Larivière *et al. BMC Microbiology* (2021) 21:168
https://doi.org/10.1186/s12866-021-02184-4

BMC Microbiology

METHODOLOGY ARTICLE Open Access

Reproducible and accessible analysis of transposon insertion sequencing in Galaxy for qualitative essentiality analyses

Delphine Larivière[1,2], Laura Wickham[1], Kenneth Keiler[1], Anton Nekrutenko[1,2*] and The Galaxy Team[2]

Abstract

Background: Significant progress has been made in advancing and standardizing tools for human genomic and biomedical research. Yet, the field of next-generation sequencing (NGS) analysis for microorganisms (including multiple pathogens) remains fragmented, lacks accessible and reusable tools, is hindered by local computational resource limitations, and does not offer widely accepted standards. One such "problem areas" is the analysis of Transposon Insertion Sequencing (TIS) data. TIS allows probing of almost the entire genome of a microorganism by introducing random insertions of transposon-derived constructs. The impact of the insertions on the survival and growth under specific conditions provides precise information about genes affecting specific phenotypic characteristics. A wide array of tools has been developed to analyze TIS data. Among the variety of options available, it is often difficult to identify which one can provide a reliable and reproducible analysis.

Results: Here we sought to understand the challenges and propose reliable practices for the analysis of TIS experiments. Using data from two recent TIS studies, we have developed a series of workflows that include multiple tools for data de-multiplexing, promoter sequence identification, transposon flank alignment, and read count repartition across the genome. Particular attention was paid to quality control procedures, such as determining the optimal tool parameters for the analysis and removal of contamination.

Conclusions: Our work provides an assessment of the currently available tools for TIS data analysis. It offers ready to use workflows that can be invoked by anyone in the world using our public Galaxy platform (https://usegalaxy.org). To lower the entry barriers, we have also developed interactive tutorials explaining details of TIS data analysis procedures at https://bit.ly/gxy-tis.

Keywords: Transposon Insertion Sequencing, TIS, TraDis, TnSeq, Bacteria, Annotation

Importance

A wide array of tools has been developed to analyze TIS data. Among the variety of options available, it is often difficult to identify which one can provide a reliable and reproducible analysis. Here we sought to understand the challenges and propose reliable practices for the analysis of TIS experiments. Using data from two recent TIS studies, we have developed a series of workflows that include multiple tools for data de-multiplexing, promoter sequence identification, transposon flank alignment, and read count repartition across the genome. Particular attention was paid to quality control procedures, such as determining the optimal tool parameters for the analysis and removal of contamination. Our work democratizes the TIS data analysis by providing open workflows supported by public computational infrastructure.

*Correspondence: anton@nekrut.org
[1]Biochemistry and Molecular Biology Department, Eberly College of Science, The Pennsylvania State University, University Park, Pennsylvania, USA
[2]The Galaxy Project https://galaxyproject.org/

Background

Transposon insertion sequencing (TIS) is based on random integration of transposons throughout a genome. These insertions knock out or alter the expression of genes and functional elements. A TIS library —a population of bacterial cells carrying transposon insertions—is divided into aliquotes subjected to different experimental conditions. Upon completing an experiment, sites flanking transposon insertions are amplified, and amplification products are subjected to high throughput sequencing (HTS). Mapping of resulting sequencing reads against the host genome reveals locations of insertions. Regions containing insertions can tolerate disruptions and thus are *non-essential*, while location void of insertions (or with underrepresented insertions) are likely under purifying selection and are considered/*essential* in the given growth conditions. These conclusions depend on the assumption that transposons' random insertion will impact every gene (library saturation). Methods used for this classification must account for the probability of genes not being impacted by any insertion. The TIS approach led to successful genome-wide identification of essential and non-essential genes in several species [1]. A more comprehensive application of TIS involves the insertion of transposon constructs carrying regulatory elements such as promoters [2]. In addition to binary readout (essential/non-essential), this approach yields information about the effects of up- and down-regulation of specific genes.

Randomly pooled transposons libraries are commonly created with Mariner or Tn5 transposons. Mariner transposons target `TA` dinucleotides. They have wide species specificity and are stable [3]. The methods based on Mariner transposons are referred to as *TnSeq* methods. Tn5 transposons, on the other hand, do not target specific sequence motifs while exhibiting a preference for GC-rich sites [4]. Tn5-based methods are called *TraDis* methods if the reads are sequenced directly after the PCR, or *HITS* if the PCR products are subjected to a size selection and a purification. Most bacteria species have TA sites equally distributed across the genome, but when it is not the case, Mariner transposons provide a biased library. The Tn5 is then a common alternative as building a Mariner-based library may be problematic for these species [1]. The downside is that the larger number of Tn5 insertions produce less saturated mutant libraries. The saturation is defined by the probability of a possible insertion site to be impacted by an insertion. As Tn5 can insert anywhere, the potential insertion sites are more numerous than Mariner transposon sites, decreasing the probability to impact each one with the same quantity of transposons. Another difference between the two types of transposons is the use of a restriction endonuclease in Mariner-based libraries. These enzymes, such as *Mme*I, cut a fixed-length sequence upstream from the insertion site generating reads of equal length. The Tn5-based methods do not use this approach and produce fragments of various lengths, potentially allowing for PCR bias [3].

In the end, each flavor of TIS experiments produces a collection of sequencing reads. Before interpreting TIS data, these reads need to be processed, mapped, filtered, and converted into a form suitable for downstream analysis tools. TIS sequencing reads have complex structures as they include fragments of transposon backbone, primer annealing sites, molecular barcodes, and other elements. For example, only ≈13 base pairs (bp) of a TnSeq read correspond to the genomic region adjacent to the integration site—the portion that is mapped against the host genome—everything else needs to be stripped away before mapping. After read trimming and mapping resulting `BAM` (Binary Alignment Map) datasets need to be filtered and converted into more compact representations such as, for example, `wig` (Wiggle) format. The `BAM` are filtered to remove reads that align outside of insertion sites, when applicable, thus removing reads that are incomplete and do not provide information on transposon insertions. Such *derived* datasets can then be paired with appropriate annotation datasets (representing the location of genes and functional elements across the host genome) and used as inputs to analysis tools.

A number of algorithmic approaches have been developed to facilitate TIS data analysis. These include Hidden Markov Model (HMM)-based methods for identification of essential sites as well as regression analyses utilizing gene saturation or runs of consecutive empty sites. These approaches are implemented in tools such as ESSENTIAL [5], Tn-seq Explorer [6], El-ARTIST [7] suite, TRANSIT [8], or Bio-Tradis [9] (Table 1). The output of these tools—lists of genes classified as essential/non-essential or coordinates of regions enriched or void of insertions—needs to be further processed by, for example, comparing results across conditions. In this paper, we focus on identifying essential genes, and the review of methods for conditional essentiality would require a larger study.

The tools such as TRANSIT and Bio-Tradis offer powerful means for the analysis of TIS data. However, while they cover the essential step of data interpretation, it is just one part of a multi-step process involving, as we demonstrated above, trimming, mapping, filtering, and additional statistical analyses. In this manuscript, we developed a set of comprehensive workflows for the analysis of TIS data that include all analyses step from initial read processing to preparation of final figures for publications, and can be adapted to different datasets. To demonstrate the utility of our approach, we reanalyzed data from two recent studies employing Tn5 and Mariner transposons. Add tools and workflows developed by us are publicly available and can

Table 1 Tools availability and maintenance metrics: frequency of updates, number of contributors, commits, and issues

Tool	Data	Availability	Essentiality	Conditional Essentiality	Release	GitHub	Last Update	Contributors	Commits	Open/Resolved Issues
Bio-Tradis [9]	TraDis TnSeq	Command line Galaxy	✓	✓	2015	sanger-pathogens/ Bio-Tradis	3.2 12-2019	12	336	1/13
El-Artist [7]	TnSeq TraDis	MatLab	✓	✓	2014	–	–	–	–	–
ESSENTIALS [5]	TraDis TnSeq	Web	✓	✓	2012	–	–	1	1	–
Magenta [10]	TraDis TnSeq	Command line Galaxy	–	✓	2017	vanOpijnenLab/ MAGenTA	–	1	28	3/1
TnseqDiff [11]	TraDis TnSeq	R library	–	✓	2017	–	0.1.2 05-2019	–	–	–
Tn-Seq Explorer [6]	TraDis TnS q	Stand-alone tool Graphic interface	✓	–	2015	sina-cb/ Tn-seqExplorer	–	1	172	2/2
TRANSIT [8]	TraDis TnSeq	Command line Galaxy	✓	✓	2015	mad-lab/ transit	3.0.2 12-2019	8	1111	1/13
TSAS [12]	TraDis TnSeq	Command line	✓	✓	2017	srimam/ TSAS	0.3.1 08-2019	1	19	0/0

These metrics are indicators of the health of the tools: number of people maintaining the tools, frequency of update, the responsivity of the developing team. Most of the tools can support both TnSeq and TraDis. The gray words indicate that the tools support the data in theory but will necessitate adapting the data. Across this list of tools, Bio-Tradis and TRANSIT seem well maintained. The other tools with git hub repositories have only one contributor, which means that the tool could stop being maintained if the one contributor change project. Among these 3, 2 have not been updated since the first release some years ago. Three tools have no Github directory, although ESSENTIAL has an svn repository. The svn logs showed a single commit in 2012

be run as-is to reproduce this study as described at https://bit.ly/gxy-tis.

Results

Our goal was to design a publicly accessible high throughput system for transparent and reproducible analysis of transposon insertion data. To devise and test our approach, we selected data from two recent studies: one performed in *Escherichia coli* [13] and another conducted in *Staphylococcus aureus* [14]. The first study [13] used the TraDIS approach [15] based on a Tn5 transposon inserting with high-frequency into arbitrary sites within the target genome [16]. The second study [14] used the TnSeq approach based on a phage-assisted Mariner-derived system that targets `TA`-sites within the host's genome [2].

Analysis of TraDIS data

Goodall et al. [13] used TraDIS to identify essential genes in *Escherichia coli* BW25113. TraDIS technique generates reads that contain experimental barcodes and segments of the transposon backbone in addition to the fragment of genomic DNA proximal to the insertion site. Before the reads can be used for mapping, required to identify locations of insertion sites, they need to be trimmed down to include only genomic DNA adjacent to the insertion site. In the case of the Goodall et al. study[13] this has been done prior to submitting the sequencing data to the Short Read Archive (SRA - BioProject PRJEB24436), and thus the reads can be used directly for mapping without any preprocessing. We mapped the reads against the *E. coli* BW25113 genome (CP009273.1) and computed read coverage (see Methods) using the 5'-end of the reads as it is immediately adjacent to the insertion site [15].

Regression on genes saturation indexes

To identify essential genes, we proceeded to carefully re-implement the analysis performed by Goodall et al. [13]. These authors conducted a regression analysis on gene saturation indices by fitting known distributions to the distribution of insertion indexes. First, we computed gene saturation index S—, a simple statistic calculated by dividing the number of insertions within a coding region (CDS) by its length. In this dataset, S is bimodally distributed with the first mode at low saturation and the second at high saturation (Fig. 1). This profile is coherent with the expected distribution of gene saturation in TIS studies with saturated libraries [1]. It is a mixture of two distinct distributions: one of essential genes and the other of non-essential genes.

To separate two distributions, we performed a regression by fitting a bimodal distribution to our data. The bimodal distribution is composed of an exponential and a gamma distribution components corresponding to essential and non-essential genes, respectively. Using distribution parameters, we then computed a probability of every gene being drawn from exponential or gamma distributions. A gene is said to be *essential* if the probability of it belonging to the essential distribution is X times higher than the probability of belonging to the non-essential distribution (and vice versa). X is a threshold that can vary between studies, and genes that do not meet this requirement are classified as "undetermined". Goodall et al. [13] used $X = 12$ as it was previously employed in another *E. coli* gene essentiality study by Phan et al.[17]. Using $X = 12$ we identified 364 essential genes (Additional file 1: Table S1). We then compared our results with the list of essential genes identified by Goodall et al.[13] as well

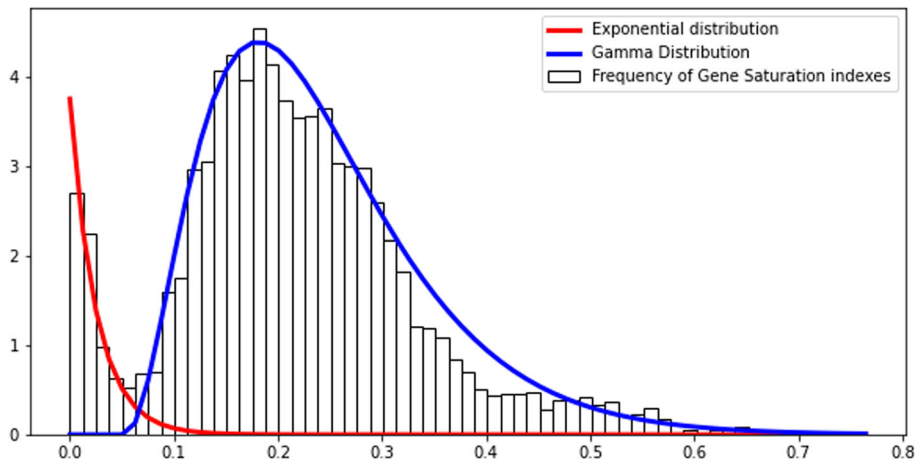

Fig. 1 Regression analysis on gene saturation indexes for TraDIS dataset. The histogram represents the frequency of saturation indexes of genes. A gene saturation index is the ratio of insertion sites with insertions on the number of potential insertion sites. The regression uses the density of frequencies, as we are using density functions to fit the data. A parametric regression algorithm is used to divide the bimodal distribution into two distributions. An exponential distribution is fit to essential genes, and a Gamma distribution to non-essential genes. Although the Gamma distribution deviated from observation at higher saturations, the two distributions fit the data nicely in the saturation range where the division happens

as those from Keio (based on BW25113; [18]) and PEC (based on MG1655; [19]) databases (Fig. 2). It showed that several essential genes predicted by us and absent from the Goodall et al. are found on the other databases.

To evaluate the effect of the threshold choice's impact, we compared results obtained with 4-fold ($X = 4$) and 12-fold differences ($X = 12$) used above. There were little differences between the two thresholds for the prediction of essential genes. (Additional file 1: Table S1). With $X = 4$ we missed eight essential genes and overpredicted 24 genes compared to the original study. $X = 12$ misses 12 essential genes and overpredicts 23. The variations in prediction from the two studies could be explained by differences in curve fitting that could be introduced by a different regression tool. The parameters of the bimodal distribution have not been specified by Goodall et al. [13] for comparison. Another reason for the different results obtained could be due to different mapping parameters. The differences obtained when trying to replicate an analysis highlight the importance of open and reproducible science. Publishing exact data and workflows allows reducing or tracking the variability of results.

Automatic regression with bio-tradis

Next, instead of performing manual fitting, we employed Bio-Tradis [9] toolkit. It identified 398 essential genes and classified the other gene as "ambiguous", failing to detect any non-essential genes. We compared the essentiality prediction of Bio-Tradis with the previous regression analysis (Fig. 3C). Among the 353 essential genes listed by [13], 351 were also identified by Bio-Tradis. It predicted 47 additional essential genes, among which 21 are also identified by the hand-fit regression. The results were very similar for gene essentiality, whether we used the automated or hand-fit method.

Classification based on rows of empty sites

In addition to a regression analysis performed to identify essential genes, Goodall et al. analyzed the density of transposon insertions in the genome and describe the probability of observing consecutive potential insertion sites void of transposon insertions [13]. This metric can be used to detect essential genes and regions as well, and such a method is implemented in the TRANSIT suite [8] with the *Tn5Gaps* tool. Comparison of *Tn5Gaps* results to the Goodall et al. predictions (Fig. 4) showed that TRANSIT predicted 124 additional essential genes, and 331 predicted essential genes were shared with the published results. We did not find evidence that the genes predicted by Transit and not by the original study were true positives. Some of the overpredicted essential genes showed some partial depletions of insertions, which could indicate a growth defect induced by the gene disruption. For other genes, the reason for their classification as essential was unclear. The list of predicted non-essential genes generated by TRANSIT, on the other end, was very close to the Goodall et al. results.

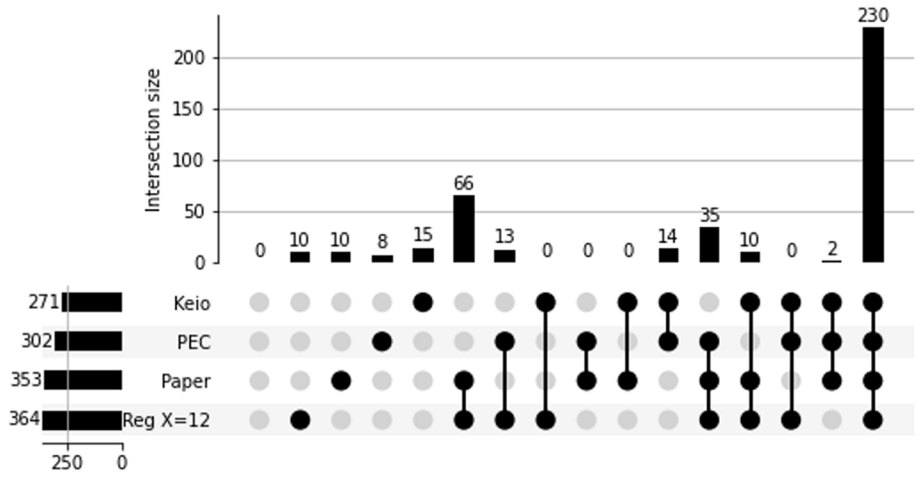

Fig. 2 TraDIS - Upset plot comparing the regression results to Goodall et al. paper, Keio [18] and PEC [19] databases. We compared the list of essential genes resulting from our reproduction of the Goodall and al. study with their results and those identified by two external databases Keio and PEC. We can see that our replication predicts slightly more essential genes than the manually curated results provides, 364 genes against 353. Out of the 232 core essential genes identified in the study, we successfully identified 230. We identified most of the genes predicted by the paper that are absent in PEC or Keio (101 genes). Our replication identifies 13 essential genes identified in PEC but not in the original study in addition to the 35 genes identified everywhere but in Keio. The genes predicted by only one of the sources are of similar magnitude, a dozen genes each. They can be explained by manual annotation, parameter change, and domain essential genes

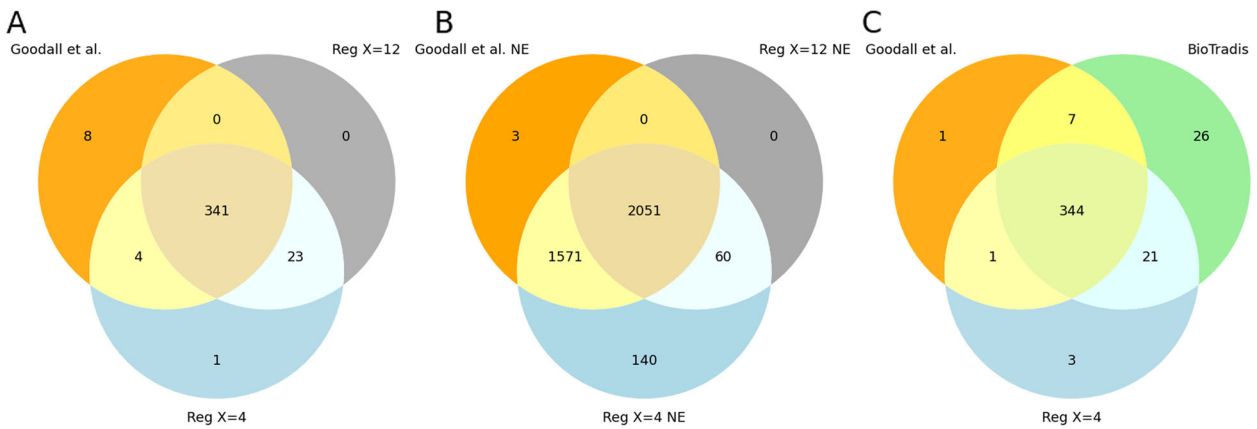

Fig. 3 TraDIS - Comparison of methods for genes classification by regression analysis. **a** UpSet plot comparing the essential genes predicted by each threshold with the curated list. There is little difference between the two thresholds for essential gene prediction. 341 essential genes are identified in the three sets. Using a threshold of 4 predicts five additional genes compared to a threshold of 12, four of which are also identified in the curated list. **b** UpSet plot comparing the non-essential genes predicted by each threshold with the curated list. When looking at the non-essential gene list, we see that using a threshold of 12 misses 1570 non-essential genes identified in the paper and a regression using a threshold of 4. The latest predicts 140 non-essential genes that are unclassified in the two other sets. **c** UpSet plot comparing the non-essential genes predicted by using hand-fit curves, automated fit with Bio-TraDis, and the curated list. The two regression methods identify 344 of the essential genes predicted by the publication. One gene is missed by the two methods, and eight are identified by only one of the regression analyses

Comparison of the regression and Tn5Gaps results

Finally, we compared the regression results, using the two thresholds and the Tn5Gaps method implemented in TRANSIT (Fig. 5). We could observe that TRANSIT over-predicted essential genes (Fig. 5A). Among the 124 genes identified by TRANSIT but not by the Goodall et al. study, 108 were identified by TRANSIT alone. Some of the over-predicted genes were domain essential genes (only part of the gene is free of insertions). Most of them, however, did not seem to contain essential regions but showed sparse insertions. When looking at non-essential genes predictions, both methods were very close to the paper (Fig. 5B). The regression predictions were overall closer to the manually curated results than TRANSIT regardless of the chosen threshold.

Read count normalization

The transposon insertion sequencing datasets can be normalized for three factors: (i) positional read bias, (ii) differences in sequencing depth, (iii) stochastic differences

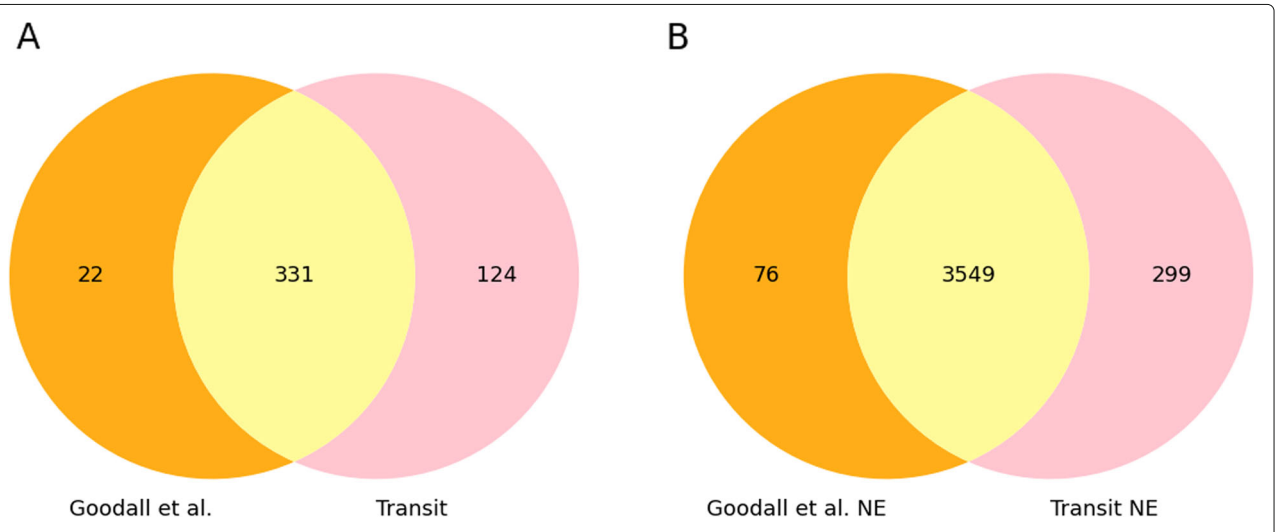

Fig. 4 TraDIS - Comparison of genes predicted by the Gumbel tool in Transit and those published in the paper. **a** Essential genes. Transit seems to largely over-predict essential genes. 124 genes are identified as essential in addition to the 331 genes also identified as essential by manual curation. **b** Non-essential genes. Transit prediction of non-essential genes overlaps well with the genes predicted as non-essential in the paper

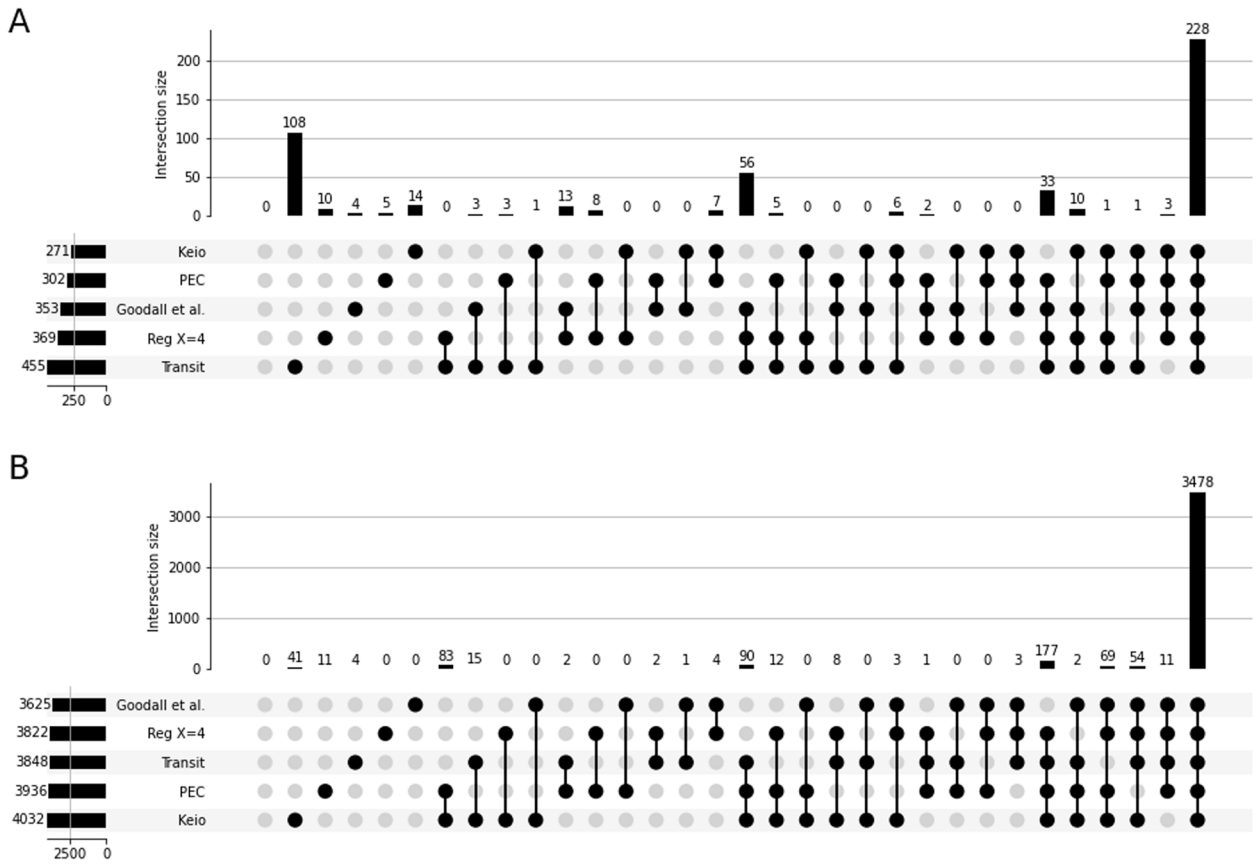

Fig. 5 TraDIS - Comparison of regression results, Transit results, external databases, and the curated list of essential genes. **a** Comparison of essential genes predictions. We can see that Transit over-predicts the number of essential genes compared to databases as well. 108 are classified as such that does not appear in any other source. **b** Comparison of non-essential genes predictions. All sources have similar results for non-essential gene identification. 3478 genes are always identified. The second-largest set of genes are genes identified as unclassified by the manual curation only. These could correspond to domain-essential genes or genes that impact growth

in library diversity [1]. To evaluate the impact of normalization on the essentiality prediction, we compared the results obtained without normalization and using one of TRANSIT toolset's normalization method. In particular, we used the "TTR" (Trimmed Total Reads) method recommended by the authors to remove outliers and normalize for the differences in saturation cited above [20]. We also compared the results obtained by ignoring reads covering gene extremities, as they may not disrupt the gene function. The results were identical regardless of the normalization choices for all three analyses and both essential and non-essential genes.

Analysis of TnSeq data

We applied the analysis approach developed on TraDIS data to a Mariner-based TnSeq dataset produced by Santiago et al. [14]. The methodology pioneered by these authors allows transposon-mediated insertion of promoters into the target genome [2]. As a result, this technique provides two types of readout. First, similarly to TraDIS, the lack of TnSeq reads mapping to a genomic locus indicates its functional importance. This readout—lack of reads—can be used for the identification of essential genes. Second, promoters contained within insertion constructs may affect neighboring genes. Because promoters act directionally, sequencing reads derived from these insertions exhibit strand bias. The second type of readout—regions where the majority of reads map to one of the two strands—allows finding genes whose expression change is beneficial given an experimental condition. In this study, we focused on the first type of TnSeq readout to identify essential genes. The data produced by Santiago et al. [14] contain raw reads (SRA BioProject PRJNA417822) for 82 experimental conditions with a varying number of replicates. We used the control condition (containing 14 replicates) to develop reproducible strategies for read preprocessing and control for noise in the data.

TnSeq data preprocessing

Sequencing reads produced by Santiago et al. [14] contain transposon backbone and auxiliary sequences that need to be removed prior to analysis (Fig. 6A). In addition, the reads contain molecular barcodes that identify constructs containing different promoters. The genomic portion of the reads, ultimately mapped against the host genome, is only 16-17 bp in length because TnSeq protocol uses *Mme*I restriction endonuclease. *Mme*I cleaves DNA 18 nucleotides downstream of the recognition site. After mapping against the *Staphylococcus aureus* (CP000253.1) we computed coverage at 3'-end of the reads only, as this corresponds to the position of the insertion site (Fig. 6B). Finally, we compared the coverage information with the position of all TA dinucleotides found in the *Staphylococcus aureus* genome.

Regression on the TnSeq dataset

The regression is the method that gave us the best results with TraDIS data. However, TnSeq data distribution of the saturation index S is different for TnSeq data (Fig. 7). To fit the profile of the TnSeq data, we used different distributions to perform the regression. Here we used exponential distribution for the essential genes and normal distribution for the non-essential genes. When correcting the method to use appropriate distribution, we could compare the different thresholds' efficiency (Fig. 8A). The threshold choice had more impact on the results of TnSeq data compared to TraDis data: the use of $X = 4$ as a threshold leads to the prediction of 480 genes, which was higher than the number of essential genes expected, and more than predicted with the use of a threshold of $X = 12$ (Additional file 1: Table S2). When compared with the results of the publication [2], most of the essential genes and half of the domain essential genes were identified by the regression analyses (Fig. 8). A threshold of $X = 12$ produced more false-negative results than $X = 4$ and as many false positives. The prediction of non-essential genes was very close to the published results when using either threshold (Fig. 8B). Overall, the choice of the threshold in this analysis appeared to change the essentiality prediction significantly. A threshold of 12 appeared too stringent in that case, perhaps due to regression curves fitting less closely to the actual distribution than TraDIS data.

Automatic regression with bio-tradis on the TnSeq dataset

We compared our regression analysis results to the automated regression performed with the Bio-Tradis toolkit (Fig. 8C and D). Bio-Tradis predicts a larger number of essential genes than Santiago et al. [2]. Our regression analysis had a better performance than the tool for the prediction of essential genes for this dataset. The tool had a slightly better rate of true-positive than the hand-fit regression but a worse false-positive rate. When we looked at the results of non-essential genes predictions, the two methods performed equally. These results are not unexpected, as Bio-Tradis has been developed for Tn5-based methods.

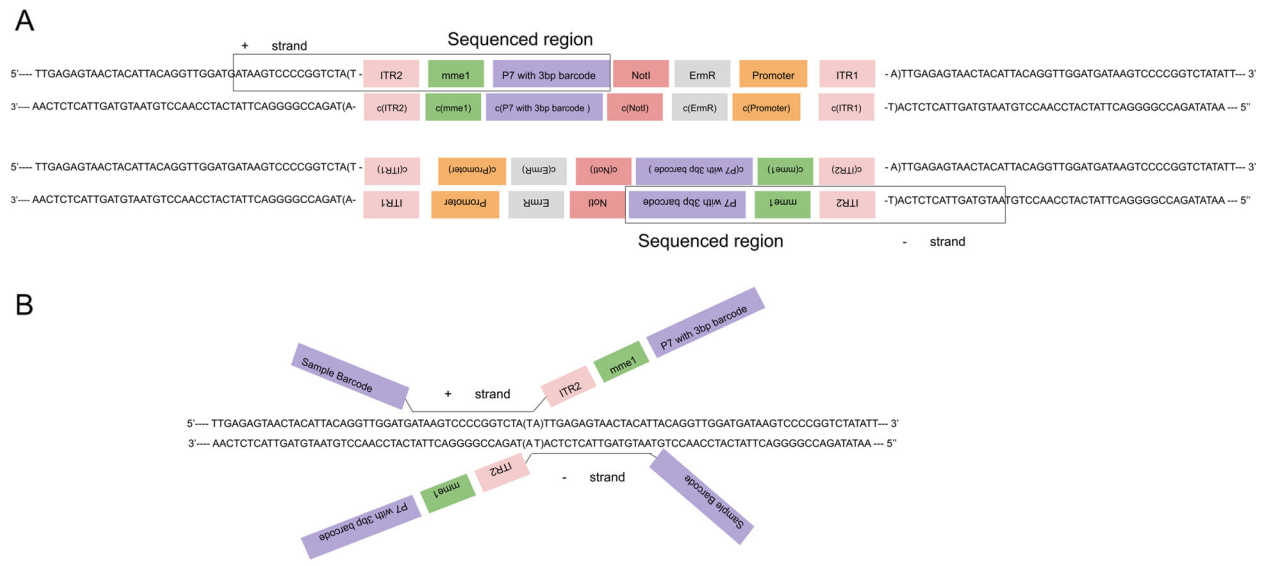

Fig. 6 TnSeq - Structure of transposon insertions and reads in Santiago et al. [2] technique. **a** Structure of a transposon construct insertion. The transposon can insert in two directions since a TA site is identical in both strands. The Mmel insertion in the transposon allows reads of uniform size by cutting DNA 20 bases upstream of the transposon. **b** Structure of the reads. In addition to the fragment of 20bp upstream of the Mmel site, the reads contain an Illumina adapter P7 containing a 3bp barcode specific to the type of construct and an Illumina adapter P5 containing a sample barcode at the 5' end. The read also contains the Mmel transcription site included in the ITR region. We can see from that structure that the insertion site's position is aligned with the 3' end of the read

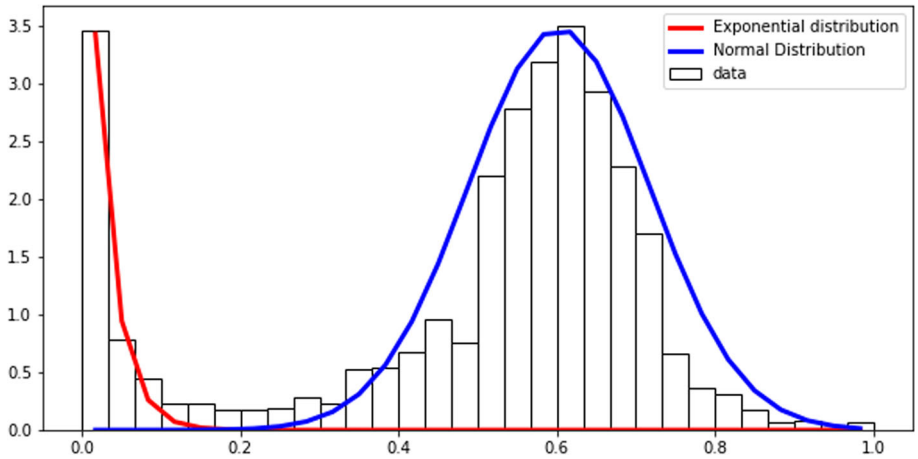

Fig. 7 TnSeq - Regression on the frequency of gene saturation indexes for HiMar transposon and S. aureus. The Exponential distribution and the Normal distribution seem to fit the general shape of the distribution, despite some lose fit around the quartiles of the normal distribution. There is no intersection between the two distributions; however, some genes have a null probability of belonging to either population

Classification of genes based on rows of empty sites on the TnSeq dataset

Using the same method as Tn5gaps, the Gumbel tool in the TRANSIT suite is designed for gene essentiality prediction based on Mariner transposons. As with the regression method, we did not differentiate essential or domain essential genes. Therefore, we compared our results both with the essential genes and essential and domain essential genes identified by the essentiality study with TnSeq data [2] (Fig. 8E). TRANSIT and the regression analysis identified half of the full essential genes, about a third of domain-essential genes. Overall, TRANSIT performs more poorly than other methods. It could be due to the higher sensitivity of TRANSIT to gene sizes compared to saturation indexes. If a gene is too small compared to regions void of insertions, then TRANSIT will not identify an essential gene. Many of the domain essential genes were not identified by any method. The identification of domain essential genes by TRANSIT or the regression method is dependent on the size of the essential domain. Too small, and it would not appear as significant for TRANSIT, and would not impact the saturation index enough to be classified into the essential category.

When looking at the non-essential genes, we could see that all the methods provided very similar results (Fig. 8F). It seemed to indicate that this variation in condition impacted only the essential genes predictions. Some genes may have been going from growth defect to entirely essential, depending on the conditions. It would have impacted their classification into essential genes, going from low insertion rate to no insertions, without them having sufficient saturation index to be classified as non-essentials.

Prediction of essential genes using HMM on TnSeq data

TnSeq analyses often use a third method in addition to the one used in the TraDIS. Hidden Markov Model (HMM) methods consider each potential insertion site independently of the previous states to predict the state of the next position and attribute costs of changing from one state to another. This method allows predicting regions with the same state. It is a standard method available in several tool suits, like EL-ARTIST, TRANSIT, ESSENTIALS, Tn-seq explorer, and MAGENTA (See methods). Our goal was to identify a tool that facilitate reproducibility and transparency of analyses. For these reason we chose TRANSIT suite to perform an HMM analysis of the TnSeq Data—it has a permissible open source license and is actively maintained (Table 1). This tool uses HMM to classify each site into four states: Essential, Non-Essential, Growth Defect, and Growth Advantage. This method returned about two thousand essential genes, which is more than the expected 350 to 400 genes. This result could be explained by the fact that Santiago data may be too sparse for the HMM, which is sensitive to zero-inflated datasets. This method could be suited for a more saturated library, with few random empty insertion sites.

Comparison of the different methods for essentiality prediction on TnSeq data

We compared the results of the regression, using a threshold of 4, and the Gumbel with the original study, and other similar essentiality studies (Fig. 9). We did not include the HMM analysis, as the results were inconclusive. We compared our predictions to two other published TnSeq studies: Chaudhuri et al. [21], and Valentino et al. [22] (Fig. 9). While the Valentino et al. [22] study used the

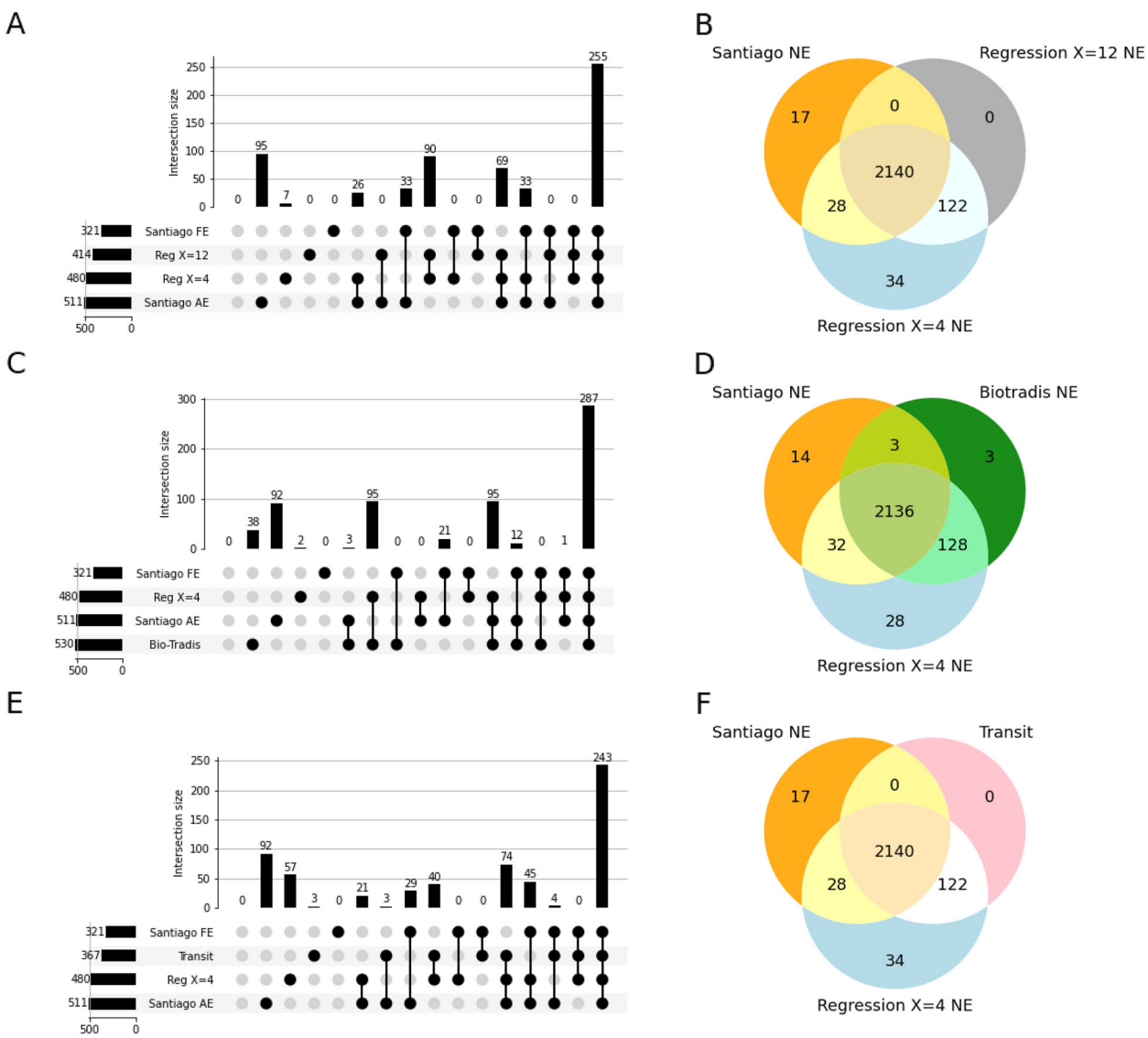

Fig. 8 TnSeq - Comparison of the essential genes prediction per method. Santiago FE refers to the genes that are identified as fully essential in the Santiago et Al. paper. Santiago AE refers to both fully essential and domain-essential genes. The classification of the regression results has been performed with thresholds of 4 and 12. **a** Comparison of regression on gene saturation index with the genes identified as essential in Santiago et al. [2]. 255 genes are shared between the regressions and the paper full essential genes list. **b** Comparison of regression on gene saturation index with the genes identified as non-essential in Santiago et al. The methods do not differ much on non-essential genes predictions. 2071 genes are identified by the paper and the regression, regardless of the chosen threshold. **c** Comparison of Bio-Tradis prediction of essential genes with the hand-fit regression and paper predictions. 287 genes are common to the three methods, and Bio-Tradis overpredicts 32 genes not identified by other methods. **d** Comparison of Bio-Tradis prediction of non-essential genes with the hand-fit regression and paper predictions. The results are very similar. 197 genes are predicted by both regressions that are not classified as non-essential by the study. **e** Comparison of Transit results, regression results, and genes identified as essential in Santiago et al. We are using the hand-fit regression results using a threshold of 4 to be more relaxed, as the threshold of 12 provides more stringent results and Bio-Tradis overpredicts essential genes. **f** Comparison of Transit results, regression results, and genes identified as non-essential in Santiago et al. The three methods are in accordance with the non-essential genes predictions. 1923 are common across all results

same strain of *S. aureus* (HG003), Chaudhuri et al. [21] used a different one(SH1000). Both strains derive from the strain NCTC8325. Chaudhuri et al. [21] performed an automatic detection of genes void of insertions and completed with a manual detection of genes with depleted insertions. Valentino et al. [22] identified genes as essential if the insertions within the genes constitute less than 1% of total reads. We could observe a core of 229 genes reported as essential across all studies. Overall, TRANSIT and the regression analysis results were coherent

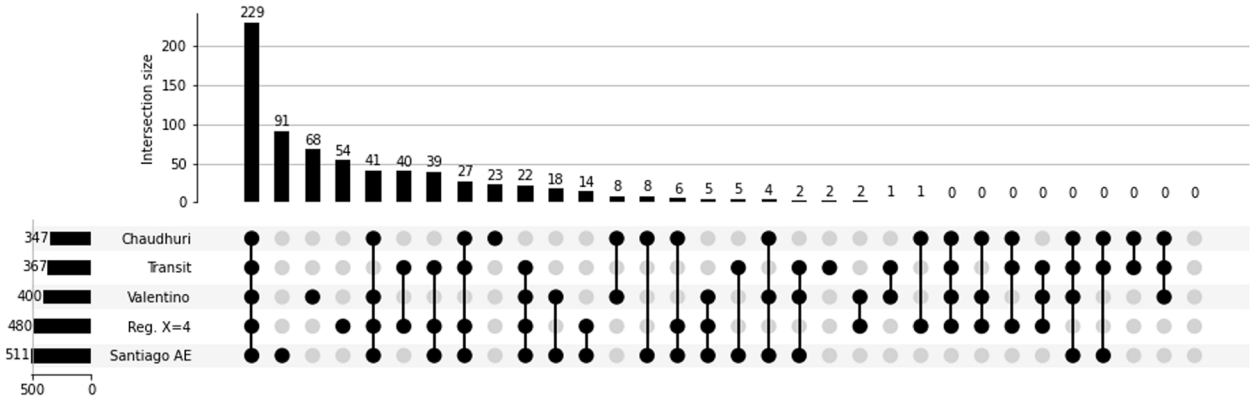

Fig. 9 TnSeq - upSet plot of the essential genes predicted by the different methods, the genes published by the reference paper [2], and other studies of gene essentiality in S. aureus. 229 genes are identified as a "core" of essential genes between studies. As Santiago et al. paper differentiates domain essential and full essential genes, it shows the largest difference compared to the other studies with 89 genes not found in other sets of essential genes. Transit is the closest to all other studies, with only two extra genes predicted, against 54 predicted by the regression analysis

with the results expected from essentiality analyses compared to other studies performed in different conditions. TRANSIT prediction was closer to the other studies, with only two reported essential genes not predicted by other methods. The regression method identified 54 additional essential genes compared to other studies. Manual inspection of these genes shows a depletion in insertion, indicating a growth depletion when the gene is disrupted. Since the genes still show regular insertions, they are not considered essential. The method used in the Santiago analysis [2], with EL-ARTIST, classify genes into three categories: Essential, domain-essential, and non-essential genes. The methods used in this study classify them only into non-essential or essential genes. We expect to observe four possible outcomes for domain-essential genes. i) The essential domain is small and does not decrease the gene saturation enough to cross the threshold between essential and non-essential genes. In that case, the regression analysis will classify the gene as non-essential. If the domain is also smaller than the threshold of detection of TRANSIT Gumbel, The gene is predicted as non-essential by TRANSIT as well. ii) The essential domain is long enough to be detected by TRANSIT but too small compared to the gene's size to be detected by the regression method. iii) The domain is large compared to the gene but smaller than the detection threshold of TRANSIT. In that case, the gene is predicted as essential by the regression analysis but not by TRANSIT. iv) The essential domain is large enough compared to the rest of the gene and compared to TRANSIT's detection threshold to be detected by both methods. The additional 89 predicted in the Santiago analysis could then be explained by its distinction between essential and domain essential genes. Some domain essential will be predicted as essential, others as non-essential, so there will not be a perfect fit for either class. For this TnSeq dataset, TRANSIT performed better than the other methods.

Implementation in Galaxy

We implemented the analysis workflows in Galaxy [23], including preprocessing, alignment, read counts, and essentiality prediction. We also developed and deployed a comprehensive tutorial explaining the application of our system to TIS data [24]. Despite being very similar, the two types of data, TnSeq and TraDIS, require distinct workflows due to the difference in read structure and coverage analysis. We constructed these workflows and made them available in the Galaxy main instance (usegalaxy.org; Additional file 1: Figure S1 and Figure S2). The TraDIS workflow includes two parallel tracks (Additional file 1: Figure S1). One track performs essentiality analysis with Bio-Tradis [9] and the other one with TRANSIT [8]. The latter include the mapping of reads with Bowtie2 [15] and the analysis of coverage with bamCoverage tool [25]. The TnSeq Workflow is more complex than the TraDIS one, in part due to the read prepossessing needed with this dataset. Finally, steps have been included to replace the tool *Reads to counts* from the Bio-Tradis suite, which is not suitable for TnSeq data. This workflow uses Bowtie to map reads, as it is more adapted to very short reads than Bowtie2. Data are imported into Galaxy using the SRA tool fasterq-dump [26]. All the steps are highly adaptable to different read structures and sequencing technology agnostic. The entire workflow runs in approximately half a day on TnSeq data from [2]. The most time-consuming steps were splitting by barcodes and alignment. The availability of Jupyter in Galaxy permitted the in-depth analysis of the final data. We included these notebooks in galaxy histories containing

this paper analyses (See Additional file 1: Supplemental Information). We used them to perform regression analyses and compare the results of different methods with other published studies.

Discussion

As there is no Gold Standard Dataset for transposon analyses, we based our result evaluation on two factors. The first factor was manual analysis. Manual analysis of TIS data means looking at the coverage for a particular gene and seeing if it differs significantly from the surrounding region. The second factor was comparing with experimental studies on the same organism. The experimental conditions being variable between studies, we expected to find different genes that are essential in different conditions. However, it provided a base to identify "core essential genes", essential in any conditions, and constitute the majority of essential genes. Using these factors, we wanted to emphasize that our results' comparison is focused on getting the most accurate results from the data. If a gene is depleted in reads, we considered it as essential in our study. When we talk about "True Essential," it means that the automatic prediction of essentiality is coherent with the data's manual analysis. Since there was no experimental validation of the gene essentiality in the given conditions, we could not ascertain that the gene was truly essential without experimental validation.

To test the performance of TIS analysis methods, we selected two different datasets. The first one is a TraDIS dataset utilizing Tn5 transposons that can potentially insert at any genomic position. The second dataset used Mariner transposons inserting at `TA` sites. The choice of transposon impacts the overall library saturations: TraDIS data is more sparse than TnSeq data because Tn5 transposons have many more potential insertion sites than Mariner transposons. The TnSeq library has been built using different types of promoter-containing constructs. We performed the analyses using either the control constructs, that do not contain any promoter, or all constructs as replicates. The second solution provided more reliable results by increasing library saturation.

Regression analysis produced results that were in good agreement with published data as well as with databases of essential genes. We performed the manual regression in a Jupyter notebook in Galaxy. We also used the Bio-Tradis [9] tool suite to perform an automatic regression. The manual method adds the burden of the choice of distribution on the user. On the other hand, it provides more flexibility as differently saturated libraries may present different profiles that would change the distributions. When the appropriate parameters are determined, the manual method was highly reliable on datasets we tested.

HMM analyses produced inconclusive results in our hands. In both cases, the data were too sparse for HMM to identify stable regions (consecutive insertion sites with the same state covering a genomic region large enough to be considered significant biologically). Zero-inflated data disrupt the continuity of "regions with insertions" with empty sites that do not provide information. HMM predict "blocs" of sites in the same state by analyzing the probability of transition between states. If consecutive sites constantly jump between empty and with insertion, the algorithm is unable to identify contiguous states. These results were expected for the TraDIS dataset due to the libraries' sparse nature but were more surprising for the TnSeq data.

Conclusion

The goal of this study was to develop end-to-end workflows for the analysis of transposon insertion sequencing data. TIS studies are used to identify genes essential to bacterial growth in specific conditions or to detect genes causing growth defect or advantage. While the experimental aspects of the technique are continually evolving, there is no consensus on the way to perform the essentiality data analysis as many different approaches are described in the literature. These can be broadly classified into three categories. The first group of methods is based on a regression analysis of the gene saturation indices. The second type of analysis uses runs of consecutive sites with no transposon insertion. Finally, the third group consists of HMM-based methods. We added tools representing each of these categories to Galaxy toolkit.

The workflows implemented in Galaxy are using robust open-source tools. The tools used explicitly for TIS analysis are open source and well maintained, ensuring reproducibility and transparency of analyses. The workflows include Jupyter notebooks for exploratory analyses without losing Galaxy history tracking that allows traceability and sharing. The combination of open-source tools and Jupyter notebooks provides a complete and flexible workflow that can be easily modified to fit any analysis need.

Methods

Selecting appropriate tools

First, we assessed the status of existing tools for the analysis of TIS data (Table 1). This information is essential for tool selection as it identifies actively maintained tools that will be supported in the future. Based on this analysis, only Bio-Tradis, Magenta, and TRANSIT are actively developed, maintained, and regularly released (Table 1).

Alignment of TraDIS data

The reads have been trimmed to remove low-quality (Phred score < 20) bases at the end of the reads with Trimmomatic [27]. The reads were mapped using Bowtie2 [15]

against the reference genome of *Escherichia coli* BW25113 (CP009273.1), used in the Goodall et al. study. The coverage of the genome was computed with BamCoverage, from the Deeptools suite [25]. In our case, we were interested in identifying only insertion points. For that reason, we computed coverage using 5'-ends of the reads only.

TRANSIT for TraDIS data

Tn5Gaps is the Gumbel tool adapted for TraDIS data. These methods perform a gene by gene analysis of essentiality based on the longest consecutive sequence of potential sites without insertions in a gene. This metric allows identifying essential domains regardless of insertion at other locations of the gene. We ran TRANSIT with default parameters but changed the normalization method and selected not to normalize the counts. TRANSIT offers two ways to deal with replicates: either the counts of all replicates at each site are summed, or they are averaged. Using the sum of counts provides a better saturation of the library: a site that might not have been impacted during the initial transfection might have been impacted in another sample. In the case of TraDIS data, where the saturation tends to be low due to a large number of potential sites, the use of the sum of counts provides a better resolution. By averaging the counts at each site, we significantly decreased the noise.

Saturation indexes

Saturation is defined by the ratio of the number of sites impacted by insertions on the total number of sites able to receive an insertion in the gene. If the library is sufficiently saturated, we should observe two distinct distributions. The essential genes distribution has a low average saturation, and the distribution of the non-essential genes has a higher average saturation.

Regression on TraDIS data

Regression is a statistical analysis aiming to estimate the relationship between variables. It provides a function modeling this relationship. In our particular case, we are trying to identify the models behind gene saturations' observed distribution. Formalizing these models then allows calculating the probability of each gene to belong to either one. We performed this regression using the python library scipy [28]. We started by defining the probability density function (*pdf*) of the gene saturation as the sum of two known *pdf*s: here, an exponential and a γ distribution. We anchor the regression by providing expected parameters. Goodall et al. [13] did not provide the parameters, and thus we had to estimate them. We attempted to plot several distributions and correcting the parameters to make them look like our data as much as we could (See Jupyter notebook). Once we approximated the parameter, the fitting function fits the data and returns the corrected parameters and their standard deviations for the two distributions.

Classification of TraDIS data

Once we identified the two distributions, we calculate the probability of genes to belong to each of them. This probability is calculated using each class's probability density function (library *scipy* in python). *pdf*s are function whose area under the curve in the interval $x = [a, b]$ is the probability of a random value of the distribution to belong to the interval $[a, b]$. When used for a single value, it also provides a relative likelihood that the value n belongs to the distribution (the absolute likelihood of n is null since our variable is continuous).

To decide between the two categories, we select the most likely category if the difference between the two probabilities is significant. Formula (1) defines the significance thresholds for gene classification:

$$log_2\left(\frac{P(ES)}{P(NE)}\right) > log_2(x) \quad (1)$$

The threshold is calculated as the $log_2(x)$, where x is the number of times the likelihood for a model must be superior to the likelihood of the other one. The use of a logarithm allows ignoring the direction of the ratio for the decision calculation. Genes whose differences had not been judged significant are classified as undetermined.

Bio-TraDis on TraDIS data

The Bio-Tradis toolkit provides a comprehensive set of tools for preprocessing reads from TIS. Our data are already exempt from transposon sequences. We used the dataset provided as input for a read mapping step, followed by the count of insertion per gene and gene essentiality predictions. The mapping is done using bwa, with a minimum mapping quality of 0 and default values for the other parameters. The toolkit does not handle replicates, so we merged the datasets after the mapping by adding read counts at every position.

Pre-processing of TnSeq data

The reads published still contain transposon sequences that need trimming. The transposon sequence includes not only primers but also barcode sequences that separate the constructs containing different promoters. We use Cutadapt [29] to separate the reads of each construct. We then trimmed the transposon sequence downstream of the construct barcode, the sequence including the ITR and the MmeI site, the sample barcode at the beginning of the reads, and finally, the low-quality bases before alignment and calculation of coverage.

Alignment of TnSeq data

The reads have been aligned to the reference genome of *Staphylococcus aureus subsp. aureus NCTC 8325*

(CP_000253.1), used in the Santiago et al. study, using Bowtie. Bowtie [30] has been selected instead of Bowtie2 [31] for this dataset, as it is recommended for very short reads (shorter than 50 bp). We enforce alignment with no mismatches. It is necessary as we are working with very short reads of 16-17 bp, which is the minimum length for the majority of reads to have a unique alignment [32].

Counts for TnSeq data

The genome coverage is computed with BamCoverage, from the Deeptools suite [25]. The coverage is calculated with an offset of −1, meaning that the read is counted only at the 3' position, at the TA site position (Fig. 6b). This study has seven datasets for each biological replicates: one dataset corresponding to control constructs and six corresponding to transposon containing a promoter. Contrary to the TraDIS data, where the Tn5 transposon inserts everywhere, Mariner transposons insert only at TA sites. To compute the gene saturation and row of empty sites, we need to use the coverage of all TA sites, whether it is null or not. It is accomplished by merging the coverage files with the positions of TA sites calculated with the Nucleotide subsequence search tool available in Galaxy [33]. The resulting file is a tabular file with a column containing the read position (leftmost site) and another containing the counts of reads aligning at this TA site.

Bio-Tradis TnSeq data

The Bio-Tradis toolkit provides a comprehensive set of tools for preprocessing reads from TIS. This preprocessing requires to provide the sequence of the transposon located at the beginning of the read. Sequences needing trimming in our reads are located on both sides of the reads and are variable between constructs. Bio-Tradis includes a script for read alignment that provides both the mapped reads and the insertion counts at each nucleotide. The scripts do not provide the option to choose which end of the read should be used to attribute the insertion at a nucleotide. We used the counts generated by our workflow as input for Bio-Tradis essentiality analysis to circumvent this problem.

TRANSIT for TnSeq data

We ran HMM and Gumbel tools on the TnSeq data. The HMM results were inconclusive, probably due to each sample's low saturation after we divided the data based on transposon constructs. We ran the Gumbel method with default data, except for the option to not normalize data and the choice of replicates handling method. For this dataset, we are using the mean instead of the sum because of the large number of samples. While using the sum increases the library saturation, it is more sensitive to artifacts than the mean. If one read align by mistake at a position in 30% of the 84 samples, the mean would be 0 when the sum would be 15. it makes a big difference in the metrics used for essentiality prediction.

Regression on TnSeq data

The regression on TnSeq data follows the same protocol as the one used for TraDIS data. The difference is that we did not have any information about the type of distribution we expected. We added a step of distribution selection to the parameter estimation. We used a book describing the different distribution to select those who appeared close to the shape of our data [34]. We first started with the distribution of non-essential genes (Fig. 7a). The distribution shows no skew and seems to correspond to a normal distribution. An exponential distribution seemed to be the most appropriate to describe essential gene saturations.

Acknowledgments
Authors are grateful to all members of the Galaxy team for help with developing and deploying software components described in this manuscript. Dave Bouvier provided critical help with wrapping necessary tools. Nate Coroar deployed production versions of tools and workflows.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12866-021-02184-4.

Additional file 1: Supplementary file.

Authors' contributions
DL performed all analyses, wrapped tools, and wrote the manuscript. AN developed initial design, guided analysis, and edited the manuscript. KK and LW provided guidance on minute details of the TnSeq approach and participated in all stages of data analysis and writing. All authors are in agreement on publishing on this work in its present form

Funding
usegalaxy.org efforts are funded by NIH Grants U41 HG006620, R01 AI134384 and NSF ABI Grant 1661497. The funding organization had no input on design of the study, and collection, analysis, and interpretation of data and in writing the manuscript

Availability of data and materials
All software developed here is open source and fully accesible to anyone:

- Workflow for Tradis Analyses : https://usegalaxy.org/u/delphinel/w/tradis-analysis
- Workflow for TnSeq Analyses : https://usegalaxy.org/u/delphinel/w/tnseq-analysis
- GitHub repository of the paper : https://github.com/galaxyproject/TIS_methods_review
- Training material for TIS https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/tnseq/tutorial.html
- Detail of Regression on TnSeq Data : https://github.com/galaxyproject/TIS_methods_review/blob/master/TnSeq/Essential%20genes%20in%20S.%20aureus.ipynb
- Detail of Regression on TnSeq Data : https://github.com/galaxyproject/TIS_methods_review/blob/master/TraDis/Essential%20genes%20in%20E.%20coli%20K12%20.ipynb

Declarations

Ethics approval and consent to participate
This manuscript uses previously published data and does not utilize any identifiable information.

Consent for publication
Not Applicable.

Competing interests
AN is founder and advisor for https://galaxyworks.io/. Other authors declare no competing interests.

Received: 8 October 2020 Accepted: 8 April 2021
Published online: 05 June 2021

References

1. Chao MC, Abel S, Davis BM, Waldor MK. The design and analysis of transposon insertion sequencing experiments. Nat Rev Microbiol. 2016;14(2):119–28.
2. Santiago M, Matano LM, Moussa SH, Gilmore MS, Walker S, Meredith TC. A new platform for ultra-high density staphylococcus aureus transposon libraries. BMC Genomics. 2015;16:252.
3. van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol. 2013;11(7):435–42.
4. Lodge JK, Weston-Hafer K, Berg DE. Transposon tn5 target specificity: preference for insertion at G/C pairs. Genetics. 1988;120(3):645–50.
5. Zomer A, Burghout P, Bootsma HJ, Hermans PWM, van Hijum SAFT. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. PLoS ONE. 2012;7(8):43012.
6. Solaimanpour S, Sarmiento F, Mrázek J. Tn-seq explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries. PLoS ONE. 2015;10(5):0126070.
7. Pritchard JR, Chao MC, Abel S, Davis BM, Baranowski C, Zhang YJ, Rubin EJ, Waldor MK. ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing. PLoS Genet. 2014;10(11):1004782.
8. DeJesus MA, Ambadipudi C, Baker R, Sassetti C, Ioerger TR. TRANSIT–A software tool for himar1 TnSeq analysis. PLoS Comput Biol. 2015;11(10):1004401.
9. Barquist L, Mayho M, Cummins C, Cain AK, Boinett CJ, Page AJ, Langridge GC, Quail MA, Keane JA, Parkhill J. The tradis toolkit: sequencing and analysis for dense transposon mutant libraries. Bioinformatics. 2016;32(7):1109–11.
10. McCoy KM, Antonio ML, van Opijnen T. MAGenTA: a galaxy implemented tool for complete Tn-Seq analysis and data visualization. Bioinformatics. 2017;33(17):2781–3.
11. Zhao L, Anderson MT, Wu W, T Mobley HL, Bachman MA. TnseqDiff: identification of conditionally essential genes in transposon sequencing studies. BMC Bioinformatics. 2017;18(1):326.
12. 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):00015–17.
13. Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, Cunningham AF, Lund PA, Cole JA, Henderson IR. The essential genome of escherichia coli K-12. MBio. 2018;9(1):e02096-17.
14. Santiago M, Lee W, Fayad AA, Coe KA, Rajagopal M, Do T, Hennessen F, Srisuknimit V, Müller R, Meredith TC, Walker S. Genome-wide mutant profiling predicts the mechanism of a Lipid II binding antibiotic. Nat Chem Biol. 14(6):601–8.
15. Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK. Simultaneous assay of every salmonella typhi gene using one million transposon mutants. Genome Res. 2009;19(12):2308–16.
16. Reznikoff WS. Transposon Tn5. Annu Rev Genet. 42:269–86.
17. Phan M-D, Peters KM, Sarkar S, Lukowski SW, Allsopp LP, Moriel DG, Achard MES, Totsika M, Marshall VM, Upton M, et al. The serum resistome of a globally disseminated multidrug resistant uropathogenic Escherichia coli clone. PLoS genetics. 2013;9(10):.
18. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of escherichia coli K-12 in-frame, single-gene knockout mutants: the keio collection. Mol Syst Biol. 2006;2:2006.0008.
19. Yamazaki Y, Niki H, Kato J-I. Profiling of escherichia coli chromosome database. Methods Mol Biol. 2008;416:385–9.
20. DeJesus MA, Ioerger TR. Capturing Uncertainty by Modeling Local Transposon Insertion Frequencies Improves Discrimination of Essential Genes. IEEE/ACM Trans Comput Biol Bioinform. 12(1):92–102.
21. Chaudhuri RR, Allen AG, Owen PJ, Shalom G, Stone K, Harrison M, Burgis TA, Lockyer M, Garcia-Lara J, Foster SJ, Pleasance SJ, Peters SE, Maskell DJ, Charles IG. Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). BMC Genomics. 2009;10:291.
22. Valentino MD, Foulston L, Sadaka A, Kos VN, Villet RA, Maria JS, Lazinski DW, Camilli A, Walker S, Hooper DC, Gilmore MS. Genes Contributing to Staphylococcus aureus Fitness in Abscess- and Infection-Related Ecologies. mBio. 2014;5(5):.
23. Afgan E, Baker D, Batut B, Van Den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46(W1):W537–W544.
24. Lariviere D. GitHub. GitHub repository. 2020. https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/tnseq/tutorial.html.
25. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deeptools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44(W1):160–5.
26. NCBI. SRA-tools Github repository. GitHub repository. https://github.com/ncbi/sra-tools. GitHub. Accessed with Galaxy, Tool version 2.10.4.
27. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–2120.
28. Oliphant TE. SciPy: Open source scientific tools for python. Comput Sci Eng. 2007;9(1):10–20.
29. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17(1):10–2.
30. Langmead B. Aligning short sequencing reads with bowtie. Curr Protoc Bioinforma. 2010;Chapter 11:11–7.
31. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9.
32. Kwon YM, Ricke SC, Mandal RK. Transposon sequencing: methods and expanding applications. Appl Microbiol Biotechnol. 2016;100(1):31–43.
33. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3.
34. Crooks GE. Field guide to continuous probability distributions. Berkeley: Berkeley Institute for Theoretical Science; 2019.

Publisher's Note