Approaches to querying bacterial genomes with transposon-insertion sequencing

Lars Barquist,1,2,* Christine J. Boinett,1 and Amy K. Cain1

1Wellcome Trust Sanger Institute; Hinxton, Cambridge, UK; 2EMBL-European Bioinformatics Institute; Hinxton, Cambridge, UK

Keywords: transposon mutagenesis, essential genes, sequencing, systems biology, bacteria, sRNA

In this review we discuss transposon-insertion sequencing, variously known in the literature as TraDIS, Tn-seq, INSeq and HITS. By monitoring a large library of single transposon-insertion mutants with high-throughput sequencing, these methods can rapidly identify genomic regions that contribute to organismal fitness under any condition assayable in the laboratory with exquisite resolution. We discuss the various protocols that have been developed and methods for analysis. We provide an overview of studies that have examined the reproducibility and accuracy of these methods, as well as studies showing the advantages offered by the high resolution and dynamic range of high-throughput sequencing over previous methods. We review a number of applications in the literature, from predicting genes essential for in vitro growth to directly assaying requirements for survival under infective conditions in vivo. We also highlight recent progress in assaying non-coding regions of the genome in addition to known coding sequences, including the combining of RNA-seq with high-throughput transposon mutagenesis.

Introduction

A common approach to identifying genomic regions involved in survival under a particular set of conditions is to screen large pools of mutants simultaneously. This can be done with defined mutants;1,2 however, the construction of defined mutant libraries is labor intensive and requires accurate genomic annotation, which can be particularly difficult to define for non-coding regions. An alternative to defined libraries is the construction and analysis of random transposon-insertion libraries. The original application of this method used DNA hybridization to track uniquely tagged transposon-insertions in Salmonella enterica serovar Typhimurium during the course of BALB/c mouse infection.3 DNA hybridization was eventually superseded by methods that used microarray detection of the genomic DNA flanking insertion sites, variously known as TraSH, MATT and DeADMAn (reviewed in ref. 4). However, these methods suffered from many of the problems microarrays generally suffer from: difficulty detecting low-abundance transcripts, mis-hybridization, probe saturation and difficulty identifying insertion sites precisely.

The application of high-throughput sequencing to the challenge of determining insertion location and prevalence solves many of these problems. Interestingly, the first application of transposon-insertion sequencing, developed by Hutchison et al., actually predates the development of microarray-based methods.5 However, this was applied to libraries of only approximately 1,000 transposon mutants in highly reduced Mycoplasma genomes, and the difficulty of sequencing at the time prevented wide spread adoption or high resolution. Modern high-throughput sequencing technology allows the methods discussed in this review to routinely monitor as many as one million mutants simultaneously in virtually any genetically tractable microorganism.

Protocols. Several methods were developed concurrently for high-throughput sequencing of transposon-insertion sites: TraDIS,6 INSeq,7 HITS8 and Tn-seq9 followed by Tn-seq Circle10 and refinements to the INSeq protocol.11 All of these protocols follow the same basic workflow with minor variations (see Fig. 1, Table 1): transposon mutagenesis and construction of pools of single insertion mutants; enrichment of transposon-insertion junctions and, finally, in some protocols a purification step either precedes or follows PCR enrichment before sequencing.

Transposon mutagenesis. Most studies have used either Tn5 or mariner transposon derivatives. Tn5 originated as a bacterial transposon which has been adapted for laboratory use. Large-scale studies have shown that Tn5, while not showing any strong preference for regional GC-content, do have a weak preference for a particular insertion motif.12-14 Transposon-insertion sequencing studies performed with Tn5 transposons in S. enterica serovars have reported a slight bias toward AT-rich sequence regions.5,15 However, this preference does not appear to be a major obstacle to analysis given the extremely high insertion densities obtained with this transposon6,15,16 (see Table 1). Additionally, Tn5 has been shown to be active in a wide range of bacterial species, though the number of transformants obtained can vary significantly depending on the transformation efficiency of the host.

Mariner/Himar1 transposons on the other hand originate from eukaryotic hosts and have an absolute requirement for TA bases at their integration site;17,18 with no other known bias besides a possible preference for bent DNA.17 This can be a disadvantage in that it limits the number of potential insertion sites, particularly in GC-rich sequence. However, this specificity can also be used in the prediction of gene essentiality in near-saturated libraries: as every potential integration site is known and the probability of
integration at any particular site can be assumed to be roughly equal, it is straight-forward to calculate the probability that any particular region lacks insertions by chance. Himar1 transposition can also be conducted in vitro in the absence of any host factors, and inserted transposons can then be transferred to the genomes of naturally transformable bacteria through homologous recombination. This can be advantageous when working with naturally transformable bacteria with poor electroportation efficiency. It is worth noting that Tn5 is also capable of transposition in vitro, and could potentially be used to increase insertion density and, hence, the resolution of the assay, particularly in GC-rich genomic regions.

Pool construction. Once mutants have been constructed, they are plated on an appropriate selective media for the transposon chosen, and colonies are counted, picked and pooled. A disadvantage of this is that the mutants must be recreated for follow up or validation studies. Goodman et al. introduced a clever way around this in the INSeq protocol: by individually archiving mutants, then sequencing combinatorial mutant pools, it is possible to uniquely characterize 2^n insertion mutants by sequencing only n pools. Each mutant is labeled with a unique binary string that indicates which pools it has been added to. These binary strings can then be reconstructed for each insertion observed in these pools by recording their presence or absence in sequencing data, providing a unique pattern relating insertions to archived mutants. The authors control false identifications due to errors in sequencing by requiring that each binary label have a minimum edit distance to every other label, allowing for a robust association of labels with insertions despite sometimes noisy sequencing data. As a proof of concept, the authors were able to identify over 7,000 Bacteroides thetaiotaomicron mutants from only 24 sequenced pools. This effectively uses methods for the generation of random transposon pools to rapidly generate defined mutant arrays, though it is heavily dependent on liquid-handling robotics.

Enrichment of transposon-insertion junctions. Once pools have been constructed, they are grown in either selective or permissive conditions, depending on the experiment, and then genomic DNA is extracted. Fragmentation proceeds either through restriction digestion in the case of transposons modified to contain appropriate sites or via physical shearing, then sequencing adapters are ligated to the resulting fragments. PCR is performed on these fragments using a transposon-specific primer and a sequencing adaptor-specific primer to enrich for fragments spanning the transposon-genomic DNA junction. Some protocols purify fragments containing transposon insertions using biotinylated primers or PAGE before and/or after PCR enrichment. The purification step from the Tn-seq Circle protocol is particularly unusual in that restriction-digested fragments containing transposon sequence are circularized before being treated with an exonuclease that digests all fragments without transposon insertions, theoretically completely eliminating background. Given the success of protocols that do not include a purification step and the lack of systematic comparisons, it is currently unclear whether including one provides any major advantages.

Reproducibility, Accuracy, and Concordance with Previous Methods

A number of studies have looked at the reproducibility of transposon-insertion sequencing. Multiple studies using different protocol variations have repeatedly shown extremely high reproducibility in the number of insertions per gene (correlations of
| Study                      | Publication date | Organism                  | Total mutants | Unique insertion sites; Density | Application                                      | Tn used | Name coined                |
|----------------------------|------------------|---------------------------|---------------|---------------------------------|-------------------------------------------------|---------|---------------------------|
| Hutchison III et al., Science | Oct 1999         | M. genitalium/M. pneumonia | 1,291         | 685; ~1/850 bp                  | Required gene sets                               | Tn4001  | Global Transposon Mutagenesis |
| Goodman et al., Cell Host and Microbe | Sep 2009      | B. thetaiotaomicron       | 2 x 35,000    | 35,000; 1/182 bp                | Establishment in human gut as a natural habitat | Mariner | INSeq                     |
| Gawronski et al., PNAS   | Sep 2009         | H. influenzae             | 75,000        | 55,935; 1/32 bp                 | Prolonged survival in lung in vivo               | Mariner | HITS                      |
| van Opijnen et al., Nature Methods | Oct 2009       | S. pneumoniae             | 6 x 25,000    | 23,875; 1/91 bp                 | Transcriptional regulation and carbohydrate transport | Mariner | Tn-seq                    |
| Langridge et al., Genome Research | Dec 2009      | S. Typhi                  | 1.1 million   | 370,000; 1/13 bp                | Gene requirements, bile tolerance                | Tn5     | TraDIS                    |
| Gallagher et al., mBio    | Jan 2011         | P. aeruginosa             | ~100,000      | 95,905; 1/65 bp                 | Tobramycin resistance                             | Mariner | Tn-seq (circle method)    |
| Eckert et al., J Bact.   | Jan 2011         | E. coli                   | ~19 x 95      | 1,645                           | Colonization of bovine intestinal tract; retrospective re-evaluation of a STM study | Tn5     | -                         |
| Christen et al., Mol Syst Biol | July 2011    | C. crescentus             | 800,000       | 428,735; 1/8 bp                 | Gene/ncRNAs/promoter requirements                | Tn5     | -                         |
| Griffin et al., PLOS Pathogens | Sept 2011   | M. tuberculosis           | 2 x 100,000   | 36,488; 1/120 bp                | Gene requirements and cholesterol utilization    | Mariner | -                         |
| Khatiwara et al., AEM     | May 2012         | S. Typhimurium            | 16,000        | ~8,000; ~1/610                   | Bile, low nutrient and heat tolerance             | Tn5     | -                         |
| Mann et al., PLOS Pathogens | July 2012    | S. pneumoniae             | ~9,000–24,000 | ~8,000–22,000                   | Determining roles of sRNAs in pathogenesis       | Mariner | -                         |
| van Opijnen and Camilli, Genome Research | July 2012     | S. pneumoniae             | ~4,000–30,000 | Varying                         | Stress response and metabolism in vitro and murine in vivo colonization | Mariner | -                         |
| Brutinel and Granick, Molecular Microbiology | Aug 2012      | S. oneidensis             | 50,000        | 26,793; ~1/191 bp               | Gene requirements and Metabolism                 | Mariner | -                         |
| Zhang et al., PLOS Pathogens | Sept 2012   | M. tuberculosis           | 2 x 100,000   | 36,488; ~1/120 bp               | Identifying genes, regulators and ncRNAs required for growth | Mariner | -                         |
| Klein et al., BMC Genomics | Oct. 2012   | P. gingivalis             | N/A           | 54,000; 1/43 bp                 | Gene requirements                                 | Mariner | -                         |
| Pickard et al., J Bact.   | Jan 2013        | S. Typhi                  | 1.1 million   | 370,000; 1/13 bp                | Bacteriophage infection                           | Tn5     | -                         |
| Barquist et al., Nuc. Acids Res | March 2013  | S. Typhi                  | 1.1 million   | 370,000; 1/13 bp                | Comparison of coding and non-coding gene requirements between two Salmonella serovars | Tn5     | -                         |

1The same mutant library was used in these studies. A collection of studies to date utilizing transposon-insertion sequencing. Columns: (1) study reference, (2) date of publication, (3) organism mutagenized, (4) number of mutants generated, (5) number of unique insertion sites recovered from sequencing, (6) brief description of the application, (7) transposon used, (8) method name coined, if any.
-90%) in replicates of the same library grown and sequenced independently,7,9,10 and good reproducibility (correlations between 70–90%) in independently constructed non-saturated libraries.9,22 Van Oprijen and Camilli23 compared traditional 1 × 1 competition experiments between wild-type and mutant Streptococcus pneumoniae to results obtained by transposon-insertion sequencing and showed that there was no significant difference in results over a range of tested conditions.

The accuracy of transposon-insertion sequencing in determining library composition has also been assessed. Zhang et al. constructed a library of identified transposon-insertion mutants in known relative quantities, and then were able to recover the relative mutant prevalence with transposon-insertion sequencing.23 Additionally, by estimating the number of PCR templates prior to enrichment, this study showed that there is a high correlation between enrichment input and sequencing output.

Two studies have evaluated concordance between results obtained with transposon-insertion sequencing and microarray monitoring of transposon insertions in order to demonstrate the enhanced accuracy and dynamic range of sequencing over previous methods. In the first, 19 libraries of 95 enterohemorrhagic Escherichia coli (EHEC) transposon mutants that had previously been screened in cattle using signature-tagged mutagenesis (STM), were pooled and re-evaluated using the TraDIS protocol.24 The original STM study had identified 13 insertions in 11 genes attenuating intestinal colonization in a type III secretion system located in the locus of enterocyte effacement (LEE).25 By applying sequencing to the same samples, an additional 41 mutations in the LEE were identified, spanning a total of 21 genes. Additional loci outside the LEE, which have been previously implicated in intestinal colonization but had not been detected by STM, were also reported by TraDIS.

The second study re-evaluated genes required for optimal growth determined by TraSH in Mycobacterium tuberculosis.26,27 The greater dynamic range of sequencing as compared with microarrays allowed easier discrimination between insertions that were nonviable and those that were only significantly underrepresented. The authors estimate that genes called as required by sequencing in their study are at least 100-fold underrepresented in the pool. In comparison, the threshold in the previous microarray experiment reported genes that had log probe ratios at least 5-fold lower than average between transposon-flanking DNA hybridization and whole genomic DNA hybridization. Additionally, the nucleotide-resolution of insertion sequencing allowed the authors to identify genes which had required regions, likely corresponding to required protein domains,23 but which tolerated insertions in other regions. Altogether, the authors increase the set of genes predicted to be required for growth in laboratory conditions in M. tuberculosis by more than 25% (from 614 to 774).

**Gene Requirements**

The earliest application of transposon-insertion sequencing was to determine the minimal set of genes necessary for the survival of Mycoplasma.7 This essential genome is of great interest to synthetic and systems biology where it is seen as a foundation for engineering cell metabolism, and in infection biology and medicine where it is seen as a promising target for therapies. However, it is important to remember that “essentiality” is always relative to growth conditions: a biosynthetic gene that is non-essential in a growth medium supplying a particular nutrient may become essential in a medium that lacks it. Traditionally, gene essentiality has been determined in clonal populations;1,28,29 since the high-throughput transposon sequencing protocols described here necessarily contain a short period of competitive growth before DNA extraction, many of these studies prefer to refer to the “required” genome for the particular conditions under evaluation.

Because of this short period of competitive growth, and because many otherwise required genes tolerate insertions in their terminus7,27,30 or outside essential domains,23 the determination of required genomic regions is not completely straightforward and a number of approaches have been taken to counter this. These include only calling genes completely lacking insertions as required,4 determining a cut-off based on the empirical or theoretical distribution of gene-wise insertion densities.6,15,27,30 Additionally, windowed methods have been developed which can be used to identify essential regions in the absence of gene annotation,23,31 and have had success in identifying required protein domains, promoter regions and non-coding RNAs (ncRNAs). The organisms that have been evaluated for gene requirements under standard laboratory conditions are summarized in Table 1.

In agreement with previous studies,1,28 many required genes identified by transposon-insertion sequencing are involved in fundamental biological processes such as cell division, DNA replication, transcription and translation,6,7,15,27 and many of these requirements appear to be conserved between genera and classes.15,36 However, a recent study defining required gene sets in Salmonella serovars has found that phage repressors, necessary for maintaining the lysogenic state of the prophage, are also required,37 even though mobile genetic elements such as phage are usually considered part of the accessory genome. This study also highlights the need for temperance when interpreting the results of high-throughput assays of gene requirements. For example, many genes in Salmonella Pathogenicity Island 2 (SPI-2) did not exhibit transposon-insertions, despite clear evidence from directed knockouts showing that these genes are non-essential for viability or growth. Under laboratory conditions, SPI-2 is silenced by the nucleoid-forming protein H-NS,32,33 which acts by oligomerizing along silenced regions of DNA blocking RNA polymerase access. A previous study has shown that transposon insertion “cold spots” can be caused by competition between high-density proteins and transposases for DNA.34 This suggests that H-NS may be restricting transposase access to DNA, though this has not previously been observed in transpo-son-insertion sequencing data, and will require additional work to confirm.

**Defining Conditional Gene Requirements**

One of the most valuable applications of the transposon-insertion sequencing method is the ability to identify genes essential in
a condition of interest, by comparing differences in the numbers of sequencing reads from input (control) mutant pools to output (test) pools that have been subject to passaging in a certain growth condition. Insertion counts are compared from cells in the input pool and those after passage, thereby identifying genes that either enhance or detract from survival and/or growth in the given condition, defined by decreased or increased insertion frequency, respectively. A further application of this method involves comparing insertions between biologically linked conditions, such as cellular stresses or different stages of a murine infection, to gain insight into complex systems.\(^{22}\)

So far, transposon-insertion sequencing has been used to investigate a number of interesting biological questions: bile tolerance in *S. Typhi*\(^{2}\) and *S. Typhimurium*,\(^{36}\) bacteriophage infection of *S. Typhi*,\(^{36}\) antibiotic resistance in *Pseudomonas aeruginosa*,\(^{10}\) cholesterol utilization in *M. tuberculosis*\(^{27}\) and a number of stress and nutrient conditions in *S. pneumoniae*.\(^{22}\) Transposon-insertion sequencing of populations passed through murine models have been used to assess genes required to establish the gut commensal *B. thetaiotaomicron* in its niche,\(^{7}\) for *Hemophilus influenzae* infection,\(^{8}\) as well as *S. pneumoniae* responses to two in vivo niches, the lung and nasopharynx.\(^{22}\) A further extension of the method examined double mutant libraries, that is transposon mutant libraries generated in a defined deletion background, to tease apart complex networks of regulatory genes.\(^{9}\)

Two studies in particular illustrate the power of using transposon-insertion sequencing to identify conditionally required genes. In the first, Goodman et al. set out to determine the genes necessary for the establishment of the commensal *B. thetaiotaomicron* in a murine model.\(^{7}\) First, the growth requirements of transposon mutant populations in the cecum of germ-free mice was assessed, and genes required for growth in monoassociation with the host were found to be enriched in functions such as energy production and amino acid metabolism. By further comparing monoassociated transposon mutant libraries with those grown in the presence of three defined communities of human gut-associated bacteria, the authors identified a locus upregulated in the absence of genome annotation. Twenty-seven small RNAs (sRNAs) had previously been detected in *C. crescentus*;\(^{38}\) six were found to be depleted in transposon insertions indicating an important role in basic cellular processes. Additionally, the well-characterized ncRNAs tmRNA and RNaseP, as well as 29 non-redundant tRNAs, were found to be required. An additional 90 unannotated non-disruptable regions were identified throughout the genome, implying an abundance of unexplored functional non-coding sequence.

While the non-coding transcripts of *M. tuberculosis* have been explored more thoroughly than those of *C. crescentus*, most remain functionally uncharacterized, though there are hints that some of these may be involved in pathogenicity.\(^{39}\) Using a marinier transposon-based assay and a windowed statistical analysis that accounted for the distribution of potential TA integration sites, 35 intergenic regions were identified as putatively required in the *M. tuberculosis* genome.\(^{23}\) In common with the *C. crescentus* study, the RNA component of RNase P, required for the maturation of tRNAs and tmRNA, involved in the freeing of stalled ribosomes, were identified as required (Fig. 2A) together with 10 non-redundant tRNAs and potential promoter regions. However, due to the lower overall insertion density and lack of TA sites in some GC-rich regions, there were some regions that could not be assayed and the resolution was limited to 250 bases.

A recent study has examined ncRNA requirements in the *S. enterica* serovars Typhi and Typhimurium.\(^{35}\) Using the tRNAs as a model set of ncRNAs, this study showed that the high transposon insertion density achieved by the TraDIS protocol is capable of assaying the requirement for genomic regions as small as 70–80 bases. *S. enterica*, together with the closely related *E. coli*, has served as a model organism for the discovery and elucidation of ncRNA function, and extensive annotations of non-coding

---

**Monitoring ncRNA Contributions to Fitness**

To date, four studies have used transposon-insertion sequencing to examine the contribution of non-coding RNAs (ncRNAs) and other non-coding regions to organismal fitness (see Table 1). Two of these examined requirements for non-coding regions in the relatively under-explored bacterial species *C. crescentus*\(^{38}\) and *M. tuberculosis*.\(^{23}\) Both utilized analytical techniques that allowed for the identification of putative required regions in the absence of genome annotation. Twenty-seven small RNAs (sRNAs) had previously been detected in *C. crescentus*;\(^{38}\) six were found to be depleted in transposon insertions indicating an important role in basic cellular processes. Additionally, the well-characterized ncRNAs tmRNA and RNaseP, as well as 29 non-redundant tRNAs, were found to be required. An additional 90 unannotated non-disruptable regions were identified throughout the genome, implying an abundance of unexplored functional non-coding sequence.

While the non-coding transcripts of *M. tuberculosis* have been explored more thoroughly than those of *C. crescentus*, most remain functionally uncharacterized, though there are hints that some of these may be involved in pathogenicity.\(^{39}\) Using a marinier transposon-based assay and a windowed statistical analysis that accounted for the distribution of potential TA integration sites, 35 intergenic regions were identified as putatively required in the *M. tuberculosis* genome.\(^{23}\) In common with the *C. crescentus* study, the RNA component of RNase P, required for the maturation of tRNAs and tmRNA, involved in the freeing of stalled ribosomes, were identified as required (Fig. 2A) together with 10 non-redundant tRNAs and potential promoter regions. However, due to the lower overall insertion density and lack of TA sites in some GC-rich regions, there were some regions that could not be assayed and the resolution was limited to 250 bases.

A recent study has examined ncRNA requirements in the *S. enterica* serovars Typhi and Typhimurium.\(^{35}\) Using the tRNAs as a model set of ncRNAs, this study showed that the high transposon insertion density achieved by the TraDIS protocol is capable of assaying the requirement for genomic regions as small as 70–80 bases. *S. enterica*, together with the closely related *E. coli*, has served as a model organism for the discovery and elucidation of ncRNA function, and extensive annotations of non-coding
sequencing. To identify sRNA loci, the authors first sequenced size-selected RNA from the wild-type and three two-component system knockouts, identifying 89 putative sRNAs, 56 of which were novel. Fifteen of these candidates, selected on the basis of high expression and low predicted folding free energy, were assayed for their ability to establish invasive disease in a murine model. Of these, eight sRNA deletions showed a significant attenuation of disease. To more broadly establish the roles of sRNAs in infecting particular organs, transposon insertion libraries were administered directly to the nasopharynx, lungs or blood of mice, and bacteria were harvested following disease progression. Twenty-six, 28 and 18 sRNAs were found to attenuate infection in the nasopharynx, lung and blood, respectively. These results were then validated with targeted deletions of 11 transcripts are available. As a result, this study was able to assay approximately 300 non-coding regions with evidence for function or transcription. Among the ncRNAs identified as required were RNase P; the RNA component of the signal recognition particle, involved in targeting proteins to the plasma membrane; and a number of known autoregulatory ribosomal protein leader sequences, as well as providing evidence for a novel leader sequence, StyR-8, which appears to be involved in the auto-regulation of the rpmB gene. In total, this study identified 15 confirmed and putative ncRNAs required for robust competitive growth on rich media in both serovars, including a number of known sRNAs involved in stress response.

A particularly exciting study has been conducted in S. pneumoniae TIGR4 combining RNA-seq with transposon-insertion sequencing. To identify sRNA loci, the authors first sequenced size-selected RNA from the wild-type and three two-component system knockouts, identifying 89 putative sRNAs, 56 of which were novel. Fifteen of these candidates, selected on the basis of high expression and low predicted folding free energy, were assayed for their ability to establish invasive disease in a murine model. Of these, eight sRNA deletions showed a significant attenuation of disease. To more broadly establish the roles of sRNAs in infecting particular organs, transposon insertion libraries were administered directly to the nasopharynx, lungs or blood of mice, and bacteria were harvested following disease progression. Twenty-six, 28 and 18 sRNAs were found to attenuate infection in the nasopharynx, lung and blood, respectively. These results were then validated with targeted deletions of 11

Figure 2. Applications of transposon-insertion sequencing to non-coding RNAs. (A) Plots of genomic regions in Mycobacterium tuberculosis containing the required non-coding RNAs RNase P (top) and tmRNA (bottom). Tracks, from top to bottom, 1. Histogram of insertion counts, 2. Comprehensive heat-map of requirement of 500-bp windows, 3. Position of annotated genes, 4. Position of TA dinucleotide sites, 5. Position of non-coding RNA. Reproduced from reference 23. (B) 1 × 1 competition assays validate attenuating Streptococcus pneumoniae sRNA mutants identified by transposon-insertion sequencing. Mice were infected with defined deletions of sRNAs identified as attenuating by Tn-seq and wild-type S. pneumoniae TIGR4 at the body site indicated and bacterial densities were compared 24 h post-infection. These plots show the derived competitive index in blood (top) and the nasopharynx (bottom). Each point represents the result of a competition experiment between an sRNA deletion mutant and wild-type TIGR4. A competitive index of 1 indicates equivalent numbers of mutants and wild-type were recovered. Modified from reference 46.
sRNAs (Fig. 2B). In addition to establishing the role of sRNAs in *S. pneumoniae* virulence, this study illustrated the power of combining RNA-seq and transposon-insertion sequencing to rapidly assign phenotypes to non-coding sequences.

**Limitations.** In this review, we have largely focused on the potential of transposon-insertion sequencing. However, this technology does have a number of important limitations, which we collect here and summarize in Table 2. As discussed previously, requirements for particular nucleotides at insertion sites, such as the TA required by the Mariner transposon, or preference for certain sequence composition, such as the AT bias exhibited by Tn5, can limit the density of observed insertions in certain genomic regions. This may impact any downstream analysis, and can potentially bias results, particularly the determination of gene requirements. Even if this bias has been accounted for, transposon-insertion screens will always over-predict gene requirements in comparison to targeted deletion libraries as discussed previously. However, this over-prediction can be controlled either through careful consideration of known insertion biases as in many Mariner-based studies, or by high-insertion densities, such as those achieved in several Tn5-based studies (Table 1). Once the library has been created, only regions that have accumulated insertions in the conditions of library creation will be able to be assayed for fitness effects in further conditions. This means that regions that lead to slow growth phenotypes when disrupted in standard laboratory conditions may be difficult to assay in other conditions. Additionally, the dynamic range of fitness effects detected will depend on the complexity of the input library(s). The absence of insertions may be a particular problem for assaying small genomic elements, such as sRNAs or short ORFs. Finally, the validation of hypotheses derived from transposon-insertion sequencing will require the construction of targeted deletions, as individual mutants cannot be recovered from pools unless specialized protocols have been followed during library construction (as in Goodman, 2009).

**Table 2. Advantages and limitations of transposon-insertion sequencing**

| Advantages | Limitations |
|------------|-------------|
| Library construction is extremely rapid in comparison to targeted deletion libraries. | Requirements for particular nucleotides at transposon-insertion sites or insertion biases can limit resolution. |
| Gene requirements and fitness effects can be quickly assayed in a wide range of conditions. | Determination of gene essentiality is dependent on insertion density, and is less conclusive than targeted gene deletion in clonal populations. |
| The precise location of transposon insertions can be determined due to the nucleotide resolution of high-throughput sequencing. | Only genomic regions that tolerate insertions under the conditions of library creation may be assayed for fitness effects in further conditions. |
| Wide dynamic range compared with older microarray-based technologies. | The dynamic range for fitness effects is dependent on mutant abundance in the initial library and may be limiting for some genes. |
| Requirements and fitness effects of genomic regions can be determined independently of annotation. | Mutants must be reconstructed for follow-up experiments in the absence of specialized protocols and robotics (see e.g., Goodman, 2009). |

The combination of transposon-insertion sequencing with other high-throughput and computational methods is already proving to be fertile ground for enhancing our understanding of bacterial systems. For instance, by using transposon-insertion sequencing in a collection of relatively simple conditions combined with a computational pathway analysis, van Opijnen and Camilli were able to provide a holistic understanding of the genetic subsystems involved in a complex process such as *S. pneumoniae* pathogenesis. In the future, methods to assay phenotype in a high-throughput manner may be combined with transposon-insertion sequencing to provide exhaustive simple genotype-phenotype associations with which to understand complex processes in a systems biology framework. We look forward to the adoption of these data sets by the community as an important tool for rapid hypothesis generation.
Acknowledgments

This work was supported by the Wellcome Trust (grant number WT098051). CJB and AKC were supported by the Medical Research Council (grant number S1710).
41. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 2013; 41(Database issue):D226-32; PMID:23125362; http://dx.doi.org/10.1093/nar/gks1005.
42. Raghavan R, Goeke EA, Ochman H. Genome-wide detection of novel regulatory RNAs in E. coli. Genome Res 2011; 21:1487-97; PMID:21665928; http://dx.doi.org/10.1101/gr.119370.110.
43. Chinni SV, Raabe CA, Zakaria R, Randau G, Hau CH, Zemann A, et al. Experimental identification and characterization of 97 novel npcRNA candidates in Salmonella enterica serovar Typhi. Nucleic Acids Res 2010; 38:5893-908; PMID:20460466; http://dx.doi.org/10.1093/nar/gkq281.
44. Sirtha A, Sharma CM, Rolle K, Vogel J. Deep sequencing of Salmonella RNA associated with heterologous Hfq proteins in vivo reveals small RNAs as a major target class and identifies RNA processing phenotypes. RNA Biol 2009; 6:266-75; PMID:19353007; http://dx.doi.org/10.4161/rna.6.3.8332.
45. Fu Y, Desirio-Haggar K, Anthony J, Meyer MM. Most RNAs regulating ribosomal protein biosynthesis in Escherichia coli are narrowly distributed to Gammaproteobacteria. Nucleic Acids Res 2013; 41:5491-503; PMID:23396277; http://dx.doi.org/10.1093/nar/gkt955.
46. Mann B, van Opipjen T, Wang J, Obert C, Wang YD, Carter R, et al. Control of virulence by small RNAs in Streptococcus pneumoniae. PLoS Pathog 2012; 8:e1002788; PMID:22807675; http://dx.doi.org/10.1371/journal.ppat.1002788.
47. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, Jansen R, et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature 1999; 402:413-8; PMID:10586881; http://dx.doi.org/10.1038/46558.
48. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 2009; 462:1056-60; PMID:20035048; http://dx.doi.org/10.1038/nature08656.
49. Bochner BR. Global phenotypic characterization of bacteria. FEMS Microbiol Rev 2009; 33:191-205; PMID:19054113; http://dx.doi.org/10.1111/j.1574-6976.2008.00149.x.
50. Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chab M, et al. Phenotypic landscape of a bacterial cell. Cell 2011; 144:143-56; PMID:21185072; http://dx.doi.org/10.1016/j.cell.2010.11.052.
51. Brutinel ED, Gralnick JA. Anomalies of the anaerobic tricarboxylic acid cycle in Shewanella oneidensis revealed by Tn-seq. Mol Microbiol 2012; 86:273-83; PMID:22925268; http://dx.doi.org/10.1111/j.1365-2958.2012.08196.x.
52. Klein BA, Tenorio EL, Lainez DW, Camilli A, Duncan MJ, Hu LT. Identification of essential genes of the periodontal pathogen Porphyromonas gingivalis. BMC Genomics 2012; 13:578; PMID:22185072; http://dx.doi.org/10.1111/j.1574-6976.2008.00149.x.
53. Raghavan R, Goeke EA, Ochman H. Genome-wide detection of novel regulatory RNAs in E. coli. Genome Res 2011; 21:1487-97; PMID:21665928; http://dx.doi.org/10.1101/gr.119370.110.
54. Chinni SV, Raabe CA, Zakaria R, Randau G, Hau CH, Zemann A, et al. Experimental identification and characterization of 97 novel npcRNA candidates in Salmonella enterica serovar Typhi. Nucleic Acids Res 2010; 38:5893-908; PMID:20460466; http://dx.doi.org/10.1093/nar/gkq281.
55. Sirtha A, Sharma CM, Rolle K, Vogel J. Deep sequencing of Salmonella RNA associated with heterologous Hfq proteins in vivo reveals small RNAs as a major target class and identifies RNA processing phenotypes. RNA Biol 2009; 6:266-75; PMID:19353007; http://dx.doi.org/10.4161/rna.6.3.8332.