\%$ of their TA sites in vitro (to determine non-essentiality in vitro), and set a threshold of < 0.30 for the survival index to determine genes who had significantly less reads in vivo.

We applied our method to both of these data sets (in vivo, SD3, and in vitro, SD2) to determine essentiality. Although Gawronski et al. (2009) considered only insertions that occurred within the 5-80% coding region of the genes, we consider the entire gene, as our method is capable of characterizing essentiality without having to rule out certain portions of the gene a priori.

Of the 135 characterized by Gawronski *et al.* (2009) as uniquely essential for growth in lung, all 135 were found to be non-essential in vitro by our method, and 69 were found to be essential in lung (out of a total of 122 genes found to be uniquely essential in vivo by our method). These include genes like *galU*, *orfH*, *opsX* and *rfaF* which are necessary for capsule synthesis and are known to be required for infection (Gawronski *et al.*, 2009).

Although our method does not label all 135 genes as essential in lung, it does identify various genes that might be essential for infection (e.g. *dppB*, *dppC*, *dppF* - components of a transporter involved in

heme uptake (Morton *et al.*, 2009)), which were not identified by the method used by Gawronski *et al.* (2009). One possible reason for the discrepancy between the two methods is that our method does not distinguish genes which are non-essential from those genes that cause growth-defects when disrupted, whereas the method used by Gawronkski et al. identifies genes with suppressed read counts. For example, *cysK* showed a survival index of 0.283 (i.e 71 reads in vitro, and only 20 reads in vivo) despite having insertions at 8 TA sites, which suggests disruption of this gene may lead to a growth defect. We adopt a stricter view of essentiality and do not include these as essential genes.

In addition, the method utilized by Gawronski et al. requires a 40% insertion frequency for a gene to qualify as non-essential in vitro. This strict cut-off could lead to some non-essential genes being treated as essential in vitro (and therefore not uniquely essential in lung) because they incurred slightly less insertions. For instance, *atpA*, *atpH*, and *atpF* (components of ATP synthase) do not meet this threshold (with insertion frequencies of 0.28, 0.36, and 0.23 respectively, in vitro) and are treated as essential in vitro, yet these genes show a significant reduction in reads in lung (survival indexes of 0.14, 0.0, and 0.0 respectively, and t hus should be counted as essential in vivo). Note that *atpD* and *atpG* were found to be uniquely essential for infection in lung, as they contained enough insertions in vitro to meet this threshold.

4.2 Blades and Broman

To evaluate our method against other statistical models proposed in the literature, we obtained the R package "*negenes*" maintained by Karl W. Broman (http://www.biostat.wisc.edu/ kbroman/software/), which implements the Gibbs Sampling procedure described by Blades and Broman (2002). The Gibbs sampler was run for 50,000 iterations on the same H37Rv glycerol dataset analyzed in the main manuscript, using the number of insertions within the N-terminal 80% of the ORF, as representative of the number of viable mutants with insertions in genes.

Figure S3 shows a plot of the posterior probability of essentiality obtained as a function of the number of TA sites in each gene. The distribution of posterior probabilities resembles the distribution shown in Figure 2 of Lamichhane *et al.* (2003), who used this method to characterize essentiality in low-density datasets of *M. tuberculosis*. Our distribution of posterior probabilities shows a sharper slope of essential genes due to the larger insertion density (38,984 insertions) in our glycerol dataset, whereas their dataset was much sparser (1,425 insertions), producing lower posterior probabilities for many essential genes.

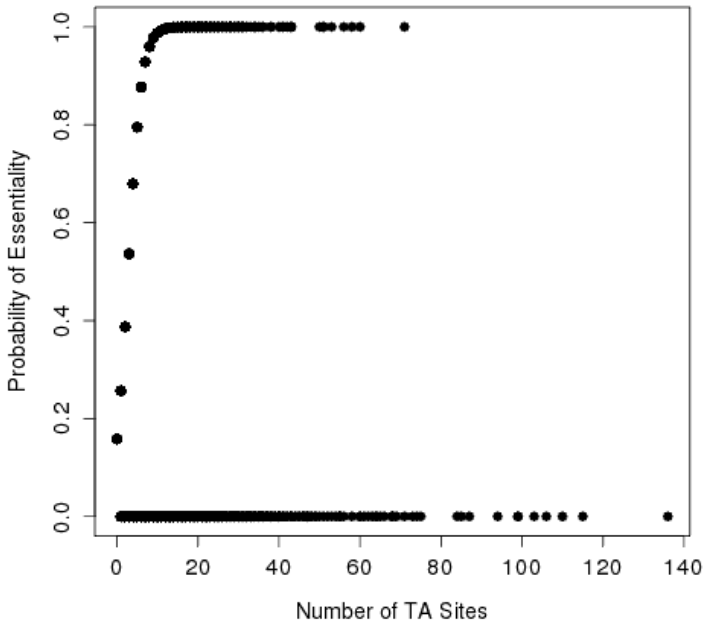


Figure S3: Plot of the posterior probability of essentiality obtained by the Blades and Broman method, and the number of TA sites within the genes. Non-Essential genes appear at the bottom of the graph.

Table S5 shows a comparison of the results obtained by both methods, and the results obtained by Sassetti et al. Genes were classified as essential, non-essential and uncertain by implementing a procedure analogous to the Benjamini and Hochberg procedure for posterior probabilities (Muller *et al.*, 2006). The Blades and Broman method predicts a significantly smaller number of essential (244 compared to 614 defined by Sassetti) and a significantly larger number of non-essential genes (3194 compared to 2520 defined by Sassetti). This is because even a small number of insertions (i.e. ≥ 1) is enough to make a gene appear non-essential under this model.

Table S5: Essentiality Predictions of the Blades and Broman method and our Gumbel method, compared to the results obtained by Sassetti et al.

		Blades and Broman / Gumbel Method				
		Essential	**Uncertain**	**Non-Essential**	**No-Data**	**Total**
Sassetti	**Essentials**	195 / 429	251 / 75	167 / 81	1 / 29	614
	Growth-Defect	2 /9	5 / 4	35 / 28	0 / 1	42
	Non-Essential	13 / 94	117 / 151	2384 / 2131	6 / 144	2520
	No-Data	34 / 135	135 / 112	608 / 453	36 / 113	813
	Total	244 / 667	508 / 342	3194 / 2693	43 / 287	3989

5 Table S1

Essentiality Analysis of H37Rv. Posterior probability of essentiality is reported for all 3989 ORFs in H37Rv, along with the number of insertions, the number of TA sites, and the length of the maximum run of non-insertions observed.

6 Table S2

Pfam Domains with Runs of Non-Insertion Matching Domain Boundaries. Pfam domain predictions where matched to the closest run of non-insertions, with 95 known domains identified that showing consistency between a significant run of non-insertions and the boundaries of the predicted domains.

7 Table S3

Genes with Essential and Non-Essential Domains. A set of 36 genes that labeled as essential by our Bayesian analysis that also contain a significant number of insertions (i.e. $p > 0.05$) according to the cumulative Binomial distribution, representing those genes most likely to contain both essential and non-essential domains.

8 Figure S4

Figure S4: The histogram of the number of read counts per TA site. The distributions of reads shows an exponential (Poisson-like) distribution, with a single mode at 0 corresponding to the abundance of TA sites lacking any reads.

9 Additional Files

-*gumbelMH.py*:	Python implementation of the Bayesian essentiality analysis.
-*MH_tools.py*:	Python module used by *gumbelMH.py*.
-*README.html*:	Instruction manual.
-*trash_glycerol.igv*:	IGV formatted counts of reads at TA sites from the glycerol dataset. (2.5MB)
-*example_output.txt*:	Example output using *trash_glycerol.igv*

References

Blades, N. J. and Broman, K. W. (2002). 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.

Gawronski, J. D., Wong, S. M. S., Giannoukos, G., Ward, D. V., and Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for haemophilus genes required in the lung. *PNAS*, **106**(38), 16422–16427.

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

Morton, D. J., Seale, T. W., Vanwagoner, T. M., Whitby, P. W., and Stull, T. L. (2009). The dppBCDF gene cluster of Haemophilus influenzae: Role in heme utilization. *BMC Res Notes*, **2**, 166.

Muller, P., Parmigiani, G., and Rice, K. (2006). Fdr and bayesian multiple comparisons rules. In *Proceedings of the ISBA 8th World Meeting on Bayesian Statistics*, Benidorm, Spain.