Bioinformatics, 2021, 1–7
doi: 10.1093/bioinformatics/btab508
Advance Access Publication Date: 13 July 2021
Original Paper
OXFORD

Sequence analysis

Probabilistic identification of bacterial essential genes via insertion density using TraDIS data with Tn5 libraries

Valentine U. Nlebedim [1,]*, Roy R. Chaudhuri [2] and Kevin Walters[1]

[1]Department of Statistics, School of Mathematics, University of Leeds, LS2 9JT, UK and [2]Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, UK

*To whom correspondence should be addressed.

Associate Editor: Can Alkan

Received on February 15, 2021; revised on June 24, 2021; editorial decision on July 2, 2021; accepted on July 23, 2021

Abstract

Motivation: Probabilistic Identification of bacterial essential genes using transposon-directed insertion-site sequencing (TraDIS) data based on Tn5 libraries has received relatively little attention in the literature; most methods are designed for mariner transposon insertions. Analysis of Tn5 transposon-based genomic data is challenging due to the high insertion density and genomic resolution. We present a novel probabilistic Bayesian approach for classifying bacterial essential genes using transposon insertion density derived from transposon insertion sequencing data. We implement a Markov chain Monte Carlo sampling procedure to estimate the posterior probability that any given gene is essential. We implement a Bayesian decision theory approach to selecting essential genes. We assess the effectiveness of our approach via analysis of both simulated data and three previously published *Escherichia coli*, *Salmonella* Typhimurium and *Staphylococcus aureus* datasets. These three bacteria have relatively well characterized essential genes which allows us to test our classification procedure using receiver operating characteristic curves and area under the curves. We compare the classification performance with that of Bio-Tradis, a standard tool for bacterial gene classification.

Results: Our method is able to classify genes in the three datasets with areas under the curves between 0.967 and 0.983. Our simulated synthetic datasets show that both the number of insertions and the extent to which insertions are tolerated in the distal regions of essential genes are both important in determining classification accuracy. Importantly our method gives the user the option of classifying essential genes based on the user-supplied costs of false discovery and false non-discovery.

Availability and implementation: An R package that implements the method presented in this paper is available for download from https://github.com/Kevin-walters/insdens.

Contact: k.walters@sheffield.ac.uk

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

1 Introduction

Bacterial essential genes are those required for growth and survival (i.e. viability). They are usually defined operationally in natural or specific conditions (Karash and Kwon, 2018) with those genes required for basic metabolism or growth in the natural host or environment identified as essential (Tateishi *et al.*, 2020). Genes required for specific biological processes like motility, drug resistance, cell division etc. can also be identified under specific conditions (Mekalanos *et al.*, 2001). The genes that are required under almost all growth conditions are known to be generally or unconditionally essential. Such genes perform essential functions that include fundamental processes like the DNA replication required in all organisms, as well as other essential functions required for the organism's particular lifestyle (Chao *et al.*, 2016). Genes that are only required for growth under some specific conditions are referred to as condition-specific essential genes or simply said to be conditionally essential. The conditionally essential genes depend on factors ranging from environmental to genetic context, and the adaptability of the organism to survive inactivation of unconditional essential genes (Chao *et al.*, 2016). Due to the implications for identifying effective and narrower biochemical drug targets, essential genes are of great interest (Kinnings *et al.*, 2010). The ever-growing concern about antibiotic resistance has coincided with revolutionary progress in the availability of genome sequences and high throughput methods to study bacteria (Su *et al.*, 2019). The development of new technologies and approaches has transformed pathogen studies. The development of genome-wide experimental approaches to identify essential bacterial or virulence genes for *in vivo* survival has seen considerable progress, which could yield potential drug targets (Friedman and Hughes, 2003).

A transposon is a sequence of bacterial DNA that can move, either by a ‘copy and paste’ (replication) or ‘cut and paste’ (relocation) mechanism, within or between DNA molecules. Some transposons have specific target sites (like the mariner family of transposons), others [e.g. transposon 5 (Tn5)] insert into almost any target sequence (Hickman and Dyda, 2016). Transposons have implications for the function of essential genes since their insertion into a region of the genome leads to disruption of the processes associated with that region (Judson and Mekalanos, 2000). Essential genes are known to be conservative and usually do not tolerate transposon insertions, except at the distal parts (near the 5’ or 3’ end) (Bourque *et al.*, 2018). This characteristic of essential genes is key to their identification. Recent approaches to the identification of essential genes are based on the development of hybrid methods that integrate transposon mutagenesis with high-throughput sequencing (Chao *et al.*, 2016).

In recent years Transposon Insertion Sequencing (TIS) technologies have been developed. The processes involved in the TIS technique include the construction of large libraries of mutants in which genes are disrupted randomly by transposon insertions. The expectation at this stage is that the created libraries will have significant numbers of insertions only in genes not required for growth. Furthermore, these libraries are grown under certain experimental conditions of interest. Organisms that grow and survive under the specified conditions are those whose disrupted genes are unnecessary for such functions. The use of next-generation sequencing to facilitate high-throughput identification of essential genes compliments experiments using random mutant libraries (Peng *et al.*, 2017). Different applications of second generation sequencing to transposon mutagenesis screens have evolved independently (Chao *et al.*, 2016). Among them are: High-throughput Insertion tracking by deep sequencing (Gawronski *et al.*, 2009), Transposon-directed Insertion-site Sequencing (TraDIS) (Langridge *et al.*, 2009), Insertion Sequencing (INSeq) (Goodman *et al.*, 2009) and Transposon Sequencing (Tn-Seq) (Van Opijnen *et al.*, 2009; Van Opijnen and Camilli, 2013). As highlighted in Chao *et al.* (2016) these TIS approaches are conceptually identical but there are significant differences in the protocols, as reviewed by Van Opijnen and Camilli (2013).

Most of the statistical methods developed to analyse TIS data utilize insertion-level approaches. They rely on information derived from the number of potential insertion sites to identify essential genes in saturated libraries. They are predominantly designed for mariner-based transposons and exploit the preference for TA site insertions in their approaches to classification. The mariner-based transposons enable densely saturated libraries, that have insertions at all/nearly all TA sites, to be constructed. A major advantage of the mariner-based transposons is that insertions sites (TA-sites) are defined and as such has the assumption of a uniform insertion probability (Barquist *et al.*, 2013). However, there exists some evidence in the literature against this uniform insertion probability assumption (Kimura *et al.*, 2016).

Bio-Tradis is a processing and analysis pipeline by (Barquist *et al.*, 2016) to support the use of TraDIS protocols for identification of essential genes. To make a prediction of gene essentiality, it calculates the insertion index as the number of insertion sites for any gene divided by its length. Based on the assumption that the distribution of the insertion indices across all genes is bimodal with a mode at zero corresponding to essential genes (Langridge *et al.*, 2009), it fits two gamma distributions to the insertion indices corresponding to the two modes of the distribution of the insertion indices. It calculates Log_2 likelihood ratios (LLR) comparing the likelihoods of the insertion index under the two fitted gamma probability densities. A gene is classified as essential if it has an LLR of less than −2, indicating that it is at least 4 times more likely under the essential gene model than the non-essential one. Similarly, a gene is classified as non-essential if it has an LLR greater than 2. AlbaTraDIS (Page *et al.*, 2020) also builds on Bio-Tradis by adopting a sliding window approach, rather than being dependent on the genome annotation.

TRANSIT (DeJesus *et al.*, 2015) applied the Bayesian method for essentiality analysis. Their Bayesian method uses information on the long consecutive sequences of TA-sites lacking insertions as the variable of interest and has the assumption that insertion gaps of TA-sites occur by chance in non-essential regions, with a geometrical decrement in the probability of a long gap. The Gumbel or Extreme Value distribution was used to model the longest consecutive sequence of TA-sites lacking insertion in a gene. Hence, they identified essential genes by unusual long gaps and using the Bayesian framework, the posterior probability of the longest gap is calculated.

ARTIST (Pritchard *et al.*, 2014) has two pipelines for the analysis of TraDIS data: the Essential Loci Analysis (EL-ARTIST) pipeline and Conditional Essential Loci Analysis (Con-ARTIST) pipeline. The EL-ARTIST pipeline identifies regions that are required for optimal growth under a given condition. It uses a sliding window method to define regions that have low read numbers after normalizing the data for incomplete DNA replication. A hidden Markov model is trained on the results of the sliding window analysis and this refines the prediction of whether each TA site is in a region required or dispensable for growth. ARTIST was developed to analyse TIS datasets generated using mariner-based transposons although the authors comment that it should be adaptable to Tn5 transposons.

Lariviere *et al.* (2020) adopted the approach developed by Goodall *et al.* (2018) which involves fitting known distributions to the distribution of saturation indices. Saturation index was computed by Lariviere *et al.* (2020) as the number of insertions within a coding region divided by the length of the coding region. They applied the Bio-TraDIS package to the distribution of the saturation indices for gene essentiality analysis.

One of the significant limitations in the existing methods is that currently-assumed distributions may not model the insertion variability seen in Tn5-based TraDIS data. The negative binomial or Poisson distributions used to model read counts (Seyednasrollah *et al.*, 2015) may not reflect characteristics of Tn5-based TraDIS data which have greater insertion density and genomic resolution due to the non-preferential insertion of Tn5 transposoons. Another drawback of current methods is the way they handle low-frequency sequencing events (Klein *et al.*, 2012; Le Breton *et al.*, 2015; Yang *et al.*, 2017). The need to develop suitable statistical approaches and computational methods to identify essential bacterial genes using Tn5 transposons is paramount.

Insertion-level-based methods that capitalize on the advantages of mariner transposons dominate the literature of current statistical methods for identification of essential genes. Gene-level methods have not been the focus of so much attention. Nonetheless, some studies have successfully used Tn5 transposons under different growth conditions to classify genes (Barquist *et al.*, 2013; Chao *et al.*, 2016; Christen *et al.*, 2011). This paper presents a novel Bayesian computational method for classifying essential bacterial genes using Tn5-based TraDIS data. It coherently accounts for noise that could lead to spurious findings during statistical analysis. Our gene-level approach uses insertion density as the sole variable in the classification. It avoids the need to arbitrarily set a threshold or to use normalization procedures before analysis like, for example, the trimmed mean method (Zomer *et al.*, 2012).

2 Materials and methods

Our approach uses gene-level Tn5 transposon data in the classification procedure. With a large number of transposon insertions, some potential insertion positions could record multiple insertions. Rather than focussing on the total number of insertions per gene we count unique insertions sites (e.g. three insertions in the same position count as one unique insertion site). Given the assumption that transposons inserts randomly within the genome, the number of unique insertion sites for any gene is assumed to increase with gene length so we scale the number of unique insertion sites by gene length. We define the insertion density for a given gene as the number of unique insertion sites divided by the gene length and use insertion density as a classifier of bacterial gene essentiality. We exploit the fact that insertions in essential genes are lethal except at the distal portions, whilst taking into account the fact that what counts as distal in this context will likely vary by gene.

We assume conditional independence of the insertion densities for any two genes given that their essentiality statuses and the values of any relevant parameters. Insertion densities are in the interval $[0, 1]$ so we chose to model the probability distribution of insertion density as a beta distribution. For each gene, we derived the posterior probability that the gene is essential using Markov Chain Monte Carlo (MCMC) via both Metropolis–Hastings (MH) and Gibbs sampling as required. MCMC is now a common tool for the analysis of a variety of applied genetic problems (Alenazi *et al.*, 2019; Boggis *et al.*, 2016) We call our model the INSDENS model.

2.1 Bayesian model

In this paper, π is used to denote the prior probability distribution; f is used to represent the likelihood, full conditional distributions and joint probability distributions; underscore is used to represent a vector or set; the shape and rate parameters are used to parameterize the gamma distribution.

Let $\underline{d}$ represent the set of insertion densities values for all genes. Let d_i represent the insertion density for the ith gene. Let Z_i represent a binary indicator variable indicating whether the ith gene is essential ($Z_i = 1$) or not ($Z_i = 0$). Let G be the total number of genes, $\underline{Z} = (Z_1, Z_2, Z_3, \ldots, Z_G)$ represent the indicator vector of essentiality for all genes. Let $\underline{d}_E$ and $\underline{d}_N$ represent the set of insertion densities for essential and non-essential genes, respectively. Furthermore, d_{Ej} is the jth element of the set d_E with similar meanings for d_{Nj}. Let G_E be the cardinality of the set $\underline{d}_E$ and G_N be the cardinality of the set $\underline{d}_N$.

2.1.1 The likelihood

For essential genes we assume that

$$d_i | Z_i = 1, \alpha_E, \beta_E \quad \sim \quad \mathrm{Beta}(d_i; \alpha_E, \beta_E) \tag{1}$$

and for non-essential genes, we assume that

$$d_i | Z_i = 0, \alpha_N, \beta_N \quad \sim \quad \mathrm{Beta}(d_i; \alpha_N, \beta_N). \tag{2}$$

Let $\underline{\Theta} = \{\alpha_E, \beta_E, \alpha_N, \beta_N\}$. Assuming conditional independence of d_m and d_n given Z_m and Z_n and the parameters, the full likelihood is

$$\mathrm{f}(\underline{d} | \underline{Z}, \underline{\Theta}) = \prod_{i=1}^{G} \mathrm{f}(d_i | Z_i, \underline{\Theta}). \tag{3}$$

2.1.2 The prior distributions

We assume the random variable Z_i, representing the essentiality status for gene i, has a Bernoulli probability distribution with parameter θ. This, assuming that Z_i is independent of Z_j, a priori, for $i \neq j$ implies

$$\pi(\underline{Z} | \theta) = \theta^{G_E} (1 - \theta)^{G_N}. \tag{4}$$

Since $0 \leq \theta \leq 1$ and since θ is uncertain, we place a Beta distribution on θ to capture our prior belief in the uncertainty.

$$\theta \sim \mathrm{Beta}(0.1, 0.9). \tag{5}$$

This choice of hyperparameters for the theta prior specifies a mean of 0.1 and a monotonically decreasing probability density. Its 10th percentile is approximately 1×10^{-10} and its 90th percentile is approximately 0.4, so that there is reasonable probability density in the range of values that, a priori, might be anticipated for the proportion of essential genes. For each dataset, we used the observed insertion densities to make a guess at the essentiality of each gene (genes with insertion density below some threshold were assumed to be essential). We calculated the mean and variance of the insertion densities separately for those genes we guessed as essential and those we guessed as non-essential genes. Equating expressions for the mean and variance of the Beta distribution with the sample means and variances of the insertion densities in both groups allows us to obtain estimates of the parameters in $\underline{\Theta}$. This is similar in spirit to an empirical Bayes approach (Spencer *et al.*, 2016) but equating moments is simpler than maximizing marginal likelihoods in this case as it avoids the need for numerical optimization. To allow for uncertainty in the parameter values in $\underline{\Theta}$, we place gamma priors on each of them. We set the rate parameter of each gamma prior to be 1 and the shape parameter equal to the value of the relevant $\underline{\Theta}$ parameter determined my moment matching. This gives a prior with a suitably large variance. For example if, by moment matching, we guessed that a parameter in $\underline{\Theta}$ was 3 our prior for that parameter would be gamma$(3, 1)$. The actual priors used are given in Table 1. We also performed a marginal sensitivity analysis for the *Staphylococcus aureus* dataset to determine whether our results were sensitive to the choice of these hyperparameter values.

2.1.3 Full joint distribution

Combining the full likelihood and prior distributions, the full joint probability distribution for our model is derived as below as:

$$\mathrm{f}(\theta, \underline{\Theta}, \underline{Z}, \underline{d}) = \pi(\theta)\pi(\Theta)\pi(\underline{Z}|\theta) \prod_{i=1}^{G} \mathrm{f}(d_i | Z_i, \underline{\Theta}). \tag{6}$$

2.1.4 The conditional distributions

To obtain posterior probabilities of each gene being essential we derived, up to proportionality, the conditional probability densities from which to sample. Using Equations (4) and (5) the conditional distribution for θ is given by

$$\mathrm{f}(\theta | \Theta, \underline{Z}, \underline{d}) \propto \pi(\theta)\pi(\underline{Z}|\theta) \tag{7}$$

$$\propto \theta^{G_E - 0.9} (1 - \theta)^{G_N - 0.1} \tag{8}$$

which gives

$$\theta | \Theta, \underline{Z}, \underline{d} \sim \mathrm{Beta}(0.1 + G_E, 0.9 + G_N). \tag{9}$$

Let $\Theta_{-\psi}$ represent the set $\underline{\Theta}$ excluding the element ψ and let $B(.,.)$ be the usual beta function. Using Equations (1) and (3) and assuming $\alpha_E \sim \mathrm{Ga}(\phi, \lambda)$, the conditional distribution for α_E is given by

$$\mathrm{f}(\alpha_E | \Theta_{\alpha_E}, \theta, \underline{Z}, \underline{d}) \propto \frac{\alpha_E^{\mathrm{f}-1} \mathrm{e}^{-\lambda \alpha_E}}{B(\alpha_E, \beta_E)^{G_E}} (\prod_{j=1}^{G_E} d_{Ej})^{\alpha_E} \tag{10}$$

$$\propto \frac{\alpha_E^{\phi-1} \exp\left[-\alpha_E \left(\lambda - \sum \log(d_{Ej})\right)\right]}{B(\alpha_E, \beta_E)^{G_E}} \tag{11}$$

where the limits on the summation are the same as those on the product. The conditional distributions of the remaining parameters in $\underline{\Theta}$ can be obtained in a similar manner and are detailed in Supplementary File S1. The full conditional for Z_i is a Bernoulli$\left(\frac{p_E}{p_E + p_N}\right)$ probability distribution where

$$p_E = \theta \prod_{j=1}^{G_E} \mathrm{f}(d_{Ej} | Z_j = 1, \Theta) \tag{12}$$

Table 1. Shape parameters of the gamma prior distributions of the parameters in $\underline{\Theta}$

Dataset	Shape parameters of			
	α_E	β_E	α_N	β_N
E. coli	2	166	2	17
S. aureus	1	182	11	83
S. Typhimurium	2	178	3	18

$$p_N = (1-\theta)\prod_{j=1}^{G_N} \mathrm{f}(d_{Nj}|Z_j = 0, \Theta) \qquad (13)$$

The posterior probability of gene i being essential is calculated as the proportion of MCMC iterations (ignoring burn-in) where it is classified as being essential.

2.1.5 Sampling from the posterior distributions

We used the a random-walk MHs algorithm to sample from the posterior distribution of the parameters in $\underline{\Theta}$ and Gibbs sampling to sample from the posterior distributions of $\underline{Z}$ and θ. We update each element of $\underline{Z}$ sequentially. After sampling Z_i both p_E and p_N are recalculated to take into account which genes are currently considered to be essential.

We use the following procedure when using the MH algorithm to update each of the parameters in $\underline{\Theta}$: Let Θ_{ic} and Θ_{ip} be the current and proposed value of the ith parameter in $\underline{\Theta}$, respectively. The proposed value is obtained by drawing a random deviate h from a $N(0, 1)$ distribution and then calculating $\Theta_{ip} = \Theta_{ic}\exp(\sigma^2 \times h)$ where σ^2 is tuned to give an appropriate acceptance rate of around 30%. Since the proposal density is symmetric in Θ_{ic} and Θ_{ip}, the MH ratio is

$$\gamma(\Theta_{ip}, \Theta_{ic}) = \frac{f(\Theta_{ip}|\Theta_{-i}, \theta, \underline{Z}, \underline{d})}{f(\Theta_{ic}|\Theta_{-i}, \theta, \underline{Z}, \underline{d})} \qquad (14)$$

We sample a realization u of a uniform[0, 1] distribution and accept the proposed value Θ_{ip} with probability $\min(1, \gamma(\Theta_{ip}, \Theta_{ic}))$. After some experimentation, we set the initial value of σ^2 of the MHs sampling procedure for the parameters α_E, β_E, α_N and β_N as 0.25, 0.30, 0.04 and 0.05, respectively and manually modified each value to get an acceptable acceptance rate.

We selected multiple sets of initial values for each chain to check for convergence. We performed a quantitative diagnostic check for convergence by calculating the potential scale reduction factor values for all the parameters in $\underline{\Theta}$. All values were sufficiently close to 1 to indicate convergence.

We specified the burn-in to be 2000 iterations. This is highly conservative, but we wanted to be sure that the retained values were from the posterior distribution in all analyses. Some of the parameters showed moderate autocorrelation. Current practice is not to thin the samples to reduce the autocorrelation but instead to run the chain for longer. We ran the MCMC sampler for a 22 000 so that after discarding burn-in we were left with 20 000 iterations to base the posterior probabilities on.

We initialized the probability of a gene being essential as $\theta = 0.1$ corresponding to a 10% chance that any given gene is essential. We initialized the values of the parameters α_E, β_E, α_N and β_N to be the values specified in Table 1. We initialized the essentiality assignment, Z_i, to be 1 if the insertion density for that gene was below 0.025 and 0 otherwise.

2.1.6 Real datasets

We apply our INSDENS model to three TraDIS datasets: *Escherichia coli*; *Salmonella* Typhimurium and *S. aureus*. The *E. coli* K-12 BW25113 data (Goodall *et al.*, 2018) were sourced from the European Nucleotide Archive under accession number ERR2249109. The *E. coli* K-12 BW25113 data has a total number of 448 854 insertions. The list of *E. coli* K-12 genes likely to be essential was obtained from Baba *et al.* (2006), and the likely non-essential genes were genes for which a deletion mutant was successfully generated in that study.

The *S.* Typhimurium data used in this study was based on the work of Barquist *et al.* (2013) and was sourced from the European Nucleotide Archive under accession numbers ERR009073 and ERR009074. The *S. Typhimurium* data has a total of 639969 distinct transposon insertions mapped to its genome. For the *S. Typhimurium* dataset, the genes we listed as essential were homologues of the genes from *S. Typhimurium* ST14028 which Porwollik *et al.* (2014) attempted to knock out using two different approaches and were unsuccessful, and also homologues of *E. coli* K-12 genes which could not be knocked out by Baba *et al.* (2006). The genes listed as non-essential were homologues of genes which had mapped knockout mutations by both approaches in *S. Typhimurium* ST14028 (Porwollik *et al.*, 2014), with the exclusion of six genes which were homologues of *E. coli* K-12 genes defined as essential by Baba *et al.* (2006).

The *S. aureus* data (Christiansen *et al.*, 2014) were sourced from the database under accession numbers SRR105406 to SRR1056422. The genes designated as essential were homologues of genes which were identified as essential both in Fey *et al.* (2013) and Chaudhuri *et al.* (2009). The non-essential genes were defined as genes which had mutations in the Fey *et al.* (2013) study.

All datasets were re-analysed using a common pipeline. Transposon tag sequences (where present) were removed using Cutadapt (Martin, 2011). The tag-free reads were mapped to the reference genome using BWA mem (Li, 2013). The mapped position of the 5' end of each read was determined using bedtools bamtobed (Quinlan and Hall, 2010), and used to infer the position of transposon insertions. The UNIX tool uniq was used to count the number of reads associated with each unique insertion site.

2.1.7 Synthetic datasets

We simulated four different datasets to mimic four scenarios of interest. Each dataset contained 4000 genes of which 200 were essential. The datasets were generated using properties of the *E. coli* TraDIS dataset. In the *E. coli* dataset, we observed that the variance in the number of insertions by gene increases with gene length (Supplementary File S2). To mimic this behaviour, we implemented the following procedure: initialize the gene-specific insertion probability to be the gene length divided by the total gene length across all genes; generate a multiplicative factor for each gene from a beta(5, 4) distribution; randomly multiply or divide the gene-specific insertion probability by this multiplicative factor. We determined the number of insertions for each gene by randomly sampling genes according to this adjusted insertion probability. Insertions in non-essential genes were inserted at random anywhere in the gene. For essential genes, the insertion probability decreased with distance from the gene ends. Specifically we sample values from an exponential distribution with rate parameter λ. These sampled values represent the distance from the gene ends relative to the gene length. We considered $\lambda = 50$ and $\lambda = 10$. When $\lambda = 50$ we expect more than 99% of insertions to be within 10% of the gene ends. When $\lambda = 10$ this value drops to 63%. We also incorporated noise into our simulations by allowing spurious insertions to insert anywhere in both essential and non-essential genes.

We label our four scenarios as HH, HL, LH and LL. The first letter specifies whether the total number of insertions across all genes is high (H) or low (L). High corresponds to 700 000 insertions and low to 20 000 insertions. The second letter specifies whether the exponential rate parameters is high (H) or low (L). High corresponds to $\lambda = 50$, low to $\lambda = 10$. In all scenarios, we specified the number of spurious reads to be approximately 5% of the number of true insertions. Table 2 shows the different combinations used. The total number of unique insertions for the *E. coli*, *S. aureus* and *S. Typhimurium* datasets are approximately 347 133 and 353 000, respectively.

2.1.8 Comparative study

We used receiver operating characteristic (ROC) curves to compare the classification performance of INSDENS in both our real and

Table 2. Numbers of true insertions, spurious insertions and exponential rate parameter, λ, for the four simulated data scenarios

Dataset	Number of Insertions	Number of spurious insertions	λ
HH	700 000	30 000	50
HL	700 000	30 000	10
LH	20 000	1000	50
LL	20 000	1000	10

simulated datasets. When comparing the performance of INSDENS versus Bio-Tradis for our real datasets we compared the true positive rates (TPRs) and false positive rates (FPRs). Rather than having to place an arbitrary threshold on the posterior probability of essentiality, we used Bayesian decision theory which allows us to assign costs to making a decision (gene is essential or non-essential), given the truth (gene is essential or non-essential). We assume that there is no cost in making the correct decision. There are, therefore, two costs to assign: false discovery cost (C_A) when INSDENS incorrectly identifies a gene as essential when it is actually non-essential and the false non-discovery cost (C_B) when INSDENS identifies a gene as non-essential when it is actually essential (Walters *et al.*, 2021). Bayesian decision theory states that gene i is declared to be essential if the posterior cost of declaring it to be essential is less than the posterior cost of declaring it be non-essential. If p_i is the posterior probability of gene i being essential then this inequality becomes $(1-p_i)C_A < p_iC_B$ which leads to the following classification rule: gene i is classified as essential if

$$p_i > R = \frac{C_A}{C_A + C_B} = \frac{1}{1+Q}. \qquad (15)$$

where $Q = C_B/C_A$ is a ratio of costs.

3 Results

3.1 Analysis of real datasets

The ROC curves for the INSDENS model applied to the *E. coli*, *S. aureus* and *S. Typhimurium* datasets are shown in Figure 1. The area under the curve (AUC) values are 0.975, 0.983 and 0.967 for the *S. aureus*, *E. coli* and *S. Typhimurium* datasets, respectively. We investigated the effect of changing one hyperparameter at a time (marginal sensitivity analysis) for the *S. aureus* dataset to determine the sensitivity of the results to the hyperparameter value. We took each of the four values for *S. aureus* in Table 1 one at a time and multiplied it by 2 and then by 0.5, keeping all the other values fixed. This require eight runs of the MCMC routine. The sensitivity analysis conducted showed that the AUCs are not very sensitive to changes in the hyperparameters with all 8 AUCs between 0.973 and 0.977 inclusive.

3.2 Analysis of synthetic datasets

The ROC curves when our model is applied to the synthetic datasets are shown in Figure 2. The AUCs are 0.994, 0.876, 0.766 and 0.541 for HH, LH, HL and LL, respectively. The AUC values in the four simulated scenarios are affected by both the number of insertions (AUC increases with the number of insertions) and how far from the gene ends insertions are tolerated in essential genes (AUC increases as this tolerance decreases). Our simulation software could be used as a tool to estimate the likely classification performance using our MCMC procedure based on simple dataset-specific characteristics.

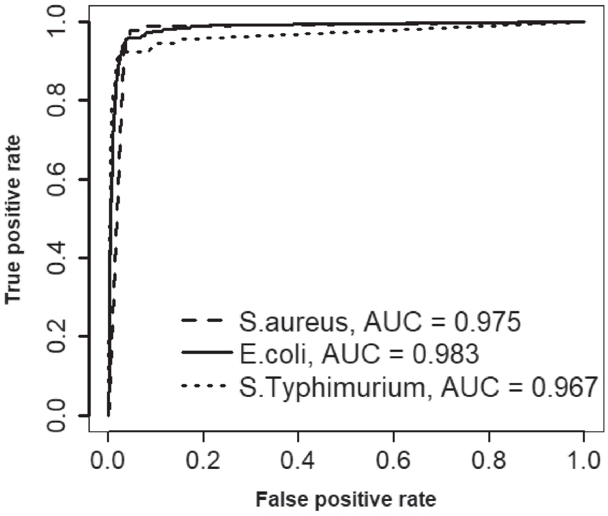


Fig. 1. ROC curves and AUCs (in the legend) for the *E. coli*, *S. Typhimurium* and *S. aureus* datasets using the posterior probability of gene essentiality as the classifier

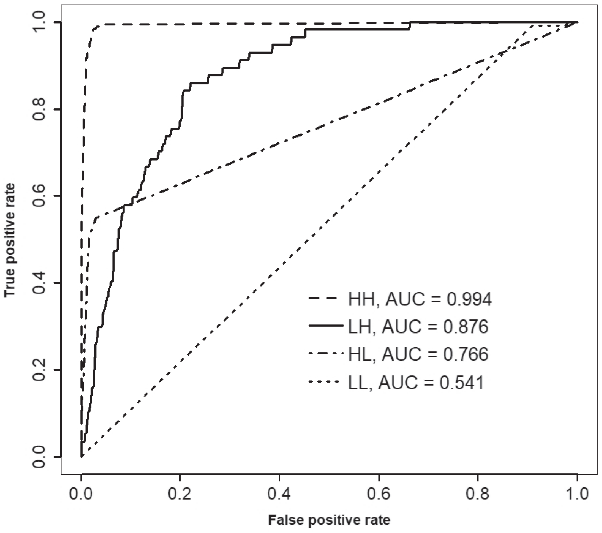


Fig. 2. The ROC curves for the simulated HH, LH, HL and LL datasets. The AUCs are given in the legend. The classification statistic is the posterior probability of gene essentiality (**a**) *E. coli*, (**b**) *S. Typhimurium* and (**c**) *S. aureus*

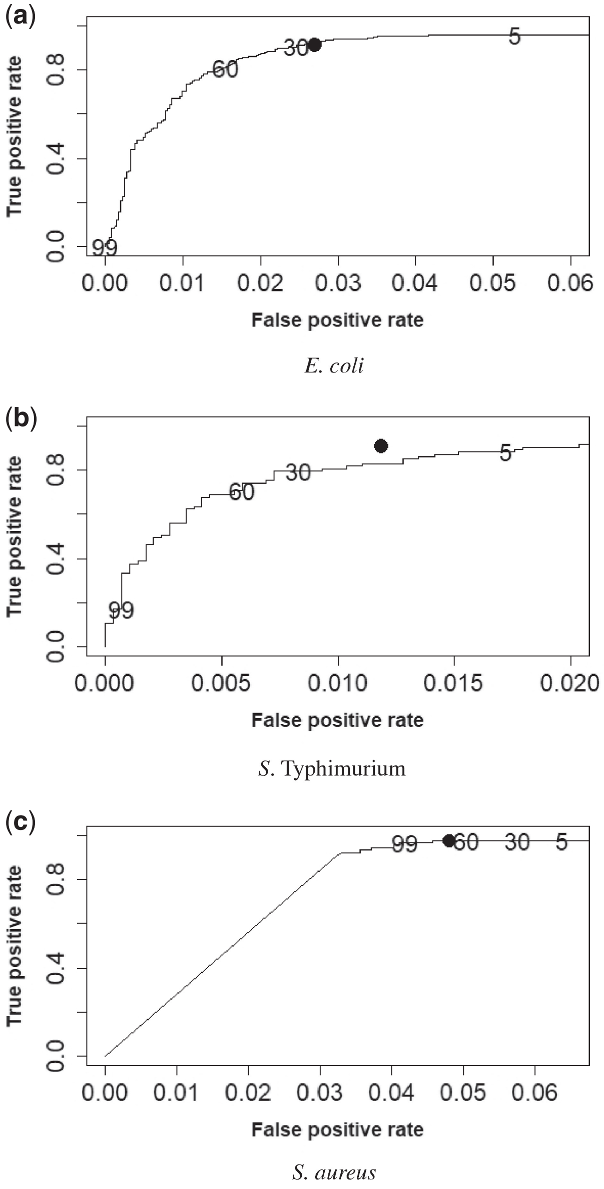
3.3 Comparison with bio-tradis

When calculating the FPRs and TPRs for INSDENS we used four values of $C_A(5, 30, 60, 99)$ and $C_B(95, 70, 40, 1)$ in the Bayesian decision theory approach. These require the posterior probability to exceed $R = 0.05, 0.30, 0.60$ and 0.99, respectively, for the gene to be declared essential. The FPRs and TPRs for INSDENS and Bio-Tradis when both methods were applied to the *E. coli*, *S. Typhimurium* and *S. aureus* datasets are shown in Figure 3. Each plot shows four points for INSDENS, one for each value of $R = C_A/(C_A + C_B)$, and a single point for Bio-Tradis. We observe from Figure 3 that the FPRs and TPRs of the Bio-Tradis analyses are on or close to the ROC curves for each of the three datasets. We also see that our method affords considerable flexibility in the TPRs and FPRs depending on the costs C_A and C_B.

4 Discussion

Our new method for classifying bacterial genes using TraDIS data avoids the need to arbitrarily set thresholds or use normalization procedures before analysis. It also avoids giving a hard classification as, for example, Bio-Tradis does. This gives the user more flexibility in determining which set of genes might be essential. Using our approach it is possible to select fewer genes by decreasing Q. This would lead to a reduction in the both the TPR and FPR. Whilst Bio-Tradis performs well, it offers no control over the number of genes classified as essential.

Choosing the costs in a Bayesian decision theory approach is not routinely undertaken and some users may feel unable to confidently attach such costs. In this case, we suggest conducting a sensitivity analysis to see how the selected essential gene set changes as the costs of making incorrect decisions vary. Conceptually, the simplest way to do this is to vary the value of Q, which measures how much more costly it is to misclassify an essential gene than to misclassify a non-essential gene. A user might choose a range of values of Q say $Q_{\min} < Q < Q_{\max}$ and monitor how the set of genes declared essential varies for values of Q in this range. Alternatively users can put their own threshold on the posterior probability or ranks genes using it.



E. coli

S. Typhimurium

S. aureus

Fig. 3. True and false positive rates for Bio-Tradis and INSDENS for four different values of $R = C_A/(C_A + C_B)$. The point in ROC space is labelled with its numerical R value for INSDENS or with a solid circle for Bio-Tradis. For convenience, the R value shown is 100 times the actual value

Running 10 000 MCMC iterations on a desktop PC with a single 1.8 GHz processor with 8 GB of RAM took 100, 80 and 30 s for the *E. coli*, *S. Typhimurium* and *S. aureus* datasets, respectively. It may be possible to reduce the run-time further using importance sampling or sampling importance re-sampling, provided a suitable choice of importance distribution can be found.

There is further information that could improve the classification of essential genes. Our current approach is potentially susceptible to noise since we do not leverage read count information in our model. Doing so would allow us to better distinguish true reads from spurious reads and might lead to further classification performance with relatively little computationally cost.

Funding

This work was supported by the Nigerian Government through the NEEDS Assessment Scholarship/Fellowship [FUT/DVC(Acad.)/GEN 101/1/18].

Conflict of Interest: none declared.

References

Alenazi,A.A. *et al.* (2019) Bayesian variable selection using partially observed categorical prior information in fine mapping association studies. *Genet. Epidemiol.*, **43**, 690–703.

Baba,T. *et al.* (2006) Construction of *Escherichia coli* k-12 in-frame, single--gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, **2**, 2006.0008.

Barquist,L. *et al.* (2013) Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biol.*, **10**, 1161–1169.

Barquist,L. *et al.* (2016) The tradis toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics*, **32**, 1109–1111.

Boggis,E. *et al.* (2016) equips: eqtl analysis using informed partitioning of snps–a fully Bayesian approach. *Genet. Epidemiol.*, **40**, 273–283.

Bourque,G. *et al.* (2018) Ten things you should know about transposable elements. *Genome Biol.*, **19**, 1–12.

Chao,M.C. *et al.* (2016) The design and analysis of transposon insertion sequencing experiments. *Nat. Rev. Microbiol.*, **14**, 119–128.

Chaudhuri,R.R. *et al.* (2009) Comprehensive identification of essential *Staphylococcus aureus* genes using transposon-mediated differential hybridisation (tmdh). *BMC Genomics*, **10**, 291.

Christen,B. *et al.* (2011) The essential genome of a bacterium. *Mol. Syst. Biol.*, **7**, 528.

Christiansen,M.T. *et al.* (2014) Genome-wide high-throughput screening to investigate essential genes involved in methicillin-resistant *Staphylococcus aureus* sequence type 398 survival. *PLoS One*, **9**, e89018.

DeJesus,M.A. *et al.* (2015) Transit-a software tool for himar1 tnseq analysis. *PLoS Comput. Biol.*, **11**, e1004401.

Fey,P.D. *et al.* (2013) A genetic resource for rapid and comprehensive phenotype screening of non-essential *Staphylococcus aureus* genes. *MBio*, **4**, e00537–12.

Friedman,R. and Hughes,A.L. (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.*, **20**, 154–161.

Gawronski,J.D. *et al.* (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for haemophilus genes required in the lung. *Proc. Natl. Acad. Sci.*, **106**, 16422–16427.

Goodall,E.C. *et al.* (2018) The essential genome of *Escherichia coli* k-12. *MBio*, **9**, e02096–e02117.

Goodman,A.L. *et al.* (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe*, **6**, 279–289.

Hickman,A.B. and Dyda,F. (2016) DNA transposition at work. *Chem. Rev.*, **116**, 12758–12784.

Judson,N. and Mekalanos,J.J. (2000) Transposon-based approaches to identify essential bacterial genes. *Trends Microbiol.*, **8**, 521–526.

Karash,S. and Kwon,Y.M. (2018) Iron-dependent essential genes in *Salmonella typhimurium*. *BMC Genomics*, **19**, 610.

Kimura,S. *et al.* (2016) The nucleoid binding protein h-ns biases genome-wide transposon insertion landscapes. *MBio*, **7**, e01351–16.

Kinnings,S.L. *et al.* (2010) The mycobacterium tuberculosis drugome and its polypharmacological implications. *PLoS Comput. Biol.*, **6**, e1000976.

Klein,B.A. *et al.* (2012) Identification of essential genes of the periodontal pathogen *Porphyromonas gingivalis*. *BMC Genom.*, **13**, 578.

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

Lariviere,D. *et al.* (2020) Reproducible and accessible analysis of transposon insertion data at scale. *bioRxiv*.

Le Breton,Y. *et al.* (2015) Essential genes in the core genome of the human pathogen *Streptococcus pyogenes*. *Sci. Rep.*, **5**, 9838.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *ArXiv*, 1303.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10–12.

Mekalanos,J.J. *et al.* (2001) Systematic identification of essential genes by in vitro transposon mutagenesis. *US Patent*, **6**, 207384.

Page,A.J. *et al.* (2020) Albatradis: comparative analysis of large datasets from parallel transposon mutagenesis experiments. *PLoS Comput. Biol.*, **16**, e1007980.

Peng,C. *et al.* (2017) A comprehensive overview of online resources to identify and predict bacterial essential genes. *Front. Microbiol.*, **8**, 2331.

Porwollik,S. *et al.* (2014) Defined single-gene and multi-gene deletion mutant collections in salmonella enterica sv typhimurium. *PLoS One*, **9**, e99820.

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

Quinlan,A. and Hall,I. (2010) Quinlan ar, hall im. bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Seyednasrollah,F. *et al.* (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.*, **16**, 59–70.

Spencer,A.V. *et al.* (2016) Incorporating functional genomic information in genetic association studies using an empirical Bayes approach. *Genet. Epidemiol.*, **40**, 176–187.

Su,M. *et al.* (2019) Genome-based prediction of bacterial antibiotic resistance. *J. Clin. Microbiol.*, **57**, e01405–18.

Tateishi,Y. *et al.* (2020) Genome-wide identification of essential genes in mycobacterium intracellulare by transposon sequencing—implication for metabolic remodelling. *Sci. Rep.*, **10**, 1–12.

Van Opijnen,T. *et al.* (2009) Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods*, **6**, 767–772.

Van Opijnen,T. and Camilli,A. (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol.*, **11**, 435–442.

Walters,K. *et al.* (2021) The utility of the laplace effect size prior distribution in Bayesian fine-mapping studies. *Genet. Epidemiol.*, **45**, 386–401.

Yang,Z.R. *et al.* (2017) A noise trimming and positional significance of transposon insertion system to identify essential genes in *Yersinia pestis*. *Sci. Rep.*, **7**, 41923.

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