Model-based identification of conditionally-essential genes from transposon-insertion sequencing data

Vishal Sarsani[1], Berent Aldikacti[2], Shai He[1], Rilee Zeinert[3], Peter Chien[2], Patrick Flaherty[1*]

[1]Department of Mathematics and Statistics, University of Massachusetts Amherst
[2]Department of Biochemistry and Molecular Biology, University of Massachusetts Amherst
[3]Division of Molecular and Cellular Biology, Eunice Kennedy Shriver National Institute of Child Health and Human Development

* pflaherty@umass.edu

Abstract

The understanding of bacterial gene function has been greatly enhanced by recent advancements in the deep sequencing of microbial genomes. Transposon insertion sequencing methods combines next-generation sequencing techniques with transposon mutagenesis for the exploration of the essentiality of genes under different environmental conditions. We propose a model-based method that uses regularized negative binomial regression to estimate the change in transposon insertions attributable to gene-environment changes without transformations or uniform normalization. An empirical Bayes model for estimating the local false discovery rate combines unique and total count information to test for genes that show a statistically significant change in transposon counts. When applied to RB-TnSeq (randomized barcode transposon sequencing) and Tn-seq (transposon sequencing) libraries made in strains of *Caulobacter crescentus* using both total and unique count data the model was able to identify a set of conditionally essential genes for each target condition that shed light on their functions and roles during various stress conditions.

Author summary

Transposon insertion sequencing allows the study of bacterial gene function by combining next-generation sequencing techniques with transposon mutagenesis under different genetic and environmental perturbations. Our proposed regularized negative binomial regression method improves the quality of analysis of this data.

Introduction

A central question in molecular genetics is, What genes are essential for life? Prior to the advent of high-throughput technology this question was addressed by mutagenesis and fine mapping [1, 2]. The simplicity of homologous recombination in *S. cerevisiae* allowed for the generation of a complete mutant library containing strains each with a complete knockout of a single gene and tagged with a unique genetic barcode [3]. Subsequent analysis of this library by custom microarrays and sequencing revealed genes essential for growth in rich media as well as *conditionally essential* genes — genes that are dispensable in rich media, but are essential in different environmental conditions [3–5]. However, generating a mutant pool from individual genetic knockout strains is labor-intensive and not feasible in organisms for which homologous recombination is inefficient. Transposon sequencing (Tn-seq) methods have alleviated this problem and provide a powerful method for identifying essential and dispensable genes under a variety of environmental conditions and genetic backgrounds.

Transposon Sequencing Transposon sequencing uses a modified transposon to generate a saturation mutant library of a background strain of interest. Each transposon has a selectable marker; a unique, random DNA barcode (in some cases); and loci for PCR amplification that can be used to identify the DNA adjacent to the transposon insertion site [6, 7]. Once the transposon mutant library is generated, it can be grown in various environmental conditions of interest. Strains that have a fitness defect due to the transposon insertion grow more slowly or not at all. The abundance of the transposon insertion mutant strain in the library can be assayed by sequencing the library after growth and counting the reads that map to a particular insertion site. For each gene, the change in the count of sequenced transposon insertions between the control and the perturbed environment can be used to identify conditionally essential or conditionally dispensable genes.

Since the introduction of the original Tn-seq method, many variations have been developed to facilitate the study of a wider range of organisms or to improve efficiency [6]. Random-barcode transposon sequencing amortizes the cost of multiple environmental perturbation experiments by doing the expensive mapping of transposon insertion site to random barcodes once and then using that mapping for all future experiments [8]. Transposon sequencing technology addresses the time consuming and often technically challenging process of generating one-at-a-time gene deletions by using parallel mutagenesis and counting-by-sequencing [9]. But, this technology has introduced a new, statistical problem. How can the transposon count data be used to test the hypothesis that a gene is essential such that all of, and only, the essential/dispensable genes are identified?

Related Work There are several existing statistical approaches for analyzing transposon sequencing data. van Opijnen et al. [10] used several normalization steps to compute a ratio of the fold-expansion of the mutant relative to the rest of the population. Then, a t-test with a Bonferroni correction was used for each gene to decide if a change in the fitness statistic is significant. This type of normalization renders the statistic independent of growth duration, but requires an additional calibration experiment to estimate an expansion factor which measures the growth of the bacterial population during library selection. Despite these benefits, the fitness effect estimator is non-linearly dependent on the calibration factor factor because it appears in both a logarithm and in the denominator of the fitness effect ratio.

Wetmore et. al. [8] dispensed with the calibration step and still found good estimates of fitness effect. They computed the log-ratio of start-time t_0 count to the stop-time t_{after} count. They added a pseudo-count term to regularize noisy estimates for low counts. These low count observations were filtered out in [10].

ESSENTIALS is a software package developed by [11] that uses Loess [12] normalization followed by the application of edgeR [13], a software package developed for identifying differentially expressed genes from RNA-seq data, to call essential genes. They demonstrated that their package is robust to transposon sequencing technology—a significant benefit as TnSeq experimental methods continue to be revised and improved.

DeJesus et. al. [14] developed a full Bayesian model for Tn-seq count data. They approached the problem by defining a Boolean variable to represent whether a gene is essential or non-essential. In their method, the data for a gene includes the number of insertions, the longest run of non-insertions, and the span of nucleotides of the longest run of non-insertions. This additional information beyond the number of insertion counts is informative and the Bayesian model elegantly incorporates all of the data into a posterior probability of essentiality.

Subramaniyam et. al. [15] focuses on fine-resolution mapping of essential regions. Their method applies to transposon libraries constructed with the *mariner* transposon family which preferentially inserts in TA dinucleotides. Their method models the number of transposon insertions at each TA dinucleotide site rather than aggregating by gene. Because many TA dinucleotide sites are unlikely to harbor any transposon insertions, they employ a zero-inflated negative binomial model to accommodate the many zero counts. It should be noted that [10] and [8] include a normalization for the number of Tn counts at the start of the experiment, but more recent model-based work does not require this normalization [14, 15].

Contributions Our work builds upon these previous works in several ways. Like [15], our approach employs a negative binomial generalized linear model to use information from the entire experimental data set rather than using only pairs of experiments. Our model employs a Bayesian prior over coefficients as in [14] that manifests as a regularization term in the regression formulation. Our work differs from these efforts in that we aggregate transposon counts at the gene level in the context of a negative binomial model with nested effects which allows our model to be robust to the transposon library creation method, and we use a joint false discovery rate approach to call essential/dispensable genes. Our contributions are: (1) a regularized negative binomial model with nested effects to estimate the effect of varying environmental conditions in the context of genetic background, (2) the use of both unique Tn insertions and total Tn insertions to improve sensitivity and specificity, (3) the use of a joint local false discovery rate control to call conditionally essential/dispensable genes.

Problem statement The goal of this work is to identify all of the genes that are essential or dispensable in the context of a particular combination of genetic background and environmental condition. Let us denote the genetic background of the experiment $g \in \mathcal{G}$ and the environmental condition $e \in \mathcal{E}$. Note that not all pairwise combinations in $\mathcal{G} \times \mathcal{E}$ may be available in a data set. For a given combination (g, e) the data set contains R_{ge} replicate experiments; we index the replicate with r. In experiment (g, e, r), there are N_{ger} observed transposon insertions that are mapped to genes (perhaps excluding some trimmed region around the start and stop codon of the gene). We reduce the raw data to two features for each gene: (1) the **total** count of insertions and (2) the count of **unique** insertions. For gene i and experiment (g, e, r), let y^{tot}_{geri} be the total count of insertions and let y^{uniq}_{geri} be the count of unique insertions.

Let the number of genes under investigation be $m = |\mathcal{G}|$. The null hypothesis (H_0) is that a gene is non-essential (not essential or dispensable) and suppose that there is a true number $m_0 \leq m$ of such genes. The goal of our method is to declare some set $\mathcal{R} \subseteq \mathcal{G}$ to be essential/dispensable such that $\mathcal{R}$ contains all of the genes that are truly essential/dispensable and none of the genes that are non-essential. This task is often too challenging and instead a more approachable task is to ensure that the rate of false discoveries in $\mathcal{R}$ is bounding in probability. Therefore, the problem is to identify a set $\mathcal{R}$ of called essential/dispensable genes such that the false discovery rate is bounded.

Materials and Methods

RB-TnSeq experimental methods

RB-TnSeq uses a randomly barcoded transposon to amortize the cost of many related experiments [8]. Barcoded transposon donor plasmids are transferred to the cell of interest by either electroporation or conjugation. Subsequently, cells containing plasmids are selected using a selection media, and small aliquots are frozen in 10% glycerol. The frozen aliquot is the mutagenesis libraries used in all experiments. In RB-Tnseq, a sequencing run is done on the libraries to assign each barcode to its genomic location. For subsequent experiments on these libraries, a simple single PCR step is required to amplify and count the barcodes.

Read mapping and pre-processing. In this study, we used RB-TnSeq data of *Caulobacter crescentus* and *Pseudomonas fluorescens FW300-N1B4* from Price et. al. [16]. As input we have downloaded `all.poolcounts` (`http://genomics.lbl.gov/supplemental/bigfit/`), and generated two different count files from it. The first, labeled "total counts", are the sum of all insertions aggregated by each gene. The second is the "unique counts", where instead of using the sum of all insertions, we have used the sum of the number of unique barcodes that have non-zero reads per gene.

Tn-seq experimental methods

Transposon mutagenesis libraries used in this study were generated as previously described [17]. Briefly, wild-type (wt) and Δ*lon Caulobacter crescentus NA1000* strains were grown until mid-log phase, pelleted, washed three times with 10% glycerol, and transformed with EzTn5 <Kan-2> transposomes (Lucigen) by electroporation. Following recovery in PYE, transformed cells were plated on PYE + Kan selection media and grown for 7 days. Colonies were scraped, pooled, and frozen in PYE + 20% glycerol in 1 ml aliquots and frozen for further experiments. For stress condition experiments, 2 aliquots of each library was thawed and separately recovered overnight in 2 x 10 ml of PYE in a 30°C shaker. These saturated cultures were then stressed as described below. All conditions were performed in quadruplicates, optical density (OD) measurements were taken at 600 nm.

Control environment. Libraries were back diluted to OD 0.008 into 7 ml of PYE and grown overnight until they reach saturation at OD ∼1.6.

Heat shock stress. One ml of the overnight culture was heat-shocked at 42°C for 45 minutes in a heat-block, then back-diluted to OD 0.008 and grown overnight until saturation.

L-canavanine. Overnight cultures of cells were back diluted to OD 0.008 in 7 ml of PYE + 100 ug/ml L-canavanine and grown at 30°C for 90 minutes. After 90 minutes of L-canavanine stress, the cells were spun for 10 minutes at 5000 rpm, washed once with PYE, spun again, then resuspended with 7 ml of PYE, and recovered overnight until they reached saturation.

Library preparation Following overnight growth, 1.5 ml of saturated culture from each Tn library was pelleted at 15,000 RPM for 1 minute and gDNA was extracted by MasterPure Complete DNA and RNA purification kit according to manufacturer's protocol. Sequencing libraries were prepared for Next-generation sequencing via three PCR steps. Indexed libraries were pooled and sequenced at the University of Massachusetts Amherst Genomics Core Facility on a NextSeq 500 (Illumina).

Read mapping and pre-processing Mapping and pre-processing of the Tnseq raw data was done as described previously with some modifications [17]. Briefly, samples were de-multiplexed, and unique molecular identifiers (UMIs) were added during PCR steps removed using `Je` [18]. Clipped reads mapped to the *Caulobacter crescentus* NA1000 genome (NCBI Reference Sequence: NC011916.1) using `bwa`, sorted with `samtools` [19, 20]. Duplicate transposon reads removed by `Je` and indexed with `samtools`. Genome positions are assigned to the $5'$ position of transposon insertions using `bedtools genomecov` [21]. Subsequently, the `bedtools` map used to count either the total number of transposon insertions per gene using the `bedtools map -o sum` argument or the unique number of insertions using the `bedtools map -o count` argument.

In-vivo validation Overnight cultures of wild-type and Δ*clpA* *Caulobacter crescentus* strains each mixed at a 1:1 ratio with a reporter strain constitutively expressing fluorescent *Venus.*(CPC798) The mixtures were kept at either 30°C or heat-shocked at 42°C for 45 minutes in a thermocycler. After the heat-shock, the mixtures were diluted to 1:4000 in PYE media and allowed to grow for 24 hours (~ 12 doublings) at 30°C. Number of fluorescent control (Venus) and nonfluorescent tester (WT or Δ*clpA*) cells were counted in both the initial mixture and after 24 hour growth using phase contrast and fluorescent microscopy. The same tester/control normalization coefficients were used for initial and 24 hour time points for each strain (normalizaton coefficient $= 1/(\text{tester}/\text{control})$ at time $= 0$). and time $= 24$ by adjusting the time $= 0$ ratios to 1 for each strain. Normalized 24 hour ratios are what we are reporting as competitive index (Figure 6). An index of > 1 means the tester condition were able to grow faster compared to the control and an index of < 1 means the tester grew slower compared to the control. Quantifications of at least 100 cells were performed for each condition with replicates when possible.

Regularized negative binomial regression

Our approach for integrating all of the experimental data to estimate the effect of the genetic background and the environmental condition is based on a generalized linear model framework. Here, we describe the negative binomial model framework, the nested effects model matrix structure, and the form and rationale for regularization.

Negative binomial model The generalized linear model consists of three components: (1) a probability distribution for the sampling error, (2) a model matrix structure, and (3) a link function connecting the expected value of the response to the covariates. It has been observed that Tn-seq count data is often overdispersed and therefore, the data is better fit by a negative binomial distribution rather than a Poisson distribution because of the additional free parameter to allow for a variance that does not directly depend on the mean parameter. The link function that is often chosen for a negative binomial distribution is a log function and we do so here. The generalized linear model takes the form $E(\mathbf{y}_i|\mathbf{x}) = f^{-1}(\mathbf{x}\beta)$, where $\mathbf{y}_i$ is the vector of observed Tn counts across all experiments in the data set for gene i, $\mathbf{x}$ is the model matrix, β is the vector of parameters, and f^{-1} is the log link function.

Nested effects in generalized linear regression model The model matrix must be designed to specifically address the questions of interest of the data. First, we are interested in the main effect of the genetic background in relation to the wild-type strain. For example, if a there is a drastic reduction in Tn counts in a mutant background relative to wild-type, it indicates that the gene is essential conditional on the strain mutation(s). Likewise if there is a drastic increase in Tn counts in a mutant background relative to wild-type, the gene is likely dispensable conditional on the strain mutation(s). Second, we are interested in the effect of the environmental condition, but only in the context of the genetic background. For example, if there is a reduction in Tn counts in the $g = \Delta$YFG background relative to the wild-type background in rich media growth conditions, but then no change when shifted to a heat-stress, the gene may be viewed as interesting in the genetic background, but not in the conditions specific to heat-stress. One would expect that if a gene is essential in the $g = \Delta$YFG background that it continues to be essential in all environmental conditions — only deviations from that expectation should be flagged as scientifically interesting. These questions of interest logically lead to the consideration of a nested effects model matrix structure:

$$E(\mathbf{y}_i|\mathbf{x}) = f^{-1}(\beta_0 + \mathbf{x}_g\beta_g + \mathbf{x}_{e|g}\beta_{e|g}), \tag{1}$$

where $\mathbf{x}_g$ and $\mathbf{x}_{e|g}$ are the standard indicator matrix encodings for the genetic background and nested environmental condition respectively. Note that this nested model matrix structure is different than the one usually employed for modeling interactions in that there is no term corresponding to the main effects of the environmental condition $\mathbf{x}_e$. Structuring the model matrix in this way allows the inferential products of the model (the model parameters) to inform the scientifically interesting questions we have of the data.

A way to interrogate this data is to observe the baseline number of total and unique insertions in the wild-type background strain with no stress (control). An excess or depletion of insertions in the Δlon background are viewed as a shift from the control. Finally, an excess or depletion of the stress conditions is viewed relative to the particular background strain the library was created in. This interpretation of the data leads to the nested effects model proposed here.

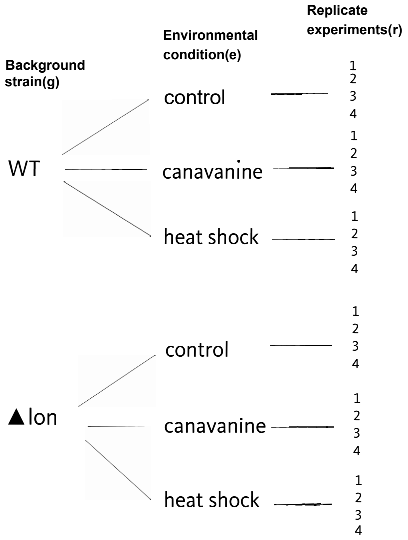


Fig 1. Example of nested experimental design of Tn-seq data. Shown are two background strains: WT and Δlon, and three nested environmental perturbations: control, canavanine, heat shock. Each perturbation experiment is replicated four times.

Regularization Estimating the model parameters when the number of transposon count is small has been noted by others and handled either by filtration [10] or the addition of pseudo-counts [8]. The low counts in response variables can result in inflated regression coefficients and are susceptible to very high variance. They also affect false discovery rate procedures increasing the risk of type-I errors.

A gene that has zero observed transposon insertions in a condition is considered frankly conditionally essential and a model is not needed to make that decision. Therefore, we filter for genes in the set $\mathcal{G} \setminus \mathcal{G}^0_{ge}$ where $\mathcal{G}^0_{ge} = \{i | \min(\bar{y}_{gei}, \bar{y}_{\text{control},i}) < 1\}$ for each genetic backgrounds (g) and environment (e), where $\bar{y}_{gei}$ is the average across replicates. The model is useful only for genes where the conditional essential decision is ambiguous and we restrict the modeling to those genes.

For genes that are not frankly essential, we employ a regularization methodology that has proven successful in many statistical contexts and has Bayesian as well as classical statistical rationale [22–24]. Regularization can be viewed as a prior distribution on the regression coefficients,

$$\beta \sim \text{Gaussian}(\lambda). \qquad (2)$$

The Gaussian prior converts the maximum likelihood estimation problem for the regression coefficients to a penalized maximum likelihood estimation problem with an L_2 norm penalty or equivalently a maximum a-posteriori estimation problem. The parameters for the penalized count regression are estimated by a combination of the iteratively reweighted least squares (IRLS) algorithm and coordinate descent algorithm as implemented in the `mpath` package [25].

We have found that this regularization effectively shrinks large coefficient estimates due to small Tn counts. However, it does not address situations where there are exactly zero counts. In those cases, our model is not necessary — the gene can be considered conditionally essential in the condition with high confidence. Therefore, we restrict our modeling to genes that have non-zero Tn counts in all experiments in the data set.

Local false discovery rate

The regularized negative binomial generalized linear model was fit to both the total count data, $\mathbf{y}_i^{\text{tot}}$, and the unique count data, $\mathbf{y}_i^{\text{uniq}}$ independently for each gene i. The next task is to decide if a gene is conditionally essential/dispensable or non-essential. In a generalized linear model the response is conditionally independent of a covariate given all the other covariates in the model if and only if the associated model coefficient is equal to zero (for proof see [26]). Therefore, under the model-based framework testing if a gene is conditionally essential or dispensable is equivalent to testing whether the model coefficient is equal to zero.

Under the assumption that a large fraction of the genes under investigation are non-essential, the local false discovery rate can be used to control the proportion of false positives in the set of called essential/dispensable genes [27]. The central idea is to fit a Gaussian distribution to the center of the empirical distribution of coefficients for a given effect across all genes. Genes that have a coefficient that is unlikely under that distribution are called essential/dispensable. There is abundant theory to support the use of this procedure to control the proportion of false discoveries [28–30].

The false discovery rate of the regression coefficient is

$$\text{Fdr}(\beta_{ic}) = \text{Prob}\{\text{gene } i \text{ is null in condition } c \mid |\beta_{ic}| \geq \bar{\beta}\} \qquad (3)$$

The local false discovery rate makes use of a mixture model framework with two components. It fits a Gaussian distribution to the center of the empirical distribution of the regression coefficients β_{ic} across genes. Genes associated with coefficients that are not attributable to the central Gaussian are called conditionally essential or dispensable [31].

Intersection of marginal local false discovery tests The standard false discovery rate approach only considers the coefficients estimated from one model, however, in our analysis, we estimate coefficients from the model fit to $\mathbf{y}^{\text{tot}}$ and the model fit to $\mathbf{y}^{\text{uniq}}$. Yet, we would like a single decision as to whether the gene is non-essential or not. Our approach is to take the intersection of the decisions from the two models. That is, only genes that are deemed essential/dispensable on the basis of both unique counts and total counts are retained. This approach has the effect of reducing the number of calls and thus the number of false positives.

Results

We generated simulated data on 4,000 genes under 3 simulated knockout backgrounds and 4 environmental conditions with 5 replicates for each combination of strain background and environment. We compared the fit of the regularized negative binomial model to a zero-inflated negative binomial model of the type used by [15] and to a unregularized negative binomial model [11].

Our method was then applied to two independent data sets using different transposon sequencing methods. First, our method was applied to RB-TnSeq data. This data set explored the essential genes in many organisms across varying carbon sources, nitrogen sources, and environmental stress conditions. We selected only the *Caulobacter crescentus* data set for this study. The background genotype for all the RB-TnSeq experiments is wild-type so no synthetic lethality combinations are identifiable. Second, our method was applied to Tn-seq data that was collected in our lab. Both wild-type and a Δlon knockout strain were used as genetic backgrounds for library preparation. These strain pools were subjected to heat-shock stress and canavanine. Each condition was replicated at least two times in biological replicates.

Simulation Experiments

We simulated samples from total of three background strains (g) with four conditions (e) and each condition having five replicates (r). First the dispersion parameter was sampled from a Gamma distribution for each condition and for 8 intervals (l) each containing 500 genes. The hyper-parameters of the Gamma distribution were drawn from uniform distributions as

$$\begin{aligned} a_g \sim U(0,5),\ b_g \sim U(0,5) \quad & \text{for } g = 1,2,3, \\ \theta_{gel} \sim \text{Gamma}(a_g, b_g) \quad & \text{for } g = 1,2,3, e = 1,\ldots,4, l = 1,\ldots,8. \end{aligned} \tag{4}$$

The number of unique insertions for each gene was sampled from a negative binomial distribution with mean parameters shared across groups of 500 genes, $\mu = (0.5, 1, 2, 4, 8, 16, 32, 64)$,

$$\mathbf{y}^{\text{uniq}}_{gerl} \sim NB(\mu_l, \theta_{sl}). \tag{5}$$

This simulation provides the number of *unique* transposon counts for each gene. For every gene, the total transposon insertion counts were obtained by sampling from a negative binomial distribution with mean $\mu = 100$ and dispersion $\theta = 1$ for each unique insertion site previously generated

$$y^{\text{tot}}_{geri} \sim \sum_{s=1}^{y^{\text{uniq}}_{geri}} NB(\mu = 100, \theta = 1). \tag{6}$$

Regularized negative binomial model reduces over-fitting Out of 4,000 simulated genes, there were 82 for which the regularized negative binomial model fitting algorithm did not converge leaving 3,918 simulated genes for comparison to other algorithms. We observed that for 3,456(86.67%) genes, the regularized negative binomial model had a better fit as measured by residual variance compared to a unregularized negative binomial model [11]. Figure 2 shows the mean counts and the residual variance for each of the 3,918 models. Clearly, the negative binomial model alone fits poorly for low mean count values. Supplementary Figure 1 shows the mean counts and the residual variance for a two condition setup. Even though the regularized negative binomial model has higher variance in the residual variance across genes, on a per-gene basis, the residual variance for the regularized negative binomial model is lower than the zero-inflated negative binomial model and the negative binomial model for the vast majority (86.67%) of genes.

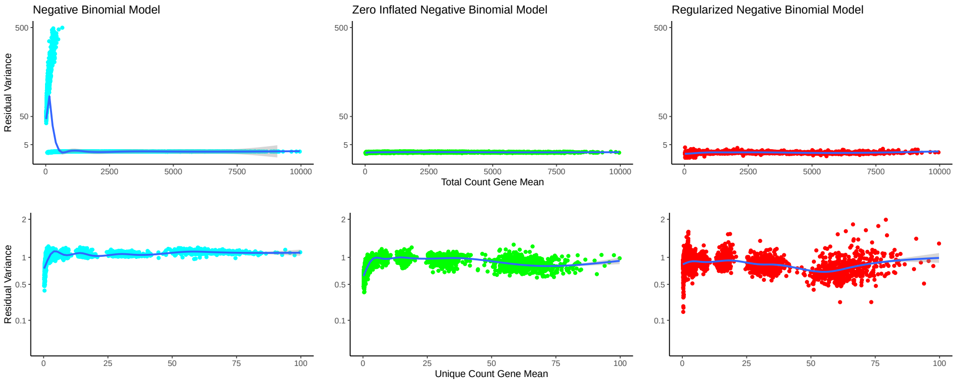


Fig 2. A simple negative binomial model (left) does not fully capture the variance in genes with low counts. Zero-inflated negative binomial (center) model overfits count data, attributing almost all variation to strain and conditional effects. As a result, almost every gene exhibits low residual variance. A regularized negative binomial error model (right) successfully captured the mean-variance relationship inherent in the data independent of gene counts. Mean-variance trendline shown in blue for each panel.

Analysis of RB-TnSeq data

We fit the regularized negative binomial model to RB-TnSeq data [8]. We selected all *Caulobacter crescentus* and *Pseudomonas fluorescens* experiments and grouped the conditions into carbon-source, nitrogen-source, and stress conditions. The stress conditions, such as heat-stress, antibiotic addition, etc, were conducted in rich media (PYE or LB), while the carbon and nitrogen source changes were conducted in minimal media. The control (wild-type, no stress) experiments were conducted in rich media. The lack of replicate experiments in this data set prevents us from inferring high-confidence conditionally essential genes in finer resolution conditions.

Caulobacter crescentus results Of the 3,312 genes with at least one transposon insertion, we identified 2/75 (total/unique) as frankly conditionally essential in carbon, 5/75 (total/unique) in nitrogen, and 5/75 (total/unique) in stress by the criteria described previously that the average transposon counts in one or more conditions is less than one. Figure 3(A-C) shows the conditionally essential and dispensable genes in each of the conditions considered for this data set (excluding the frankly conditionally essential genes). Each data point is a gene and genes labeled as green diamonds are called conditionally essential/dispensable by the local FDR criterion. It is clear that many genes are called conditionally essential (decrease in both total and unique transposon insertions) in both the carbon and nitrogen shift conditions. Our hypothesis is that these genes are required for general biosynthetic processes necessary to survive in minimal media conditions. Figure 4(A) shows the intersection of the gene sets identified in these two conditions and the high degree of overlap and the identities of the genes supports this hypothesis.

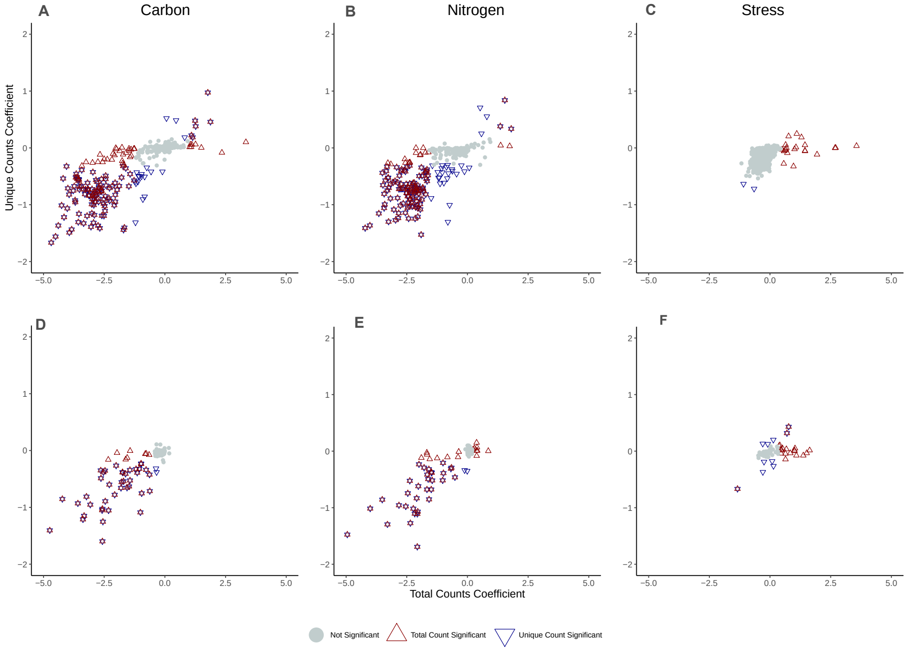


Fig 3. Conditionally essential/dispensable genes in the published RB-TnSeq data set for *Caulobacter crescentus NA1000* (A-C) and *Pseudomonas fluorescens FW300-N1B4* (D-F) [8].

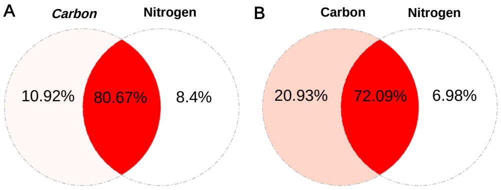


Fig 4. Venn diagram showing a high degree of overlap between genes identified in carbon and nitrogen shift conditions in *Caulobacter crescentus NA1000* (A) and *Pseudomonas fluorescens FW300-N1B4* (B) indicating genes involved in the shift to minimal media are identified.

In total there are 21 conditionally essential/dispensable genes by total insertion counts and 2 conditionally essential/dispensable genes by unique insertion counts for the stress condition. The two genes that are conditionally essential by unique counts: CCNA_03859 (*cenR*), known to be critical for envelope maintenance [32], and CCNA_03346 *ruvC*, a nuclease important for homologous recombination. Because so many of the tested stresses involve the cell envelope either directly (ethanol, polymyxin, etc) or indirectly rely on components in the cell envelope (drug transporters), it is not surprising that a cell envelope maintenance gene like *cenR* would be important for many of these stresses. Because many stresses also lead to DNA damage (cisplatin, metals, etc) we reason that the conditional essential nature of *ruvC* stems from its crucial role in resolving crossover junctions, a critical step for DNA damage repair by homologous recombination [33].

***Pseudomonas fluorescens* results** No genes were identified as frankly conditionally essential by the criteria described previously. Figure 3(D-E) shows the conditionally essential and dispensable genes as identified by the mode in each of the conditions considered for this data set. There are two conditionally dispensable genes and one conditionally essential gene by both measures of total insertion counts and unique insertion counts for the stress condition. First, a conditionally dispensable gene, Pf1N1B4_2858 (*CbrB*), is a two-component sensor needed for cells to use a variety of carbon or nitrogen sources [34]. The second conditionally dispensable gene is Pf1N1B4_1906, a Shikimate 5-dehydrogenase which would influence synthesis of aromatic amino acids. The conditionally essential gene is Pf1N1B4_2106, also known as OxyR, a hydrogen peroxide-inducible transcriptional activator which controls expression of oxidative stress response proteins. The conditional essentiality of this gene makes mechanistic sense because many of the stress conditions impact the oxidative stress system.

Analysis of Tn-seq data

We fit the regularized negative binomial model to our own Tn-seq data on *Caulobacter crescentus* experiments described previously. Of the 4,084 genes with at least one transposon insertion, we identified 337/395 (total/unique) as frankly conditionally essential in heat shock, 285/323 (total/unique) in canavanine, 304/348 (total/unique) in Δlon, 296/334 (total/unique) in Δlon+heat shock, and 311/346 (total/unique) in Δlon+canavanine by the criteria described previously that the average transposon counts in in the condition or control is less than one. Figure 5 shows the conditionally essential and dispensable genes in each of the conditions considered for this data set (excluding the frankly conditionally essential genes). Each data point is a gene and genes labeled as triangles are called conditionally essential/dispensable by the local FDR criterion.

Conditionally Essential Genes In wildtype strains under heat stress conditions, we found four genes that are conditionally essential. One of these (*metK*) is a known substrate of the chaperone GroEL [35], suggesting that during heat stress prolific misfolding of MetK could result in a higher need for the metK gene in *Caulobacter*. Under canavanine conditions, there is only one gene found essential in this condition, the *katG* gene, a peroxidase-catalase gene that is critical for oxidative tolerance in stationary phase [36]. Neither of these genes seem to be conditionally essential during stress conditions for cells lacking the Lon protease, suggesting that these mutant strains respond differently to protein homeostasis stresses.

Conditionally Dispensable Genes We found twelve genes that were conditionally dispensable during heat stress and eight during canavanine stress in wildtype strains. For those dispensable during heat stress, we were intrigued to find *katG* as well, suggesting that while *katG* is important for tolerating canavanine induced protein misfolding, its presence confers less fitness when cells are subject to heat stress. We note that during canavanine stress, the *dksA* gene becomes dispensable. *dksA* was identified as a multicopy suppressor of growth defects stemming from loss of the DnaK

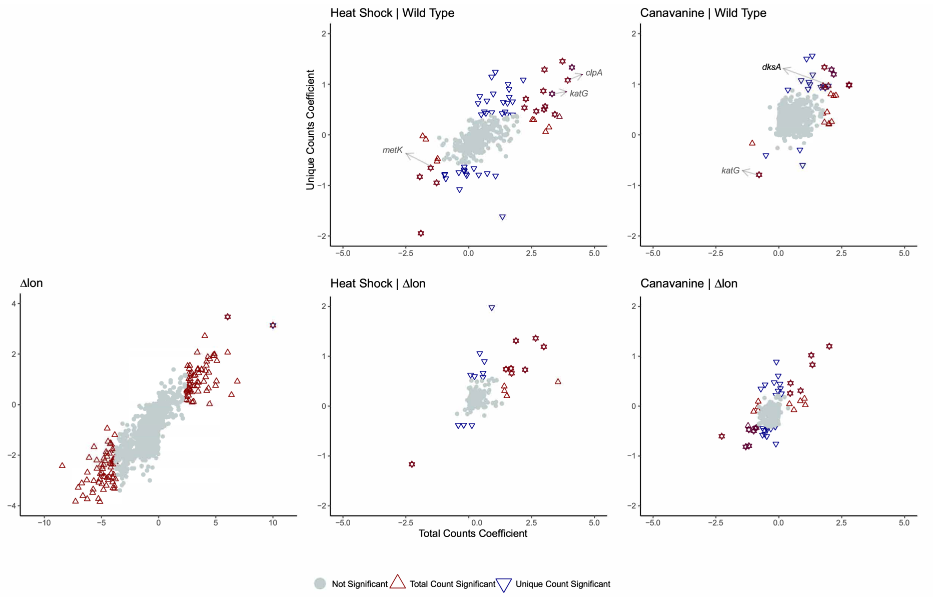


Fig 5. Analysis of transposon sequencing data of *Caulobacter crescentus*. Shown are the regularized model coefficients for the genetic background effect for Δlon and the nested environmental conditions: heat shock and canavanine for total count and unique count data.

chaperone and it is known to inhibit ribosome synthesis [37], suggesting a strong role in protoestasis. We speculate that loss of *dksA* may guard against protein misfolding stress resulting from canavanine misincorporation, or improve ribosome capacity which is taxed due to misincorporation of canavanine. Again, while we see similar numbers of genes being conditionally dispensable in cells lacking Lon under these stress conditions, there is no overlap in the sets, suggesting a different program in place for stress response.

Validation Experiments Our model has identified *clpA* to be conditionally dispensable by both measures of total and unique counts in the *wt* background under heat stress.To validate this we performed competitive mutant fitness assays comparing the growth rate of *wt* and $\Delta clpA$ in competition with a *wt* strain constitutively expressing the fluorescent reporter, Venus. The competition assay results in Figure 6 shows that heat-stress (42°C) compensates for the fitness defect caused by the loss of *clpA* under normal conditions (30°C) across three biological replicates.

Conclusion

We have presented a model-based method that uses regularized negative binomial regression to estimate the change in transposon insertions attributable to gene-environment changes without transformations or uniform normalization. Simulation experiments indicate that the regularized negative binomial model performs well without over-fitting. When applied to RB-TnSeq and Tn-Seq using both total and unique data. The model able to identify sets of conditionally essential/dispensible genes for each perturbation that shed light on their functions and roles during various stress conditions.

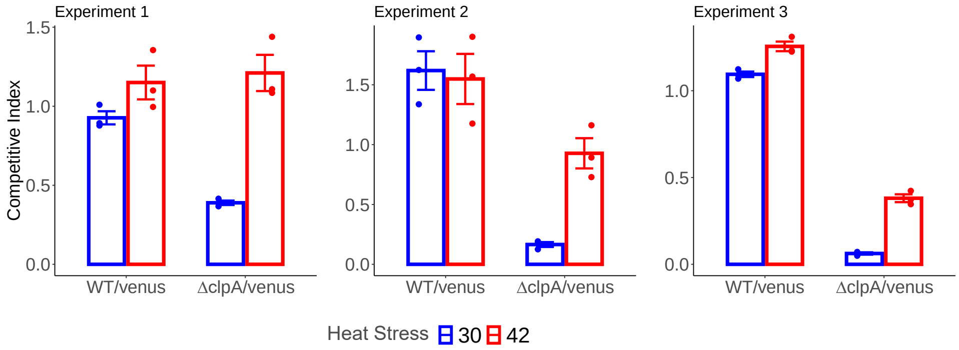


Fig 6. Competitive mutant fitness experiment comparing fitness of wild-type and Δ*clpA* under heat stress. Y-axis corresponds to the ratio of cells after 24 hours growth either with no heat stress (30) or after a transient 42 degrees C heat stress (42) compared to the Venus reporter strain. Ratios of initial mixtures normalized to 1. Error bars indicate the standard error of the mean of each group.

Acknowledgements

This work is supported by NIH R01 GM135931 and NSF HDR TRIPODS 1934846.

Supporting information

S1 Table. Tn-seq data analysis results. List of all significant conditionally essential/dispensable genes found in the Tn-seq data by both measures of the total and unique count, total count, and unique count.

S2 Table. RB-TnSeq data analysis of *Caulobacter crescentus* List of all significant conditionally essential/dispensable genes found in the RB-TnSeq data of of *Caulobacter crescentus* by total count, and unique count.

S3 Table. RB-TnSeq data analysis of *Pseudomonas fluorescens* List of all significant conditionally essential/dispensable genes found in the RB-TnSeq data of of *Pseudomonas fluorescens* by total count, and unique count.

Availability

The data used in this project is available at https://doi.org/10.6084/m9.figshare.14879514. The code used in this paper is part of `rnbtn` R package and will be freely available online as part of the Bioconductor project (http://www.bioconductor.org).

References

1. Kuspa A, Loomis WF. Tagging Developmental Genes in Dictyostelium by Restriction Enzyme-Mediated Integration of Plasmid DNA. Proceedings of the National Academy of Sciences of the United States of America. 1992;89(18):8803–8807.

2. Lander ES, Botstein D. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. Genetics. 1989;121(1):185–199. doi:10.1093/genetics/121.1.185.

3. Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, et al. Chemogenomic Profiling: Identifying the Functional Interactions of Small Molecules in Yeast. Proceedings of the National Academy of Sciences. 2004;101(3):793–798. doi:10.1073/pnas.0307490100.

4. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, et al. Functional Profiling of the Saccharomyces Cerevisiae Genome. Nature. 2002;418(6896):387–391. doi:10.1038/nature00935.

5. Flaherty P, Giaever G, Kumm J, Jordan MI, Arkin AP. A Latent Variable Model for Chemogenomic Profiling. Bioinformatics. 2005;21(15):3286–3293. doi:10.1093/bioinformatics/bti515.

6. Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, van Opijnen T. A Decade of Advances in Transposon-Insertion Sequencing. Nat Rev Genet. 2020;21(9):526–540.

7. van Opijnen T, Camilli A. Transposon Insertion Sequencing: A New Tool for Systems-Level Analysis of Microorganisms. Nature Reviews Microbiology. 2013;11(7):435–442. doi:10.1038/nrmicro3033.

8. 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). doi:10.1128/mBio.00306-15.

9. Fan HC, Quake SR. Sensitivity of Noninvasive Prenatal Detection of Fetal Aneuploidy from Maternal Plasma Using Shotgun Sequencing Is Limited Only by Counting Statistics. PLOS ONE. 2010;5(5):e10439. doi:10.1371/journal.pone.0010439.

10. van Opijnen T, Bodi KL, Camilli A. Tn-Seq: High-Throughput Parallel Sequencing for Fitness and Genetic Interaction Studies in Microorganisms. Nat Methods. 2009;6(10):767–772.

11. Zomer A, Burghout P, Bootsma HJ, Hermans PWM, van Hijum SAFT. ESSENTIALS: Software for Rapid Analysis of High Throughput Transposon Insertion Sequencing Data. PLOS ONE. 2012;7(8):e43012. doi:10.1371/journal.pone.0043012.

12. Cleveland WS, Devlin SJ. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association. 1988;83(403):596–610. doi:10.1080/01621459.1988.10478639.

13. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. Bioinformatics. 2010;26(1):139–140. doi:10.1093/bioinformatics/btp616.

14. 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 (Oxford, England). 2013;29(6):695–703. doi:10.1093/bioinformatics/btt043.

15. Subramaniyam S, DeJesus MA, Zaveri A, Smith CM, Baker RE, Ehrt S, et al. Statistical Analysis of Variability in TnSeq Data across Conditions Using Zero-Inflated Negative Binomial Regression. BMC Bioinformatics. 2019;20(1):603. doi:10.1186/s12859-019-3156-z.

16. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, et al. Mutant Phenotypes for Thousands of Bacterial Genes of Unknown Function. Nature. 2018;557(7706):503–509. doi:10.1038/s41586-018-0124-0.

17. Zeinert RD, Baniasadi H, Tu BP, Chien P. The Lon Protease Links Nucleotide Metabolism with Proteotoxic Stress. Molecular Cell. 2020;79(5):758–767.e6. doi:10.1016/j.molcel.2020.07.011.

18. Girardot C, Scholtalbers J, Sauer S, Su SY, Furlong EEM. Je, a Versatile Suite to Handle Multiplexed NGS Libraries with Unique Molecular Identifiers. BMC Bioinformatics. 2016;17(1):419. doi:10.1186/s12859-016-1284-2.

19. Li H, Durbin R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. Bioinformatics (Oxford, England). 2009;25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map Format and SAMtools. Bioinformatics. 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

21. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012;337(6099):1190–1195. doi:10.1126/science.1222794.

22. Tusher VG, Tibshirani R, Chu G. Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. Proceedings of the National Academy of Sciences. 2001;98(9):5116–5121. doi:10.1073/pnas.091062498.

23. Bhattacharya A, Pati D, Pillai NS, Dunson DB. Dirichlet–Laplace Priors for Optimal Shrinkage. Journal of the American Statistical Association. 2015;110(512):1479–1490. doi:10.1080/01621459.2014.960967.

24. Golub GH, Hansen PC, O'Leary DP. Tikhonov Regularization and Total Least Squares. SIAM Journal on Matrix Analysis and Applications. 1999;21(1):185–194. doi:10.1137/S0895479897326432.

25. Wang Z, Ma S, Wang CY, Zappitelli M, Devarajan P, Parikh C. EM for Regularized Zero-Inflated Regression Models with Applications to Postoperative Morbidity after Cardiac Surgery in Children. Statistics in Medicine. 2014;33(29):5192–5208. doi:10.1002/sim.6314.

26. Candès E, Fan Y, Janson L, Lv J. Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2018;80(3):551–577. doi:10.1111/rssb.12265.

27. Efron B. Size, Power and False Discovery Rates. Annals of Statistics. 2007;35(4):1351–1377. doi:10.1214/009053606000001460.

28. Donoho D, Jin J. Asymptotic Minimaxity of False Discovery Rate Thresholding for Sparse Exponential Data. Annals of Statistics. 2006;34(6):2980–3018. doi:10.1214/009053606000000920.

29. Efron B. Microarrays, Empirical Bayes and the Two-Groups Model. Statistical Science. 2008;23(1):1–22. doi:10.1214/07-STS236.

30. Storey JD. A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64(3):479–498. doi:10.1111/1467-9868.00346.

31. Efron B. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. 1st ed. No. 1 in Institute of Mathematical Statistics Monographs. Cambridge: Cambridge Univ. Press; 2013.

32. Skerker JM, Prasol MS, Perchuk BS, Biondi EG, Laub MT. Two-Component Signal Transduction Pathways Regulating Growth and Cell Cycle Progression in a Bacterium: A System-Level Analysis. PLoS Biology. 2005;3(10):e334. doi:10.1371/journal.pbio.0030334.

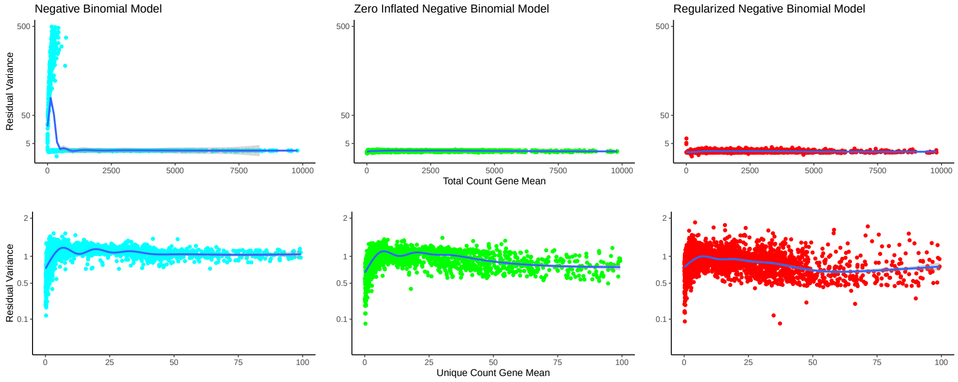
33. West SC, Connolly B. Biological Roles of the Escherichia Coli RuvA, RuvB and RuvC Proteins Revealed. Molecular Microbiology. 1992;6(19):2755–2759. doi:10.1111/j.1365-2958.1992.tb01454.x.

34. Nishijyo T, Haas D, Itoh Y. Mol MicrobiolThe CbrA-CbrB two-component regulatory system controls the utilization of multiple carbon and nitrogen sources in Pseudomonas aeruginosa. Mol Microbiol. 2001;40(4):917–931.

35. Kerner MJ, Naylor DJ, Ishihama Y, Maier T, Chang HC, Stines AP, et al. Proteome-Wide Analysis of Chaperonin-Dependent Protein Folding in Escherichia Coli. Cell. 2005;122(2):209–220. doi:10.1016/j.cell.2005.05.028.

36. Steinman HM, Fareed F, Weinstein L. Catalase-Peroxidase of Caulobacter Crescentus: Function and Role in Stationary-Phase Survival. Journal of bacteriology. 1997;179(21):6831–6836. doi:10.1128/JB.179.21.6831-6836.1997.

37. Lemke JJ, Sanchez-Vazquez P, Burgos HL, Hedberg G, Ross W, Gourse RL. Direct Regulation of Escherichia Coli Ribosomal Protein Promoters by the Transcription Factors ppGpp and DksA. Proceedings of the National Academy of Sciences. 2011;108(14):5712–5717. doi:10.1073/pnas.1019383108.



Supplementary Figure 1. Comparison of regularized negative binomial model(right) for the two-condition setup with a simple negative binomial model (left) and Zero-inflated negative binomial (center).