Published in final edited form as:

Nat Rev Microbiol. 2016 February ; 14(2): 119–128. doi:10.1038/nrmicro.2015.7.

The Design and Analysis of Transposon-Insertion Sequencing Experiments

Michael C. Chao^{1,2,3}, Sören Abel^{1,2,3,4}, Brigid M. Davis^{1,2,3}, and Matthew K. Waldor^{1,2,3,§}

¹ Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA, USA 02115

² Division of Infectious Disease, Brigham and Women's Hospital, Boston, MA, USA 02115

³ Howard Hughes Medical Institute, Boston, MA US 02115

⁴ Department of Pharmacy, UiT, The Arctic University of Norway, Tromsø, Norway

Preface

Transposon-insertion sequencing (TIS) is a powerful approach that can be widely applied to genome-wide definition of loci that are required for growth in diverse conditions. However, experimental design choices and stochastic biological processes can heavily influence the results of TIS experiments and affect downstream statistical analysis. Here, we discuss TIS experimental parameters and how these factors relate to the benefits and limitations of the various statistical frameworks that can be applied to computational analysis of TIS data.

Transposon-insertion sequencing (TIS), the most recent incarnation of transposon-based genomic analyses, is revolutionizing microbiology, by enabling researchers to carry out genome-wide studies in a wide range of bacterial species, under a multitude of conditions, with unprecedented depth (reviewed in ^{1,2}). Several variations of the TIS approach have been independently developed e.g. transposon insertion site sequencing (TnSeq), insertion sequencing (INSeq), high-throughput insertion tracking by deep sequencing (HITS) and transposon-directed insertion site sequencing (TraDIS)³⁻⁶; however, they all share the same fundamental methodology (Figure 1). First, transposon mutagenesis is used to create a pool of insertion mutants in which ideally all genomic loci have been disrupted at multiple sites. In principle, the starting transposon pool will not contain disruptions in any of the genomic regions that are required for growth (i.e. 'essential' regions). This library can then

[§] To whom correspondence should be addressed: Tel: 617-525-4646, Fax: 617-525-4660, mwaldor@research.bwh.harvard.edu.

ONLINE ONLY

Links

[CE: please add.]

Subject categories

Biological sciences / Biological techniques / Sequencing / DNA sequencing

[URI /631/1647/514/1948]

Biological sciences / Genetics / Genome / Interspersed repetitive sequences / DNA transposable elements

[URI /631/208/726/2001/1428]

Biological sciences / Genetics / Functional genomics / Mutagenesis

[URI /631/208/191/1908]

Biological sciences / Biological techniques / Sequencing / RNA sequencing

[URI /631/1647/514/1949]

be grown under selective conditions, so that mutants that are attenuated under a specific condition of interest are outcompeted and mutants with increased growth and survival become overrepresented in the population. High-throughput sequencing is used to quantify all transposon junctions, the prevalence of each of which should be proportional to the corresponding mutant's fitness in each growth condition.

Experimental parameters that vary among TIS studies include the transposon selected, the complexity of the TIS libraries generated (i.e., the number of independent mutants per library), the constraints imposed by the experimental conditions chosen and the reliability with which representative DNA libraries are created and sequenced. A key challenge for the field is to understand how these parameters may influence the choice of downstream data normalization and statistical methods for TIS analysis. Here, we discuss the benefits and limitations of several statistical frameworks that have been applied to the two most common types of TIS analyses—identification of ‘essential’ loci using a single TIS library, and definition of ‘conditionally essential’ loci through the comparison of two or more datasets derived from a single ancestral library. Slightly different approaches may be used for comparative analyses of independently derived libraries, such as libraries that have been generated in distinct genetic backgrounds (Box 1).

Experimental design

The structure of TIS data and its optimal compatibility with particular downstream analysis methods are influenced by experimental factors such as transposon insertion bias, the complexity of the TIS libraries created, the presence of experimental bottlenecks, and technical sources of noise during DNA library preparation, and during the amplification and sequencing of transposon junctions.

Choosing a transposon

The majority of TIS studies have been carried out using libraries of randomly pooled transposon mutants, although defined transposon libraries can also be used and are preferred in certain circumstances (Box 1). Randomly pooled transposon libraries are commonly created using either mariner-based transposons, which target TA dinucleotides^{7, 8}, or Tn5-based vectors, which insert at random sites⁹. Due to the reported absence of target sequence requirements apart from a TA dinucleotide, the number of reads from each mariner insertion is thought to reflect that mutant's fitness. For most bacteria, TA dinucleotides are distributed relatively evenly across the genome, although TA frequency can vary locally within the genome, e.g. in horizontally acquired DNA regions. In *Escherichia coli*, which has a GC content of ~50%, there are on average >30 potential insertion sites per 1 kilobase gene. Although loci with fewer potential insertion sites (such as those in species with high GC content) may have less statistical power, mariner-based transposons have been used successfully to define genes that are essential for growth in the high GC organism *Mycobacterium tuberculosis*¹⁰ as well as essential loci as small as 200 nucleotides in *Vibrio cholerae*¹¹. In general, we favor the use of mariner-based transposons, both for their lack of insertion bias and because the construction of dense libraries that are saturated or near saturated (i.e. have insertions at all or nearly all dispensable TA sites) can be more easily

attained than with other transposons, which facilitates discrimination between essential and non-essential loci.

Tn5 transposons have no target sequence requirement for insertion, which enables potentially greater insertion density and thus genomic resolution than mariner-based studies. However, Tn5 transposons possess a preference for high GC regions^{12, 13}, and read counts at these hotspots can reflect insertion bias instead of reporting the fitness of the corresponding mutant. Preferential insertion at hotspots may also reduce overall library complexity. Nonetheless, Tn5 libraries have been successfully used for TIS studies in many conditions^{14, 15}, and may be preferable for species in which mariner-based transposition proves difficult to optimize.

Library complexity

The power of any TIS study is greatly influenced by the complexity of the initial library generated and the maintenance of this complexity during selection. A library with a high degree of complexity contains multiple unique insertions in every locus, which together serve as independent indicators (akin to ‘biological replicates’) of a locus’ fitness. Libraries with a relatively high density of insertions also enable greater genomic precision in mapping the boundaries of regions of interest. However, TIS studies are still feasible even when highly complex libraries cannot be created (e.g. due to technical or biological constraints), although in some cases alternative library designs or methods for analysis are preferred in these cases (Box 1).

To illustrate the advantage of higher insertion frequency, one can separate genes based on the percentage of potential insertion sites disrupted. A high-density library (Figure 1b, orange graph) shows a separation between presumed non-essential genes with a high rate of insertion (right peak) and genes likely to be essential with few insertions (left peak). In contrast, a less saturated library (Figure 1b, purple graph) will be characterized by overlapping peaks, suggesting that many non-essential genes—especially those that are small or contain few potential insertion sites—were not disrupted by chance and may be incorrectly classified as essential. As such, these genes cannot be meaningfully investigated during subsequent conditional essentiality studies. A major benefit of transposons that target specific sequences (e.g. mariner-based transposons) is that they may only integrate at fixed number of potential insertion sites, enabling the generation of highly saturated libraries. Recent approaches utilizing Bayesian statistical models and insertion density as a measure of library saturation have enabled the definition of essential genes from libraries created using mariner-based (and probably also randomly inserting) transposons^{2, 4, 16, 17}.

The ability to create complex transposon libraries depends on the methods by which transposons are introduced into bacterial genomes. Transformation with commercially available pre-assembled transposon complexes (e.g., EZ-Tn5)⁹, in which the purified transposase enzyme is pre-loaded with the transposon, involves little genetic engineering and can be a powerful technique for the generation of libraries of strains in which genetic engineering is difficult. Alternatively, TIS libraries can be generated using suicide constructs, in which the transposon and transposase are separately encoded on a replication-deficient vector; once this construct is introduced into the host strain (through diverse

mechanisms such as transduction^{8, 18}, transformation¹⁹ and conjugation¹¹), the transposon is mobilized and stably inserts itself into the genome, while the vector is lost during subsequent bacterial growth. Mutants can also be generated in naturally competent species following the introduction of purified genomic DNA that has been previously subjected to *in vitro* transposition^{3, 20}.

To maximize diversity during library construction, it is beneficial to choose conditions that offer high transposition rates while limiting outgrowth before marker selection. For example, transformation of cells with pre-assembled transposon complexes²¹ or transduction with suicide constructs⁷ typically involves little bacterial doubling before mutants are plated, producing a direct relationship between the number of colonies obtained and the number of unique insertions in the library (assuming that there is little non-integrative carriage of the transposon or insertion at genomic hotspots). In contrast, protocols that involve lengthy incubation times prior to plating, such as extended conjugation or transformation, may lead to large numbers of mutant colonies but low library complexity. In general, it is preferable to increase the number of independent transposition reactions rather than the incubation time when using these approaches.

Experimental Bottlenecks

In theory, only mutants with attenuated survival should decrease in frequency under selective growth conditions. However, insertion mutants can be lost during selective growth due to fitness-independent stochastic processes, such as population constrictions due to experimental bottlenecks (Box 2). If experimental bottlenecks are significantly smaller than the complexity of the library being tested, then many insertion mutants will be lost entirely due to chance. Failure to consider and account for the effects of such bottlenecks can lead to spurious findings during statistical analysis.

Bottlenecks are often present in TIS infection studies in which a transposon library is used to infect an animal host²²⁻²⁶. For example, it was recently determined that typically only 10^4 - 10^5 *V. cholerae* of 10^9 bacteria inoculated establish infection in a rabbit host, suggesting that >99.99% of bacteria are lost due to infection bottlenecks²⁷. Much of this population loss may result from the acidic pH of the stomach, which is a severe but non-specific infection barrier for this pathogen²⁸. Several sequencing-based approaches for the detection and quantification of bottlenecks are presented in Box 2.

Both experimental and computational approaches can be used to mitigate the impact of bottlenecks on TIS data. Bottleneck effects can be reduced experimentally by using subpools of transposon libraries (at the same dose) that are less complex, such that representatives of all fit genotypes in the input have a high chance of passing the bottleneck^{4, 23, 25}. However, this approach may require more experimental replicates with different transposon subpools to survey the fitness of all genomic insertions in the full library. In the face of extremely narrow bottlenecks, the number of biological replicates required to screen all mutants of a highly saturated library becomes experimentally infeasible. In such cases, the use of a defined transposon library (Box 1) with lower complexity can facilitate screening of all desired genotypes.

Experimental conditions can also be changed to ‘widen’ the bottleneck and reduce the severity of the population constriction. For example, in experimental *V. cholerae* infections, antacids can be given to the host²⁹, the inoculum can be suspended in a buffered solution²⁴, and the pathogen can be pre-conditioned by acid-tolerizing treatment prior to intragastric inoculation³⁰. While all these measures minimize the stomach acid-mediated bottleneck, they also change the selective conditions and may affect the fitness of specific mutants in non-physiological ways, for example retaining mutants with increased acid sensitivity that would be otherwise lost during infection.

Computational approaches can be used to predict the likelihood that insertions will disappear from a library due to bottlenecks, and aid accurate interpretation of the experiment, as mutants expected to disappear solely by chance cannot be classified with confidence as conditionally essential. In general, computational approaches to compensate for bottleneck effects^{24-26, 31} determine the fractional loss of insertions in presumed neutral genes during the experiment, then remove a similar fraction of reads from the control library, which enables more accurate comparison between control and experimental datasets. In conditional essentiality analyses, computational compensation for bottlenecks results in substantially fewer false positive assignments and greater confidence in the candidates identified; however, it may also obscure the identification of legitimate conditionally essential insertions (i.e., produce false negatives) if such loci could also be lost solely by chance.

Insertion site amplification and sequencing

Experimental differences related to the amplification and sequencing of transposon insertion sites can affect the quality of TIS data. For example, prior to amplification of transposon junctions, adaptors must be attached to genomic DNA prepared from the transposon pool, either by ligation³, by enzymatic homopolymer nucleotide extension³² or through polymerase chain reaction (PCR) amplification³³. The goal is to attach adaptors onto every unique transposon junction, thereby generating a DNA library that is representative of the transposon mutant pool. Enzymatic homopolymer nucleotide extension³² has been shown to be more efficient than DNA ligation, and thus may be preferred when only low amounts of starting material are available. Alternatively, a circular amplification method was developed³³ in which transposon junctions are amplified as single-stranded DNA (ssDNA) and background double-stranded DNA (dsDNA) is degraded, thereby enriching for sequences of interest from complex samples. DNA amplification using arbitrary PCR with degenerate primers¹⁴ is an alternative to adaptor-based methods; however, uneven annealing efficiencies at different insertion sites can distort the relationship between a mutant's fitness and its read count (Figure 2A). Bias may also be introduced through unequal amplification of DNA junctions due to differences in sequence content or fragment size. Several groups^{3, 4} avoid the latter difficulty by using a mariner-based transposon that contains a *MmeI* restriction site in its inverted repeat, which enables cleavage of the genomic DNA downstream of the transposition site to produce a uniform-sized fragment for all insertions. However, this approach precludes the use of fragment boundaries as the basis for assessing the extent of amplification jackpots (discussed below).

Regardless of the amplification method used, certain insertions may be highly overrepresented in the sequencing data, producing a long tail when the average read counts per insertion is graphed for all genes (Figure 2B). Overrepresentation of fragments may result from amplification jackpot events (e.g. early or preferred amplification of some transposon junctions over others) or reflect mutants that are truly more prevalent in the library³⁴ (e.g., due to insertion biases or a selective growth advantage of these mutants). These possibilities can be distinguished by the incorporation of a unique, randomly generated sequencing tag in the adaptor primer, which leaves every DNA template in the library uniquely marked after adaptor ligation¹⁰, an approach that has also been used to the same effect in RNA sequencing (RNA-seq) studies^{35, 36}. Each sequence read will consequently be associated with a unique tag, and any sequence that is associated with an overrepresented tag is likely to be abundant as a result of PCR amplification bias. Alternatively, when sheared DNA is used for library generation, paired-end sequencing may help identify jackpot amplification products, as reads with the same site of shearing (and thus mapping) are likely to represent clonal amplification products; these duplicate reads can be detected using several sequencing analysis programs (e.g. SAMtools³⁷ and Picard (<http://broadinstitute.github.io/picard/>)). As PCR jackpots do not reflect the actual frequency of the associated insertion mutants in the library, it is advisable to remove all duplicated reads from the data before further analysis.

The level of sequencing required to identify all mutants in a library can be estimated by randomly sampling the sequencing output in sequentially larger pools and assessing how many unique insertion mutants are detected²⁴. Sequencing saturation is achieved when the discovery of new insertion mutants nears a plateau (Figure 2C, dotted line). Beyond this point, deeper sequencing will largely increase the counts for previously identified mutants. In general, deeper sequencing promotes accurate discrimination between dispensable and essential sites and facilitates identification of loci with relatively subtle effects on growth or survival. However, depending on the desired read count range and the number of samples, sufficient sequencing may be achieved on a lower throughput sequencing platform than is generally used. Empirically, in a *V. cholerae* library containing ~120,000 unique insertion mutants that was sequenced to saturation at a depth of 5 million mapped reads, non-attenuated insertions had 10-50 read counts²⁴, which represents the study's ceiling for detecting differences between fit and attenuated mutants.

Normalization based on chromosomal location

The numbers of reads that map to replication origin-proximal insertions may be inflated by incomplete rounds of DNA replication relative to insertions near the replication terminus (Figure 2D). Regardless of the downstream statistical analysis method being employed, positional bias should be normalized, such that on average, insertions with comparable fitness possess similar read counts, regardless of their genomic position. The local positional bias along the chromosome can be modeled utilizing a LOESS regression function, followed by reads normalization based on the extrapolated slope of this curve³¹. Alternatively, positional bias can be corrected by dividing the genome into large genomic windows (e.g. 100 kilobase), and using the average read count of each window to locally scale each insertion³³. Both methods produce normalized data that no longer possess a visible skew

along the chromosome (Figure 2D), which enables more accurate downstream statistical testing between loci. Although the genomes of some species contain multiple replicons (chromosomes, plasmids, etc.), it is advisable to normalize reads across a single replicon, as the rate and completeness of replication may vary among an organism's replicons.

Essential Loci Analysis

Essential loci are typically identified as regions (e.g. genes) in a high-density transposon library that lack or have a significantly lower than average frequency of insertions even when the library is grown on optimal laboratory media. Below, we discuss the benefits and limitations of different strategies that have been used to identify essential loci (Figure 2E). In general, meeting three criteria makes identification of essential loci easier: use of a highly saturated transposon library, in which non-essential genes that escape transposition by chance are rare; sequencing at a depth sufficient to detect the vast majority of transposon mutants in the library; and normalization of read counts based on chromosomal location. Essential loci analysis approaches fall into two broad categories—annotation-dependent and annotation-independent methods. Annotation-dependent methods aggregate datapoints across known genetic units (e.g. within predicted genes) to conduct gene-level analysis, while annotation-independent methods analyze data without considering known genetic boundaries and consequently offer subgenic resolution. Annotation-independent methods also enable the characterization of intergenic regions, which are often ignored in annotation-dependent analyses and may contain novel genomic features of interest. Finally, as annotation-dependent analyses compare all genes to each other, genes with different numbers of potential insertion sites may be analyzed with different statistical power and small genes may be classified less accurately than larger loci.

Annotation-dependent analysis

Several studies^{3, 5, 14, 38, 39} defined essential genes as those in which very few or no reads were mapped (Figure 2E). As insertions within the 5' and 3' ends of coding regions sometimes do not mimic true gene disruption, insertions in the terminal 5-15% of genes were generally ignored in these analyses. Defining essential genes as those lacking any associated reads is a particularly stringent criterion, and the use of total reads per locus as the basis for identifying essential loci is particularly appropriate for relatively low complexity libraries in which most genes contain only a few insertions.

Alternatively, annotated loci can be classified based on the percentage of potential insertion sites that have been disrupted^{6, 15, 40-42}, which provides some resistance to noise introduced by rare reads (Figure 2E), since genes with a low but non-zero percentage of sites disrupted can still be distinguished from genes defined as non-essential. As insertion density alone is less powerful in non-saturated libraries, which may lack a clear distinction between genes that have high and low insertion frequencies (Fig. 1 graphs), Bayesian statistical frameworks^{4, 17} have been employed to classify loci based on both the number of potential insertion sites in a gene and the overall saturation of the library. The use of Bayesian statistics has been extended¹⁶ to identify essential genes in more saturated libraries by utilizing overall library saturation to define the largest contiguous interval of potential

insertion sites that would not sustain transposition by chance. Genes containing undisrupted sequence larger than this interval were thus defined as essential.

Annotation-independent analysis

Annotation-independent approaches enable the identification of subgenic and intergenic essential regions. One such approach utilizes a sliding window¹⁰; the genome is divided into overlapping windows of a specific size and the read counts within each window are used to determine if the region is significantly underrepresented in read counts when compared to the reads in all genomic windows (Fig. 2F). The use of small, standardized windows enables all loci to be equivalently analyzed and can potentially allow high-resolution identification of essential sequences; it also permits classification of genes that contain both essential and non-essential windows as “domain essential” (Fig. 2F). However, the increased number of windows that accompanies decreases in window size can also limit statistical power, due to the need for multiple testing corrections. Additionally, a non-parametric test¹⁰ has been used to reduce the negative impacts of sequencing jackpots on the accuracy of gene classification.

Finally, hidden Markov models (HMM) were adapted independently for essential loci analysis in *V. cholerae*¹¹ and *M. tuberculosis*⁴³. The HMM is a probabilistic statistical model that uses the read count at individual insertion sites to predict whether the site belongs to a biological category of interest (e.g. resides in a region that is required for growth). The HMM predicts the likeliest category for each insertion in the genome in an annotation-independent manner, which enables sub-genic classification of insertion sites and subsequent definition of domain essential loci (Fig. 2F). However, unlike sliding window analysis, the HMM statistical framework incorporates read counts at adjacent sites as well as read counts at the insertion of interest into its predictions. Compared to sliding window analysis, HMM-based approaches are less sensitive to noise arising from PCR jackpots and mismapping events, and they provide potentially single-insertion level resolution in mapping essential sequences. Furthermore, as the number of reads at a given site can be assigned to multiple biological categories, the HMM has been used successfully in stratifying genes into different biological classes^{11, 43} (e.g., essential genes, genes required for optimal growth, and genes that impede growth). A major drawback of the HMM is that it evaluates all potential insertion sites, and consequently requires a relatively high-density transposon library in order to determine whether a region is truly essential or simply stochastically underrepresented for reads.

Conditional Essentiality Analysis

Conditional essentiality analysis involves the comparison of two TIS datasets and enables the definition of loci and pathways that are required for optimal growth under only a subset of conditions. The two datasets can be derived from libraries grown in different environments, or from libraries that were generated in different genetic backgrounds and grown similarly (Box 1, ^{3, 25}). Previous studies have used various analysis methods to unravel bacterial pathways required for nutrient utilization^{18, 44}, define factors that promote host colonization and infection^{23, 24, 30, 45, 46} and map regulatory pathways required for environmental adaptation⁴⁷⁻⁴⁹.

Library generation

Comparative TIS studies utilize either sequentially selected or independently generated libraries. In sequential experiments, an unselected library is subsequently grown under more stringent, “selective” conditions. The set of insertions in the secondary, selected library are therefore a subset of the insertions that were present in the primary library, permitting the pairwise comparison between libraries of reads at each insertion site and the exclusion of sites that were not disrupted in the initial library from analysis, so that they do not dilute statistical power.

Alternatively, there are specific applications that require the construction and comparison of independent libraries: e.g. synthetic lethal studies that dissect genetic interactions in specific pathways compare TIS libraries made in wildtype and mutant strain backgrounds^{3, 25, 49} (Box 1). These libraries will contain a different compilation of insertions, which prevents direct pairwise comparison of insertion sites between each library.

Library normalization

Datasets to be analyzed for conditional essentiality can be normalized for three potential confounding factors: positional read bias, differences in sequencing depth, and stochastic differences in library diversity.

First, cells isolated from different conditions may have different growth rates, which may affect the DNA replication rate and thus introduce uneven positional biases between datasets. These biases should be assessed and remedied, as discussed above during essential loci analysis. Next, control and experimental datasets should also possess the same overall read counts. Most TIS studies equalize libraries by multiplicatively scaling reads from one sample to match the total read counts from the more deeply sequenced sample. This scaling, however, assumes that control and experimental datasets have the same complexity, which is not always true, especially in experimental systems with significant bottlenecks. As noted above, several methods have been used to computationally address this problem^{24-26, 31} by using the difference in library complexity between TIS datasets to estimate the severity of technical and experimental bottlenecks; the more complex library is then computationally passed through a bottleneck of similar size to model which insertion mutants are likely to be lost entirely by chance during the experiment. Loss resulting from bottlenecks can be simulated by proportionally removing reads equal to the overall difference in diversity between datasets³¹, proportionally removing reads based on the difference in unique reads in a set of known neutral genes^{25, 26, 30} or using the difference in overall complexity to simulate reads using a multinomial distribution²⁴. Use of these simulated datasets for comparison with the less complex experimental library enables more robust statistical comparisons and reduces false positive assignments.

Annotation-dependent analysis

Several studies have reported conditional essentiality analysis^{50, 51} using tools originally developed for RNA-Seq. In the RNA-Seq workflow (reviewed in^{52, 53}), reads are mapped to annotated genes, and the reads for every gene are normalized to the total reads sequenced and to the size of the gene. The normalized value, RPKM (reads per kilobase per million

mapped reads), is then compared for each gene between multiple samples, and genes that are significantly under- or overrepresented are determined, typically by identifying genes whose read counts lie outside the counts that would be predicted by chance using a negative binomial distribution. The negative binomial distribution aligns well with RNA-seq data, and is preferentially used when only a small proportion of the total reads maps to any given locus (e.g., when there are many loci represented by relatively similar expression levels) and when the read count variability between biological replicates is high⁵⁴. In principle, the negative binomial distribution should also suit TIS data, but it is important that researchers establish that this is the case for their datasets.

Another factor that can complicate the use of RNA-seq pipelines is that such analyses typically presume that overrepresented and underrepresented loci will be balanced, yet comparative TIS analyses typically have a preponderance of underrepresented loci. Additionally, RNA-seq pipelines do not typically adjust for experimental bottlenecks, since stochastic changes in population structure that can have major effects on TIS analysis (Box 2) do not typically confound transcriptomic data. Thus, unless the impact of bottlenecks has been first addressed, RNA-Seq-based pipelines should not be used 'out of the box' for TIS analysis. Recently, a publicly available TIS analysis pipeline, ESSENTIALS³¹, adapted RNA-seq analysis tools for TIS analysis. Importantly, this pipeline includes methods to mitigate many of the potential complications discussed above (e.g. correction of genome position and adjustment for differences in diversity between libraries). Additionally, ESSENTIALS offers an integrated and standalone web-based program to conduct TIS analysis that does not require previous computation experience.

In addition to RNA-Seq-based methods, many studies have used methodologically diverse custom models for data analysis. For example, some studies use arbitrary cutoffs in read counts differences to identify candidates of interest^{44, 55}. These types of approaches can yield valuable candidate genes, but are highly sensitive to experimental noise (e.g., mismapping, amplification jackpots). Standard statistical tests can also be used to analyze the read count distribution of insertions across genes^{22, 45, 56}. Parametric tests, such as the t-test, assume that TIS read count data follows a normal distribution, which is not usually the case (Figure 2B), and may be less sensitive at identifying candidates from noisy data. In contrast, nonparametric tests, such as the rank sum Mann Whitney U test, make no assumption regarding the distribution of the data and are more resistant to noise, but are also less robust at identifying candidate genes with subtle changes in fitness or those with few insertions. Finally, several studies have created customized models that may better tailor the analysis to specific TIS data by explicitly modeling the effect of amplification bias, jackpot events and library saturation into the statistical framework^{15, 39}. These models may offer a more nuanced assessment of conditional essentiality compared to more standardized statistical tests, but generally require programming ability and may not be entirely applicable to unrelated datasets.

Another approach^{2, 3} identifies conditionally essential genes by converting read counts into a generalizable growth or death rate for each mutant and gene. These growth or survival rates (i.e., fitness) may be compared for each gene and for growth under different selection conditions. Consequently, this fitness-based approach can also be utilized for the definition

of genes required for optimal intrinsic (i.e., not conditional) bacterial growth (Box 1) and loci that hinder growth, and for identification of loci that promote growth yet are not conditionally essential.

In principle, TIS experiments are thousands of competition experiments conducted simultaneously, where the change in a mutant's abundance before and after selection can be reported as a competitive index (CI). However, CIs depend on the number of generations that have passed in an experiment, making it difficult to compare the CIs of mutants grown under different conditions. To address this issue, the number of replication cycles that have occurred for each library in different growth conditions can be calculated³, enabling conversion of CIs to normalized fitness values (i.e., growth or death rates) that can be compared between different datasets. With such normalization, it becomes possible to define the extent to which mutants exhibit generalized vs. environment -dependent growth deficiencies and advantages.

In *in vitro* assays, the extent of population expansion can be enumerated by counting colony-forming units (CFUs) before and after selection. A recent study²⁶ reported that enumerating CFUs from infected animals at several time points could also be used to assess replication rate *in vivo*. Alternatively, an unstable plasmid with a loss rate per generation that is known^{3, 30} can be used to monitor the number of doublings occurring in each experimental system. The latter approach may be preferable when bacteria are subjected to environments in which replication and death occur simultaneously, as high levels of cell death could mask the extent of replication detected through CFU enumeration.

One potential complication for the fitness method is that it presumes that fitness defects or advantages reflect a uniform and constant rate of loss or enrichment during selection. For many mutants, calculated fitness has been confirmed to be very near their actual growth rate^{3, 25, 26, 30}. However, it is possible that assays with complex selection conditions (e.g., assays spanning different growth phases) may yield fitness values for some genes (e.g., those with growth phase-dependent phenotypes) that do not precisely correspond to the mutants' growth.

Annotation-independent analysis

A hidden Markov model (HMM) was recently adapted for TIS conditional essentiality analysis as part of the Con-ARTIST pipeline²⁴. Unlike most of the analyses described above, which group and analyze insertions at the gene level, this pipeline allows annotation-independent characterization of transposon mutants at the single insertion level, and consequently enables the discovery of subgenic or intergenic regions of conditionally essentiality. The Con-ARTIST pipeline also includes a module that normalizes for bottleneck effects by generating multiple simulated control libraries, based on a multinomial distribution. This modeling approach compensates for the stochastic effects of bottlenecks on population composition, and allows identification of loci with consistently different representation between two datasets following normalization. This normalization approach can reduce false positive assignments approximately 3 fold²⁴ (e.g. we found that only 37% of genes classified as conditionally essential *in vivo* based on multiplicative scaling of a *M. tuberculosis* input library were reproducibly classified in this way after application of our

modeling approach), although the extent to which these findings are generalizable is not known. However, Con-ARTIST is not appropriate for analysis of all TIS datasets, as it requires high library saturation (insertions at >50% of all potential insertion sites), and the use of related—rather than independently generated—transposon libraries for determination of conditional essentiality.

Conclusions

While TIS methods have the power to capture the genomic bases for diverse phenotypes in many different bacterial species, understanding how TIS design choices as well as stochastic biological processes impact the data will aid in choosing the most appropriate downstream analysis methods and lead to optimal extraction of biologically meaningful information from these experiments. While experimental metrics can vary between studies, and there is no ‘one size fits all’ approach for TIS analysis, addressing certain experimental factors will allow researchers to maximize the yield from their experiments while considering the biological constraints of their system. First, greater library diversity can enhance statistical rigor and increase the pool of loci that can be screened in a given experiment. However, the optimal library complexity for an experiment is dependent on the size of the experimental bottleneck, which can alter the TIS population structure stochastically and introduce noise. Thus, experimental and biological limitations may require use of libraries with relatively low complexity, for which gene-level analytical approaches are most appropriate. Furthermore, normalizations for bottleneck effects (and positional bias) are important steps before statistical analysis. Lastly, statistical methods that analyze TIS data at the insertion level (as opposed to combining reads per gene) can enable fine scale genomic analyses but are not appropriate unless library complexity is extremely high.

Finally, future TIS tool development and biological discovery will be aided by the centralization of insertion-level TIS data in an online database like the Sequencing Read Archive (SRA) database (<http://www.ncbi.nlm.nih.gov/sra>), as is currently done for whole-genome sequencing and transcriptomic data. While published TIS datasets often include read counts for each gene, insertion-level data is necessary to assess important quality metrics and is most compatible with other analysis methods, enabling additional information to be mined from the data (e.g. from intergenic regions). Thus, a repository of TIS data in the form of raw read files (such as FASTQ files), alignment files in which reads have been mapped to a genome of interest (e.g. BAM and SAM files), or even custom tables that include the number of reads mapped to every insertion site in the library will serve as an important resource and expand the utility of TIS-based approaches for discovery in microbiology.

Biography

Michael Chao received his Ph.D. in biomedical sciences from Harvard University in 2011. As a postdoctoral fellow in the laboratory of Matthew Waldor at Harvard Medical School and Brigham and Women's hospital, he studied epigenetic regulation in *Vibrio cholerae* using various genome-based tools. He currently serves as an editor at Nature Microbiology.

Sören Abel studied Biochemistry and Molecular Biology in Kiel, Germany. During his Ph.D. he worked on bacterial second messenger signaling at the University of Basel, Switzerland. He then moved to Harvard Medical School and Brigham and Women's Hospital, where he became interested in next-generation sequencing technologies and applied them to study *Vibrio cholerae* infections. In 2015, he started his own group as NCMM young associate investigator at the University of Tromsø, Norway and is also a Visiting Assistant Professor in Medicine at the Harvard Medical School. He and his group investigate the population dynamics of enteric pathogens: Sören Abel's homepage: https://en.uit.no/forskning/forskningsgrupper/gruppe?p_document_id=418147

Brigid Davis received her Ph.D. in Biology from MIT in 1998 for studies of the molecular consequences of extended trinucleotide repeats in non-coding RNA sequences associated with myotonic dystrophy. As a post-doctoral fellow in the Waldor lab, she investigated replication, regulation, and secretion of *Vibrio cholerae*'s toxin-encoding bacteriophage. Subsequently, she focused on sRNA-mediated regulation in *V. cholerae* and other bacterial species, as well as on *V. cholerae*'s σ^E -dependent outer membrane stress response pathway.

Matthew Waldor received his MD and PhD degrees from Stanford and subsequently trained in clinical infectious disease. He is currently a Professor of Medicine in the Division of Infectious Diseases at the Brigham & Women's Hospital/Harvard Medical School in Boston. His lab has studied the evolution and virulence of bacterial enteric pathogens for nearly 20 years. The chief goal of his laboratory is to enhance understanding of pathogen-host interactions, particularly in the context of animal models relevant to human infection. In recent years, his group has been keen to develop and implement new (primarily DNA sequencing-based) approaches for analyses of bacterial pathogenesis.

References

1. Barquist L, Boinett CJ, Cain AK. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biol.* 2013; 10
2. van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol.* 2013
3. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods.* 2009; 6:767–72. [PubMed: 19767758]
4. Goodman AL, et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe.* 2009; 6:279–89. [PubMed: 19748469]
5. Gawronski JD, Wong SM, Giannoukos G, Ward DV, Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci U S A.* 2009; 106:16422–7. [PubMed: 19805314]
6. Langridge GC, et al. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res.* 2009; 19:2308–16. [PubMed: 19826075]
7. Chiang SL, Rubin EJ. Construction of a mariner-based transposon for epitope-tagging and genomic targeting. *Gene.* 2002; 296:179–85. [PubMed: 12383515]
8. Rubin EJ, et al. In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria. *Proc Natl Acad Sci U S A.* 1999; 96:1645–50. [PubMed: 9990078]
9. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS. Tn5/IS50 target recognition. *Proc Natl Acad Sci U S A.* 1998; 95:10716–21. [PubMed: 9724770]
10. Zhang YJ, et al. Global assessment of genomic regions required for growth in *Mycobacterium tuberculosis*. *PLoS Pathog.* 2012; 8:e1002946. [PubMed: 23028335]

11. Chao MC, 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 Res.* 2013; 41:9033–48. [PubMed: 23901011]
12. Lodge JK, Weston-Hafer K, Berg DE. Transposon Tn5 target specificity: preference for insertion at G/C pairs. *Genetics.* 1988; 120:645–50. [PubMed: 2852135]
13. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mob DNA.* 2012; 3:3. [PubMed: 22313799]
14. Christen B, et al. The essential genome of a bacterium. *Mol Syst Biol.* 2011; 7:528. [PubMed: 21878915]
15. Barquist L, et al. A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. *Nucleic Acids Res.* 2013; 41:4549–64. [PubMed: 23470992]
16. DeJesus MA, et al. Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. *Bioinformatics.* 2013; 29:695–703. [PubMed: 23361328]
17. Lamichhane G, et al. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A.* 2003; 100:7213–8. [PubMed: 12775759]
18. Griffin JE, et al. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog.* 2011; 7:e1002251. [PubMed: 21980284]
19. Zemansky J, et al. Development of a mariner-based transposon and identification of *Listeria monocytogenes* determinants, including the peptidyl-prolyl isomerase PrsA2, that contribute to its hemolytic phenotype. *J Bacteriol.* 2009; 191:3950–64. [PubMed: 19376879]
20. Johnson CM, Grossman AD. Identification of host genes that affect acquisition of an integrative and conjugative element in *Bacillus subtilis*. *Mol Microbiol.* 2014; 93:1284–301. [PubMed: 25069588]
21. Shevchenko Y, et al. Systematic sequencing of cDNA clones using the transposon Tn5. *Nucleic Acids Res.* 2002; 30:2469–77. [PubMed: 12034835]
22. Troy EB, et al. Understanding barriers to *Borrelia burgdorferi* dissemination during infection using massively parallel sequencing. *Infect Immun.* 2013; 81:2347–57. [PubMed: 23608706]
23. Fu Y, Waldor MK, Mekalanos JJ. Tn-Seq analysis of *vibrio cholerae* intestinal colonization reveals a role for T6SS-mediated antibacterial activity in the host. *Cell Host Microbe.* 2013; 14:652–63. [PubMed: 24331463]
24. Pritchard JR, et al. ARTIST: High-Resolution Genome-Wide Assessment of Fitness Using Transposon-Insertion Sequencing. *PLoS Genet.* 2014; 10:e1004782. [PubMed: 25375795]
25. van Opijnen T, Camilli A. A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res.* 2012; 22:2541–51. [PubMed: 22826510]
26. Carter R, et al. Genomic analyses of pneumococci from children with sickle cell disease expose host-specific bacterial adaptations and deficits in current interventions. *Cell Host Microbe.* 2014; 15:587–99. [PubMed: 24832453]
27. Abel S, et al. Sequence tag-based analysis of microbial population dynamics. *Nat Methods.* 2015
28. Kaper JB, Morris JG Jr, Levine MM. Cholera. *Clin Microbiol Rev.* 1995; 8:48–86. [PubMed: 7704895]
29. Ritchie JM, Rui H, Bronson RT, Waldor MK. Back to the future: studying cholera pathogenesis using infant rabbits. *MBio.* 2010; 1
30. Kamp HD, Patimalla-Dipali B, Lazinski DW, Wallace-Gadsden F, Camilli A. Gene fitness landscapes of *Vibrio cholerae* at important stages of its life cycle. *PLoS Pathog.* 2013; 9:e1003800. [PubMed: 24385900]
31. Zomer A, Burghout P, Bootsma HJ, Hermans PW, van Hijum SA. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS One.* 2012; 7:e43012. [PubMed: 22900082]
32. Lazinski DW, Camilli A. Homopolymer tail-mediated ligation PCR: a streamlined and highly efficient method for DNA cloning and library construction. *Biotechniques.* 2013; 54:25–34. [PubMed: 23311318]

33. Gallagher LA, Shendure J, Manoil C. Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. *MBio*. 2011; 2:e00315–10. [PubMed: 21253457]
34. Yamaichi Y, et al. High-resolution genetic analysis of the requirements for horizontal transmission of the ESBL plasmid from *Escherichia coli* O104:H4. *Nucleic Acids Res*. 2014
35. Fu GK, et al. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc Natl Acad Sci U S A*. 2014; 111:1891–6. [PubMed: 24449890]
36. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci U S A*. 2012; 109:1347–52. [PubMed: 22232676]
37. Li H, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25:2078–9. [PubMed: 19505943]
38. Baugh L, et al. Combining functional and structural genomics to sample the essential *Burkholderia* structome. *PLoS One*. 2013; 8:e53851. [PubMed: 23382856]
39. Valentino MD, et al. Genes contributing to *Staphylococcus aureus* fitness in abscess- and infection-related ecologies. *MBio*. 2014; 5:e01729–14. [PubMed: 25182329]
40. Klein BA, et al. Identification of essential genes of the periodontal pathogen *Porphyromonas gingivalis*. *BMC Genomics*. 2012; 13:578. [PubMed: 23114059]
41. Deng J, Su S, Lin X, Hassett DJ, Lu LJ. A statistical framework for improving genomic annotations of prokaryotic essential genes. *PLoS One*. 2013; 8:e58178. [PubMed: 23520492]
42. Remmele CW, et al. Transcriptional landscape and essential genes of *Neisseria gonorrhoeae*. *Nucleic Acids Res*. 2014; 42:10579–95. [PubMed: 25143534]
43. 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; 14:303. [PubMed: 24103077]
44. Brutinel ED, Gralnick JA. Anomalies of the anaerobic tricarboxylic acid cycle in *Shewanella oneidensis* revealed by Tn-seq. *Mol Microbiol*. 2012; 86:273–83. [PubMed: 22925268]
45. Zhang YJ, et al. Tryptophan biosynthesis protects mycobacteria from CD4 T-cell-mediated killing. *Cell*. 2013; 155:1296–308. [PubMed: 24315099]
46. Mann B, et al. Control of virulence by small RNAs in *Streptococcus pneumoniae*. *PLoS Pathog*. 2012; 8:e1002788. [PubMed: 22807675]
47. McDonough E, Lazinski DW, Camilli A. Identification of in vivo regulators of the *Vibrio cholerae* xds gene using a high-throughput genetic selection. *Mol Microbiol*. 2014; 92:302–15. [PubMed: 24673931]
48. Moll A, et al. Cell separation in *Vibrio cholerae* is mediated by a single amidase whose action is modulated by two nonredundant activators. *J Bacteriol*. 2014; 196:3937–48. [PubMed: 25182499]
49. Dorr T, et al. A novel peptidoglycan binding protein crucial for PBP1A-mediated cell wall biogenesis in *Vibrio cholerae*. *PLoS Genet*. 2014; 10:e1004433. [PubMed: 24945690]
50. Skurnik D, et al. A comprehensive analysis of in vitro and in vivo genetic fitness of *Pseudomonas aeruginosa* using high-throughput sequencing of transposon libraries. *PLoS Pathog*. 2013; 9:e1003582. [PubMed: 24039572]
51. Turner KH, Everett J, Trivedi U, Rumbaugh KP, Whiteley M. Requirements for *Pseudomonas aeruginosa* acute burn and chronic surgical wound infection. *PLoS Genet*. 2014; 10:e1004518. [PubMed: 25057820]
52. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009; 10:57–63. [PubMed: 19015660]
53. Trapnell C, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. 2013; 31:46–53. [PubMed: 23222703]
54. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010; 11:R106. [PubMed: 20979621]
55. Khatiwara A, et al. Genome scanning for conditionally essential genes in *Salmonella enterica* Serotype Typhimurium. *Appl Environ Microbiol*. 2012; 78:3098–107. [PubMed: 22367088]

56. Santa Maria JP Jr. et al. Compound-gene interaction mapping reveals distinct roles for *Staphylococcus aureus* teichoic acids. *Proc Natl Acad Sci U S A*. 2014; 111:12510–5. [PubMed: 25104751]
57. Hsiao A, et al. Members of the human gut microbiota involved in recovery from *Vibrio cholerae* infection. *Nature*. 2014; 515:423–6. [PubMed: 25231861]
58. Moxon ER, Murphy PA. *Haemophilus influenzae* bacteremia and meningitis resulting from survival of a single organism. *Proc Natl Acad Sci U S A*. 1978; 75:1534–6. [PubMed: 306628]
59. Barnes PD, Bergman MA, Mecsas J, Isberg RR. *Yersinia pseudotuberculosis* disseminates directly from a replicating bacterial pool in the intestine. *J Exp Med*. 2006; 203:1591–601. [PubMed: 16754724]
60. Kaiser P, Slack E, Grant AJ, Hardt WD, Regoes RR. Lymph node colonization dynamics after oral *salmonella typhimurium* infection in mice. *PLoS Pathog*. 2013; 9:e1003532. [PubMed: 24068916]
61. Grant AJ, et al. Modelling within-host spatiotemporal dynamics of invasive bacterial disease. *PLoS Biol*. 2008; 6:e74. [PubMed: 18399718]

Box 1. Alternative TIS experiments

Although this Review focuses on the analysis of randomly generated high-density insertion libraries that are sequentially selected on different conditions, TIS studies can also be performed by comparing independently derived transposon libraries or using defined transposon libraries.

Comparison of independently derived transposon libraries made in wildtype and mutant backgrounds (i.e., single and double mutants, respectively) enables dissection of genetic interactions in pathways of interest^{3, 25, 48, 49}. For synthetic lethal interactions (in which disruptions in two loci lead to loss of bacterial viability), statistical methods that use pairwise comparisons should not be employed, as the insertion sites in one library are unlikely to be shared in the second. Nonlethal genetic interactions can readily be detected with an approach^{3, 25} for the calculation of mutants' fitness^{3, 25}, which permits comparisons of fitness between libraries and thus can reveal fitness effects limited to a particular genetic background.

Genome-wide screens can also be carried out using defined transposon libraries, which are made by pooling previously sequenced individual transposon mutants, typically with at least one insertion per dispensable gene²³. It has been shown⁴ that arrayed insertion mutants grown as clonal cultures can be pooled in specific patterns to maximize mutant multiplexing, then sequenced *en masse* to quickly identify the insertion site of each arrayed mutant; thus, independent sequencing of individual mutants is not essential for construction of defined libraries. Defined libraries provide several benefits. First, insertion mutants that might otherwise be rarely represented (e.g., insertions in small genes) can be included in a study. Second, defined libraries can cover more diverse genotypes in a smaller pool of cells, which can be preferable in systems containing severe experimental bottlenecks. Finally, because defined transposon libraries are derived from arrayed single insertion mutants, it is simple to retrieve and characterize specific mutants of interest.

Even so, the creation, sequencing and arraying of an ordered transposon library can be relatively time consuming and cost intensive, and may be impractical if multiple libraries (e.g. in different genetic backgrounds) are desired. Furthermore, because defined libraries tend to have few insertions (sometimes one) per gene, more experimental replicates are generally required to reach the statistical power obtained with saturated libraries. Additionally, single insertions may not be representative of true null mutations (e.g., insertions at the extreme N and C terminal coding regions may not be fully disruptive), which may lead to inaccurate gene classification.

Box 2. Bottlenecks in TIS experiments

In biology, a bottleneck—which is an event that constricts the size of a population—can alter population composition (a process known as ‘genetic drift’) by causing stochastic fluctuation in the abundance of certain genotypes or even their loss (see the figure, part A). Genetic drift can arise from both biological and technical sources. Technical bottlenecks can occur at every step of the TIS experiment, if sample sizes are too small to reflect the composition of the initial population. For example, genotypes may be missed during library amplification or insertion mutants may go undetected due to insufficient (sub-saturating) sequencing depth. Technical bottlenecks can be especially confounding when comparing two independently generated libraries that have different complexity. Biological bottlenecks, which result from processes inherent in the experimental system, often reflect bacterial death in adverse growth conditions (e.g. widespread bacterial killing by the host immune system) or limited resource availability (e.g. lack of niches due to competition with endogenous microflora⁵⁷). Biological bottlenecks are especially problematic in infection studies, since often only a small fraction of an inoculum may survive host barriers and successfully colonize a host⁵⁸⁻⁶¹.

TIS experiments aim to identify transposon mutants whose representation in a population changes in response to selection, but the stochastic experimental ‘noise’ arising from bottlenecks can mask these changes²⁴. The size of the bottleneck in *in vitro* TIS experiments can be determined through careful colony forming unit (CFU) enumeration, but this is difficult in more complex experiments such as host infection experiments (unless the infection model tested reproduces unrestrained exponential growth²⁶), as population constrictions are often masked by subsequent expansions²⁷. The presence and magnitude of host bottlenecks can be assessed using a population of differentially tagged organisms (e.g. containing sequence barcodes) that have equal fitness, so that changes in the population size and tag composition result solely from genetic drift^{27, 58-61}. TIS datasets themselves also contain information about bottlenecks: the number of unique insertion mutants recovered after selection defines a lower limit for bottleneck size, while the presence of a bottleneck impacting TIS data is reflected by loss of insertion mutants in non-essential genes in the recovered library²⁴⁻²⁶ (see the figure, part B; red line). In contrast, purely selective pressures will only eliminate insertions in the subset of genes that are conditionally essential (see the figure, part B; blue line). Thus, bottlenecks should be considered when designing TIS experiments and their impact should be factored into downstream analyses.

Figure 1. Transposon insertion sequencing (TIS) workflow

A. A high-density transposon insertion library containing multiple insertions in every non-essential genomic locus is created and then grown under different conditions (e.g. selective and non-selective), and mutants that are viable under each condition are recovered. The transposon junctions in the selected and non-selected pools are attached to sequencing adaptors and amplified, then subjected to high throughput sequencing. **B.** The sequences are mapped to the genome and the read counts at each insertion site are subjected to statistical analysis to define genomic loci that appear significantly underrepresented in the selective

growth condition (conditionally essential loci, blue box). The insertions in the non-selected library (Condition A) can also be statistically evaluated to define essential loci (orange box) that are required for growth under optimal conditions. In highly saturated libraries (orange graph), non-essential and essential genes are easily distinguishable; conversely, many non-essential genes are likely not to have been disrupted by chance in less saturated libraries (purple graph).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 2. Experimental considerations for TIS experiments

A. *V. cholerae* TIS DNA libraries were constructed either using an adaptor ligation¹¹ or arbitrary PCR protocol¹⁴. Plotting of sequenced reads from multiple libraries reveals a substantially lower correlation coefficient when an arbitrary PCR protocol is used. **B.** Plotting of the average read count per insertion in all non-essential genes of a *V. cholerae* TIS library (right peak from Figure 1, red graph) shows that read abundance has a non-normal distribution with a positive skew. **C.** Pools of reads were sampled from a sequenced *M. tuberculosis* TIS library⁴⁵ that was grown *in vitro* (input) or recovered from a mouse

infection (mouse output), and the number of unique transposon insertions detected in each pool was then plotted. At 500,000 reads (dotted line), >90% of all unique insertions were detected for each dataset, indicative of near-saturation sequencing. Figure adapted from Ref 24. **D.** Due to ongoing DNA replication, insertion sites (triangles) adjacent to the origin from a *V. cholerae* TIS dataset¹¹ have higher read counts than those near the terminus. Positional correction removes this bias (right). Figure adapted from Ref 11. The red arrows indicate the direction of chromosome replication. **E.** Annotation-dependent essential loci analysis methods compile all reads or insertion sites disrupted for a gene and determine whether fewer reads or insertion sites are represented compared to the genome-wide distribution. **F.** Sliding window analysis is one form of annotation-independent analysis. It compares reads in subgenic windows instead of genes, enabling the definition of domain essential genes. Hidden Markov-model (HMM)-based methods, which are also annotation-independent, predict essentiality/non-essentiality for individual insertion sites, which can subsequently form the basis for gene-level or subgenic predictions. *Differences in the distribution of reads (blue or orange bars) across windows due to experimental sources of noise may alter significance cutoffs (dashed lines) and consequently the classification of genes.