ConNIS and labeling instability: new statistical methods for improving the detection of essential genes in TraDIS libraries

Moritz Hanke[1*], Theresa Harten[2], Ronja Foraita[1]

1 Department Statistical Methods in Epidemiology, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Bremen, Germany
2 No affiliation

* hanke@leibniz-bips.de

Abstract

The identification of essential genes in *Transposon Directed Insertion Site Sequencing (TraDIS)* data relies on the assumption that transposon insertions occur randomly in non-essential regions, leaving essential genes largely insertion-free. While intragenic insertion-free sequences have been considered as a reliable indicator for gene essentiality, so far, no exact probability distribution for these sequences has been proposed. Further, many methods require setting thresholds or parameter values *a priori* without providing any statistical basis, limiting the comparability of results.

Here, we introduce *Consecutive Non-Insertions Sites* (*ConNIS*), a novel method for gene essentiality determination. *ConNIS* provides an analytic solution for the probability of observing insertion-free sequences within genes of given length and considers variation in insertion density across the genome. Based on an extensive simulation study and different real world scenarios, *ConNIS* was found to be superior to prevalent state-of-the-art methods, particularly when libraries had only a low or medium insertion density. In addition, our results showed that the precision of existing methods can be improved by incorporating a simple weighting factor for the genome-wide insertion density. To set methodically embedded parameter and threshold values of *TraDIS* methods a subsample based instability criterion was developed. Application of this criterion in real and synthetic data settings demonstrated its effectiveness in selecting well-suited parameter/threshold values across methods. A ready-to-use R package and an interactive web application are provided to facilitate application and reproducibility.

Author summary

Identifying essential genes in bacteria is key to understanding their ability to survive, which can, for example, be applied to the development of new treatments. One way to do identify these genes is by creating libraries where small DNA fragments ("insertions") are randomly placed in the genome: essential genes tend to remain insertion-free because insertions disrupt their function. The challenge is to determine whether a (long) uninterrupted sequence is due to chance or because the gene is truly essential. Here, we present *Consecutive Non-Insertion Sites (ConNIS)*, a statistical method that calculates the probability of such insertion-free sequences. Extensive comparisons on simulated and real datasets show that *ConNIS* outperforms existing methods, especially when a library is rather sparse in terms of the total number of insertion sites. Since many

analysis methods rely on parameter values that have to be set before the analysis and can heavily influence the final results, we also propose a data-driven approach to set these values, making results more comparable across studies. Our methods are freely available as an R package and all results are presented in a web app.

Introduction

Determination of genes essential for the growth and survival of bacteria has been of major interest in genetic research as it provides a deeper understanding of lifestyle and adaptation [1–3]. While site-directed mutagenesis approaches determine essential genes accurately, such methods are laborious and time-consuming when performed globally. Consequently, whole genome analyses have only been conducted for well-known model organisms such as *Escherichia coli*, i.e., the Keio library [4]. In the last decade, wider availability of high-throughput sequencing methods initiated a shift from single-gene to whole-genome analysis, resulting in the development of *transposon insertion sequencing (TIS)* methods. Employing transposons that are randomly inserted into the genome enables researchers to generate large mutant libraries and to characterize them by the location of insertion sites (IS) via high-throughput sequencing. *Transposon directed insertion site sequencing (TraDIS)* is a widely applied *TIS* method [5–8] and has been established for the determination of essential genes in various scientific set-ups [9–15].

A key challenge in *TIS* studies resides in the statistical analysis, which typically aims to maximize the detection of true positives (essential genes) while minimizing the number of false positives (non-essential genes incorrectly identified as 'essential'). Although there are multiple software suits and packages, not every method embedded therein will be equally suitable for the analysis of the obtained data set. For example, sliding window approaches [16, 17] and Hidden Markov Models [18, 19] have been proposed for the analysis of high-density libraries which regularly originate from *mariner* transposon based mutagenesis [20, 21]. However, many *TIS* studies utilize *Tn5*-based transposons, which have different underlying assumptions and constraints in terms of the data generating process: Unlike *mariner* transposons, *Tn5* insertions do not depend on the presence of specific motifs and can theoretically occur in any non-essential region of the genome [6, 22]. Nevertheless, *Tn5*-based libraries reported so far tend to be less dense than *mariner*-based libraries. Consequently, observing larger genomic regions deployed of IS just by chance becomes more likely in *Tn5* based libraries. Furthermore, the entity of detected IS across the genome rarely displays a uniform distribution of gene-wise insertion densities. Reasons may be transposon-driven preferences for GC- or AT-rich regions [23–26] and genomic hot- or coldspots, i.e., genomic regions of notably higher or lower insertion densities, respectively. [24, 27–29].

So far, a couple of *Tn5*-based statistical methods for identifying essential genes have been proposed. Burger et al. [30] suggest estimating the probability of observing several IS within a gene of a given length based on a binomial distribution using the genome-wide insertion density as success probability. The *Tn5Gaps* method of the `Transit` package [31] uses a Gumbel distribution to approximate the probabilities of observed IS-free gaps along the genome. Essentiality is determined by the largest gap within or partially overlapping a gene. Since both methods rely on p-values derived from thousands of genes, the authors recommend correcting for multiple testing. However, they do not evaluate how different correction approaches might affect the identification of essential genes. Alternatively, the `Bio-TraDIS` software package [32] avoids the multiple testing problem by heuristically leveraging an often observed bimodal distribution of gene-wise insertion densities. Combining an exponential distribution (for essential genes) with a gamma distribution (for non-essential genes), genes are labeled either 'essential', 'non-essential', or 'ambiguous' based on an *a priori*

set $\log_2$ likelihood ratio threshold. In practice, a clear distinction between the two distributions is not always guaranteed [6], and the threshold values are usually set arbitrarily. The recently proposed Bayesian method *InsDens* calculates the posterior probability of a gene being essential [33] and the authors suggest to use Bayesian decision theory to set a posterior probability threshold. Although this method offers a clear interpretation, it requires choosing an *a priori* probability distribution parameter, too, which can influence the outcome.

Some but rather limited comparative studies of *TIS* methods are available. Based on high-density library data, Larivire et al. [20] draw a comparison between the bimodal approach of `Bio-TraDIS`, the *Tn5Gaps* method and a custom modification of the bimodal approach [11]. However, only two threshold values were applied and no performance analysis under different controlled data-generating processes was described. While Nlebedim et al. [33] used four different parameter combinations to generate synthetic data, the only method applied in the analysis was their method *InsDens*. The additional analysis of three real-world datasets using `Bio-TraDIS` suffers from the application of only one and rather low threshold value. Similarly, Ghomi et al. [34] proposed to use the non-parametric clustering algorithm embedded in the `DBSCAN` R package for the sole reason that it allows omitting the heuristic setting of threshold values required by `Bio-TraDIS`. Again, only a single and rather low threshold value for comparison using `Bio-TraDIS` was applied. However, the `DBSCAN` clustering algorithm itself requires setting two *a priori* parameter values but the authors did not report how different values might affect labeling performance, nor did they provide guidance on how to choose appropriate values.

The widespread use of *TIS*, particularly *Tn5* statistical analysis methods, contrasts with the lack of systematic reviews of these methods, especially when considering different data-generating processes. Furthermore, a transparent comprehensive statistical method for setting threshold or parameter values in *TIS* methods is missing. As a consequence, publications often only justify the choice of methods and parameters by citing prior studies that used similar approaches and parameters. In addition, most studies lack sensitivity analyses for their threshold and parameter values. A common practice is truncating the 5'- and/or 3' ends of genes by several base pairs or up to 20%, to align with the assumption that the gene ends are generally non-essential [6, 13, 35–38], yet this approach is never investigated in sensitivity analyses.

At this scientific stage, we provide the following contributions to the statistical detection of essential genes. First, we introduce *Consecutive Non-Insertion Sites (ConNIS)*, a novel method that determines gene essentiality based on insertion-free sequences within genes. *ConNIS* provides an analytical solution for the probability of observing the longest insertion-free sequence within a gene, based on its length and the number of IS under the assumption of being non-essential. Second, we performed an extensive simulation study with 160 parameter combinations mimicking different data-generating processes. Using these synthetic data, four semi-synthetic datasets and three real datasets, *ConNIS* demonstrated its superiority against five state-of-the-art *Tn5* analysis methods, especially, in settings with low and medium-dense libraries. In this context, we propose to use a weighting factor when applying genome-wide insertion density values to better represent genomic regions with low insertion densities. This modification also improved three competing methods by reducing the number of false positives without losing too many true positives in many settings. Third, we provide for the first time a data-driven instability criterion for setting thresholds and parameter values in *TIS* methods, thereby making the results from different studies and methods comparable and more transparent. Applications of this approach to real and synthetic datasets clearly demonstrate its suitability for setting parameter values for all methods considered. *ConNIS* and the instability criterion for all competing methods have been

made available as a ready-to-use R package. To further explore our results, we provide an interactive web app and publicly available code.

Materials and methods

Consecutive Non-Insertion sites (ConNIS)

The transcription of an essential gene is, by definition, vital for an organism's survival and its hindrance due to transposon insertion will result in the mutant's removal from the population. Consequently, IS are expected to be exclusively detected in the non-essential genome, which comprises genes, intergenic regions and smaller fractions of essential genes [7]. Based on the assumption that a *Tn5* transposon can occur at any position in the *non-essential* genome, we propose *ConNIS*, a method that classifies a gene as 'essential' or 'non-essential' by analyzing its largest insertion-free sequence in terms of base pairs. Therefore, we derive a novel probability distribution (see S1 File) that we used to determine the probability of observing an insertion-free sequence in a gene of given length and number of IS.

Let b denote the length of a genome in terms of base pairs that contains p genes with corresponding lengths $b_1, b_2, \ldots, b_p$. We define $\theta = h/b$ as the genome-wide insertion density, where h is the genome-wide number of observed IS. For a gene $j = 1, \ldots, p$ let $\hat{h}_j = b_j \cdot \theta$ be the expected number of IS within gene j under the assumption that gene j is not essential. Furthermore, let l_j be the length of the *longest* observed consecutive sequence of non-insertions of gene j (Fig 1A).

Fig 1. Overview of *ConNIS*. A For a gene j, determine its length b_j and its longest insertion-free sequence l_j. **B** Set a weight w for the genome-wide insertion density θ reflecting rather low density regions of the genome. **C** *ConNIS*: The probability of observing l_j within gene j due to random chance.

Based on Theorem 1 (see Section 1 in S1 File), we calculate the probability of observing l_j as

$$\mathbb{P}(l_j = i) = f(i, b_j, h_j) = \frac{\binom{b_j - i - 1}{b_j - \hat{h}_j - i}}{\binom{b_j - 1}{b_j - \hat{h}_j - 1}}.$$

ConNIS further takes into account that IS might not be uniformly distributed across the genome (see Fig S1 in S2 File for examples). Simply using θ to estimate the number of IS in a gene under the assumption of non-essentiality bears the risk of increasing the number of false positives in genomic regions with insertion densities well below θ because non-insertion sequences are more likely to occur by chance. While normalization methods for non-uniformly distributed *insertion counts* are available, these cannot handle non-uniformly distributed IS. We address this issue by applying a weight factor $0 < w \leq 1$ to θ to represent better low-density regions (Fig 1B). ConNIS is then defined as the probability of observing an insertion-free consecutive sequence of at least length l_j in gene j:

$$\mathcal{C}\left(l_j, b_j, \hat{h}_j, w\right) := 1 - \mathbb{P}(l_j < i) = 1 - \sum_{i=1}^{l_j - 1} \frac{\binom{b_j - i - 1}{b_j - \hat{h}_j w - i}}{\binom{b_j - 1}{b_j - \hat{h}_j w - 1}}. \tag{1}$$

Given a significance level of $0 < \alpha < 1$, we declare a gene j to be essential if $\mathcal{C}\big(l_j, b_j, \hat{h}_j, w\big) \leq \alpha$. To control the global type I error when applying *ConNIS*, we suggest using the Bonferroni(-Holm) method [39, 40] to control the family-wise error

rate (FWER) or the Benjamini-Hochberg method [41] to control the false discovery rate (FDR).

Competing state-of-the-art methods with proposed weighting strategy

We compared *ConNIS* with five popular state-of-the-art *Tn5* analysis methods for determining essential genes:

1. the *Binomial* distribution approach in the `TSAS 2.0` package [30],
2. the approach of fitting a bimodal distribution based on gene-wise insertion densities included in the `Bio-TraDIS` package [32] (referred to as *Exp. vs. Gamma* method throughout this paper),
3. the *InsDens* method[1] [33],
4. the *Tn5Gaps* method of the `TRANSIT` package [31] and
5. the *Geometric* distribution.

Although the geometric distribution has not been published as a stand-alone method, we included it as a competitor because it is the limiting distribution of *ConNIS* (see Theorem 2 in the S1 File) and it has been part of an analysis pipeline for determining the probability of insertion-free regions in the genome [42].

The *Binomial* and *Geometric* methods use the genome-wide insertion density θ as success probability, and the *Tn5Gaps* method uses it as a location parameter. However, for the reason outlined above, this naive use of θ can increase the number of false positives within genomic regions with a relatively low insertion density compared to the rest of the genome. To address this potential pitfall, we introduce a weight w to adjust θ when applying these methods. We then use these modified methods, as well as the original versions, for comparison. See Section 2 in S1 File for methodological details.

A labeling instability criterion for tuning parameter selection

TIS methods often require *a priori* set parameter or threshold values that will influence the final number of genes labeled 'essential' and therefore the methods' performances in terms of correct classification. The setting of an 'appropriate' parameter/threshold value in a given data scenario can be interpreted as a *tuning* problem. In this context our data-driven tuning approach selects a parameter of threshold value for a TIS method from a set of candidate values.

Therefore, we consider observed IS as a realizations of an (unknown) probability distribution across the genome, that is, if a *TIS* experiment is repeated multiple times, we would expect to observe each time different IS positions across the non-essential parts of the genome. Consequently, gene-wise IS metrics, such as the longest insertion-free sequence or the IS density, would also exhibit variation, thus, the set of genes labeled as 'essential' could change (at least in parts) across experiments. The main idea of our selection criterion is to leverage these variations by quantifying the average variation of gene labeling based on m subsample for given a tuning value. Inspired by stability selection approaches in linear regression and graph estimation problems [43–45], a 'good' tuning value should give rather stable results in terms of gene classification, i.e. the gene labeling should be less sensitive to the random occurrence of

[1] The method requires several prior hyperparameters. In line with the authors' claim, our tests on selected simulation settings showed minimal impact from these choices. Thus, we used the default settings of the `R` package `insdens` for all simulations and real data analyses. See `https://github.com/Kevin-walters/insdens`, commit: `286f114`.

IS across the genome. In the following, we detail the procedure using the selection of a weight value w for *ConNIS* as an example. A transfer to other *TIS* methods or to threshold based filters in the data pre-processing steps is straight forward.

Let $\mathbf{w} = w_1, w_2, \ldots, w_z$ be a sequence of ordered weights w with $0 < w_q < w_r \leq 1$ for $q < r$ and $w_q, w_r \in \mathbf{w}$. Assume further that m subsamples of size $h_{sub} < h$ have been drawn from the h observed IS (Fig 2A).

Fig 2. Overview of the labeling instability criterion to select the weight parameter for *ConNIS*. A Drawing m subsamples of the original observed IS. **B** Calculation of the instability values for all weights $w_y \in \mathbf{w}$ based on the estimated variances of a Bernoulli variable. **D** Selecting the weight w_s with the lowest instability $\phi(w_y)$. **D** Application of *ConNIS* using w_s followed by a multiple testing correction to identify putative essential genes.

For $w_y \in \mathbf{w}$, gene labeling is then performed on each of the m subsets using *ConNIS*. By modeling the gene labeling as a Bernoulli process, we estimate the probability of labeling a gene j as 'essential' using

$$\hat{\pi}_j(w_y) = \frac{g_j(w_y)}{m},$$

where $g_j(w_y)$ represents the number of times gene j is labeled as 'essential' across m subsamples (Fig 2B). The Bernoulli variance is estimated by $\hat{\pi}_j(w_y) \cdot \big(1 - \hat{\pi}_j(w_y)\big)$. We can then define the instability of labeling genes for a given weight w_y as

$$\phi(w_y) := \sum_{j=1}^{p} \frac{\hat{\pi}_j(w_y) \cdot \big(1 - \hat{\pi}_j(w_y)\big)}{q(w_y)}. \tag{2}$$

Here, $q(w_y) = \sum_{j=1}^{p} \mathbb{1}\big(\hat{\pi}_j(w_j) > 0\big)$ is the total number of genes that have been labeled at least once as 'essential' in the m subsamples. This normalization factor ensures $0 \leq \phi_j(w_y) \leq 0.25$. A value of $\phi(w_y) = 0$ indicates complete consistency in gene labeling across subsamples, whereas $\phi(w_y, d) = 0.25$ reflects total instability, equivalent to randomly assigning labels by flipping a fair coin for each gene in each subsample.

Calculating the instability for all $w_y \in \mathbf{w}$, we get a sequence of instability values $\boldsymbol{\phi} = \{\phi(w_1), \phi(w_2), \ldots, \phi(w_p)\}$ and select

$$w_s = \underset{w_y \in \mathbf{w}}{\arg\min}\, \phi(w_y),$$

that is the weight w_y that minimizes the labeling instability (Fig 2C). Finally, *ConNIS* is applied to the original data with w_s (Fig 2D).

Depending on the range of $\mathbf{w}$ and the number of IS, very small weight values may lead to instability values approaching or even reaching zero with (nearly) all genes being labeled as 'non-essential'. Following other stability approaches [43–45], these values are excluded from the set of candidate tuning values because they provide no useful information. For our instability criterion, we propose to omit all weights smaller than the smallest weight that maximizes the function $\phi(w)$, i.e.,

$$w_{max} = \min \left\{ w_y \in \mathbf{w} \mid \underset{w_y \in \mathbf{w}}{\arg\max}\, \phi(w_y) \right\}$$
$$w_s = \underset{\tilde{w}_y \in \mathbf{w} : \tilde{w}_y \geq w_{max}}{\arg\min} \phi(\tilde{w}_y).$$

Fig 3. MCC and PRC performance based on synthetic data. The plots show the LOESS-smoothed curves with 95% confidence intervals for the MCC and PRC in four synthetic data settings. Dots on the curves indicate the average performance using no weighting ($w = 1$) or the least stringent threshold ($t = 2$ for *Exp. vs. Gamma* and $r = 0.1$ for *InsDens*). In all settings, 5 % of both ends of each gene were trimmed.

Results

Comparison of ConNIS with state of the art methods

To evaluate the performance of *ConNIS* and its competitors we applied all methods to synthetic, semi-synthetic and real-world data. A detailed explanation of our simulation schemes for generating synthetic data that covers different assumptions about the data generating process can be found in S3 File. For the application to real-world data we used three publicly available *Tn5* libraries of different organisms and insertion densities. The semi-synthetic datasets were generated from a high-density *Tn5* library by randomly deleting IS.

Since the number of essential genes is small compared to the number of non-essential genes, we used the Mathew's Correlation Coefficient (MCC), a metric suitable for imbalanced data [46–48], as main performance measure. For comparison the MCC is plotted given the number of genes labeled 'essential'. The latter can be controlled by the methods' different parameter and threshold values. In addition, the precision-recall-curve (PRC) are shown to investigate two desirable yet occasionally conflicting objectives: selecting as many true positives as possible while avoiding an inflated number of false positives.

In all applications the weight value of *Binomial*, *ConNIS*, *Geometric* and *Tn5Gaps* was set to $w = 0.1, 0.2, \ldots, 1$ with $w = 1$ being the original version. For *Exp. vs. Gamma*, we set the $\log_2$-likelihood ratio threshold $t \in \{2, 3, \ldots, 12\}$, covering the range of values commonly reported in the literature. The posterior probability threshold of *InsDens* was set at $r \in \{0.01, 0.1, 0.2, \ldots, 0.9, 0.99\}$. Furthermore, we truncated genes by excluding the distal ends by either 0%, 5% or 10% and applied Bonferroni(-Holm) and Benjamini-Hochberg procedures for multiple testing correction.

Synthetic data settings

We present four illustrative synthetic data settings offering an overview of performance the methods. The results of all 160 different settings and additional classification metrics can be interactively explored at `https://connis.bips.eu`, supporting our findings.

Synthetic data example 1 (SDE1) had 200,000 IS randomly distributed along the genome in a sinusoidal shape. 'Essential' genes were defined to contain insertion-free sequences of at least 75% of the gene's length. This represented scenarios where essential genes can contain IS relatively far from their distal ends. Fig 3A shows *ConNIS* clearly outperforming the other methods with regard to the MCC. The PRC demonstrates that *ConNIS* effectively enhanced the identification of true essential genes without substantially inflating false positives by maintaining high precision. Notably, all methods showed their peak performance when the number of genes labeled 'essential' was about the number of true essential genes (dashed orange line). The plot also highlights the beneficial effect of applying a weight $w < 1$ to *Binomial* and *Geometric* (the points indicate the average performance if no weight is applied, i.e., $w = 1$).

Synthetic data example 2 (SDE2) mimicked high-density libraries with a relatively high number of 'noise IS', i.e., spurious observations of IS due to technical

Fig 4. MCC and PRC performances based on real-world data. Dots indicate the performance of the original methods ($w = 1$). **A** *E. coli* BW25113 strain with ≈102,000 IS [49]. **B** *E. coli* K-12 MG1655 strain with ≈390,000 IS [50]. **C** *Salmonella enterica* serovar Typhimurium 14028S strain with ≈186,000 IS [37]. Note, in applications A and B, most of the results of *Exp. vs. Gamma* are covered by those of *InsDens*.

issues. 400,000 IS were randomly distributed along the genome in a sinusoidal shape, and the essential genes had insertion-free sequences of at least 85% of their length. In addition, 2% of the total number of IS was added in a final step as noise IS. These 8000 IS could also appear in originally insertion-free sequences of essential genes, i.e., interfere with methods like *ConNIS* and *Tn5Gaps* which are based on insertion-free sequences. As expected, all methods benefit from the higher number of IS for the MCC and showed string PRC performances (Fig 3B). *ConNIS*, *Exp. vs. Gamma* and *InsDens* achieved the highest MCC values with rather small differences.

Synthetic data example 3 (SDE3) was reversed to SDE2 by mimicking scenarios where only a low insertion density can be achieved, e.g., due to bottleneck effects by environmental pressure. Therefore, 50,000 IS were randomly distributed along the genome in a sinusoidal shape, and the essential genes had insertion-free sequences of at least 85% of their length. All methods suffered from the sparse library (Fig 3C). *ConNIS* achieved the highest MCC value if the number of genes labeled 'essential' was close to the number of true essential genes. *Binomial* and *Tn5Gaps* could only achieve mediocre values at best. For *Exp. vs. Gamma*, *InsDens* and *Tn5Gaps*, the range of the number of genes labeled 'essential' never contained the number of true genes. However, the first two could achieve MCC values that were slightly worse than *ConNIS*. With respect to the PRC, all methods induced false positives due to larger non-insertion sequences occurring by chance compared to denser libraries. *ConNIS* tended to have a rather high precision while *Exp. vs. Gamma* and *InsDens* achieved rather high recall values, but at the price of an inflation of false positives.

Synthetic data example 4 (SDE4) covered scenarios with so-called 'cold-spots' along the genome, which have a much lower chance of containing IS. 'Essential' genes were defined to contain insertion-free sequences of at least 80%. In non-essential sections of the genome, each base pair had the same probability to contain one of the 200,000 IS, yet, in 25 randomly placed sections of size 10,0000bp these probabilities were lowered by factor 10. *ConNIS* achieved clearly the best MCC values and PRC performance (Fig 3D). However, all methods tended to overestimate the number of essential genes (indicated by the rather low precion values) due to the higher chance of false positives in cold spots.

Real-world data

In the first example, we applied all methods to an *E. coli* BW25113 strain library comprising approximately 102,000 IS at time point T0 [49]. As ground truth, we used the results of the single-knockout study by Baba et al. [4], which is often considered the gold standard. Fig 4A shows *ConNIS* outperforming the other method by reaching MCC values up to 0.65. *InsDens* and *Exp. vs. Gamma* labeled too many genes as 'essential' (at least 575) even for their strictest thresholds ($t = 12$ and $r = 0.99$). However, the thresholds of *Exp. vs. Gamma* had only a marginal influence on the number of genes labeled 'essential', which resembles in parts the results of the simulation study. The PRCs reveal that all methods achieved mediocre precision values at best with *ConNIS* having relatively stable precision values for rising recall values.

A high-density *E. coli* K-12 MG1655 library characterized by approximately 390,000

IS [50] was used as seccond example. As truth, we used the gene essentiality classification from the Profiling of *E. coli* Chromosome (PEC) database [51]. Here, results were consistent with the simulation study, i.e., all methods benefited from the high number of IS and performed best when the number of selected genes approached the number of true essential genes (Fig 4B). Weighting ($w < 1$) was highly beneficial for all methods, as the original versions ($w = 1$, indicated by the points) had high recall values at the cost of low precision values.

In the third example, all methods were applied to a *Salmonella enterica* serovar Typhimurium 14028S library comprising approximately 186,000 IS [37]. Following Nlebedim et al. [33], we used as truth the combined set of essential genes provided by Baba et al. [4] and Porwollik et al. [52]. In this scenario all methods showed at best mediocre performance. The removal of IS with low read counts[2] improved the performances of the methods slightly and might be a sign of the presence of spurious IS [31, 53]. *ConNIS* achieved the best MCC and precision values. *Exp. vs. Gamma* and especially *InsDens* performed very poorly (Fig 4C), with the latter labeling far too many genes as 'essential', resulting in negative MCC values. *Binomial*, *ConNIS*, and *Tn5Gaps* benefited from a fairly low weight w, which seemed to reduce the number of false positives comapred to the original versions ($w = 1$).

Semi-synthetic data settings

To further investigate the influence of the number of IS on the methods' performances we generated semi-synthic data by drawing IS subsamples of sizes 50,0000, 100,000, 200,000 and 400, 000 from a very high density *Tn5* library [11]. The Kaio library [4] was used as a reference for true gene essentiality. In low to medium density libraries (subsamples of 50,000 to 200,000 IS) *ConNIS* outperformed the other methods, clearly (see Fig 5). Similar to the synthetic and real-world data settings, *ConNIS* showed its best performance in terms of MCC when the number of selected genes was about the number of true essential genes. In case of the rather high-density library (subsample size of 400,000 IS), all methods were on par.

Fig 5. MCC performance for semi-synthetic data. Subsamples were generated by randomly drawing IS from a high-density library of *E. coli* BW25113 [11]. The Kaio library [4] was used as reference for 'true' gene essentiality.

Application of gene labeling instability for tuning parameter selection

The performance of applying the gene labeling instability criterion for tuning parameter selection was investigated on the previous mentioned three real-world datasets as well as three randomly chosen dataset examples from the simulation study[3]. We applied the criterion to *ConNIS* and the other methods to select the respective tuning parameters w, t, and r for each method. For each setting, $m = 500$ subsamples were drawn without replacement, with each subsample containing 50% ($h_{sub} = 0.5$) of the IS from the original data. Genes were truncated by 5 % at each distal end. The MCC was used to assess the performance of the labeling instability criterion and was compared to the best possible MCC across all candidate tuning values.

[2] We tried minimum read count thresholds $1, 2, 3, 5, 10$.

[3] **Synthetic dataset 1**: 400,000 sinusoidal distributed IS, essential genes contained an insertion free sequence of $\geq 80\%$. 8000 IS where added as noise. **Synthetic dataset 2**: 100,000 sinusoidal distributed IS and essential genes contained an insertion-free sequence of $\geq 75\%$. **Synthetic dataset 3**: 200,000 uniformly distributed IS with 25 cold spots, 4000 noise IS and essential genes contained an insertion-free sequence of $\geq 75\%$.

Table 1. Tuning performance of the gene labeling instability criterion.

Method	BW25113	MG1655	14028S	Synthetic data 1	Synthetic data 2	Synthetic data 3
Binomial	0.55 (0.58)	0.71 (0.79)	0.51 (0.53)	0.83 (0.89)	**0.73 (0.73)**	0.61 (0.69)
ConNIS	**0.64 (0.64)**	0.67 (0.79)	**0.57 (0.57)**	0.90 (0.94)	**0.88 (0.88)**	0.76 (0.77)
Exp. vs. Gamma	0.47 (0.49)	0.78 (0.79)	0.39 (0.41)	0.91 (0.92)	**0.81 (0.81)**	0.72 (0.74)
Geometric	**0.52 (0.52)**	0.58 (0.74)	0.46 (0.48)	0.86 (0.91)	0.59 (0.75)	0.63 (0.70)
InsDens	**0.51 (0.51)**	0.74 (0.78)	-0.26 (0.00)	0.91 (0.92)	**0.81 (0.81)**	**0.74 (0.74)**
Tn5Gaps	0.55 (0.56)	0.63 (0.77)	**0.52 (0.52)**	**0.91 (0.91)**	**0.77 (0.77)**	**0.70 (0.70)**

Tuning was applied on three real-world and three synthetic datasets. For each method, the MCC is presented when the weight was chosen using the gene labeling instability criterion. The best possible MCC over all tuning values is displayed in the parentheses. Bold entries indicate that the instability criterion selected the tuning value with the best possible MCC.

The application of the gene labeling instability criterion for determining a weight for *ConNIS* shows to be highly beneficial. For the real-world datasets *E. coli* BW25113 and *Salmonella enterica* serovar Typhimurium 14028S, as well as the synthetic dataset 2, the application of instability criterion successfully identified the 'optimal' weights based on the corresponding MCC values (Table 1). In these examples, *ConNIS* also achieved the highest MCC with its selected w compared to all other methods. Furthermore, for synthetic datasets 1 and 3, the MCCs obtained by the instability criterion were close to its highest possible MCCs in these settings (0.90 vs. 0.94 and 0.76 vs. 0.77). Only for the MG1655 strain data, our tuning approach was less successful for *ConNIS* (0.67 vs. 0.79).

When applied to the other methods, the criterion was also successful for tuning, resulting in many cases where the best possible MCC was achieved (see Table 1). For *Exp. vs. Gamma* in all six settings $\log_2$ thresholds were selected that resulted in MCC values close to the best possible MCC values. Applying labeling instability to the posterior probability threshold r in *InsDens* yielded favorable results in five settings and achieved the highest possible MCC in three of them. Only in the *Salmonella enterica* serovar Typhimurium 14028S real-world dataset, the selected r had a bad MCC value. However, since *InsDens* generally showed a weak performance in this setting before (Fig 4C), the result is not surprising.

Discussion and Conclusion

In this work, we addressed three main challenges inherent in statistical analysis in *TraDIS* studies. The first challenge arises from the fact that in *Tn5* datasets every base pair of the genome serves a potential insertion site, while reported insertion densities often remain far below saturation levels. Considering this, *ConNIS* gives an analytic solution for the probability of observing an insertion-free sequence within a gene of a given length and number of insertion sites. The second challenge is the often observed non-uniform distribution of IS across the genome. Neglecting this factor can lead to an increased number of (nearly) insertion-free genes being incorrectly labeled as essential in regions with relatively low insertion densities. Addressing non-uniformity, *ConNIS* contains a weighting parameter that increases the precision by making it more difficult to label genes as 'essential' in low-density regions. We extended this idea to three state-of-the-art methods to improve their precision. The third challenge lies in the fact that many *TIS* methods rely on *a priori* set threshold or parameter values, which can substantially influence labeling performance. However, an 'objective' criterion for setting these values has been lacking, often resulting in arbitrarily chosen values. By introducing the concept of gene labeling instability based on subsamples of observed IS, we proposed a data-driven approach to select appropriate parameters and threshold

values.

An extensive simulation study and application to three real-world datasets and four semi-synthetic datasets was conducted to compare the performance of *ConNIS* to multiple state-of-the-art *Tn5* analysis methods. In most settings, *ConNIS* outperformed these methods or was at least on par with the best of them. Unlike its competitors, *ConNIS* showed usually robust performances for (arbitrarily) chosen truncation and filter values. The results also confirm our idea of weighting the genome-wide insertion density when applied to existing methods: it could reduce the number of false positives without sacrificing too many true positives. Applying our proposed gene labeling instability criterion for tuning parameter selection in various real-world and multiple synthetic data scenarios demonstrated its potential to select favorable weight and threshold values for all methods.

Given that *ConNIS* demonstrated superior performance, especially in low and medium insertion density settings, its application is expected to improve the precision of results in experimental settings characterized by high selective pressure or observation of bottleneck effects. While we have investigated *ConNIS*' ability to select essential genes, it is tempting to speculate that its application might be similarly beneficial for the determination of *conditionally* essential genes, e.g., by comparing its gene-wise values between conditions via a fold change. The application of a weight could be further improved by incorporating multiple weighting values to target different genomic regions more effectively.

Our gene labeling instability criterion was originally developed for selecting threshold and parameter values of *Tn5* analysis methods, but it may also be applicable to other *TIS* methods that employ alternative transposons, such as the popular *mariner* transposon. It might also serve as a criterion in pre-processing steps like quality filters or trimming of distal gene ends. Last but not least, our work showed the crucial role of the underlying data-generating process on the performance of all methods. Future work could expand the range of scenarios considered, helping researchers choose the most appropriate method for analyzing their data.

Supporting information

S1 File. Proofs and methodological extensions. PDF file with proofs for *ConNIS* and formal definition of the extension of existing methods.

S2 File. Additional plots. PDF file with plots of additional analysis results.

S3 File. Detailed description of the generation of synthetic data.

Acknowledgments

The authors would like to thank Ian Henderson, Emily Goodall and Ash Robinson for providing the list of insertion sites of their high-density library of the *E. coli* BW25113 strain.

References

1. Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A. Essential genes on metabolic maps. Current Opinion in Biotechnology. 2006;17(5):448–456.

2. Zhang Z, Ren Q. Why are essential genes essential? - The essentiality of Saccharomyces genes. Microbial Cell. 2015;2(8):280–287.

3. Shang W, Wang F, Fan G, Wang H. Key elements for designing and performing a CRISPR/Cas9-based genetic screen. Journal of Genetics and Genomics. 2017;44(9):439–449.

4. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology. 2006;2(1).

5. Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, et al. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Research. 2009;19(12):2308–2316.

6. Chao MC, Abel S, Davis BM, Waldor MK. The design and analysis of transposon insertion sequencing experiments. Nature Reviews Microbiology. 2016;14(2):119–128.

7. Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, van Opijnen T. A decade of advances in transposon-insertion sequencing. Nature Reviews Genetics. 2020;21(9):526–540.

8. Liang Y, Luo H, Lin Y, Gao F. Recent advances in the characterization of essential genes and development of a database of essential genes. iMeta. 2024;3(1).

9. Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Coller JA, et al. The essential genome of a bacterium. Molecular Systems Biology. 2011;7(1).

10. Rubin BE, Wetmore KM, Price MN, Diamond S, Shultzaberger RK, Lowe LC, et al. The essential gene set of a photosynthetic organism. Proceedings of the National Academy of Sciences. 2015;112(48).

11. Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, Cunningham AF, et al. The Essential Genome of Escherichia coli K-12. mBio. 2018;9(1).

12. Sternon JF, Godessart P, Gonçalves de Freitas R, Van der Henst M, Poncin K, Francis N, et al. Transposon Sequencing of Brucella abortus Uncovers Essential Genes for Growth In Vitro and Inside Macrophages. Infection and Immunity. 2018;86(8).

13. Poulsen BE, Yang R, Clatworthy AE, White T, Osmulski SJ, Li L, et al. Defining the core essential genome of Pseudomonas aeruginosa. Proceedings of the National Academy of Sciences. 2019;116(20):10072–10080.

14. Luo H, Lin Y, Liu T, Lai FL, Zhang CT, Gao F, et al. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Research. 2020;49(D1):D677–D686.

15. Rivas-Marin E, Moyano-Palazuelo D, Henriques V, Merino E, Devos DP. Essential gene complement of Planctopirus limnophila from the bacterial phylum Planctomycetes. Nature Communications. 2023;14(1).

16. Pritchard JR, Chao MC, Abel S, Davis BM, Baranowski C, Zhang YJ, et al. ARTIST: High-Resolution Genome-Wide Assessment of Fitness Using Transposon-Insertion Sequencing. PLoS Genetics. 2014;10(11):e1004782.

17. Zhang YJ, Ioerger TR, Huttenhower C, Long JE, Sassetti CM, Sacchettini JC, et al. Global Assessment of Genomic Regions Required for Growth in Mycobacterium tuberculosis. PLoS Pathogens. 2012;8(9):e1002946.

18. Chao MC, Pritchard JR, Zhang YJ, Rubin EJ, Livny J, Davis BM, et al. High-resolution definition of the Vibrio cholerae essential gene set with hidden Markov model–based analyses of transposon-insertion sequencing data. Nucleic Acids Research. 2013;41(19):9033–9048.

19. DeJesus MA, Ioerger TR. A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. BMC Bioinformatics. 2013;14(1).

20. Larivière D, Wickham L, Keiler K, Nekrutenko A. Reproducible and accessible analysis of transposon insertion sequencing in Galaxy for qualitative essentiality analyses. BMC Microbiology. 2021;21(1).

21. Ioerger TR. In: Analysis of Gene Essentiality from TnSeq Data Using Transit. Springer US; 2021. p. 391–421.

22. Kwon YM, Ricke SC, Mandal RK. Transposon sequencing: methods and expanding applications. Applied Microbiology and Biotechnology. 2015;100(1):31–43.

23. Zhang H, Lu T, Liu S, Yang J, Sun G, Cheng T, et al. Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification. NAR Genomics and Bioinformatics. 2021;3(4).

24. van Opijnen T, Levin HL. Transposon Insertion Sequencing, a Global Measure of Gene Function. Annual Review of Genetics. 2020;54(1):337–365.

25. Lluch-Senar M, Delgado J, Chen W, Lloréns-Rico V, O'Reilly FJ, Wodke JA, et al. Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. Molecular Systems Biology. 2015;11(1).

26. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu, Tn5, and Tn7 transposons. Mobile DNA. 2012;3(1).

27. Mahmutovic A, Abel zur Wiesch P, Abel S. Selection or drift: The population biology underlying transposon insertion sequencing experiments. Computational and Structural Biotechnology Journal. 2020;18:791–804.

28. Kimura S, Hubbard TP, Davis BM, Waldor MK. The Nucleoid Binding Protein H-NS Biases Genome-Wide Transposon Insertion Landscapes. mBio. 2016;7(4).

29. Manna D, Porwollik S, McClelland M, Tan R, Higgins NP. Microarray analysis of Mu transposition in Salmonella enterica, serovar Typhimurium: transposon exclusion by high-density DNA binding proteins. Molecular Microbiology. 2007;66(2):315–328.

30. Burger BT, Imam S, Scarborough MJ, Noguera DR, Donohue TJ. Combining Genome-Scale Experimental and Computational Methods To Identify Essential Genes in Rhodobacter sphaeroides. mSystems. 2017;2(3).

31. DeJesus MA, Ambadipudi C, Baker R, Sassetti C, Ioerger TR. TRANSIT - A Software Tool for Himar1 TnSeq Analysis. PLOS Computational Biology. 2015;11(10):e1004401.

32. Barquist L, Mayho M, Cummins C, Cain AK, Boinett CJ, Page AJ, et al. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. Bioinformatics. 2016;32(7):1109–1111.

33. Nlebedim VU, Chaudhuri RR, Walters K. Probabilistic identification of bacterial essential genes via insertion density using TraDIS data with Tn5 libraries. Bioinformatics. 2021;37(23):4343–4349.

34. A Ghomi F, Jung JJ, Langridge GC, Cain AK, Boinett CJ, Abd El Ghany M, et al. High-throughput transposon mutagenesis in the family Enterobacteriaceae reveals core essential genes and rapid turnover of essentiality. mBio. 2024;15(10).

35. Zhang C, Phillips APR, Wipfler RL, Olsen GJ, Whitaker RJ. The essential genome of the crenarchaeal model Sulfolobus islandicus. Nature Communications. 2018;9(1).

36. Jana B, Cain AK, Doerrler WT, Boinett CJ, Fookes MC, Parkhill J, et al. The secondary resistome of multidrug-resistant Klebsiella pneumoniae. Scientific Reports. 2017;7(1).

37. Mandal RK, Kwon YM. Global Screening of Salmonella enterica Serovar Typhimurium Genes for Desiccation Survival. Frontiers in Microbiology. 2017;8.

38. Sarmiento F, Mrázek J, Whitman WB. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon Methanococcus maripaludis. Proceedings of the National Academy of Sciences. 2013;110(12):4726–4731.

39. Dunn OJ. Multiple Comparisons among Means. Journal of the American Statistical Association. 1961;56(293):52–64.

40. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. 1979;6(2):65–70.

41. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.

42. Goodall ECA, Azevedo Antunes C, Möller J, Sangal V, Torres VVL, Gray J, et al. A multiomic approach to defining the essential genome of the globally important pathogen Corynebacterium diphtheriae. PLOS Genetics. 2023;19(4):e1010737.

43. Meinshausen N, Bühlmann P. Stability Selection. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2010;72(4):417–473.

44. Liu H, Roeder K, Wasserman LA. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. Advances in neural information processing systems. 2010;24 (2):1432–1440.

45. Müller CL, Bonneau R, Kurtz Z. Generalized Stability Approach for Regularized Graphical Models; 2016.

46. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining. 2023;16(1).

47. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1).

48. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975;405(2):442–451.

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

50. Ma Y, Pirolo M, Jana B, Mebus VH, Guardabassi L. The intrinsic macrolide resistome of Escherichia coli. Antimicrobial Agents and Chemotherapy. 2024;68(8).

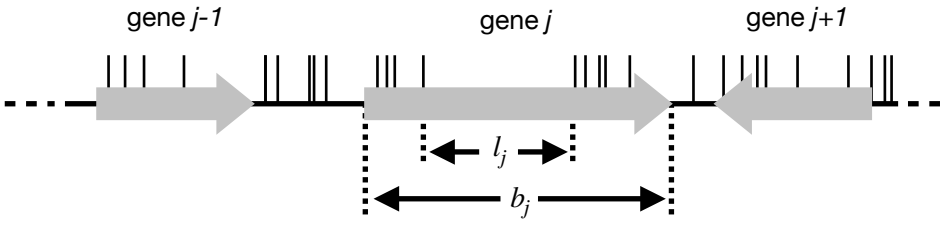
51. Yamazaki Y, Niki H, Kato, Jun-ichi. Profiling of Escherichia coli Chromosome Database. In: Osterman AL, Gerdes SY, editors. Microbial Gene Essentiality: Protocols and Bioinformatics. Totowa, NJ: Humana Press; 2008. p. 385–389.

52. Porwollik S, Santiviago CA, Cheng P, Long F, Desai P, Fredlund J, et al. Defined Single-Gene and Multi-Gene Deletion Mutant Collections in Salmonella enterica sv Typhimurium. PLoS ONE. 2014;9(7):e99820.

53. Bai J, Dai Y, Farinha A, Tang AY, Syal S, Vargas-Cuebas G, et al. Essential Gene Analysis in Acinetobacter baumannii by High-Density Transposon Mutagenesis and CRISPR Interference. Journal of Bacteriology. 2021;203(12).

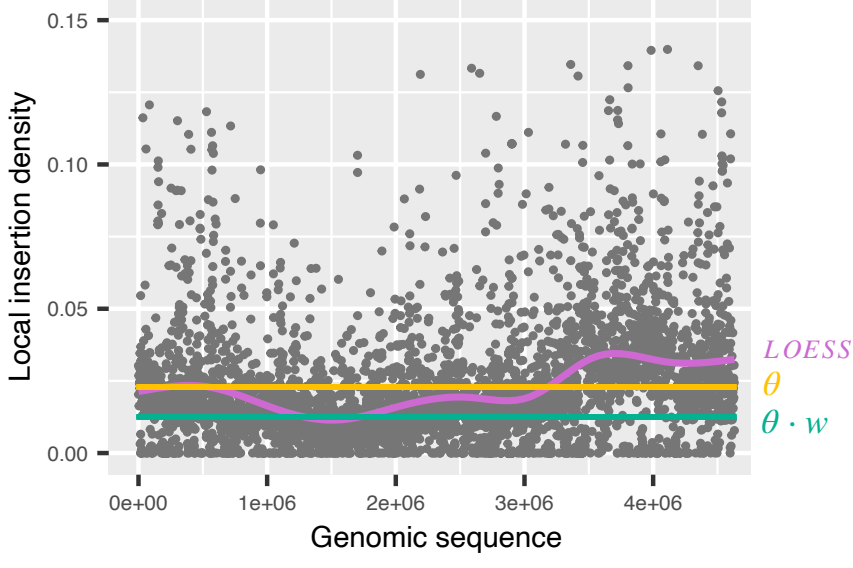
A

Determination of longest insertion-free sequence for gene *j*



B

Determination of the insertion density θ and weight value w



C

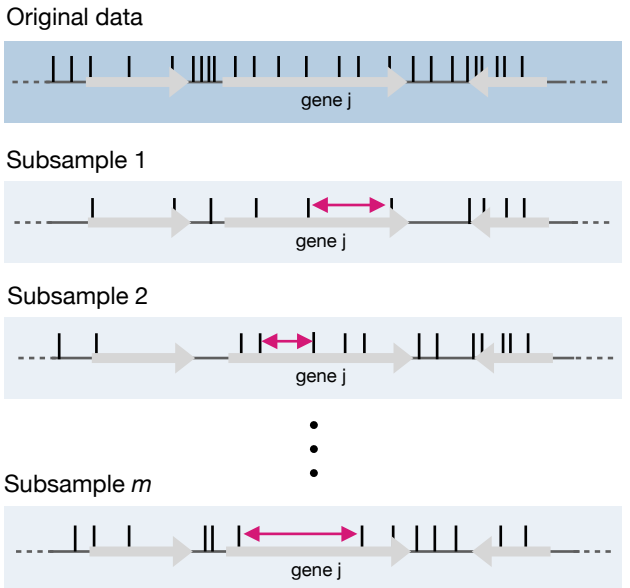
ConNIS with estimated number of insertion sites and weight

$$C\left(l_j, b_j, \hat{h}_j, w\right) := 1 - \sum_{i=1}^{l_j-1} \frac{\binom{b_j - i - 1}{b_j - \hat{h}_j w - i}}{\binom{b_j - 1}{b_j - \hat{h}_j w - 1}}$$

with $\hat{h}_j = b_j \cdot \theta$ and $0 < w \leq 1$

A

Drawing of insertion site subsamples



Subsample 1

gene	l	b	$\hat{h}$
...	...	...	...
j	500	1(	
...	...		

Subsample 2

gene	l	b	$\hat{h}$
...	...	...	...
j	300	1000	30
...	...	...	

Subsample m

gene	l	b	$\hat{h}$
...	...	...	...
j	800	1000	30
...	...	...	...

B

Instability values based on subsamples

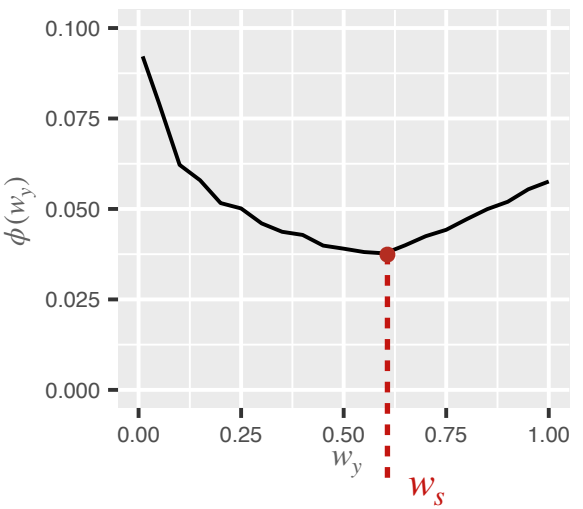
$$\hat{\pi}_j(w_y) = \frac{1}{m}\sum_{d=1}^{m} \mathbf{1}\left[C\left(l_j^{(d)}, b_j, \hat{h}_j, w_y\right) \leq \alpha\right]$$

$$\phi(w_y) := \sum_{j=1}^{p} \frac{\hat{\pi}_j(w_y)\cdot\left(1-\hat{\pi}_j(w_y)\right)}{q(w_y)}$$

for $w_y \in \mathbf{w}$ with $\mathbf{w} = \{w_1, w_2, \ldots, w_z\}$

C

Determination of $w_s \in \mathbf{w}$

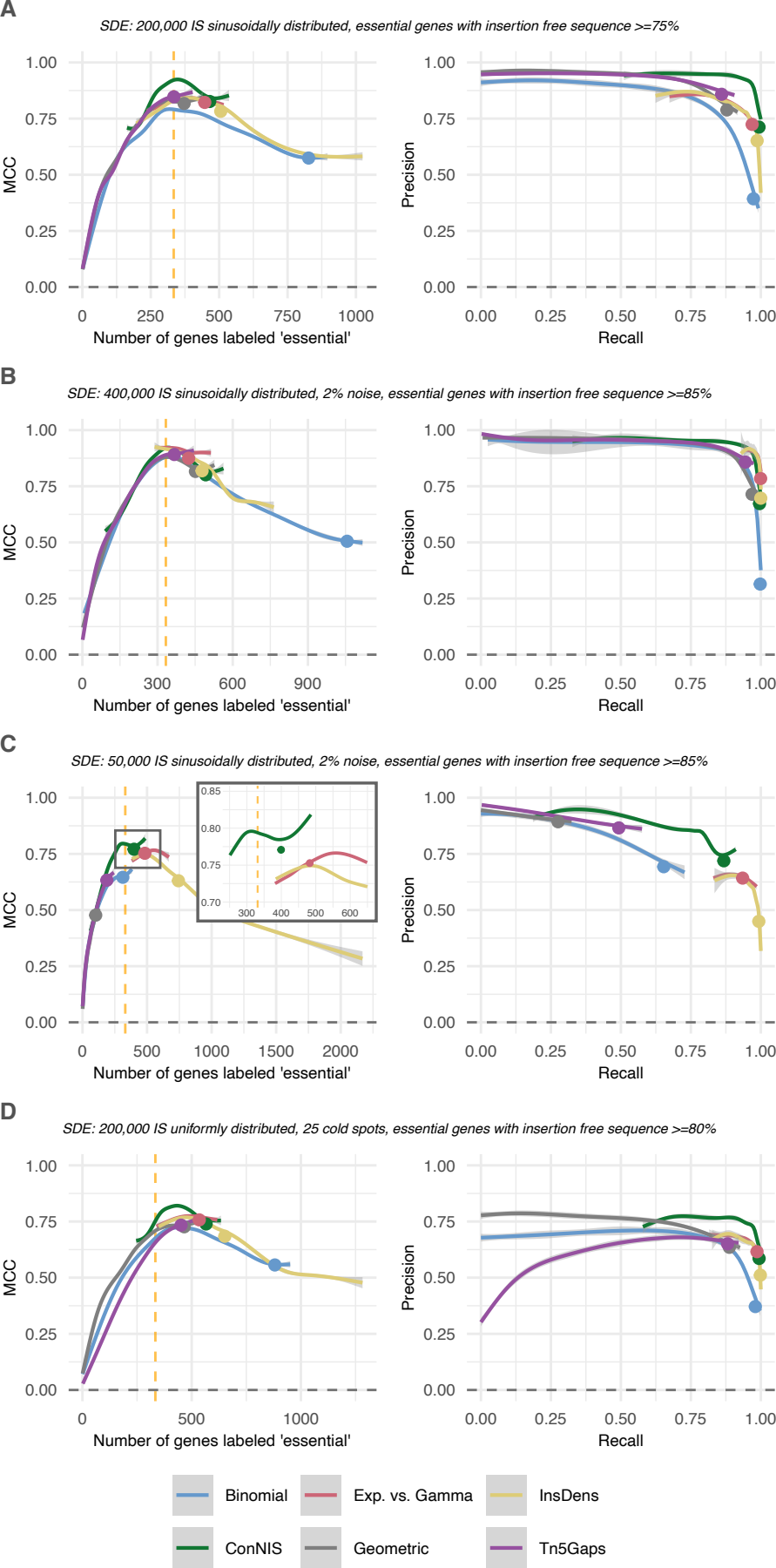


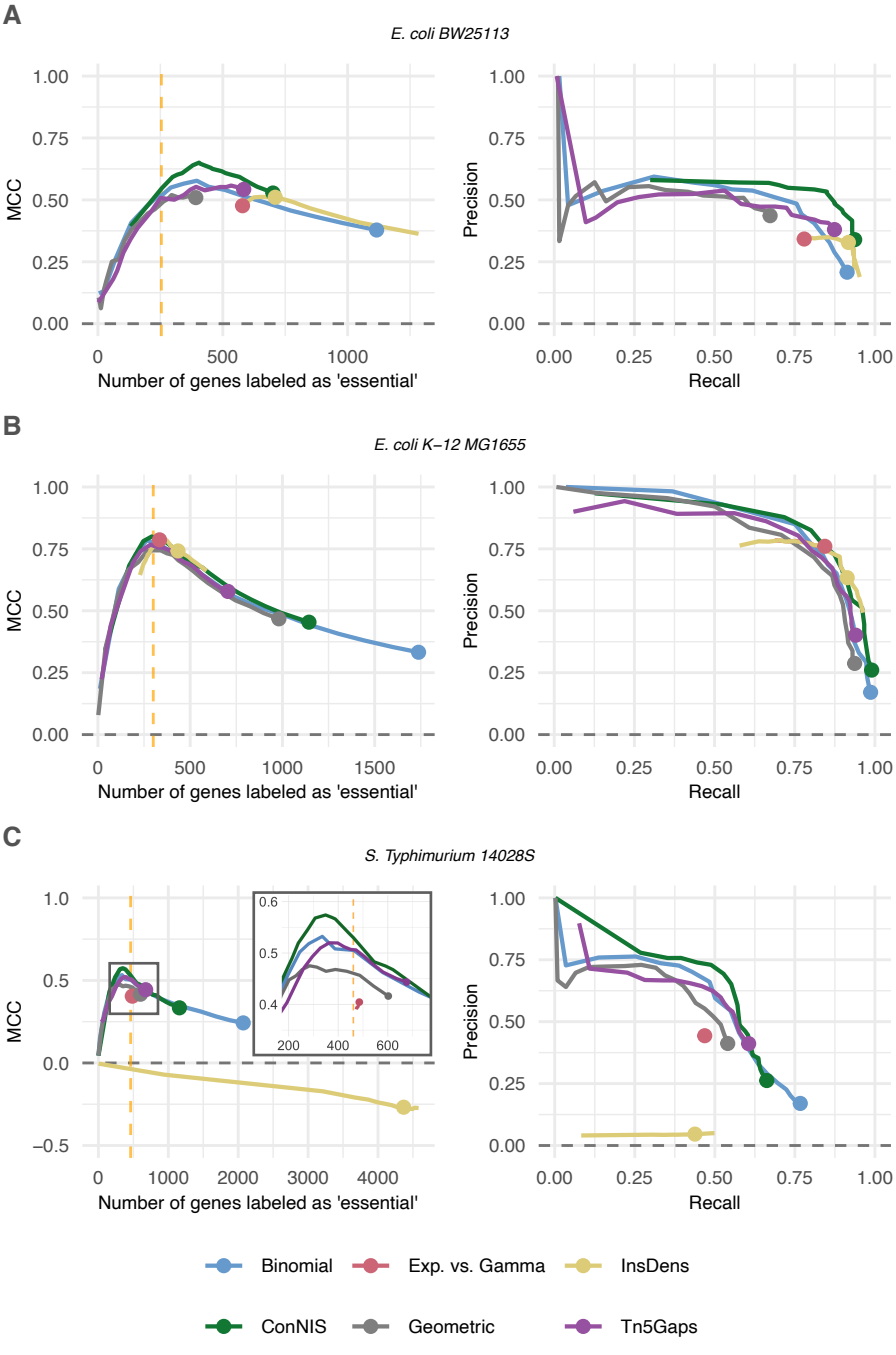
D

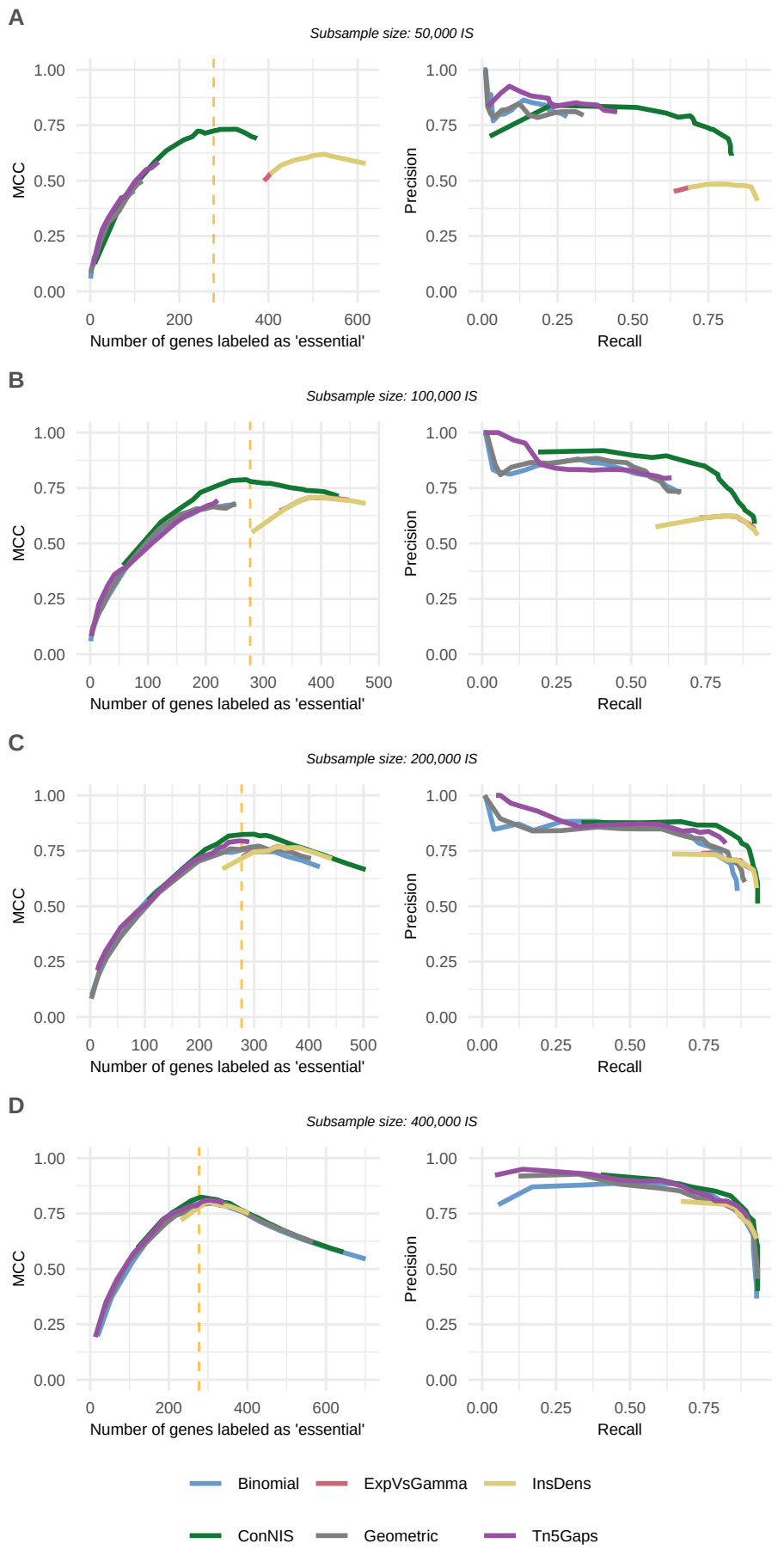
Application of ConNIS with selected weight

$$C\left(l_j, b_j, \hat{h}_j, w_s\right) \quad \text{for } j = 1,2,\ldots,p$$

gene	probability gap l	adjusted α	putative essential
..	...	...	...
j-1	0.0045	0.00005	FALSE
j	0.000007	0.00005	TRUE
...	...		...


A
SDE: 200,000 IS sinusoidally distributed, essential genes with insertion free sequence >=75%
MCC
Number of genes labeled 'essential'
Precision
Recall
B
SDE: 400,000 IS sinusoidally distributed, 2% noise, essential genes with insertion free sequence >=85%
MCC
Number of genes labeled 'essential'
Precision
Recall
C
SDE: 50,000 IS sinusoidally distributed, 2% noise, essential genes with insertion free sequence >=85%
MCC
Number of genes labeled 'essential'
Precision
Recall
D
SDE: 200,000 IS uniformly distributed, 25 cold spots, essential genes with insertion free sequence >=80%
MCC
Number of genes labeled 'essential'
Precision
Recall
Binomial
Exp. vs. Gamma
InsDens
ConNIS
Geometric
Tn5Gaps


A
E. coli BW25113
MCC
Number of genes labeled as 'essential'
Precision
Recall
B
E. coli K−12 MG1655
MCC
Number of genes labeled as 'essential'
Precision
Recall
C
S. Typhimurium 14028S
MCC
Number of genes labeled as 'essential'
Precision
Recall
Binomial
Exp. vs. Gamma
InsDens
ConNIS
Geometric
Tn5Gaps


A
Subsample size: 50,000 IS
MCC
Number of genes labeled as 'essential'
Precision
Recall
B
Subsample size: 100,000 IS
MCC
Number of genes labeled as 'essential'
Precision
Recall
C
Subsample size: 200,000 IS
MCC
Number of genes labeled as 'essential'
Precision
Recall
D
Subsample size: 400,000 IS
MCC
Number of genes labeled as 'essential'
Precision
Recall
Binomial
ExpVsGamma
InsDens
ConNIS
Geometric
Tn5Gaps