Bioinformatics, 32(11), 2016, 1701–1708
doi: 10.1093/bioinformatics/btw061
Advance Access Publication Date: 1 February 2016
Original Paper

OXFORD

Genome analysis

A zero-inflated Poisson model for insertion tolerance analysis of genes based on Tn-seq data

Fangfang Liu[1], Chong Wang[1,2], Zuowei Wu[3], Qijing Zhang[3] and Peng Liu[1,*]

[1]Department of Statistics, Iowa State University, [2]Department of Veterinary Diagnostic and Production Animal Medicine and [3]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA 50010, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on 21 May 2015; revised on 17 January 2016; accepted on 25 January 2016

Abstract

Motivation: Transposon insertion sequencing (Tn-seq) is an emerging technology that combines transposon mutagenesis with next-generation sequencing technologies for the identification of genes related to bacterial survival. The resulting data from Tn-seq experiments consist of sequence reads mapped to millions of potential transposon insertion sites and a large portion of insertion sites have zero mapped reads. Novel statistical method for Tn-seq data analysis is needed to infer functions of genes on bacterial growth.

Results: In this article, we propose a zero-inflated Poisson model for analyzing the Tn-seq data that are high-dimensional and with an excess of zeros. Maximum likelihood estimates of model parameters are obtained using an expectation–maximization (EM) algorithm, and pseudogenes are utilized to construct appropriate statistical tests for the transposon insertion tolerance of normal genes of interest. We propose a multiple testing procedure that categorizes genes into each of the three states, hypo-tolerant, tolerant and hyper-tolerant, while controlling the false discovery rate. We evaluate the proposed method with simulation studies and apply the proposed method to a real Tn-seq data from an experiment that studied the bacterial pathogen, *Campylobacter jejuni*.

Availability and implementation: We provide R code for implementing our proposed method at http://github.com/ffliu/TnSeq. A user's guide with example data analysis is also available there.

Contact: pliu@iastate.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transposons are genetic elements that can be inserted into genomes mediated by enzymes called transponsases. An application of transposons in genomic studies is insertional mutagenesis which refers to the construction of a library of bacterial strains, each being a mutant generated by random insertion of a transposon. Transposon mutagenesis has facilitated our understanding of gene functions through identifying the phenotypes of the corresponding mutants. With the recent advent of next-generation sequencing (NGS) technologies, it is possible to sequence many transposon insertional mutants simultaneously, and the resulting technique that combines transposon mutagenesis with NGS technologies has been referred to as transposon sequencing (Tn-seq). The Tn-seq technique generates huge amount of data and allows genome-wide analysis of gene functions through mutagenesis.

Four different transposon sequencing technologies were proposed independently in 2009 and have been reviewed in van Opijnen and Camilli (2014). They all share similar ideas of sequencing the DNA

sequences flanking the transposon insertion site. We conducted a Tn-seq experiment to identify genes involved in the growth of a highly virulent strain (IA3902) of the bacteria *Campylobacter jejuni*, which has been reported to be the major cause of sheep abortion and implicated in human gastroenteritis in the United States (Sahin *et al.*, 2008, 2012). By randomly inserting Tn5 transposons into locations within the genome, we generated a library of mutant bacterial strains. Details of our Tn-seq experimental procedure are described in Section S-1 of the online supplementary material. In summary, the genomic DNA was extracted from the library of mutants after growth in a liquid medium. After shearing the extracted DNA into fragments and amplifying the fragments with transposon-specific primers, we applied Illumina sequencing to obtain sequence reads flanking the transposon insertion sites. By mapping the sequence reads to the genome of *C.jejuni* IA3902, we identified locations of transposon insertions and measured the relative abundance of mutants containing a transposon at each possible insertion site by enumerating the sequence reads mapped to the corresponding site. The transposon used in our experiment, Tn5, can randomly insert into any locations of the target DNA, but was reported to insert at the $3'$ end of the sequence A-GNTYWRANC-T with a slight bias (Gerdes *et al.*, 2003; Goryshin *et al.*, 1998; Langridge, 2009). Other transposons may have different sequence preferences of the insertion sites. For example, the Himar1 transposon (Lampe *et al.*, 1996) inserts randomly between T and A in TA dinucleotides.

An important research objective in a Tn-seq experiment such as ours is to classify genes according to their tolerance to transposon insertion mutagenesis, or equivalently, to classify genes with respect to their effects on bacteria fitness/growth. A gene can be tolerant of disruption if the transposon insertion into the gene does not affect the growth of bacteria. On the other hand, a gene is hypo-tolerant of transposon insertion if the inactivation of the gene by transposon insertion suppresses the growth of bacteria. Such a gene may provide an 'essential' or 'core' function in bacteria survival. A gene can also be hyper-tolerant of transposon insertion, suggesting that the inactivation of the gene may provide a growth advantage. A successful antimicrobial treatment needs to avoid inactivation of such hyper-tolerant genes. The above classification has been discussed in Wiles *et al.* (2013) and is adopted throughout our article.

In Tn-seq data, the count of reads corresponding to each possible insertion site provides a measurement of the abundance of mutants containing a transposon at that location. A gene may have zero counts of sequence reads either because the gene is hypo-tolerant to transposon insertion and essential in bacteria survival, or because no transposons were inserted into any location in this gene just by chance. An appropriate statistical model should take into consideration both the probability of transposon insertion and the abundance of bacterial mutants in case that a transposon insertion happens. Another challenge in Tn-seq data analysis is the high dimensionality. Usually, thousands of genes are simultaneously studied in one Tn-seq experiment, whereas each gene contains up to thousands of locations for potential transposon insertions. The outcomes of Tn-seq experiments, the read counts, are measured at the level of insertion locations, which suggests the development of statistical models at the same level to avoid loss of information. Besides, statistical models need to take into account of the variation among insertion locations within a gene, as well as variation among genes.

There have been a few statistical methods proposed for the analysis of Tn-seq data and they deal with different levels of data summary. Some researchers proposed to first reduce the data from the level of locations to the level of genes using the count of locations with zero mapped reads for each gene, and then model such counts using discrete distributions such as Poisson distributions (Deng *et al.*, 2013) or negative-binomial distributions (Zomer *et al.*, 2012). Such models ignore the magnitudes of the read counts for locations with non-zero counts, which measures the abundance of the corresponding mutants and hence contains important information about gene tolerance status. Zhang *et al.* (2012) divided the genome into contiguous overlapping windows (with width 400–600 base pair) and then employed a non-parametric test to assess tolerance status to bacterial growth for each of these windows. Dejesus and Loerger (2013) described a four-state hidden Markov model and used geometric distributions for the count data at each location to model the conditional distribution of read counts given different states, and then each gene was assigned to a tolerance state according to the most frequent state found within the boundaries of the gene sequence. Pritchard *et al.* (2014) presented a Tn-seq data analysis pipeline named ARTIST. One arm of ARTIST, EL-ARTIST, implements the method proposed by Chao *et al.* (2013) that uses the results from the method proposed by Zhang *et al.* (2012) as training set for the HMM algorithm and reports the estimate of state of genes as essential, domain essential, sick or neutral. All of these methods (Chao *et al.*, 2013; Dejesus and Loerger, 2013; Zhang *et al.*, 2012) are applicable to Tn-seq data with transposons inserted only into specific target locations (such as transposon Himar1 that only inserts into TA dinucleotides). When using transposons (say, EZ::TN) that could insert into any genome locations, the resulting Tn-seq datasets typically are unsaturated and contain excess zeros. These methods have difficulty in dealing with such Tn-seq datasets because they do not model the possibilities that no transposons are inserted into certain locations of genes during the construction of transposon mutant libraries.

In this article, we propose a zero-inflated Poisson (ZIP) model to deal with the excess of zeros for the analysis of Tn-seq data at the level of locations. To account for the two possible reasons that give rise to zero read counts at a given location, we apply a mixture model with a point mass at zero and a Poisson distribution. The component of point mass at zero corresponds to the case when there is no transposon inserted at the location, and hence there should be zero reads mapped to the corresponding location. This component is associated with a mixing probability that models the chance of not having a transposon insertion at a location, and this probability may vary across different locations depending on the genomic sequence or potentially other covariates that may affect the chance of insertion. The Poisson distribution component models the count of reads mapped to the given location when a transposon is inserted there, and the mean of the Poisson distribution provides a measurement of the abundance of mutants for the corresponding gene. This model reflects how the count data of Tn-seq are generated at the location level, and allows us to estimate the chance of having transposon insertion and the abundance of mutants corresponding to different genes, and the latter allows us to identify genes involved in bacterial growth.

Our Tn-seq experiment studied one specific growth condition of the mutant library. To classify genes with respect to their effects on bacterial growth, we utilize data of pseudogenes observed in the same experiment. Pseudogenes are dysfunctional relatives of genes that have lost their protein-coding abilities. Pseudogenes are still important in genomic studies because they provide a record of how the genomic DNA has been changed without evolutionary pressure and can be used as a model for determining the underlying rates of nucleotide substitution, insertion and deletion in the genome. In our

model, we consider pseudogenes to be tolerant of disruption as they have already lost their protein coding abilities before mutagenesis. Thus comparing the normal (non-pseudo) genes with pseudogenes provides information about the tolerance status of the normal genes. More specifically, the genes with mean abundance of mutants not significantly different from the mean abundance of pseudogenes' mutants will be classified as tolerant, and those genes with mean abundance of mutants significantly below (above) the mean abundance of mutants for pseudogenes are hypo-tolerant (hyper-tolerant).

We apply an expectation–maximization (EM) algorithm to estimate parameters of the ZIP model, including the mixing probability of insertion and mean abundance of mutants for thousands of genes. Then we estimate the asymptotic variance–covariance matrix of all estimated parameters and perform a Wald test for each gene to assess whether each gene is tolerant of transposon insertion or not by comparing the mean abundance of mutants of this gene with that of the pseudogenes. The method by Benjamini and Hochberg (1995) is applied to control the false discovery rate (FDR) as an error criterion in the multiple testing problem. At the end, we classify all normal genes into three different classes: 'hypo-tolerant', 'tolerant' and 'hyper-tolerant', which reflects their tolerance to transposon insertion.

The remainder of the article is organized as follows. Section 2 presents our ZIP model for Tn-seq data, our method to estimate model parameters and our procedure for the multiple hypothesis testing to identify the tolerance status for genes. In Section 3 and Section 4, we present a simulation study and a real data analysis, respectively. Section 5 concludes the article with discussions.

2 Method

2.1 Tn-seq data at the level of locations

To study the gene functions on bacterial growth using Tn-seq experiments, we analyze the count of reads at each possible insertion site within all gene-coding regions. A read is mapped to a location if the corresponding transposon is inserted right before the 5′-end of the nucleotide at that location along the genome. Suppose there are G genes in total, without loss of generality, the genes are ordered so that the first $N(N < G)$ genes are normal genes, the last $(G - N)$ genes are pseudogenes. For the gth gene, $g = 1, ..., G$, the number of possible transposon insertion sites is n_g, and the gene length is l_g. For transposons without specific recognition of target sequences, such as the EZ::TN transposon, any location within the genome can be a potential transposon insertion site, and thus $n_g = l_g$. For transposons that recognize specific sequences, only locations satisfying the sequence requirement would be potential insertion sites and hence $n_g < l_g$. Let Y_{gi} denote the count of sequence reads mapped to the ith location in gene g where $i = 1, ..., n_g$.

2.2 The ZIP model

The transposon mutagenesis is viewed as a process that transposons are randomly inserted into the genome. Let Z_{gi} denote whether there is a transposon inserted for the ith location within gene g for $g = 1, ..., G$ and $i = 1, ..., n_g$. That is, $Z_{gi} = 1$ if there are transposons inserted at the i^{th} location of the g^{th} gene, and $Z_{gi} = 0$ if no transposons are inserted at this location. Obviously, if $Z_{gi} = 0$, then $Y_{gi} = 0$, and we refer to this state as the perfect zero state as in Lambert (1992). When $Z_{gi} = 1$, i.e. transposon insertion occurs at the corresponding location, we model the count of reads mapped to the location with a Poisson distribution with mean μ_g, where μ_g can be interpreted as the mean abundance of the mutants for gene g. The state of $Z_{gi} = 1$ will be referred to as the Poisson state.

We model the insertion indicators Z_{gi}'s with independent Bernoulli distribution with probability of insertion p_{gi}. The parameter p_{gi} models the chance of having a transposon insertion at the ith location of the gth gene, and this probability may vary across different locations depending on the genomic sequence or potentially other covariates that may affect the chance of insertion. Let $\mathbf{x}'_{gi} = (x_{gi0}, ..., x_{gi(m-1)})'$ denote the covariates for the ith location of gth gene that may affect p_{gi} and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{m-1})'$ denote the corresponding coefficients, we apply a logistic regression to model p_{gi} through $\mathbf{x}'_{gi}$. In summary, our hierarchical model is:

$$\text{logit}(p_{gi}) = \mathbf{x}'_{gi}\boldsymbol{\beta}, \tag{1}$$

$$Z_{gi} \sim \text{Bernoulli}(p_{gi}), \tag{2}$$

and

$$Y_{gi}|Z_{gi} \sim (1 - Z_{gi}) \cdot \delta_{\{0\}} + Z_{gi} \cdot \text{Poisson}(\mu_g), \tag{3}$$

where $\delta_{\{0\}}$ denotes the point mass at 0. Although $Z_{gi} = 0$ implies $Y_{gi} = 0$, the reverse statement is not true. When we observe $Y_{gi} = 0$, it could be the result of no transposon insertion ($Z_{gi} = 0$), or, it is also possible that the gene is essential to bacterial growth and disruption of this gene by having a transposon insertion is not tolerated and such bacteria cannot grow. Hence, the probability mass function of Y_{gi} can be written as:

$$f(Y_{gi}; p_{gi}, \mu_g) = \begin{cases} 1 - p_{gi} + p_{gi}\exp(-\mu_g) & Y_{gi} = 0, \\ p_{gi} \cdot f_{pois}(Y_{gi}; \mu_g) & Y_{gi} > 0, \end{cases} \tag{4}$$

where $f_{pois}(\cdot; \mu_g)$ denotes the probability mass function for Poisson distribution with the mean parameter μ_g.

Because pseudogenes are not functional, we expect their disruption have no effect on bacterial growth, i.e. they are tolerant of transposon insertion. Hence we expect the same level of growth for mutants of pseudogenes, and the same level of abundance for all pseudogenes. We denote the common mean abundance of pseudogene mutants with μ_0, that is,

$$\mu_{N+1} = \cdots = \mu_G \triangleq \mu_0. \tag{5}$$

Therefore we have $(N + 1)$ mean parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N, \mu_0)'$ in total.

The goal of our experiment is to determine the status of tolerance of each normal gene, and we compare the mean abundance of mutants for each normal gene with that of the pseudogenes. So after our estimation of model parameters, we will test the following hypotheses for each gene g:

$$H_{0g} : \mu_g = \mu_0 \quad \text{versus} \quad H_{1g} : \mu_g \neq \mu_0. \tag{6}$$

2.3 Parameter estimation

In this subsection, we describe how we obtain the maximum likelihood estimates for the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{m-1})'$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N, \mu_0)'$. Provided that the resource for bacterial growth is enough, the measurements of mutant abundance, Y_{gi}'s, are independent of each other. For notational convenience, we have $\mathbf{Y} = (Y_{11}, ..., Y_{1n_1}, ..., Y_{G1}, ..., Y_{Gn_G})'$, and $\mathbf{Z} = (Z_{11}, ..., Z_{1n_1}, ..., Z_{G1}, ..., Z_{Gn_G})'$. Based on (4) and assuming independence between locations

and genes, the log-likelihood function with observable data $\boldsymbol{Y}$ is given by

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\mu}|\mathbf{Y}) &= \sum_{g=1}^{G}\sum_{i=1}^{n_g}\log f(Y_{gi};p_{gi},\mu_g) \\ &= \sum_{g=1}^{G}\left\{\sum_{i\in S_g}\log[1-p_{gi}+p_{gi}\exp(-\mu_g)]\right\} \\ &+\sum_{g=1}^{G}\left\{\sum_{i\notin S_g}\log[p_{gi}\cdot f_{Pois}(Y_{gi};\mu_g)]\right\}, \end{aligned} \tag{7}$$

where $p_{gi}=\exp(\mathbf{x}'_{gi}\boldsymbol{\beta})/(1+\exp(\mathbf{x}'_{gi}\boldsymbol{\beta}))$, and the set $S_g=\{i: Y_{gi}=0\}$ is defined as the set of indices corresponding to those locations with zero read count for gene g.

For our ZIP model, we have $N+1$ mean parameters and m coefficients in the logistic regression for p_{gi}'s. In total, there are $(N+m+1)$ parameters to be estimated. For Tn-seq data, N is in thousands and hence, it is a high-dimensional estimation problem. Commonly used numerical optimization procedures such as Newton–Raphson and gradient conjugate method can be computationally complicated and burdensome to maximize the log-likelihood $L(\boldsymbol{\beta},\boldsymbol{\mu}|\mathbf{Y})$ in terms of the high-dimensional parameter space. Instead, we develop an EM algorithm (Dempster *et al.*, 1977) using the complete data denoted as $(\mathbf{Y},\mathbf{Z})$, where $\mathbf{Z}$ is an unobserved vector of latent variables that store the transposon insertion states. As explained below, given $\mathbf{Z}$, the likelihood function can be written into two parts and then allows separate maximization with respect to β and μ which results in an efficient way to estimate the MLEs.

The distribution of Y_{gi} conditional on Z_{gi} is

$$f(Y_{gi}|Z_{gi})=[\delta_{\{0\}}(Y_{gi})]^{1-Z_{gi}}[f_{pois}(Y_{gi};\mu_g)]^{Z_{gi}}.$$

And the joint distribution of (Y_{gi},Z_{gi}) is

$$f(Y_{gi},Z_{gi})=[(1-p_{gi})\delta_{\{0\}}(Y_{gi})]^{1-Z_{gi}}[p_{gi}f_{pois}(Y_{gi};\mu_g)]^{Z_{gi}}, \tag{8}$$

where $Z_{gi}|p_{gi}\sim$ Bernoulli(p_{gi}), and logit$(p_{gi})=\mathbf{x}'_{gi}\boldsymbol{\beta}$.

Suppose we know which zeros come from the perfect zero state and which come from the Poisson state. In other words, we could observe $Z_{gi}=0$ when Y_{gi} is from the perfect zero state and $Z_{gi}=1$ when Y_{gi} is from the Poisson state. Then the log-likelihood with the complete data $(\mathbf{Y},\mathbf{Z})$ would be written as

$$L_c(\boldsymbol{\beta},\boldsymbol{\mu}|\mathbf{Y},\mathbf{Z})=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\log f(Z_{gi}|p_{gi})+\log f(Y_{gi}|Z_{gi},\mu_g)), \tag{9}$$

where

$$\log f(Z_{gi}|p_{gi})=Z_{gi}\mathbf{x}'_{gi}\boldsymbol{\beta}-\log(1+\exp(\mathbf{x}'_{gi}\boldsymbol{\beta})),$$

$$\log f(Y_{gi}|Z_{gi},\mu_g))=Z_{gi}(Y_{gi}\log\mu_g-\mu_g)-Z_{gi}\log(Y_{gi}!).$$

This implies that

$$L_c(\boldsymbol{\beta},\boldsymbol{\mu}|\mathbf{Y},\mathbf{Z})=L_c(\boldsymbol{\beta}|\mathbf{Y},\mathbf{Z})+L_c(\boldsymbol{\mu}|\mathbf{Y},\mathbf{Z}), \tag{10}$$

where

$$L_c(\boldsymbol{\beta}|\mathbf{Y},\mathbf{Z})=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(Z_{gi}\mathbf{x}'_{gi}\boldsymbol{\beta}-\log(1+\exp(\mathbf{x}'_{gi}\boldsymbol{\beta}))),$$

$$L_c(\boldsymbol{\mu}|\mathbf{Y},\mathbf{Z})=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(Z_{gi}(Y_{gi}\log\mu_g-\mu_g)-Z_{gi}\log(Y_{gi}!)).$$

Hence, the complete log-likelihood function in (10) can be written into two parts, $L_c(\boldsymbol{\beta}|\mathbf{Y},\mathbf{Z})$ and $L_c(\boldsymbol{\mu}|\mathbf{Y},\mathbf{Z})$ that could be maximized separately. Therefore the complete log-likelihood is easier to maximize than simultaneously maximizing all model parameters with marginal likelihood based on $\boldsymbol{Y}$.

The EM algorithm used to maximize $L(\boldsymbol{\beta},\boldsymbol{\mu}|\mathbf{Y})$ alternates between an expectation (E) step in which the latent variables $\mathbf{Z}$ are calculated as their expectations under the current estimates of parameters $(\boldsymbol{\beta},\boldsymbol{\mu})$, and a maximization (M) step in which the complete log-likelihood $L_c(\boldsymbol{\beta},\boldsymbol{\mu}|\mathbf{Y},\mathbf{Z})$ at values of $\mathbf{Z}$ from the E step is maximized with respect to both $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. The EM algorithm begins with starting values $(\boldsymbol{\beta}^0,\boldsymbol{\mu}^0)$ and proceeds iteratively between the E step and the M step. At the $(k+1)$th iteration, the E step and M step are as follows.

(1) *The E step* Compute the conditional expectation of Z_{gi}

$$\begin{aligned} z_{gi}^{(k+1)} &= E(Z_{gi}|\mathbf{Y},\boldsymbol{\beta}^{\mathrm{k}},\boldsymbol{\mu}^k) \\ &= P(Z_{gi}=1|Y_{gi},\boldsymbol{\beta}^{\mathrm{k}},\boldsymbol{\mu}^k) \\ &= \begin{cases} 1 & Y_{gi}>0 \\ (1+\exp(\mu_g^k-\boldsymbol{x}'_{gi}\beta^k))^{-1} & Y_{gi}=0, \end{cases} \end{aligned}$$

This indicates that the conditional distribution of $Z_{gi}|Y_{gi},\boldsymbol{\beta}^k,\boldsymbol{\mu}^k$ is

$$Z_{gi}|Y_{gi},\ \boldsymbol{\beta}^k,\boldsymbol{\mu}^k\sim\begin{cases}\delta_{\{1\}} & Y_{gi}>0\\ \text{Bernoulli}(z_{gi}^{k+1}) & Y_{gi}=0,\end{cases}$$

where $\delta_{\{1\}}$ denotes the point mass at 1.

(2) *The M step* The $\boldsymbol{\beta}^{k+1}$ that maximizes $L_c(\boldsymbol{\beta}|\mathbf{Y},\mathbf{Z}^{k+1})$,

$$L_c(\boldsymbol{\beta}|\mathbf{Y},\mathbf{Z})=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(Z_{gi}\mathbf{x}'_{gi}\boldsymbol{\beta}-\log(1+\exp(\mathbf{x}'_{gi}\boldsymbol{\beta}))),$$

can be obtained by performing an unweighted binomial logistic regression of $\mathbf{Z}^{k+1}$ on the design matrix X using a binomial denominator of one for each observation, where the column vectors of X from left to right are $\mathbf{x}_{11},\ldots,\mathbf{x}_{1n_1},\ldots,\mathbf{x}_{G1},\ldots,\mathbf{x}_{Gn_G}$ respectively.

The $\boldsymbol{\mu}^{k+1}$ that maximizes $L_c(\boldsymbol{\mu}|\mathbf{Y},\mathbf{Z}^{k+1})$ has closed-form expression. The estimate for μ_g where $g=1,2,\cdots,N$, is

$$\mu_g^{k+1}=\frac{\sum_{g=1}^{n_g}Y_{gi}z_{gi}^{k+1}}{\sum_g^{n_g}z_{gi}^{k+1}}.$$

And we also get the estimate

$$\mu_0^{k+1}=\frac{\sum_{g=N+1}^{G}\sum_{i=1}^{n_g}Y_{gi}z_{gi}^{k+1}}{\sum_{g=N+1}^{G}\sum_{i=1}^{n_g}z_{gi}^{k+1}}.$$

As shown in Lambert (1992), the EM algorithm converges for the ZIP model. Good initial values for the EM algorithm can facilitate convergence. In this article, we use the MLE for the positive Poisson log-likelihood as the initial value for parameters μ. The positive Poisson distribution is the ordinary Poisson distribution with zero truncated from the support. The log-likelihood of positive Poisson model for the Tn-seq data is

$$L_+\propto\sum_{g=1}^{G}\sum_{i:Y_{gi}>0}[Y_{gi}\log\mu_g-\mu_g-\log(1-\exp(-\mu_g))],$$

For each $g=1,\ldots,N$, the score equation is

$$\partial L_+/\partial\mu_g=\sum_{i:Y_{gi}>0}\left(\frac{Y_{gi}}{\mu_g}-1\right)-\sum_{Y_{gi}>0}\frac{\exp(-\mu_g)}{1-\exp(-\mu_g)}\overset{\Delta}{=}0.$$

Equivalently, the equation can be written as

$$\frac{\sum_{i:Y_{gi}>0} Y_{gi}}{\mu_g} - (1 + \frac{1}{\exp(\mu_g) - 1}) \cdot \sum_{i=1}^{n_g} I(Y_{gi} > 0) = 0,$$

where $I(\cdot)$ denotes the indicator function. Solving this equation (using 'uniroot' function in R) leads to the MLE for the positive Poisson log-likelihood and our initial value for μ_g. Similarly, solving the following equation gives the initial value for μ_0.

$$\frac{\mu_0 \exp(\mu_0)}{\exp(\mu_0) - 1} \cdot \sum_{g=N+1}^{G} \sum_{i=1}^{n_g} I(Y_{gi} > 0) = \sum_{g=N+1}^{G} \sum_{i:Y_{gi}>0} Y_{gi}.$$

The initial values of elements in $\boldsymbol{\beta}$ are set to be zeros except for the intercept. The intercept is initialized as the estimated log odds of transposon insertion. The probability of having a transposon insertion is calculated by considering the two possible reasons to have locations with zero reads counts:

$$\hat{p}_0 = 1 - \frac{\sum_{g=1}^{G} \sum_{i=1}^{n_g} I(Y_{gi} = 0) - \sum_{g=1}^{G} \sum_{i=1}^{n_g} \exp(-\mu_g)}{\sum_{g=1}^{G} n_g}.$$

And then the estimated log odds is calculated as $\log(\hat{p}_0/(1 - \hat{p}_0))$.

2.4 Hypothesis testing to classify the tolerance status of bacterial genes

We apply multiple testing procedure to classify genes into different groups of transposon insertion tolerance. Let $d_g = \mu_g - \mu_0$. For each $g = 1, ..., N$, we test the null hypothesis $H_{0g} : d_g = 0$. For notational convenience, set $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\mu}')'$. In large samples, the distribution of the MLE $\hat{\boldsymbol{\theta}}$ is approximately normal with means $\boldsymbol{\theta}$ and variance–covariance matrix equal to the inverse of the observed Fisher information matrix $\hat{I}(\mathbf{Y}, \boldsymbol{\theta})^{-1}$. The observed Fisher information matrix can be directly computed by

$$\hat{I}(\mathbf{Y}, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\frac{\partial^2}{\partial \boldsymbol{\theta} \boldsymbol{\theta}^T} \boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\mu}|\boldsymbol{Y})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \quad (11)$$

Although the Fisher matrix is of extremely high dimension, it contains a big block of diagonal matrix and we are able to obtain its inverse using the block matrix inversion formula. The derivation of the Fisher information matrix as a block matrix is given in Section S-3 of the online supplementary material.

For large samples as in Tn-seq data at the level of locations, the MLE's and regular functions of the MLE's are consistent. Let the estimate for d_g be $\hat{d}_g = \hat{\mu}_g - \hat{\mu}_0$ and the estimated variance be

$$\widehat{var}(\hat{d}_g) = \widehat{var}(\hat{\mu}_g) + \widehat{var}(\hat{\mu}_0) - 2\widehat{cov}(\hat{\mu}_g, \hat{\mu}_0).$$

Then, for each $g = 1, .., N$, we apply the Wald test to test the null hypothesis, and calculate the p-value by

$$p_g = 2\Phi\left(\left|\frac{\hat{d}_g}{\sqrt{\hat{var}(\hat{d}_g)}}\right|\right), \quad (12)$$

where $\Phi(.)$ denotes the cumulative distribution function for the standard normal distribution.

A Tn-seq experiment simultaneously examines thousands of genes. Hence, we need to control the multiple testing error. The false discovery rate (FDR) proposed by Benjamini and Hochberg (1995) has been argued to be a reasonable error rate to control in different genomic studies (Storey and Tibshirani, 2003). In this article, we apply the procedure of Benjamini and Hochberg (1995) to the set of p-values to control FDR. Other FDR controlling procedures such as the q-value method (Storey and Tibshirani, 2003) can also be applied.

We use a two-step procedure to identify the transposon tolerance status of genes. The first step is to perform multiple testing procedure while controlling FDR as described above. Genes whose corresponding null hypotheses are not rejected are classified to the group of genes tolerant to transposon insertion. In the second step, we examine the estimates of $\hat{d}_g$ for genes whose corresponding null hypotheses are rejected. If the estimate of $\hat{d}_g$ is smaller than zero for gene g, it is classified as hypo-tolerant to transposon insertion. That is, the impairment of such genes negatively impact the bacterial growth and may serve as anti-bacterial target. If the estimate of $\hat{d}_g$ is larger than zero for gene g, it is classified as hyper-tolerant.

3 A simulation study

A simulation study is conducted to evaluate the performance of our proposed method as described in Section 2. In this simulation study, we set the transposon insertion rate P to be 0.01 and not affected by covariates. There were a total of 1820 genes, among which 20 were pseudogenes. The number of possible insertion sites n_g was simulated from discrete Uniform distribution $U(500, 2000)$ for each gene g. The mean abundance of mutants for pseudogenes was set be to $\mu_0 = 8$, and the means (μ_g's) for the 1800 normal genes were set as follows: $\mu_{1,...,200} = 0.5, \mu_{201,...,400} = 1, \mu_{401,...,600} = 2, \mu_{601,...,800} = 4, \mu_{801,...,1000} = 8, \mu_{1001,...,1200} = 16, \mu_{1201,...,1400} = 32, \mu_{1401,...,1600} = 64$ and $\mu_{1601,...,1800} = 128$. The response $\boldsymbol{Y}$ was obtained by first generating a Uniform(0, 1) random vector $\boldsymbol{U}$ and then simulating Y_{gi} from Poisson (μ_g) if $U_{gi} \leq p$ or assigning $Y_{gi} = 0$ otherwise. A total of 100 datasets were simulated.

After applying our method with $\beta = \beta_0$ while controlling FDR at 5%, we categorized each simulated normal gene into the three groups: tolerant, hypo-tolerant and hyper-tolerant. For each μ_g value, the proportion of genes classified into each of the three groups was calculated for each simulated dataset and then averaged over 100 simulated datasets. Such proportions are plotted against the true values of μ_g for normal genes in Figure 1. This figure demonstrates that our proposed method in Section 2 correctly identified the tolerance status for most genes. The estimation of $d_g = \mu_g - \mu_0$, the mean difference between normal genes and pseudogenes, is evaluated in Table 1. The estimates are very close to the corresponding true values.

4 Analysis of real datasets

We conducted a Tn-seq experiment that studied a highly virulent strain (IA3902) of *C.jejuni*, which was reported to be the major cause of sheep abortion in the United States and could zoonotically transmit to humans (Sahin *et al.*, 2008, 2012). The strain IA3902 is resistant to tetracyclines, the only antibiotic currently approved in the United States for the treatment of *Campylobacter* abortion in sheep (Delong *et al.*, 1996). The rise of antibiotic resistance calls for knowledge-led approaches to identify new interventions and prevention strategies. A library of Tn5 transposon insertion mutagenesis was prepared according to the experimental procedure described in Section S-1 of the online supplementary file. There are 1631 genes in total, among which 18 are known pseudogenes. In total, we have data for 1 544 034 possible locations within the gene-coding

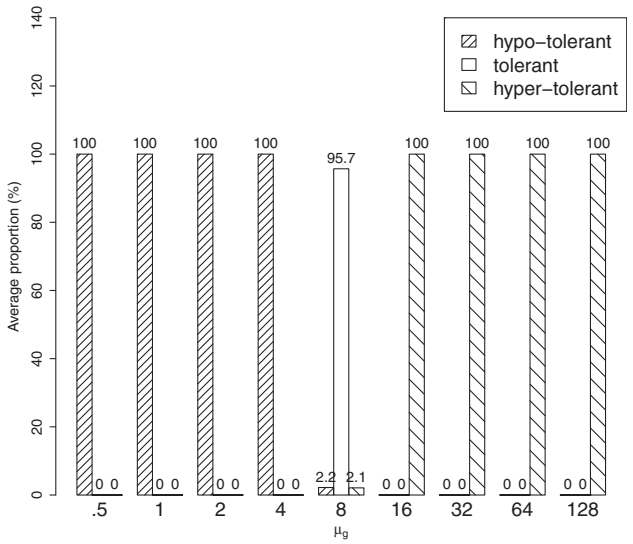

Fig. 1. Proportions of genes that were classified into each of the three different groups (tolerant; hypo-tolerant and hyper-tolerant). These proportions were averaged over the 100 simulated datasets for each true level of μ_g. The numbers above each of the bars are the averaged percentage value. The mean for pseudogenes (μ_0) was set to be 8

Table 1. Summary statistics for the simulation study)." Est Diff" is the estimated difference averaged over the 100 simulated datasets, and" SE" is the corresponding standard error

True diff	Est diff	SE
−7.5	−7.5123	0.3031
−7	−7.0363	0.4314
−6	−6.0588	0.5824
−4	−4.0279	0.7353
0	−0.0155	0.9634
8	7.9988	1.3635
24	23.9986	1.9623
56	55.9530	2.9852
120	119.9082	4.3193

'True Diff' is the true difference between mean abundance of normal genes and pseudogenes ($d_g = \mu_g - \mu_0$). 'Est Diff' is the estimated difference averaged over the 100 simulated datasets, and 'SE' is the corresponding standard error.

regions. Summary statistics are presented in Table 2 for the IA3902 dataset in terms of (1) gene length (l_g: for Tn5, the number of possible transposon insertion locations $n_g = l_g$) and (2) the proportion of locations with zero read count in a gene. Averaged over all genes, the proportion of locations with zero read count is 97.03%, which indicates that zero-inflation is common among genes. The summary statistics also indicates large variability of gene length among the genes.

Although Tn5 transposon can be inserted into any location, as seen in our dataset, it is also reported that Tn5 has a target consensus sequence, A-GNTYWRANC-T (Goryshin *et al.*, 1998). This suggests that the transposon insertion rate tends to be higher for sites matching the sequence A-GNTYWRANC-T. In addition, Herron *et al.* (2004) and Green *et al.* (2012) mentioned there is bias in transposon insertions towards genes with richer GC content. We thus included two covariates in the logistic regression model for the insertion rate p_{gi}. One is named 'MATCHING', which takes value 1 if

Table 2. Summary of the IA3902 Tn-seq data in terms of gene length and the proportion of locations with zero read count within a gene

	Gene length (l_g)	Proportion of 0
Min	93	0.6000
1st Quartile	528	0.9709
Median	808.5	0.9804
3rd Quartile	1209	0.9865
Max	4554	1.0000
Mean	946.1	0.9703

Table 3. Statistical analysis results for the regression coefficients (β's) in the logistic regression model for the insertion rate

Variable	MLE	SE	*P*-value	95% CI
Intercept (β_0)	−6.904	0.051	0.000	(−7.004, −6.803)
Matching (β_1)	1.565	0.286	4.65e−8	(1.003,2.126)
GC percentage (β_2)	0.108	0.002	0.000	(0.105, 0.111)

P-values were obtained by testing the null hypothesis that the corresponding regression coefficient equals zero.

the site matches the target sequence A-GNTYWRANC-T, and takes value 0 otherwise. The other is the percentage of GC content within each gene sequence. Table 3 presents the MLEs and their standard errors of the estimated parameters involved in the logistic regression model for the insertion rate, and *P*-values for testing the inclusion of the corresponding coefficient in the model based on the analysis of the IA3902 dataset. Table 3 shows that the covariate 'MATCHING' and 'GC percentage' are both important factors that influenced the probability of insertions. Within a gene, the odds of having transposon insertion for sites with $'matching' = 1$ over the odds of having transposon insertion for sites with $'matching' = 0$ is $\exp(1.565) = 4.78$. For one percent increase in the GC percentage of a gene, the odds of transposon insertion is expected to increase about 11.4%. The above results indicate that although Tn5 could be inserted into locations not matching the target sequence, the insertion rate was much lower than that for the sites matching the target sequence. And also there is preference for Tn5 insertion into GC-rich DNA. Such quantification of insertion rate is useful in understanding the performance of transposons and the Tn-seq experiment.

The major goal of our Tn-seq experiment is to identify genes involved in bacterial growth. Applying the method we describe in Section 2 resulted in estimates of the mean abundances for all genes, and *P*-values for testing gene tolerance of transposon insertion. The volcano plot in Figure 2 shows the relationship between *P*-values and $\hat{d}_g$'s, the estimated differences in mean abundance of mutants between normal genes and pseudogenes. As expected, there is a positive relationship between the absolute value of $\hat{d}_g$ and the significance of the test, which is indicated by the value of $-\log(P\text{-value} + 2^{-20})$ plotted as *y*-coordinate in Figure 2. We categorized all normal genes into three different states: hypo-tolerant, tolerant and hyper-tolerant while controlling FDR, and the results at different levels of FDR are presented in Table 4. When FDR was controlled at 5%, about 22% of the normal genes were classified as hypo-tolerant, and about 46% of genes were classified as hyper-tolerant. Figure 3 shows the estimated mean abundance of mutants for all normal genes ($\hat{\mu}_g$'s). The estimated mean abundance of mutants for pseudogenes is $\hat{\mu}_0 = 7.0552$, which provides a reference for

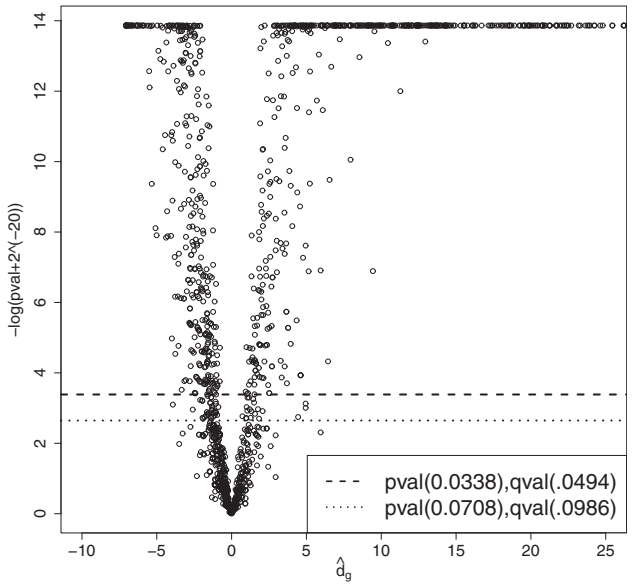

Fig. 2. Volcano plot for the relationship between *P*-values for testing the tolerance status of normal genes and the corresponding $\hat{d}_g$'s, the estimated differences in mean abundance of mutants between normal genes and pseudogenes. The y axis corresponds to $-\log(P-\text{value}+2^{-20})$ while the *x* axis corresponds to estimated d_g's

Table 4. The number of genes classified into each tolerance group at different FDR levels

FDR level α	Hypo-tolerant	Tolerant	Hyper-tolerant	Total
0.001	192	765	656	1613
0.01	265	652	696	1613
0.05	361	507	745	1613
0.1	395	454	764	1613

classifying the normal genes. The classification of genes is presented with the results when FDR was controlled at 5%.

We also applied EL-ARTIST (Chao *et al.*, 2013; Pritchard *et al.*, 2014) to this IA3902 dataset. Detailed comparison between results obtained from EL-ARTIST and our proposed methods are described in Section S-2.2 of the online supplementary material. In summary, the two methods agreed on the classifications of tolerance status for 75% of genes. For those genes whose tolerance status differed between our proposed method and the EL-ARTIST method, a closer investigation of the read count distribution across gene-coding regions indicated that our classifications of tolerance status were more consistent with the raw data observed at the insertion site level.

In addition, we implemented our method for the *M.tuberculosis* dataset (Zhang *et al.*, 2013) that was obtained from an experiment using Himar1 transposon that can only insert into TA dinucleotides. The comparison between EL-ARTIST and our proposed method is described in Section S-2.3 of the online supplementary material. Similar conclusions were obtained as those for IA3902 data analysis.

5 Discussion

A ZIP model is proposed in this article to model the excess of zeros observed in the Tn-seq data. The mixture model of point mass at zero and a Poisson distribution are constructed specifically according to the way Tn-seq data are generated. Multiple testing procedures are applied to classify genes into different categories with respect to

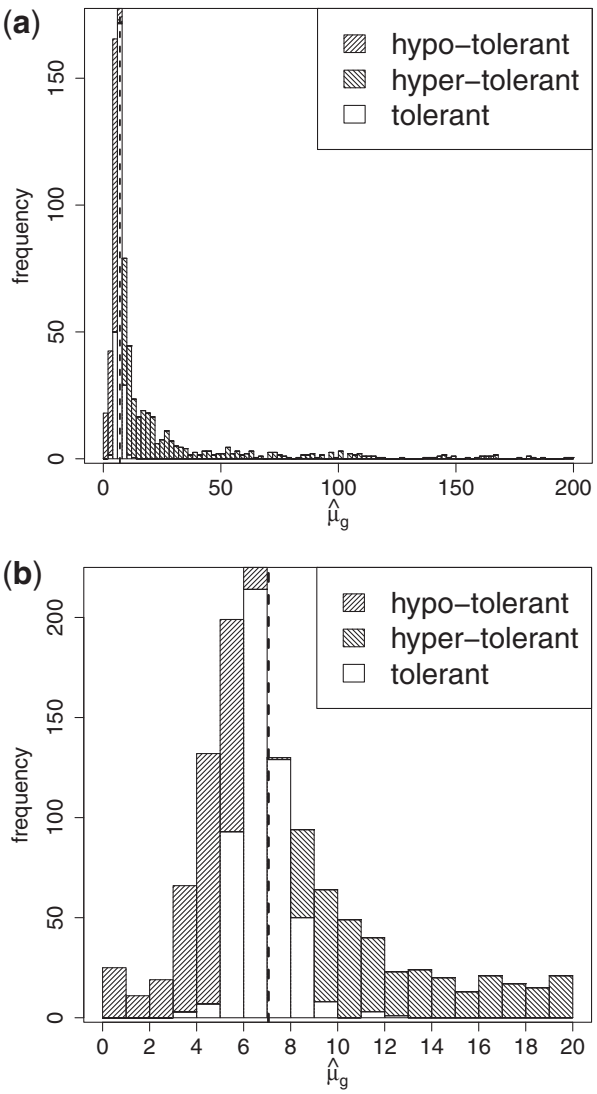

Fig. 3. The histogram of the estimated μ_g's for the normal genes. The upper panel is for all normal genes, and the lower panel is for the subset of normal genes with estimated mean less than 20. The vertical dotted line in both plots denotes the estimated mean abundance for mutants of pseudogenes (7.0552)

transposon insertion tolerance, hence identify genes whose functions are closely related to bacterial growth. We evaluated the proposed method using both simulation studies and analysis of two real datasets. Simulation studies in Section 3 demonstrates that our proposed method performs well in terms of both the estimation of parameters and the classification of genes. In the main text, we discuss an application of our proposed model to the Tn-seq data with the Tn5 transposons which could be inserted possibly into any site within the genome. Our model can also be used in the analysis of Tn-seq data from experiments where transposon insertions are restricted, for example, experiments using Himar 1 transposons that insert into TA dinucleotides only. An application of our method to a Himar 1 dataset is presented in Section S-2.3 of the online supplementary material. Investigation of the distributions of read counts across gene-coding regions indicates that the resulting tolerance statuses classified by the proposed method are reasonable for both datasets. To validate the results, experiments with knock-out genes can be conducted. If a gene is truly hypo-tolerant or even essential, disruption of this gene would reduce growth fitness or even cause death of the strain. Similarly,

knockout of a hyper-tolerant gene would induce advantages for corresponding bacterial growth.

The real data we obtained only studied one growth condition, and we propose to use pseudogenes as an internal standard to set the threshold that classifies normal genes into hypo-tolerant, tolerant and hyper-tolerant categories. The threshold depends on the reliability of the list of pseudo-genes and gene annotations. Pseudogenes are a regular feature of bacterial genomes, which are DNA sequences that are closely related to functional genes but have mutations that destroy the function (Kuo and Ochman, 2010; Lerat and Ochman, 2005). They are usually identified by comparison with the functional sequence in a close relative ('allele' if in the same species, 'homlog' if in another species; Lerat and Ochman, 2005). The mutations in pseudogenes introduce premature stop codon or frame-shift, and then prevent normal translation of a functional protein. In terms of nucleotide composition and length, they are similar to other genes in their genomes. Among different strains of *C.jejuni*, we found that the number of pseudogenes ranges from 18 (i.e. in IA3902) to a few hundreds, which account for about 1–10% of the open reading frames of their genomes. In other taxa, pseudogenes were commonly found to be at the similar ratio, as revealed in *Staphylococcus pyogenes*, *Vibrio vulnificus*, *Vibrio parahaemolyticus*, *Yersinia pestis* and *Salmonella* (Kuo and Ochman, 2010; Lerat and Ochman, 2005). In experiments where several growth conditions are included, our method can be easily generated to those applications by involving growth conditions as factors in modeling the insertion probability and the Poisson mean.

Our proposed method allows incorporation of covariates into the ZIP model and accommodates factors that might affect the insertion rate. In the real data analysis, we quantified the increased insertion probability for the insertion sites matching the target sequence of the Tn5 transposon. Such quantification provides insights into the mechanism of transposon insertion. Besides matching with target sequence, other factors that may potentially affect the transposon insertion rate can also be fit into the ZIP model and tested for significance. Factors that may affect the mean abundance (μ_g) may also be incorporated into the model through Poisson regression.

Large sample theory is implemented in our method while applying the Wald test. In our Tn-seq data, we have more than one million locations, which justifies the large sample assumption is satisfied. The estimated values of mean abundance have been used to classify normal genes into the three tolerance groups. In addition, the magnitude of the estimated values of mean abundance measures how important the corresponding genes are with respect to bacterial growth. If the estimated mean mutants abundance of a normal gene is far above that of pseudogenes, this suggests that the inactivation of this normal gene by transposon insertion provides an obvious growth advantage. On the other hand, if the estimated mean mutants abundance of a normal gene is far below that of pseudogenes or even close to zero, this normal gene may have essential function on bacterial growth.

To test for finite sample behavior of our method, we also tried a simulation study in which only several genes were simulated and we varied the number of possible insertion locations for each gene to check when our method is applicable, and it turns out that the result is good enough as long as the number of possible insertion locations is 50 or more.

Conflict of Interest: none declared.

References

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Chao,M.C. *et al.* (2013) High resolution definition of the *Vibrio cholerae* essential gene set with hidden Markov model-based analyses of transposon-insertion sequencing data. *Nucleic Acids Res.*, **41**, 9033–9048.

Dempster,A.P. *et al.* (1977) Maximum likelihood estimation from incomplete data via the EM Algorithm (with discussion). *J. R. Stat. Soc. Ser. B*, **9**, 1–38.

Dejesus,M.A. and Loerger,T.R. (2013) A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. *BMC Bioinf.*, **14**, 303.

Delong,W.J. *et al.* (1996) Antigenic and restriction enzyme analysis of *Campylobacter* spp. associated with abortion in sheep. *Am. J. Vet. Res.*, **57**, 163–167.

Deng,J. *et al.* (2013) A statistical framework for improving genomic annotations of prokaryotic essential genes. *Plos One*, **8**, e58178.

Gerdes,S.Y. *et al.* (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, **185**, 5673–5684.

Goryshin,I. *et al.* (1998) Tn5/IS50 target recognition. *Proc. Natl. Acad. Sci.*, **95**, 10716–10721.

Green,B. *et al.* (2012) Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mob. DNA*, **3**.

Herron,P.R. *et al.* (2004) Transposon Express, a software application to report the identity of insertions obtained by comprehensive transposon mutagenesis of sequenced genomes: analysis of the preference for in vitroTn5 transposition into GC-rich DNA. *Nucleic Acids Res.*, **32**, e113.

Kuo,C.H. and Ochman, H. (2010) The extinction dynamics of bacterial pseudogenes. *PLoS Genet.*, **6**.

Lambert,D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lampe,D.J. *et al.* (1996) A purified mariner transposase is sufficient to mediate transposition in vitro. *Eur. Mol. Biol. Organ. J.*, **15**, 5470–5479.

Langridge,G.C. *et al.* (2009) Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. *Genome Res.*, **19**, 2308–2316.

Lerat,E. and Ochman, H. (2005) Recognizing the pseudogenes in bacterial genomes. *Nucl. Acids Res.*, **33**, 3125–3132.

Pritchard,J.R. *et al.* (2014) ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing. *PLoS Genet.*, **10**, e1004782.

van Opijnen,T. and Camilli,A. (2014) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol.*, **11**, 435–442.

Sahin,O. *et al.* (2012) Molecular evidence for zoonotic transmission of an emergent, highly pathogenic *Campylobacter jejuni* clone in the United States. *J. Clin. Microbiol.*, **50**, 680–687.

Sahin,O. *et al.* (2008) Emergence of a tetracycline-resistant *Campylobacter jejuni* clone associated with outbreaks of ovine abortion in the United States. *J. Clin. Microbiol.*, **46**, 1663–1671.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.*, **100**, 9440–9445.

Wiles,T.J. *et al.* (2013) Combining quantitative genetic footprinting and trait enrichment analysis to identify fitness determinants of a bacterial pathogen. *PLOS Genet.*, **9**, e1003716.

Zhang,Y.J. *et al.* (2012) Global assessment of genomic regions required for growth in *Mycobacterium tuberculosis*. *PLOS Pathog.*, **8**, e1002946.

Zhang,Y.J. *et al.* (2013) Tryptophan biosynthesis protects mycobacteria from CD4 T-cell-mediated killing. *Cell*, **155**, 1296–1308.

Zomer,A. *et al.* (2012) ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *Plos One*, **7**, e43012.