Statistical Analysis of Variability in TnSeq Data Across Conditions Using Zero-Inflated Negative Binomial Regression

Siddharth Subramaniyam,[1] Michael A. DeJesus,[2]
Anisha Zaveri,[3] Clare M. Smith,[4] Richard E. Baker,[4] Sabine Ehrt,[3]
Dirk Schnappinger,[3] Christopher M. Sassetti,[4] and Thomas R. Ioerger[1]

[1]Department of Computer Science and Engineering, Texas A&M University
[2]Rockefeller University, New York, NY
[3]Department of Microbiology & Immunology, Weill Cornell Medical College, New York, NY
[4]Department of Microbiology & Physiological Systems, University of Massachusetts Medical School

Abstract

Deep sequencing of transposon mutant libraries (or TnSeq) is a powerful method for probing essentiality of genomic loci under different environmental conditions. Various analytical methods have been described for identifying conditionally essential genes whose tolerance for insertions varies between two conditions. However, for large-scale experiments involving many conditions, it would be useful to have a method for identifying genes that exhibit significant variability in insertions across multiple conditions. In this paper, we introduce a novel statistical method for identifying genes with significant variability of insertion counts across multiple conditions based on Zero-Inflated Negative Binomial (ZINB) regression. Using likelihood ratio tests, we show that the ZINB fits TnSeq data better than either ANOVA or a Negative Binomial (as a generalized linear model). We use ZINB regression to identify genes required for infection of *M. tuberculosis* H37Rv in C57BL/6 mice. We also use ZINB to perform a retrospective analysis of genes conditionally essential in H37Rv cultures exposed to multiple antibiotics. Our results show that, not only does ZINB generally identify most of the genes found by pairwise resampling (and vastly out-performs ANOVA), but it also identifies additional genes where variability is detectable only when the magnitudes of insertion counts are treated separately from local differences in saturation, as in the ZINB model.

1 Introduction

Deep sequencing of transposon mutant libraries (or TnSeq) is a powerful method for probing essentiality of genomic loci under different environmental conditions [1]. In a transposon (Tn) mutant library (such as

made with the Himar1 transposon), insertions generally occur at random locations throughout the genome (restricted to TA dinucelotides for Himar1 [2]). The absence (or reduction) of insertions in a locus is used to infer conditional essentiality, based on killing (or growth impairment) that depletes those clones from the population. While the abundance of clones with insertions at different sites can be profiled efficiently through deep sequencing, there are a number of sources of noise that induce a high degree of variability in insertion counts at each site, including: variations in mutant abundance during library construction, stochastic differences among samples, biases due to sample preparation protocol and sequencing technology, and other effects. Previous statistical methods have been developed for quantitative assessment of essential genes in single conditions, as well as pairwise comparisons of conditional essentiality. Statistical methods for characterizing essential regions in a genome include those based on tests of sums of insertion counts in genes [3], gaps [4], bimodality of empirical distributions [5], non-parametric tests of counts [6], Poisson distributions [7], and Hidden Markov Models [8]. Statistical methods for evaluating *conditional essentiality* between two conditions include: estimation of fitness differences [9], permutation tests on distribution of counts at individual TA sites (resampling in TRANSIT [10]), Mann-Whitney U-test [11], comparing combined cumulative distribution functions [12], and another Hidden Markov Model [13].

Recently, more complex TnSeq experiments are being conducted involving larger collections of conditions (such as assessment of a library under multiple nutrient sources, exposure to different stresses like a panel of antibiotics, or passaging through multiple animal models with different genetic backgrounds) [14, 15, 16, 17]. Yang et al. [18] has also looked at temporal patterns of changes in insertion counts over a time-course. A fundamental question in such large-scale experiments is to determine which genes exhibit statistically significant variability across the panel of conditions. A candidate approach might be to perform an ANOVA analysis of the insertion counts to determine whether there is a condition-dependent effect on the means. However, ANOVA analyses rely on the assumption of Normality, and Tn insertion counts are clearly not Normally distributed. First, read-counts are are non-negative integers; second, there are frequently outliers or sporadic sites with high counts that influence the means; third, most Tn libraries are sub-saturated, with a high fraction of TA sites not being represented, even in non-essential regions. This creates an excess of zeros in the data (sites were no insertion was observed), and this makes it ambiguous whether sites with a count of 0 are biologically essential (i.e. depleted during growth/selection) or simply missing from the library. An alternative method for assessing variability might be to use a non-parametric test of the differences between means by permuting the counts and generating a null distribution (as in the "resampling test" in TRANSIT [10]). However, this is limited to pairwise comparisons, and attempting to run resampling for all pairwise comparisons between conditions to identify genes that show some variation does not scale up well as the number of conditions grows.

In this paper, we introduce a new statistical method for identifying genes with significant variability of insertion counts across multiple conditions based on Zero-Inflated Negative Binomial (ZINB) regression. The ZINB distribution is a mixture model of a Negative Binomial distribution (for the magnitudes of insertion counts at sites with insertions) combined with a "zero" component (for representing the proportion of sites without insertions). ZINB regression fits a model for each gene that can be used to test whether there is a condition-dependent effect on the magnitudes of insertion counts or on the local level of saturation in each gene. Separating these factors increases the statistical power that ZINB regression has over resampling for identifying varying genes (since resampling just tests the differences in the means between conditions - zeros included). Importantly, our model includes terms to accommodate differences in saturation among the datasets to prevent detecting false positives (e.g. due to artifacts of normalization).

Another advantage of the ZINB regression framework is that it allows incorporation of additional factors as covariates in analyzing variability across multiple conditions, to account for effects dependent on relationships among the conditions, such as similar treatments, time-points, host genotypes, etc.

Using several TnSeq datasets from *M. tuberculosis* H37Rv, we show that, in pairwise tests (between two conditions), the genes detected by ZINB regression are typically a superset of those detected by resampling and hence is more sensitive. More importantly, ZINB regression can be used to identify varying genes across multiple (≥ 3) conditions, which contains most of the genes identified by pairwise resampling between all pairs (and is more convenient and scalable). Furthermore, ZINB regression vastly out-performs ANOVA, which often identifies only around half as many genes with significant variability in read-counts.

2 Methods

2.1 ZINB Model

Essential genes are likely to have no insertions or very few counts (because mutants with transposon insertions in those regions are not viable), while non-essential genes are likely to have counts near the global average for the dataset. Insertion counts in non-essential regions are typically observed to follow a "geometric-like" distribution, with many more sites having low counts, and very few with high counts (exponentially decreasing frequency). Thus, one could potentially fit the observed read-counts for these different phenotypes using a Negative Binomial (NB) distribution.

Let $Y_g^{j,k}$ represents the read-counts for gene, g, in replicate j of condition k, for conditions $C \in \{c_1 ... c_k\}$. The NB distribution can be thought of as a Poisson distribution with over-dispersion (i.e. more variability than would be expected), thus allowing greater freedom to fit the count data:

$$NB(Y = y \mid r, p) = \binom{y + r - 1}{y} p^y (1 - p)^r$$

where p is a success probability between 0 and 1, and r, often called a size parameter, represents the dispersion as it can be shown that the NB distribution converges to a Poisson as r tends to infinity [19]. A common parameterization of the NB distribution is to specify the distribution based on the mean, μ, and the dispersion parameter, r, and determine the success probability, p, through the following relationship:

$$p = \frac{\mu}{\mu + r}$$

In practice, TnSeq data often has an excess of empty sites (with counts of 0), exceeding those that would be expected under a typical NB distribution. A library with transposon insertions at 50% of its sites (i.e. 50% saturation) would mean that even non-essential genes will have a large portion of sites missing (i.e. equal to zero). Thus, while the NB distribution may be sufficient to model counts in other domains, TnSeq requires more careful consideration.

One way to solve this problem is to model the read-counts for a condition C, Y_g^C, as coming from a Zero-Inflated Negative Binomial distribution (ZINB) instead:

$$Y_g^C \sim ZINB(\pi_g^C, r_g^C, \mu_g^C)$$

where

$$ZINB(Y = y \mid \pi, r, \mu) = \begin{cases} \pi + (1 - \pi) \times NB(0 \mid r, \mu) & y = 0 \\ (1 - \pi) \times NB(y \mid r, \mu) & y > 0 \end{cases}$$

Here the π parameter represents the probability that a count of zero is extraneous (i.e. doesn't belong to the NB distribution), and can be interpreted as similar to the probability that an empty site is essential (i.e. empty due to fitness costs incurred through its disruption, rather than stochastic losses). In this way, both read-counts (through the r and μ parameters of the NB distribution) and insertion density (through π) can be used to differentiate genes that are essential in one condition and non-essential in another.

2.2 Generalized Linear Model

To capture the conditional dependence of the ZINB parameters (μ^g, r^g, π^g) on the experimental conditions, we adopt a linear regression (GLM) approach, using a log-link function. The estimated mean, μ_g, of the NB

distribution for each gene is,

$$ln\ \mu_i^g = X^g \vec{\alpha}_g$$

where X^g is a binary matrix (using a one-hot encoding) indicating the condition (as a covariate) for each observation Y_i for gene g. Hence, for n conditions, there will be n coefficients for each gene, one for estimating mean non-zero count for condition A, $\alpha_g[A]$, an independent coefficient $\alpha_g[B]$ for condition B, etc. The estimated mean non-zero count (conditional expectation) for each condition can be recovered as: $\mu_A^g = exp(\alpha_g[A])$, $\mu_B^g = exp(\alpha_g[B])$, etc.

If additional covariates distinguishing the samples are available, such as library, timepoint, or genotype, they may be conveniently incorporated in the linear model with an extra matrix of covariates, W, to which a vector of parameters $\vec{\beta}_g$ will be fit:

$$ln\ \mu_i^g = X^g \vec{\alpha}_g + W\vec{\beta}_g$$

For the dispersion parameter of the NB, τ (or size parameter $r = 1/\tau$), we assume that each gene could have its own dispersion, but for simplicity, we assume that it does not differ between conditions. Hence, it is fitted by a common intercept:

$$ln\ r^g = \rho_g$$

Finally, for the zero-inflated (Bernoulli) parameter, π, we fit a linear model depending on condition, with a logit link function:

$$logit(\pi_i^g) = ln\Big(\frac{\pi_i^g}{1-\pi_i^g}\Big) = X^g \vec{\gamma}_g$$

Thus each gene will have its own local estimate of insertion density in each condition, $\pi_A^g = exp(\gamma_g[A])/(1 + exp(\gamma_g[A]))$, etc. In the case of covariates, $logit(\pi_i^g) = X^g\vec{\gamma}_g + W^g\vec{\delta}_g$.

The parameters of the GLM can be solved by maximum-likelihood using iteratively re-weighted least squares (IWLS). In this work, we use the *pscl* package in R [20].

2.3 Correcting for Saturation Differences Among TnSeq Datasets

An important aspect of comparative analysis of TnSeq data is the normalization of datasets. Typically, read-counts are normalized such that the total number of reads is balanced across the datasets being compared.

Assuming read-counts are distributed as a mixture of a Bernoulli distribution (responsible for zeros) and another distribution, $g(x)$, responsible for non-zero counts i.e.,

$$f(x) = \begin{cases} \theta \times \mathrm{g}(x) & x > 0 \\ (1-\theta) \times \mathrm{Bern}(x|p=0) & x = 0 \end{cases}$$

then the expected value of this theoretical read-count distribution (with mixture coefficient θ) is given by:

$$\mathrm{E}\left[f(x)\right] = \theta \times \mathrm{E}\left[g(x)\right]$$

The expected value of such a distribution can be normalized to match that of a reference dataset, $f_r(x)$, (such as another condition) by multiplying it by a factor, w, defined in the following way:

$$\mathrm{E}\left[f_r(x)\right] = w \times \mathrm{E}\left[f(x)\right]$$
$$\theta_r \times \mathrm{E}\left[g_r(x)\right] = w \times (\theta \times \mathrm{E}\left[g(x)\right])$$
$$w = \frac{\theta_r \times \mathrm{E}\left[g_r(x)\right]}{\theta \times \mathrm{E}\left[g(x)\right]}$$

This guarantees that the expected value in read-counts is the same across two datasets. TTR normalization (i.e. total trimmed read count, the default in TRANSIT [10]) estimates $\mathrm{E}\left[g(x)\right]$ in a robust manner (excluding the top 1% of sites with highest counts, to reduce the influence of outliers, which can affect normalization and lead to false positives).

While TTR works well by itself for methods like resampling (which only depend on the expected counts being equivalent under the null-hypothesis), it does not work well for methods that simultaneously detect differences in both the local magnitudes of counts (non-zero mean) and the saturation (fraction of non-zero sites) such as ZINB. This is because TTR in effect inflates the counts at non-zero sites in datasets with low saturation, in order to compensate for the additional zeros (to make their expected values equivalent). This would cause genes to appear to have differences in (non-zero) mean count (μ_A^g vs μ_B^g), while also appearing to be less saturated (π_A^g vs π_B^g), resulting in false positives.

To correct for differences in saturation, we incorporate offsets in the linear model. Let D_i be a binary $n \times d$ matrix of dataset indicators for each observation (and D^g be the restriction of this matrix to sites in gene g). Furthermore, let N be a vector of non-zero means for each dataset and Z be a vector of the fraction

of sites with a count of zero for each dataset. Then the linear equations can be modified to incorporate these offsets (a specific offset for each observation depending on which dataset it comes from).

$$ln(\mu_i^g) = X^g \vec{\alpha}_g + ln(D^g N)$$

$$logit(\pi_i^g) = X^g \vec{\gamma}_g + logit(D^g Z)$$

Note that these are just constants in the linear equation, without coefficients to be fit. Hence the fitted coefficients ($\vec{\alpha}_g$ and $\vec{\gamma}_g$) are effectively estimating the *deviations* in the local insertion counts in a gene, relative to the global mean and saturation for each dataset:

$$exp(X^g \vec{\alpha}_g) = \mu_i^g / (D^g N) = \mu_i^g / \eta_{d(i)}$$

$$exp(X^g \vec{\gamma}_g) = \frac{logit(\pi_i^g)}{logit(D^g Z)} = \Big(\frac{\pi_i^g}{1-\pi_i^g}\Big)\Big(\frac{\phi_{d(i)}}{1-\phi_{d(i)}}\Big)$$

where $\eta_{d(i)}$ is the mean count for dataset $d(i)$, and $\phi_{d(i)}$ is the saturation of dataset $d(i)$.

2.4 Statistical Significance

Once the ZINB model is fit to the counts for a gene, it is necessary to evaluate the significance of the fit. T-tests could be used to evaluate the significance of individual coefficients (i.e. whether they are significantly different from 0). However, for assessing whether there is an overall effect as a function of condition, we compare the fit of the data to a simpler model - ZINB without conditional dependence - and compute the difference of log-likelihoods (or log likelihood ratio):

$$-2\{\mathcal{L}_0(Y_i^g|\Theta_0) - \mathcal{L}_1(Y_i^g|\Theta_1)\} = -2\ ln\Big(\frac{L_0(Y_i^g|\Theta_0)}{L_1(Y_i^g|\Theta_1)}\Big)$$

where the two models are given by:

$$\begin{aligned} M_1: \quad & L_1(Y_i^y|X^g, \Theta_1) = ZINB(Y_i^g|\mu^g, r^g, \pi^g) \\ & ln\ \mu^g = X^g \vec{\alpha}_g,\ ln\ r^g = \rho_g,\ logit(\pi^g) = X^g \vec{\gamma}_g \end{aligned}$$

$$\begin{aligned} M_0: \quad & L_1(Y_i^y|\Theta_0) = ZINB(Y_i^g|\mu^g, r^g, \pi^g) \\ & ln\ \mu^g = \alpha_g^0,\ ln\ r^g = \rho_g,\ logit(\pi^g) = \gamma_g^0 \end{aligned}$$

where $\Theta_1 = \langle \vec{\alpha}_g, \rho_g, \vec{\gamma}_g \rangle$ and $\Theta_0 = \langle \alpha_g^0, \rho_g, \gamma_g^0 \rangle$ are the collections of parameters for the two models, and where α_g^0 and γ_g^0 in M_0 are just constants fitted to the grand mean and saturation of the gene over all conditions.

The likelihood ratio statistic above is expected to be distributed as χ^2 with degrees of freedom equal to the difference in the number of parameters (Wilks' Theorem):

$$-2\ ln\Big(\frac{L_0(Y_i^g|\Theta_0)}{L_1(Y_i^g|\Theta_1)}\Big) \sim \chi^2_{df=df(M_1)-df(M_0)}$$

For the condition-dependent ZINB model (M_1), the number of parameters is $2c+1$ (for length of $\vec{\alpha}_g$ and $\vec{\gamma}_g$ plus ρ_g). For the condition-independent ZINB model (M_0), there are only 3 scalar parameters ($\alpha_g^0, \rho_g, \gamma_g^0$) used to model the counts pooled across all conditions. Hence $df = 2c+1-3 = 2(c-1)$. The point of the test is to determine whether the additional parameters, which should naturally fit the data better, are justified by the extent of increase in the likelihood of the fit. The inverse of the χ^2 distribution is used to calculate p-values, which are then adjusted by the Benjamini-Hochberg procedure [21] to correct for multiple tests (to limit the false-discovery rate to 5% over all genes in the genome being tested in parallel).

Importantly, if a gene is detected to be conditionally-essential, it could reflect either a difference in the mean (non-zero) counts, the saturation, or both. Thus the method is capable of detecting genes that have insertions in approximately the same fraction of sites but with a systematically lower count (e.g. reduction by X%), possibly reflecting a fitness defect. Similarly, genes where most sites become depleted but where the remaining sites (perhaps at the terminal, or in a non-essential domain) remain the same would also be detectable as conditional-essentials.

If the data include additional covariates, then the W terms will be included in the regressions for *both* models M_1 and M_0:

$$\begin{aligned} M_1: \quad & L_1(Y_i^y|X^g, W^g, \Theta_1) = ZINB(Y_i^g|\mu^g, r^g, \pi^g) \\ & ln\ \mu^g = X^g\vec{\alpha}_g \underline{+W^g\vec{\beta}_g},\ ln\ r^g = \rho_g,\ logit(\pi^g) = X^g\vec{\gamma}_g \underline{+W^g\vec{\delta}_g} \end{aligned}$$

$$\begin{aligned} M_0: \quad & L_1(Y_i^y|W^g, \Theta_0) = ZINB(Y_i^g|\mu^g, r^g, \pi^g) \\ & ln\ \mu^g = \alpha_g^0 \underline{+W^g\vec{\beta}_g},\ ln\ r^g = \rho_g,\ logit(\pi^g) = \gamma_g^0 \underline{+W^g\vec{\delta}_g} \end{aligned}$$

In this way, the covariates W will increase the likelihoods of both models similarly, and the LRT will be evaluating only the improvement of the fits due to the conditions of interest, X, i.e. the residual variance explained by X after taking known factors W into account. Although the number of parameters in both models will increase, the difference in degrees of freedom will remain the same.

2.5 Availability

The methods described in this paper have been implemented in TRANSIT [10], which is publicly available on GitHub (http://saclab.tamu.edu/essentiality/transit/) and can be installed as a python package (*tnseq-transit*) using *pip*.

3 Results

3.1 Likelihood Ratio Tests for Suitability of ZINB as a Model for TnSeq Data

To establish the suitability of ZINB as a model for TnSeq data, we compared it to ANOVA and Negative Binomial (without special treatment of zeros) using likelihood ratio tests. The data we used for these tests consisted of 2 replicates of an *M. tuberculosis* H37Rv TnSeq library grown on glycerol compared to 3 replicates grown on cholesterol [22]. This data was originally used to identity genes in the H37Rv genome that are necessary to catabolize cholesterol, a unique carbon source available within the restricted intracellular environment of macrophages, on which growth and survival of the bacilli depends [23]. The data (insertion counts at TA sites) were normalized by the TTR method [10].

First, we compared ZINB regression to simple ANOVA (based on a generalized linear model using Gaussian likelihood functions). Both models were used to fit the insertion-count observations at the TA sites in each gene, conditioned on the carbon source (glycerol vs. cholesterol). ZINB had higher likelihood than ANOVA for all genes (expect five, for which they were nearly equal). Because ZINB and ANOVA are not nested models, we used the Vuong test [24] to evaluate statistical significance of the difference in likelihoods. Furthermore, we applied the Benjamini-Hochberg procedure to adjust the p-values for an overall false-discovery rate (FDR) of 5%. ZINB was found to produce a significantly better fit than ANOVA for 3185 out of 3282 genes (97%, using $p_{adj} < 0.05$ as a criterion).

Next, we performed a likelihood ratio test (LRT) of ZINB regression compared to regular NB (as a generalized linear model). Because ZINB has more parameters (and these are nested models), the likelihood for ZINB was again higher than NB for nearly every gene. To evaluate which differences were significant, correcting for the different number of parameters, we computed p-values of the log-likelihood ratio using the χ^2 distribution, with degrees of freedom equal to the difference in number of model parameters ($df = 5-3 = 2$). After FDR-correction, ZINB fit the data significantly better than NB for 2796 genes out of 3282 (85%) genes evaluated. For the rest of the genes, the likelihoods of the two models were indistinguishable. This supports the hypothesis that modeling the fraction of sites with no insertions ("zeros") separately from the magnitudes of counts at sites with insertions enables ZINB to fit TnSeq data better.

3.2 Pairwise Comparisons of Conditional Essentiality Using ZINB

We evaluated ZINB, resampling, and ANOVA on data from an *M. tuberculosis* TnSeq library grown in-vitro compared to infections in a mouse model. A high-saturation Himar1 Tn library generated in H37Rv was inoculated into six C57BL/6 mice (8-12 week old males, obtained from Jackson Laboratory, Bar Harbor, ME) via the intravenous route at a dose that deposits a representative sample of the library (>100,000 CFU) in the spleen. After four weeks, the bacteria present in the spleen of each animal were recovered by plating on 7H10 agar (with kanamycin). As a control, the original library was replated in parallel. A total of 0.4-1.5 million reads was mapped to TA sites for each sample, and all samples had $\sim$ 50% saturation (all but one were in the 42-58% range; see Table 1). The data was normalized using TTR (Trimmed Total Read-count) normalization [10], and the mean count of all datasets after normalization was uniform, around 100.

Table 1: Statistics of TnSeq datasets

dataset	condition	mapped reads	density*	NZmean**
A1	in vitro	989413	0.55	23.9
A2	in vitro	1376613	0.58	31.9
A3	in vitro	1531598	0.58	35.3
A4	in vitro	547902	0.47	15.5
A5	in vitro	1450383	0.57	33.9
A6	in vitro	475126	0.55	11.6
B1	in vivo	1500646	0.47	42.5
B2	in vivo	601683	0.45	17.7
B3	in vivo	1245065	0.51	32.5
B4	in vivo	1472365	0.49	39.9
B5	in vivo	909394	0.42	29.3
B6	in vivo	409018	0.34	16.2

* fraction of TA sites with insertions
** mean count at TA sites with insertions before normalization

When ZINB regression method was run on the two conditions (in vitro vs. in mice), 237 conditional essentials were identified. This included genes well-known to be essential in vivo, including the Mce4 cluster, biotin biosynthesis (*bioABDF1*), ESX-1, the NRPS (non-ribosomal peptide synthase) cluster (Rv0096-Rv0101), and cholesterol catabolism genes (e.g. FadE5, *bpoC, hsaD*). Some genes involved in mycobactin-dependent iron acquisition (*irtAB, mmpL4/S4*) were essential in vivo, though none of the 14 subunits of mycobactin synthase (Mbt) were. A possible explanation is that mutants with insertions in Mbt genes are importing extracellular mycobactin produced by other mutants at the site of infection, with insertions in genes other than Mbt synthase. In contrast to infections with a homogeneous knockout mutant of genes like *MbtD*, mycobactin synthase transposon mutants in the Tn library can survive in vivo because it is a heterogeneous pool. However, individual clones with defects in mycobactin secretion/uptake (e.g. Tn insertions in *irtAB*

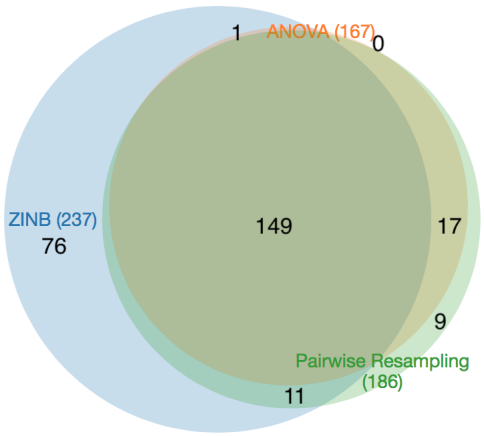

Figure 1: Venn diagram of conditional essentials ($qval < 0.05$) for three different methods: resampling, ANOVA, and ZINB.

and *mmpL4/S4*) cannot survive, despite the availablility of mycobactin in the environment.

The results of ZINB can be compared to the permutation test ('resampling' in TRANSIT), which is a non-parameteric comparison of the difference in mean counts for each gene between the two conditions. Resampling yielded 186 hits. (p-values for all tests were corrected for a false-discovery rate of $< 5\%$ using the Benjamini-Hochberg procedure [21]). Almost all of these (160, 86%) were contained in the hits from ZINB (see Figure 1). Only 26 genes identified by resampling were not detected by ZINB. Many of these were marginal cases; 21 of 26 had ZINB adjusted p-values between 0.05 and 0.2.

ANOVA was also applied to the same data, and it only detected 166 genes with significant variability between the two conditions. The genes detected by ANOVA were almost entirely contained in the set of genes detected by resampling (166 out of 167). But resampling found 20 more varying genes. In comparison, ANOVA only finds 63% of the varying genes detected by ZINB (150 out of 237). The lower sensitivity of ANOVA is likely due to the fact that ANOVA depends on the assumption of Normality of the counts, whereas resampling does not.

The advantage of ZINB is that it is capable of detecting more conditional essentials because it can take into account changes in either the local magnitude of counts or local insertion density. It detects 76 more conditional essentials than resampling, and 88 more than ANOVA. Among these are genes in the Mce1 cluster (specifically *mce1B*, *mce1C*, and *mce1F*). Mce1 is a membrane transport complex that has been shown to be essential in vivo [25]. The Mce1 genes show a modest reduction in counts (log-fold changes of -0.31 to -0.86), which was not sufficient to meet the adjusted p-value cutoff for resampling. However, the genes also exhibit a noticable reduction in local saturation in this locus (from $\sim 88\%$ to $\sim 61\%$; see Figure 2), and the combination of these two depletion effects is sufficient to make them significant in the ZINB model.

Similar examples include *esxB*, a secreted virulence factor, *fcoT* (thioesterase for non-ribosomal peptide synthase NRPS), *lysX* (lysinylation of cell-wall glycolipids [26]]), *pitA* (involved in phosphate transport [27]), and *fadE33*, *hsaB* and *kshB*, which are involved in cholesterol catabolism [22]. All of these genes have been previously shown to be essential for infection in an animal model, but did not meet the threshold for significance based on resampling. The reason that several of these genes (like *fadE33* and *esxB*, shown in Figure 3) are detected by ZINB but not resampling is due primarily to changes in saturation; the non-zero mean (NZmean) changes only slightly, but the saturation drops significantly in each case; greater depletion of insertion mutants indicates reduced fitness. This highlights the value in treating the saturation parameter separately in the ZINB model. Another gene that exhibits this effect is SecA2. SecA2 is an alternative ATPase component of the *Sec* secretion pathway and is thought to help secrete other virulence factors inside the macophage [28]. SecA2 mutants have a weak phenotype in vitro ("growth defect" gene; [29]), so that the mean counts and saturation are low compared to other genes in-vitro (only 20% saturation, compared to ∼50% globally); however, it becomes almost completely devoid of insertions in-vivo (Figure 3). While SecA2 was not detected as significant by either resampling or ANOVA, it was identified as conditionally essential by ZINB.

Although ZINB identifies more genes (76) to be statistically significant than resampling on this dataset, it is unlikely that this excess is attributable to a large number of false positives. To evaluate the susceptibility of ZINB to generate false positives, we performed a comparison between replicates from the same condition by dividing the 6 in-vitro datasets into 2 groups (3+3). In this case, we expect to find no hits because there are (presumably) no biological differences. ZINB analysis identified only 15 genes as significantly different ($p_{adj} < 0.05$), which suggests that the overall false positive rate for ZINB is quite low and probably reflects noise inherent in the data itself. Even resampling, when run on the same data (3 in-vitro vs. 3 in-vitro) for comparison, yielded 9 significant genes, which are presumably false positives.

3.3 Adjustment for Differences in Saturation Among Datasets

In real TnSeq experiments, it frequently happens that some datasets are less saturated than others. For example, there is often loss of diversity when passaging a Tn library through an animal model, possibly due to bottlenecking during infection or dissemination to target organs. TTR normalization was developed to reduce the sensitivity of the resampling method to differences in saturation levels of datasets. However, this type of normalization would be expected to exacerbate the detection of differences by ZINB. To compensate for this, we include offsets in the models that take into account the global level of saturation and non-zero mean for each dataset.

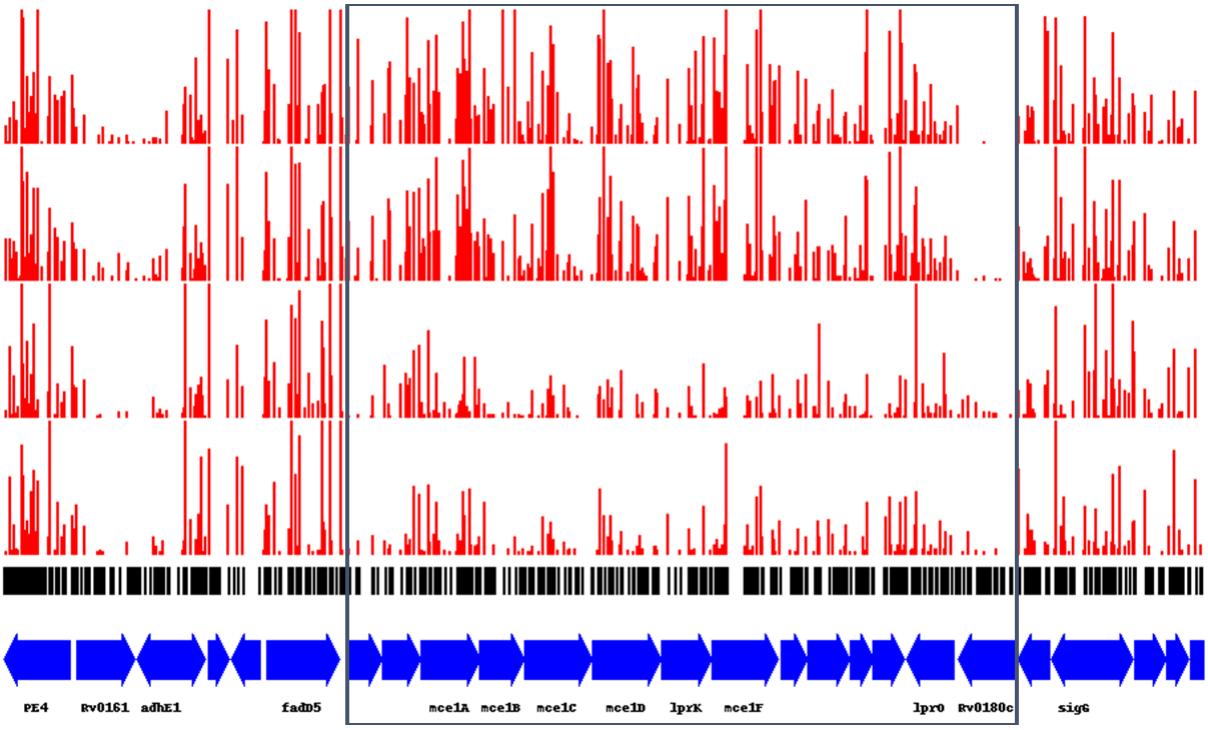

Figure 2: Normalized insertion counts in the Mce1 locus for 2 in-vitro replicates (upper tracks) and 2 in-vivo replicates (lower).

Table 2: Comparison of ZINB regression with and without saturation adjustment, for artificially-depleted samples.

scenario	datasets compared	resampling	ZINB (SA^+)	ZINB (SA^-)
1	(A1,A2) vs (B1,B2)	64	73	162
2	(A1,A2) vs (B1[50%],B2[50%])	108	121	**2668**
3	(A1[50%],A2[50%]) vs (B1,B2)	17	112	**1139**
4	(A1[50%],A2[50%]) vs (B1[50%],B2[50%])	8	30	37

To evaluate the effect of the correction for saturation of datasets, we created artificially-depleted versions of some of the replicates analyzed in the previous Section (see Table 1). Specifically, for A1, A2, B1, and B2, we created "half-saturated" versions of each by randomly (and independently) setting 50% of the sites to 0. Since each of the original datasets had around 50% saturation to begin with, the half-saturated version have a saturation of approximately 25%.

Initially, we compared the original versions of A1 and A2 to B1 and B2 (scenario 1), with their observed level of saturation. The number of hits detected by ZINB (73) is similar to resampling (64). Recall that resampling with all 12 datasets yielded 186 significant genes; the number of hits is lower overall in this experiment because only 2 replicates of each were used, instead of 6. Then we compared fully-saturated versions of A1 and A2 to half-saturated B1 and B2 (scenario 2). ZINB-SA^+ (with adjustment for saturation)

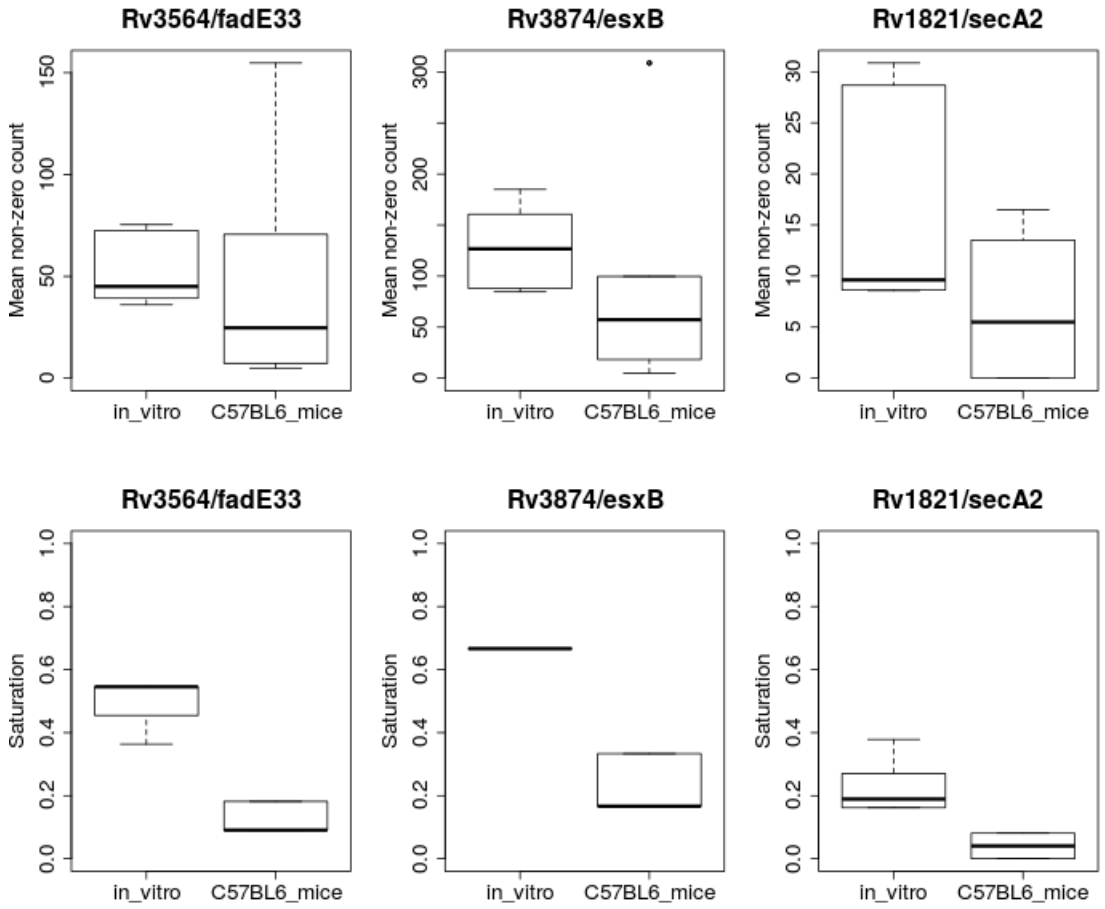

Figure 3: Statistics for three genes detected to vary significantly in mice compared to in-vitro based on ZINB regression, but not by resampling. The upper panels are the Non-Zero Mean (among insertion counts at TA sites with counts > 0), and the lower panels show the Saturation (percent of TA sites with counts > 0). Each box represents a distribution over 6 replicates.

identified nearly the same number of conditional essentials as resampling: 121 vs. 108. (see Table 2). The results are similar when half-saturated version of datasets A1 and A2 are used (scenario 3). However, when saturation adjustment is turned off, ZINB-SA$^-$ produces dramatically more hits in case of wide saturation differences (2668 and 1139). The reason for this is that, by artificially reducing the saturation of either datasets A1 and A2 or B1 and B2, it amplifies the apparent difference in local saturation for many genes, to which ZINB is sensitive. The number of significant hits (conditional essentials) detected when half-saturated versions of all four datasets are used (scenario 4) is naturally lower (8 and 30), because there is much less information (fewer observations) available, making it more challenging for many genes to achieve statistical significance. Interestingly, when half-saturated versions of all four datasets are used, ZINB-SA$^-$ works as expected, finding 37 hits (scenario 4), similar to resampling.

3.4 Application to Datasets with Multiple Conditions

In a prior study, a Himar1 transposon-insertion library in H37Rv was treated with sub-inhibitory concentrations of 5 different drugs [17]: rifampicin (RIF), isoniazid (INH), ethambutol (EMB), meropenem (MERO), and vancomycin (VAN), all grown in 7H9 liquid medium. Combined with the untreated control, this makes 6 conditions, for which there were 3 replicate TnSeq datasets each (except INH; see Table 3). The TnSeq datasets had a high saturation of 60-65% (percent of TA sites with insertions). In the original analysis, each drug-treated sample was compared to the control using resampling [17]. Several conditionally essential genes were identified for each drug. Some genes were uniquely associated with certain drugs (for example, *blaC*, the beta-lactamase, was only required in the presence of meropenem), and other genes were shared hits (i.e. conditionally essential for more than one drug). Only one gene, *fecB*, was essential for all drugs, and its requirement for antibiotic stress tolerance was validated through phenotyping of a knock-out mutant.

We observed that there are a large number of outliers in this data, consisting of isolated TA sites with counts in one sample > 10 times higher than the others (even in other replicates of the same condition). Outliers were sporadic and not associated with any particular dataset. To mitigate the effect of these outliers, we applied the beta-geometric correction (BGC) to all of the datasets [30]. The BGC is a non-linear transformation that reduces the skew in read-count distributions, and makes each dataset look more uniform (similar to a geometric distribution).

We then did resampling of each drug condition against the untreated control, and the number of conditional essentials is shown in Table 3. *fecB* was again observed to be the only hit in the intersection of all tests. We also observe other hits that can be rationalized, such as conditional essentiality of *blaC* (beta-lactamase) in presence of meropenem.

Table 3: TnSeq datasets in different antibiotic treatments. MICs for H37Rv were obtained from [31].

drug	concentration (μg/ml)	MIC_{50} (μg/ml)	# replicates	# of essentials vs. untreated*
untreated			3	
isoniazid (INH)	0.027	0.04	2	50
rifampicin (RIF)	0.004	0.02	3	68
ethambutol (EMB)	0.5	0.6	3	58
meropenem (MER)	1.2	30	3	106
vancomycin (VAN)	16	11	3	93

* number of conditional essentials by comparison to the untreated condition using resampling

Variability among all 6 conditions was analyzed using several different methods for comparison. First, a simplistic but practical approach was taken by performing pairwise analyses of conditional essentiality using resampling (the permutation test for significant differences per gene in TRANSIT). For six conditions, there are 15 pairwise comparisons. Resampling was run independently on each pair of conditions, and the p-values were adjusted independently each time. By taking the union of conditionally-essential genes over all 15 pairwise comparisons, a total of 276 distinct genes was identified to have varying counts between at least one pair of conditions (Table 4).

However, this straightforward approach is unfair because the p-values were adjusted independently. A more rigorous approach would be to perform resampling on all ~ 4000 genes for all 15 pairs of conditions, and then apply the p-value adjustment once on the pool of all $\sim 60,000$ p-values. When this is done, there are 267 significantly varying genes (using the lowest adjusted p-value for each gene). Thus, proper use of FDR correction results in a slightly more conservative list of hits.

The main problem with this approach is that it requires resampling to be run separately for all pairs of conditions, which does not scale up well as the number of conditions increases. As an alternative, ANOVA can be used to compare the counts across all six conditions simultaneously. When ANOVA is run (and the p-values are adjusted using the Benjamini-Hochberg procedure), only 234 significantly varying genes are identified. The 234 genes identified by ANOVA are almost completely contained in the set of those identified by pairwise resampling. Thus, ANOVA has lower sensitivity and under-reports genes with significant variability.

Finally, to identify genes that exhibit variability across all 6 conditions, we used ZINB regression. 307 genes were found to exhibit significant variation by ZINB, including genes identified in the original study, such as *fecB*, *blaC*, *pimE* (mannosyltransferase), and *secA*2 (protein translocase) [17]. Another example of a gene found by both ZINB and pairwise resampling is *cinA* (Rv1901), which was specifically required for cultures exposed to sub-MIC concentrations of INH (Figure 5a). *cinA* is thought to be an NAD-dependent enzyme that plays a role in nucleoside recycling [32, 33], and thus it could confer tolerance to INH through

Table 4: Identification of genes with significant variability across six conditions in antibiotic treatment data.

method	# of varying genes
union of hits from 15 pairwise resampling tests	276
genes significant in any pairwise resampling after pooled adjustment p-values	267
ANOVA	234
ZINB	307

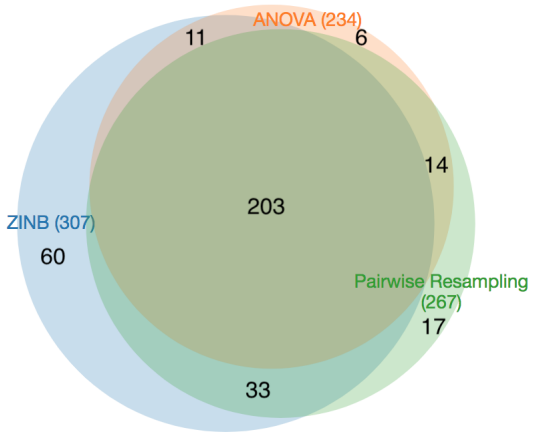

Figure 4: Venn diagram of genes with significant variability in different antibioitic treatments of transposon insertion counts evaluated by three different methods.

a mechanism involving maintaining the intracellular NADH/NAD$^+$ ratio [34].

Compared to ANOVA, ZINB finds significantly more varying genes (307 compared to 234, 31% more). Put another way, ANOVA only identifies 76% of the genes with variability identified by ZINB. The lower number of hits for ANOVA is likely due to the non-Normality of the distribution of Tn insertions counts. ZINB identified slightly more varying genes than pairwise resampling (71 additional genes). Many of these genes are on the margin and have adjusted p-value just slightly over the cutoff for resampling; 50% (36 out of 71 genes) have $0.05 < p_{adj} < 0.2$ for resampling. Among the remaining genes, one interesting case detected uniquely by ZINB is *sigE* (Figure 5b). While the mean insertion counts do not vary much for this gene (ranging between 17 and 27), the saturation level varies significantly among drug exposures, from nearly fully saturated in the control and INH conditions (88-97%), to highly depleted of insertions for RIF, MER and EMB ($29 - 52\%$). This reduction suggests that *sigE* is required for tolerance of certain drugs. Indeed, this recapitulates the growth defects observed in a $\Delta sigE$ mutant when exposed to various drugs [35]. *sigE* is an alternative sigma factor that is thought to play a regulatory role in response to various stresses. This effect was only observable with a model that treats variations in saturation separately from magnitiudes of insertions.

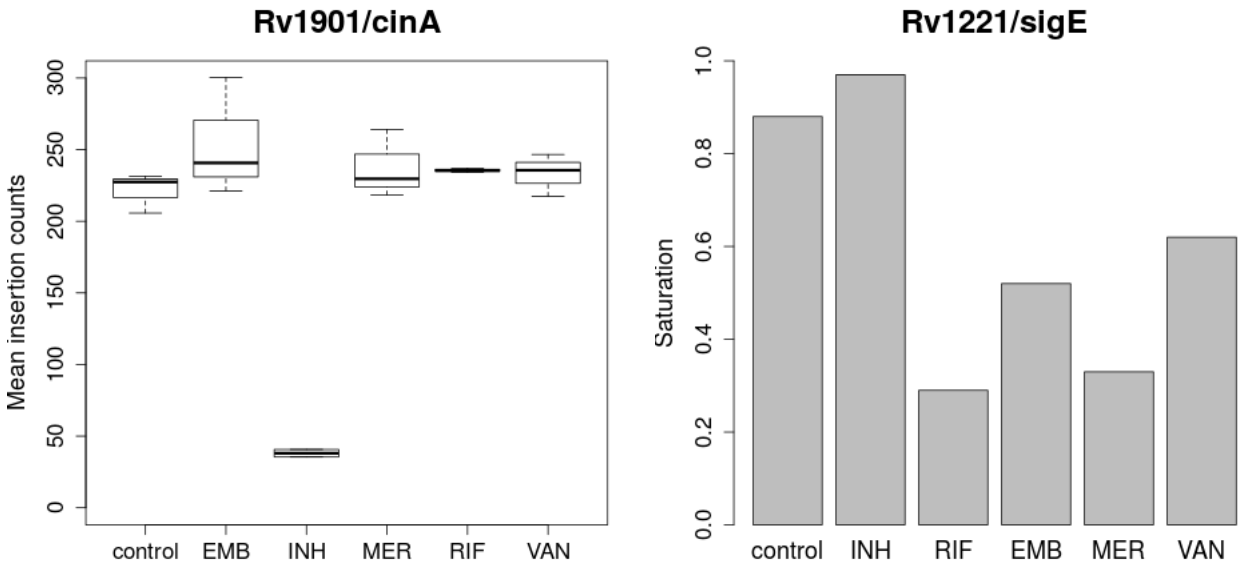

Figure 5: Significantly varying genes in cultures exposed to antibiotics. a) Mean insertion counts in CinA. b) Saturation in SigE (percent of TA sites with one or more insertions.

4 Discussion

TnSeq has proven to be such an effective tool for genome-wide assessment of functional requirements and genetic interactions in a wide range of prokaryotes that it is now being expanded to larger-scale experiments, such as profiling growth in media supplemented with an array of carbon sources or nutrients, or exposure to a variety of antibiotics/inhibitors, growth in a panel of different cell types, or infections in a collection of model-animals with different genetic backgrounds. Indeed, recent methods like BarSeq make such experiments efficient through barcoding of libraries, enabling highly multiplexed sequencing [36]. Statistical tools are needed to answer basic questions, such as what genes exhibit significant variability of insertion counts across the conditions. Resampling (or the permutation test) is designed for two-way comparisons. Attempting to perform resampling between all pairs of conditions does not scale up well, as the number of comparisons increase quadratically with number of conditions (for example, 20 conditions require 190 pairwise comparisons). In addition to the computational cost, there is a risk of loss of significance due to the p-value adjustment at the end, to control the overall false discovery rate. A classic statistical test for conditional-dependence among observations from multiple groups is ANOVA. But ANOVA relies on an assumption of Normality, and TnSeq insertion counts are not Normally distributed.

A more natural model for count data in other domains is the Negative Binomial distribution. NB would appear to be reasonable choice for insertion counts, given the "geometric-like" distribution of counts (i.e.

many sites with low counts; a few sites with high counts) observed in most datasets. Another advantage of NB is that it has an extra dispersion parameter (e.g. compared to Poisson) that provides an additional degree of freedom for fitting the variance in a set of counts. While it can often be difficult to obtain Maximum-Likelihood estimates of the dispersion parameter, that can be ameliorated by assuming the dispersion is the same across different conditions. This simplifies the calculations, and provides more data that can be used to estimate dispersion, without seriously affecting the identification of significantly varying genes.

While the NB distribution is a good choice to capture differences in mean counts, a specific problem with TnSeq data is the over-abundance of zeros, because most datasets are not fully saturated. While a subset of TA are expected to lack insertions because they are in essential genes, Tn libraries typically have insertions at only a fraction of the remaining sites (due to lack of diversity of the library, bottlenecking, etc.). This creates ambiguity over sites with counts of 0, as they could either be due to real biological effects, or just stochastically missing from the library.

To address this issue, we introduced the ZINB model, which is a mixture model of a zero-component (representing missing sites) and an NB component (for the distribution of insertion counts in a gene). For a given set of counts, the ZINB distribution has 3 parameters: μ (the mean of NB), r, the dispersion, and π, the degree of (un)saturation (percent of TA sites with counts of 0). Furthermore, the ZINB model is implemented as a regression model (with a log link function), with vectors or coefficients to represent how the insertion counts vary across conditions. Thus the zero-component captures the changes in the level of saturation of a gene across conditions, and the NB component captures how the magnitudes of counts vary across conditions. We compare the condition-dependent ZINB regression model to a condition-independent ZINB model, where the counts are all pooled, to determine whether there is a statistically significant improvement in the fit attributable to the conditions. Because of the the zero-component included in the ZINB model, there is a risk that comparisons among datasets with different levels of saturation could result in a systematic inflation of the number of false positives (i.e. genes that look like they vary because of differences in the fraction of TA sites hit). In fact, depending on the normalization procedure used, there can be a similar bias in the magnitudes of read counts that also causes more false positives when comparing datasets with widely-varying saturation. To compensate for this, we include terms in the regression ("offsets") for the overall saturation and non-zero mean count for each dataset. Thus the coefficients learned in the model are actually representing deviations (local to each gene) in count magnitudes and saturation, relative to the genome-wide averages for each dataset.

One advantage of the ZINB regression framework is that it can take into account additional information about samples as covariates. This is commonly done in RNA-seq for experiments with more complex design matrices [37]. Examples include relationships among the conditions, such as class of drug, concentration,

time of treatment/exposure, medium or nutrient supplementation, or genotype (for animal infections). By incorporating these in the model (with their own coefficients), it allows the model to factor out known (or anticipated) effects and focus on identifying genes with residual (or unexplained) variation. It can also be useful for eliminating nuisances like batch effects.

We showed with Likelihood Ratio Tests that ZINB is a much more suitable model for TnSeq data (insertion counts) than ANOVA or NB (even when taking into account differences in the number of parameters). We also showed in a synthetic experiment that failing to adjust for saturation differences (by incorporating the offset terms in the model) leads to a great increase in the false-positive rate when comparing datasets with unbalanced levels of saturation spanning a two-fold range (30-60%). On two real TnSeq datasets, ZINB identified nearly all the genes detected as significantly varying by resampling among all pairs of conditions. ZINB also identifies additional varying genes (10-20% in our tests). Some of these represent cases where the magnitudes of insertions do not vary significantly enough, but there is a significant depletion effect in the proportion of TA sites hit in different conditions. Notable examples include SecA2 in mice (a recognized virulence gene in *M. tuberculosis*), and SigE in the presence of antibiotics (regulator involved in stress response).

Of course, as with ANOVA, identifying genes that vary significantly is often just the first step and requires follow-up analyses to determine specific condition-dependent effects. For example, we observed (anecdotally) that the NAD-dependent, nucleoside-recycling gene *cinA* was required for tolerance of isoniazid. One could employ methods such as Tukey's range test [38] to drill down and identify significantly different pairs of conditions. Another approach would be to use principle-component analysis (PCA) to uncover trends/patterns among TnSeq profiles and identify clusters of conditions producing similar effects genome-wide [39].

Still, our results establish the suitability of ZINB as a model for TnSeq data (insertion counts). Examples of genes where the phenotype is primarily observed in the saturation of the read-counts, such as SecA2 and SigE, highlight the advantage of modeling condition-dependent effects on both the magnitudes of counts in a gene and local level of saturation independently. Thus, ZINB regression is an effective tool for identifying genes whose insertion counts vary across multiple conditions in a statistically significant way.

Supplemental Tables

- **Supplemental Table 1** - ZINB output file (spreadsheet, tab-separated format) containing results from Section 3.2 on comparison of a TnSeq library for *M. tuberculosis* H37Rv grown in-vitro versus in C57BL/6 mice. For each ORF in the genome, an analysis of the mean, NZ-mean, and saturation in

each condition is given, along with p-value and adjusted p-value. Significantly varying genes are those with $p_{adj} < 0.05$.

- **Supplemental Table 2** - ZINB output file containing results for Section 3.4 on comparison of a TnSeq library for *M. tuberculosis* H37Rv grown in presence of five antibiotics, isoniazid (INH), rifampicin (RIF), ethambutol (EMB), meropenem (MER), and vancomycin (VAN).

- **ZINB Test Example (zip file)** - Experimental data for the mouse experiment in Section 3.2 (.wig files with transposon insertion counts at TA sites), along with Instructions (README.docx) on how to do the ZINB analysis in TRANSIT (commands and output files).

Funding

This work was supported by NIH grant U19 AI107774 (TRI, SE, DS, and CMS).

References

[1] van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol. 2013 Jul;11(7):435–442.

[2] Lampe DJ, Churchill ME, Robertson HM. A purified mariner transposase is sufficient to mediate transposition in vitro. the The European Molecular Biology Organization Journal. 1996;15(19):5470–5479.

[3] Zomer A, Burghout P, Bootsma HJ, Hermans PW, van Hijum SA. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. PLoS ONE. 2012;7(8):e43012.

[4] DeJesus MA, Zhang YJ, Sassetti CM, Rubin EJ, Sacchettini JC, Ioerger TR. Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. Bioinformatics. 2013 Mar;29(6):695–703.

[5] Solaimanpour S, Sarmiento F, Mrazek J. Tn-seq explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries. PLoS ONE. 2015;10(5):e0126070.

[6] Zhang YJ, Ioerger TR, Huttenhower C, Long JE, Sassetti CM, Sacchettini JC, et al. Global assessment of genomic regions required for growth in Mycobacterium tuberculosis. PLoS Pathog. 2012 Sep;8(9):e1002946.

[7] Deng J, Su S, Lin X, Hassett DJ, Lu LJ. A statistical framework for improving genomic annotations of prokaryotic essential genes. PLoS ONE. 2013;8(3):e58178.

[8] DeJesus MA, Ioerger TR. A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. BMC Bioinformatics. 2013;14:303.

[9] van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods. 2009 Oct;6(10):767–772.

[10] DeJesus MA, Ambadipudi C, Baker R, Sassetti C, Ioerger TR. TRANSIT–A Software Tool for Himar1 TnSeq Analysis. PLoS Comput Biol. 2015 Oct;11(10):e1004401.

[11] Santiago M, Matano LM, Moussa SH, Gilmore MS, Walker S, Meredith TC. A new platform for ultra-high density Staphylococcus aureus transposon libraries. BMC Genomics. 2015;16.

[12] Zhao L, Anderson MT, Wu W, T Mobley HL, Bachman MA. TnseqDiff: identification of conditionally essential genes in transposon sequencing studies. BMC Bioinformatics. 2017 Jul;18(1):326.

[13] Pritchard JR, Chao MC, Abel S, Davis BM, Baranowski C, Zhang YJ, et al. ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing. PLoS Genet. 2014 Nov;10(11):e1004782.

[14] Zhang YJ, Reddy MC, Ioerger TR, Rothchild AC, Dartois V, Schuster BM, et al. Tryptophan biosynthesis protects mycobacteria from CD4 T-cell-mediated killing. Cell. 2013 Dec;155(6):1296–1308.

[15] Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, et al. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. mBio. 2015;6(3):e00306–15.

[16] DeJesus MA, Nambi S, Smith CM, Baker RE, Sassetti CM, Ioerger TR. Statistical analysis of genetic interactions in Tn-Seq data. Nucleic Acids Res. 2017 Feb;.

[17] Xu W, DeJesus M, Rcker N, Engelhart C, Wright M, Healy C, et al. Chemical Genetic Interaction Profiling Reveals Determinants of Intrinsic Antibiotic Resistance in Mycobacterium tuberculosis. Antimicrobial Agents Chemotherapy. 2017;61(22):e01334–17.

[18] Yang G, Billings G, Hubbard T, Park J, Yin-Leung K, Liu Q, et al. Time-Resolved Transposon Insertion Sequencing Reveals Genome-Wide Fitness Dynamics during Infection. mBio. 2017;8(5):e01581–17.

[19] Lloyd-Smith JO. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS ONE. 2007 Feb;2(2):e180.

[20] Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. Journal of Statistical Software. 2007;27(8):1–25.

[21] Benjamini Y, Yekutieli D. False discovery rate controlling confidence intervals for selected parameters. Journal of the American Statistical Association. 2005;100(469):71–81.

[22] Griffin JE, Gawronski JD, DeJesus MA, Ioerger TR, Akerley BJ, Sassetti CM. High-Resolution Phenotypic Profiling Defines Genes Essential for Mycobacterial Growth and Cholesterol Catabolism. PLoS Pathog. 2011 09;7(9):e1002251.

[23] Pandey AK, Sassetti CM. Mycobacterial persistence requires the utilization of host cholesterol. Proc Natl Acad Sci, USA. 2008;105(11).

[24] Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989;57:307–333.

[25] Stavrum R, Stavrum AK, Valvatne H, Riley LW, Ulvestad E, Jonassen I, et al. Modulation of transcriptional and inflammatory responses in murine macrophages by the Mycobacterium tuberculosis mammalian cell entry (Mce) 1 complex. PLoS ONE. 2011;6(10):e26295.

[26] Maloney E, Stankowska D, Zhang J, Fol M, Cheng QJ, Lun S, et al. The Two-Domain LysX Protein of Mycobacterium tuberculosis Is Required for Production of Lysinylated Phosphatidylglycerol and Resistance to Cationic Antimicrobial Peptides. PLoS Pathogens. 2009;5(7):e1000534.

[27] Tischler AD, Leistikow RL, Kirksey MA, Voskuil MI, McKinney JD. Mycobacterium tuberculosis Requires Phosphate-Responsive Gene Regulation To Resist Host Immunity. Infection and Immunity. 2013;81:317–328.

[28] Sullivan JT, Young EF, McCann JR, Braunstein M. The Mycobacterium tuberculosis SecA2 system subverts phagosome maturation to promote growth in macrophages. Infection and Immunity. 2011;80(3):996–1006.

[29] Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Molecular Microbiology. 2003;48(1):77–84.

[30] DeJesus MA, Ioerger TR. Normalization of transposon-mutant library sequencing datasets to improve identification of conditionally essential genes. J Bioinform Comput Biol. 2016 Jan;p. 1642004.

[31] Lakshminarayana SB, Huat TB, Ho PC, Manjunatha UH, Dartois V, Dick T, et al. Comprehensive physicochemical, pharmacokinetic and activity profiling of anti-TB agents. Journal of Antimicrobial Chemotherapy. 2015;70(3):857–867.

[32] Karuppiah V, Thistlethwaite A, Dajani R, Warwicker J, P DJ. Structure and mechanism of the bifunctional CinA enzyme from Thermus thermophilus. Journal of Biological Chemistry. 2014;289(48):33187–97.

[33] Stehr M, Elamin AA, M S. Pyrazinamide: the importance of uncovering the mechanisms of action in mycobacteria. Expert Review of Anti-Infective Therapy. 2015;13(5):593–603.

[34] Vilchze C, Torin R, Weisbrod BC, Kremer L, Manzour H, Hazbn FW, et al. Altered NADH/NAD+ Ratio Mediates Coresistance to Isoniazid and Ethionamide in Mycobacteria. Antimicrobial Agents and Chemotherapy. 2005;49(2):708–720.

[35] Pisu D, Provvedi R, Espinosa DM, Payan JB, Boldrin F, Pal G, et al. The Alternative Sigma Factors SigE and SigB Are Involved in Tolerance and Persistence to Antitubercular Drugs. Antimicrobial Agents and Chemotherapy. 2017;61(12):e01596–17.

[36] Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, et al. Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. mBio. 2015;6(3):e00306–15.

[37] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 2012;40(10).

[38] Tukey J. omparing Individual Means in the Analysis of Variance. Biometrics. 1949;5(2):99–114.

[39] Hubbard TP, D'Gama JD, Billings G, Davis BM, Waldor MK. Unsupervised Learning Approach for Comparing Multiple Transposon Insertion Sequencing Studies. mSphere. 2019 02;4(1).