# Reducing Type I Errors in Tn-Seq Experiments by Correcting the Skew in Read Count Distributions

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

## Abstract

Sequencing of transposon-mutant libraries using next-generation sequencing (Tn-Seq) has become a popular method for determining which genes and non-coding regions are essential for growth under various conditions in bacteria. For methods that rely on comparison of read-counts at transposon insertion sites, proper normalization of Tn-Seq datasets is vitally important. Real Tn-Seq datasets often exhibit a significant skew and can be dominated by high counts at a small number of sites (often for non-biological reasons). If two datasets that are not appropriately normalized are compared, it might cause the artifactual appearance of conditionally essential genes in a statistical test, constituting type I errors (false positives). In this paper, we propose a novel method for normalization of Tn-Seq datasets that corrects for the skew in read count distributions by fitting them to a Beta-Geometric distribution. We show that this read-count correction procedure reduces the number of false positives when comparing replicate datasets grown under the same conditions (for which no genuine differences in essentiality are expected).

## 1 Introduction

Sequencing of transposon-mutant libraries using next-generation sequencing has become a popular method for determining which genes and non-coding regions are essential for growth under various conditions in bacteria [8, 5, 6]. Briefly, a transposon-mutant library is made by transfecting in a vector carrying a transposable element, such as the Himar1 transposon, which can insert at random locations throughout the genome (Himar1 can insert at any TA dinucleotide; for reference, there are ∼75,000 TA sites distributed throughout the *M. tuberculosis* genome, spaced ∼60 bp apart on average). Each mutant has an insertion at a single location, but in a saturating library, nearly all of the potential insertion sites are represented. However, when grown under selective conditions, mutants with transposon insertions in essential regions will fail to survive. The abundance of the remaining insertion sites can be determined by using PCR to amplify the junctions between the transposon and the surrounding genome, and the position of each insertion can be efficiently determined using a next-generation sequencer such as an Ilumuna HiSeq. This experiment typically yields several million reads, and the number of reads associated with each TA site is tabulated. While TA sites in non-essential regions have stochastically varying read counts, essential genes and non-coding regions (such as tRNAs, rRNAs, and ncRNAs) can be identified as regions where the TA sites are uniformly devoid of insertions (i.e. read counts are 0).

Determining which genes are essential is a difficult problem. The primary challenge is in lower-density datasets, where the fraction of TA sites represented in the library is in the 20-30% range. The lower the density of the dataset, the more difficult it is to determine whether a region lacks insertion due to essentiality, or just due to random statistical fluctuations. In addition, not all TA sites in an essential gene must lack insertions, as insertions can sometimes be tolerated in the N- or C-terminus of an essential gene, or in non-essential domains or linkers between domains.

Despite these challenges, several statistical methods have been developed for quantifying the significance of essential genes. One method uses a non-parametric test to identify regions with significantly suppressed numbers of insertions (sums of read counts within a sliding window of fixed size) [9]. A Bayesian model employing a likelihood function based on the extreme value distribution has been used to quantify the statistical significance of essential regions based on the length of 'gaps', or consecutive TA sites lacking insertions [3]. Hidden Markov Models have also been developed for analyzing Tn-Seq data [2]. For comparisons between growth conditions, the Negative Binomial distribution has been used to compare read counts in genes between conditions, to determine a $p$-value for assessing significance of conditionally essential genes [10].

For methods that rely on comparisons of read-counts,

proper scaling (or normalization) of Tn-Seq datasets is vitally important. If two datasets are compared that are not appropriately normalized, it might result in erroneous detection of differentially essential genes (false positives), even between replicate datasets grown under the same condition.

In this paper, we propose a novel method that corrects for the skew in read count distributions observed in many Tn-Seq datasets by fitting them to a Beta-Geometric distribution. We show that this read-count correction procedure reduces the number of false positives when comparing replicate datasets grown under the same conditions (for which no genuine differences in essentiality are expected).

## 2 Normalization of Tn-Seq Datasets

The most common method for normalization is to divide the read counts at each TA site by the overall number of reads in a dataset, which factors out gross differences due to the amount of data collected. A refinement of this approach is to scale the read counts to have the same mean over non-zero sites (which we call 'Non-Zero Mean' normalization or NZMean), since different datasets can have widely varying levels of saturation, and distributing the same number of reads over fewer TA sites will naturally inflate the mean read count among them.

Despite these attempts at normalization, Tn-Seq datasets can still display quite different statistical profiles. In practice, some datasets appear well-behaved, where the distribution of read counts tends to fit a simple geometric distribution, while other datasets are skewed, with a few highly over-represented sites dominating the read-count distribution. While there is not a rigorous argument for why the distribution of read counts must be geometric, it is clear that in most datasets, TA sites with only a few reads (1-10) are highly abundant, while sites with high counts ($> 1000$) are much less abundant.

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

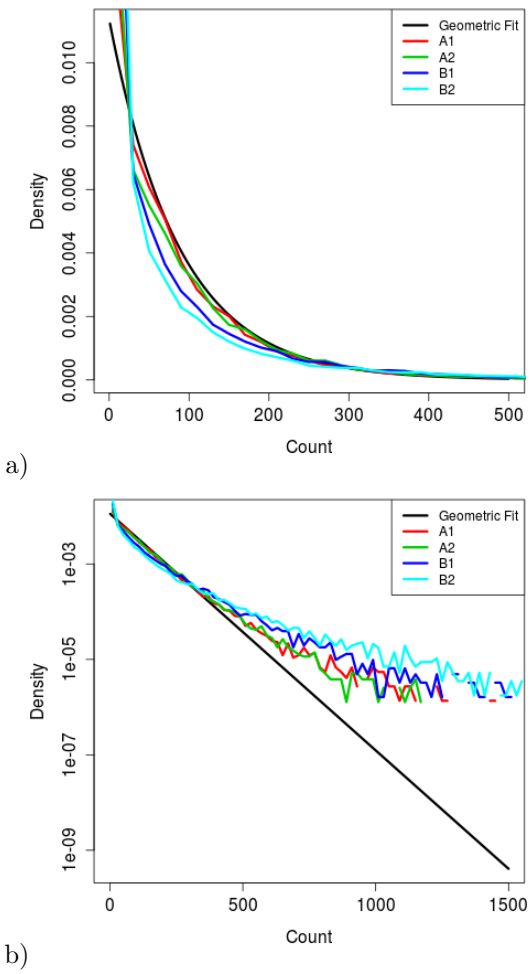



Figure 1: a) Histogram of read counts for four Tn-Seq datasets from an *M. tuberculosis* Tn-mutant library. The black line represents an ideal Geometric fit. b) Histogram of read counts on a log scale.

The effect of the skew observed in datasets like B1 and B2 (which is a common phenomenon in Tn-Seq) is that it can bias the statistical analysis of essential regions, especially for methods that depend on the read counts. Certainly, for genes containing TA sites with high spikes in read counts, they will appear highly non-essential, and this could cause those genes to appear differentially essential in other conditions by comparison. Conversely, the spikes in read counts at some TA sites will suppress the apparent level of reads at other sites, potentially making them appear

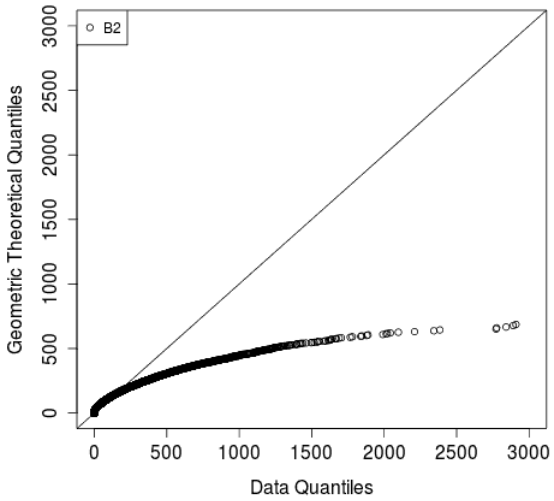



Figure 2: QQ-plot of non-zero read counts for dataset B2 versus a geometric distribution with optimal (MLE) parameter. The skew indicates there is still lack of fit.

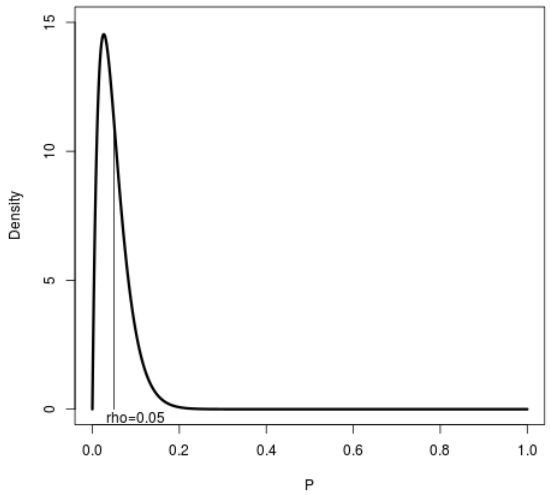
relatively more essential.

We propose a novel method for correcting for this skew in read-count distributions by fitting each dataset to a modified distribution called a **Beta-Geometric** distribution (Equation 1), and using this to adjust the observed read counts so they more closely fit a geometric. This approach is based on the observation that the skewed Tn-Seq datasets actually appear to fit not a constant geometric with a single Bernoulli parameter, $p$, but the weighted sum (integral) of multiple geometric distributions with different values of $p$. As weights on $p$, we choose the beta distribution, with parameters $\rho$ and $\kappa$ set so that the peak is around $p$.

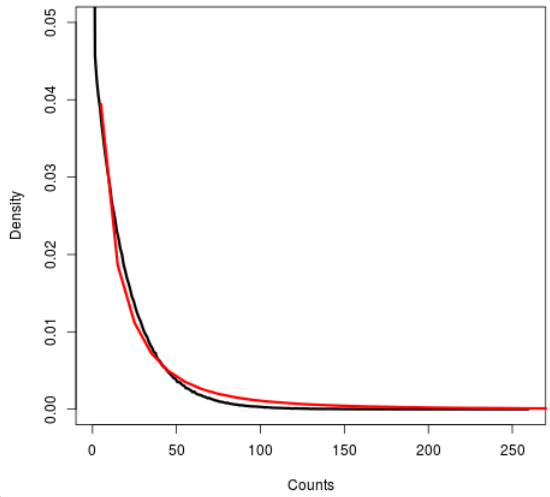
$$pdf(c;\rho,\kappa) = \int_0^1 Beta(p \mid \rho,\kappa) \times Geometric(c \mid p)\ dp \tag{1}$$

The beta distribution has an extra degree of freedom representing dispersion around $p$ (see Figure 3). This was chosen to reflect a *generative model* in which individual mutants in the Tn-insertion library have different growth rates, some growing slightly faster and some slightly slower than wild-type cells, depending on the location of the transposon insertion in their genome. This variability in growth rates will broaden out the apparent abundance of read counts after selection (i.e. several rounds of doubling in selective conditions). In this model, the spikes in read counts would come from clones that had higher-than-average growth rates.

One empirical measure we can use to evaluate whether the propsed correction method helps is to




a)




b)

Figure 3: a) Example of a beta distribution with $\rho = 0.05$ and $\kappa = 40$. The mode of the peak is near the value of $\rho$, indicated by the vertical line, but other values of $p$ (0.0-0.15) also have significant probability density. The width of the peak is controlled by $\kappa$. b) Histogram of counts sampled from a regular geometric distribution with $p = 0.05$ (black curve), and a Beta-Geometric distribution (red) with $\rho = 0.05$, $\kappa = 40$.

compare replicate datasets. In two datasets selected from the same Tn-mutant library under the same growth conditions, no differences are expected in essentiality of genes. However, in practice, there is high variability observed in Tn-Seq datasets, even between biological replicates. Any method for statistical analysis of Tn-Seq data must be conservative

enough not to detect many differentially essential genes between replicates. Yet, when using a permutation test (described below) on multiple pairs of replicates, we observe several differentially essential genes, in some cases beyond what would be expected from random statistical sampling differences. We attribute many of these false positives to the skew present in individual datasets. Our goal in this paper is to show that, by fitting each dataset to a Beta-Geometric distribution, we can correct for the skew in read counts, and thereby reduce many of these false positives. This enhanced normalization method could be applied to other Tn-Seq datasets to improve the detection of statistically significant differentially essential genes between conditions.

# 3 Methods

Given a set of read counts, $X_i$, for TA sites $i \in 1, 2, 3, ..., N$, we assume a hierarchical model in which read counts are geometrically distributed with a variable parameter, $p$, governed by the beta distribution:

$$\begin{aligned} X_i &\sim Geometric(p) \\ p &\sim Beta(\kappa\rho, \kappa(1-\rho)) \end{aligned}$$

where the beta distribution is parameterized using $\rho$ and $\kappa$, such that $\rho$ represents the mean of the parameter $p$, and $\kappa$ can be thought of as analogous to a "sample size", effectively proportional to the inverse of the variance.

We seek to estimate the parameters $\rho$ and $\kappa$, that minimize the sum of squared errors ($\epsilon$) between the observed read-counts and the quantiles of the distribution:

$$\begin{aligned} \epsilon(X;\rho,\kappa) &= \sum_i^N \left(X'_i - F^{-1}(q_i;p_i)\right)^2 \\ &= \sum_i^N \left(X'_i - \frac{\log(-q_i+1)}{\log(1-p_i)}\right)^2 \\ &= \sum_i^N \left(X'_i - \frac{\log(-q_i+1)}{\log(1-\frac{\kappa\rho-1}{\kappa-2})}\right)^2 \end{aligned} \tag{2}$$

Here, $X'$ represents the read counts in ascending order, $F^{-1}$, represents the quantile function of the geometric distribution, and $q_i \in [0,1]$ represents the quantiles.

To facilitate the parameter estimation, the parameter $\rho$ is estimated as $\rho = \left(\sum_i^N X_i\right)^{-1}$, which is the maximum likelihood estimate of the parameter $p$ of the geometric distribution. The remaining parameter, $\kappa$ is found by determining the root of the gradient. The gradient with respect to $\kappa$ is defined as follows:

$$\frac{\partial\epsilon}{\partial\kappa} = \frac{\sum_i^N 2(2\rho-1)\log(1-q_i)\left(\log(1-q_i) - X_i\log\left(\frac{-\rho\kappa+\kappa-1}{\kappa-2}\right)\right)}{(\kappa-2)((\rho-1)\kappa+1)\log^3\left(\frac{-\rho\kappa+\kappa-1}{\kappa-2}\right)} \tag{3}$$

The root of this gradient has the following analytic solution:

$$\kappa = \frac{2\times\exp\left[\frac{\sum_i^N \log^2(1-q_i)}{\sum_i^N X_i\log(1-q_i)}\right]-1}{\exp\left[\frac{\sum_i^N \log^2(1-q_i)}{\sum_i^N X_i\log(1-q_i)}\right]+\rho-1}$$

Once parameters $\rho$ and $\kappa$ have been estimated, capturing the skew in the dataset, the original read counts are corrected by mapping each of them to the equivalent quantile in an ideal geometric distribution as follows:

$$c' = F^{-1}(Q(c;\rho,\kappa);p) \tag{4}$$

where $Q(c;\rho,\kappa)$ is the cumulative distribution function for the Beta-Geometric (obtained by sampling), and $F^{-1}(q;p)$ is the quantile function for the geometric distribution.

NZMean normalization is applied to each dataset after correcting the counts using this method.

## 3.1 Analysis of Differential Essentiality

To evaluate the differential essentiality of a gene between two conditions, given multiple replicates of each, we use a non-parametric permutation test on the corrected and normalized counts at TA sites within the gene. Briefly, the counts are summed over all sites and averaged over replicates to estimate the mean read count of the gene in each condition. The difference is compared to a background (or 'null') distribution of means from many random permutations of the counts among the sites. The $p$-value is calculated from the fraction of times the observed mean is greater than one of the samples (Monte Carlo estimate).

In more detail, suppose we have $m_1$ replicate datasets in condition A, and $m_2$ replicates in condition B. Let $\mathbf{C}_{ij}$ be a $(m_1+m_2)\times n$ matrix of counts at each of $n$ TA sites $i$ within the gene $g$, for each dataset $j$. The difference of means is calculated as:

$$\Delta_g = \frac{1}{n|A|}\sum_{j\in A}\sum_i^n \mathbf{C}_{ij} - \frac{1}{n|B|}\sum_{j\in B}\sum_i^n \mathbf{C}_{ij} \tag{5}$$

Next, 10,000 random permutations of the counts in matrix $\mathbf{C}_{ij}$ are generated, and the differences $\Delta'$ are calculated for each permutation. The $p$-value is estimated as the number of times $\Delta > \Delta'$ (or $\Delta < \Delta'$ for negative differences).

# 4 Results

A set of 64 Tn-seq datasets was obtained from the same *M. tuberculosis* Tn-mutant library grown under different conditions. Each condition was tested in duplicate, yielding 32 pairs of replicates. The raw read counts were reduced to unique template counts using sequencing barcodes [7], though we will continue to refer to them generically as 'read counts' throughout this paper. Each dataset had an average of 2.4M total counts, with a range of 1.1-5.4M.

The Beta-Geometric correction was applied to each of the 64 datasets, followed by NZMean normalization. As an example, Table 1 contains statistics for the original datasets A1, A2, B1 and B2 (corresponding to the 'in vitro' and 'Trans02c' datasets among the 32 pairs), as well as the values of $\rho$ and $\kappa$ estimated by the BGC method. The dispersion parameter $\kappa$ is lower for the B1 and B2 datasets, consistent with the greater variability that is observed in those datasets. A QQ-plot of the corrected values for dataset B2 is shown in Figure 4, displaying a much better fit to the geometric distribution, with the skew removed (compare to Figure 2).

Table 1: Fitting of parameters for example datasets.

| Data-set | Total Reads | Inser. Dens. | Mean Count | Max Count | $\rho$ | $\kappa$ |
|---|---|---|---|---|---|---|
| A1 | 3.12M | 49.3% | 84.7 | 1,693 | 0.0118 | 911.1 |
| A2 | 1.93M | 52.6% | 49.2 | 1,175 | 0.0203 | 493.9 |
| B1 | 2.78M | 41.1% | 89.8 | 6,009 | 0.0111 | 422.0 |
| B2 | 3.65M | 38.1% | 128.4 | 16,146 | 0.0078 | 434.7 |

To assess the value of the Beta-Geometric Correction (BGC) for statistical tests of differential essentiality, we compared pairs of replicate datasets against each other. Because the datasets in each pair of replicates are selected under the same condition, the expectation is that there should be no differentially essential genes between them. Replicates were compared using a permutation test to detect significant differences in mean read counts in a gene. The observed difference was compared to a background distribution from resampling (permuting) the counts between replicates to estimate a $p$-value, as described above. A *false positive* is defined as a gene that has $p < 0.05$, since no differences in essentiality are expected between replicates in the same condition.

Table 2 presents the number of false positives obtained in the permutation test with and without the BGC. Note that due to the large number of genes in the *M. tuberculosis* genome (3989), the permutation test could be expected to incorrectly reject the null hypothesis on ∼5% of the genes through chance alone.

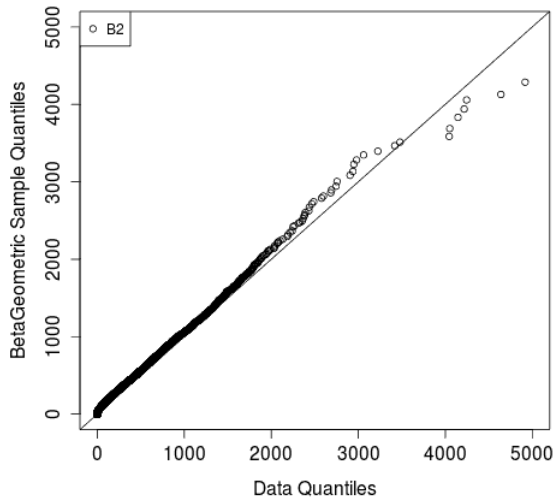



Figure 4: QQ-plot of the raw read counts for dataset B2, and the Beta-Geometric variables obtained by sampling the parameter $p$ from a Beta distribution with estimated parameters $\rho$ and $\kappa$.

However, applying BGC method reduces the Type I error in more than two-thirds of the cases (22 out of 32 pairs), achieving an average reduction of 21 false positives overall.

# 5 Discussion

Analysis of Tn-Seq data has become a valuable tool for identifying conditionally essential genes. However, the large amount of variability that is observed in these datasets makes direct comparison problematic. Common ways of normalizing the datasets have focused primarily on equilibrating the average read counts between datasets. While important, normalization of the means alone is not enough to correct for the large skew that is observed in some datasets.

The approach we propose assumes that the skew in read counts comes from dispersion in the parameter $p$ underlying a geometric distribution. The skew is captured by fitting the data to a Beta-Geometric distribution. The original read counts are then corrected back to an ideal geometric distribution by matching quantiles. We showed that one benefit of this correction is that it reduces the number of false positives in the permutation test when comparing replicates from the same condition. This approach to normalizing Tn-Seq datasets should also make fewer mistakes identifying differentially essential genes between conditions.

The BGC method is similar to quantile normalization [1, 10], except traditional quantile normalization scales datasets together based on an empirical distribution

Table 2: Number of Type I errors (genes with $p < 0.05$) obtained by comparing replicates against each other with the permutation test. All datasets were normalized with the Non-Zero Mean (NZMean) method, with or without the Beta-Geometric correction (BGC) applied first.

**Type I Errors for Permutation Test**

| Condition | without BGC | with BGC | difference |
|---|---|---|---|
| BXD04 | 535 | 248 | -287 |
| BXD08 | 241 | 87 | -154 |
| Trans07 | 158 | 97 | -61 |
| GP01 | 74 | 28 | -46 |
| BXD07 | 78 | 36 | -42 |
| Trans02c | 84 | 56 | -28 |
| BL6 | 74 | 49 | -25 |
| Trans05 | 142 | 123 | -19 |
| Trans11c | 85 | 67 | -18 |
| DS04 | 49 | 32 | -17 |
| DS0c | 42 | 27 | -15 |
| Trans01 | 32 | 19 | -13 |
| BXD06 | 91 | 81 | -10 |
| Trans11 | 22 | 18 | -4 |
| DS01 | 12 | 8 | -4 |
| Trans09c | 78 | 75 | -3 |
| CAST | 17 | 14 | -3 |
| AJ | 13 | 10 | -3 |
| PWK | 100 | 97 | -3 |
| BXD09 | 6 | 3 | -3 |
| Trans09 | 32 | 31 | -1 |
| DS02 | 22 | 21 | -1 |
| BXD01 | 2 | 2 | 0 |
| Trans07c | 70 | 71 | 1 |
| BXD03 | 0 | 2 | 2 |
| in vitro | 2 | 4 | 2 |
| Trans03 | 46 | 49 | 3 |
| CCcont | 2 | 5 | 3 |
| Trans03c | 52 | 57 | 5 |
| Trans05c | 30 | 42 | 12 |
| BXD05 | 33 | 46 | 13 |
| Trans01c | 62 | 85 | 23 |

function, without making assumptions about the form of the distribution. We choose to correct read counts back to an ideal geometric distribution, since the profile of abundances (i.e. vast majority of TA sites with low counts, comparatively few with high counts) probably reflects real biological effects, as would be expected from sampling from a population of cells growing with different levels of fitness (for example, see Motomura's model of species abundance distributions, [4]).

# 6 Acknowledgements

This work was supported by NIH grant U19 AI107774 (TRI). We thank Chris Sassetti (Univ. Mass. Medical School) for providing the datasets used in this study.

# References

[1] B.M. Bolstad, R.A. Irizarry, M. Astrand, and Speed T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2:185–193, 2003.

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

[3] 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.

[4] B.J. McGill et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10:995–1015, 2007.

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

[6] 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.

[7] J.E. Long, M. DeJesus, D. Ward, R.E. Baker, T.R. Ioerger, and C.M. Sassetti. Identifying essential genes in *Mycobacterium tuberculosis* by global phenotypic profiling. In Long Jason Lu, editor, *Methods in Molecular Biology: Gene Essentiality*, volume 1279. Springer, 2015.

[8] 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.

[9] 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.

[10] 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.