AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

Comprehensive Essentiality Analysis of the *Mycobacterium tuberculosis* Genome via Saturating Transposon Mutagenesis

Michael A. DeJesus,[a] Elias R. Gerrick,[b] Weizhen Xu,[c] Sae Woong Park,[c] Jarukit E. Long,[d] Cara C. Boutte,[b] Eric J. Rubin,[b] Dirk Schnappinger,[c] Sabine Ehrt,[c] Sarah M. Fortune,[b] Christopher M. Sassetti,[d,e] Thomas R. Ioerger[a]

Department of Computer Science and Engineering, Texas A&M University, College Station, Texas, USA[a]; Department of Immunology and Infectious Diseases, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA[b]; Department of Microbiology and Immunology, Weill Cornell Medical College, New York, New York, USA[c]; Department of Microbiology and Physiological Systems, University of Massachusetts Medical School, Worcester, Massachusetts, USA[d]; Howard Hughes Medical Institute, Chevy Chase, Maryland, USA[e]

ABSTRACT For decades, identifying the regions of a bacterial chromosome that are necessary for viability has relied on mapping integration sites in libraries of random transposon mutants to find loci that are unable to sustain insertion. To date, these studies have analyzed subsaturated libraries, necessitating the application of statistical methods to estimate the likelihood that a gap in transposon coverage is the result of biological selection and not the stochasticity of insertion. As a result, the essentiality of many genomic features, particularly small ones, could not be reliably assessed. We sought to overcome this limitation by creating a completely saturated transposon library in *Mycobacterium tuberculosis*. In assessing the composition of this highly saturated library by deep sequencing, we discovered that a previously unknown sequence bias of the *Himar1* element rendered approximately 9% of potential TA dinucleotide insertion sites less permissible for insertion. We used a hidden Markov model of essentiality that accounted for this unanticipated bias, allowing us to confidently evaluate the essentiality of features that contained as few as 2 TA sites, including open reading frames (ORF), experimentally identified noncoding RNAs, methylation sites, and promoters. In addition, several essential regions that did not correspond to known features were identified, suggesting uncharacterized functions that are necessary for growth. This work provides an authoritative catalog of essential regions of the *M. tuberculosis* genome and a statistical framework for applying saturating mutagenesis to other bacteria.

IMPORTANCE Sequencing of transposon-insertion mutant libraries has become a widely used tool for probing the functions of genes under various conditions. The *Himar1* transposon is generally believed to insert with equal probabilities at all TA dinucleotides, and therefore its absence in a mutant library is taken to indicate biological selection against the corresponding mutant. Through sequencing of a saturated Himar1 library, we found evidence that TA dinucleotides are not equally permissive for insertion. The insertion bias was observed in multiple prokaryotes and influences the statistical interpretation of transposon insertion (TnSeq) data and characterization of essential genomic regions. Using these insights, we analyzed a fully saturated TnSeq library for *M. tuberculosis*, enabling us to generate a comprehensive catalog of *in vitro* essentiality, including ORFs smaller than those found in any previous study, small (noncoding) RNAs (sRNAs), promoters, and other genomic features.

Received 29 November 2016 **Accepted** 7 December 2016 **Published** 17 January 2017

Citation DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, Rubin EJ, Schnappinger D, Ehrt S, Fortune SM, Sassetti CM, Ioerger TR. 2017. Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. mBio 8:e02133-16. https://doi.org/10.1128/mBio.02133-16.

Editor Christina L. Stallings, Washington University in St. Louis School of Medicine

Address correspondence to Thomas R. Ioerger, ioerger@cs.tamu.edu.

M.A.D. and E.R.G. contributed equally to this article.

This article is a direct contribution from a Fellow of the American Academy of Microbiology. External solicited reviewers: Colin Manoil, University of Washington; David Lampe, Duquesne University.

Deep sequencing of transposon (Tn) insertion (TnSeq) libraries has become a powerful tool for evaluating the essentiality of genomic features in bacterial organisms. Random Tn insertions can disrupt the functions of genes and regulatory regions. The consequent effects on growth can be efficiently quantified by amplifying and sequencing the transposon-chromosome junctions from a complex library, a technique known as TnSeq (1–3). Comparative analysis of genes that are essential under different environmental conditions has been used to dissect specific metabolic pathways (4) and mechanisms of pathogenesis (5, 6). Similarly, comparing the effects of Tn insertion in different genetic backgrounds can reveal genetic interactions that imply functional relationships between genes (7–9).

Comparative analyses to identify conditionally essential genes are fairly robust because they require only the accurate estimation of each mutant's relative abundance in two libraries (10–12). In contrast, identifying the primary functions necessary for growth even under favorable *in vitro* conditions represents a greater statistical challenge. These analyses rely on the characterization of a single transposon library to identify regions that are devoid of transposon insertions and therefore likely to encode functions that are necessary for growth (6, 13). Many of these procedures rely on the *Himar1* transposon, which is used because it is thought to lack sequence specificity except for the required TA dinucleotide insertion site (14). Previous analyses of *Himar1* libraries of *Mycobacterium tuberculosis* suggest that there are approximately 600 genes that are essential for growth of this bacterium in standard laboratory medium (4, 15). This number represents approximately 15% of this organism's genomic content, which is consistent with estimates from studies of other bacteria with similarly sized genomes (16). The identified functions include many well-known housekeeping genes involved in DNA replication, protein translation, cell growth and division, and core metabolic pathways. There are several widely used reference TnSeq data sets for *M. tuberculosis*, including those of Griffin et al. (4) and Zhang et al. (6). The essential genes identified in these studies are broadly consistent with the original hybridization-based technique (known as "TraSH" [15]). However, there are differences between the studies in the predicted essentiality of some genes. Some of these discrepancies may reflect differences in both growth conditions and methodology, such as sequencing depth (number of reads) and integration of barcodes to reduce the effect of PCR bias (17).

A less-appreciated limitation of the published TnSeq studies of *M. tuberculosis* is that they relied on libraries that were saturated only moderately (50% to 60%). In this situation, there is a relatively high probability that a nonessential region would be devoid of insertions due to chance rather than selection. As a result, the published TnSeq analyses had difficulty confidently assessing the essentiality of small genes with only a few TA sites. For instance, no genes with fewer than 6 TA sites were classified as essential by Griffin et al. (4) or Zhang et al. (6), despite the fact that there are 538 small open reading frames (ORFs) with 1 to 5 TA sites in the *M. tuberculosis* genome. Hence, ~13% of the ORFs of *M. tuberculosis* (538 of 3,990 genes) were effectively excluded from TnSeq profiling in those prior studies.

Low saturation levels have, to some extent, limited the statistical power of all previous attempts to define essential regions of a bacterial genome. As the *M. tuberculosis* genome contains only 74,602 TA insertion sites, and as millions of random mutants can be generated, we sought to overcome the uncertainty inherent in previous studies by saturating the possible *Himar1* insertion sites. In this work, we report a large-scale analysis of 14 independent TnSeq libraries in *M. tuberculosis* H37Rv, representing a combined total of 35,314,576 independent insertion events. All libraries were treated uniformly, grown in standard laboratory medium, and sequenced to a great depth using molecular barcodes. This unprecedented level of saturation enabled us to analyze sequence preferences for insertion, which revealed an unexpected sequence motif that is less permissive for Tn insertion. We used a hidden Markov model (HMM) which accounts for the reduced insertability of TA sites matching this motif to identify essential genes and genomic regions. This allowed us to assess the essentiality of small genomic regions, including small ORFs, promoters, and small (noncoding) RNAs

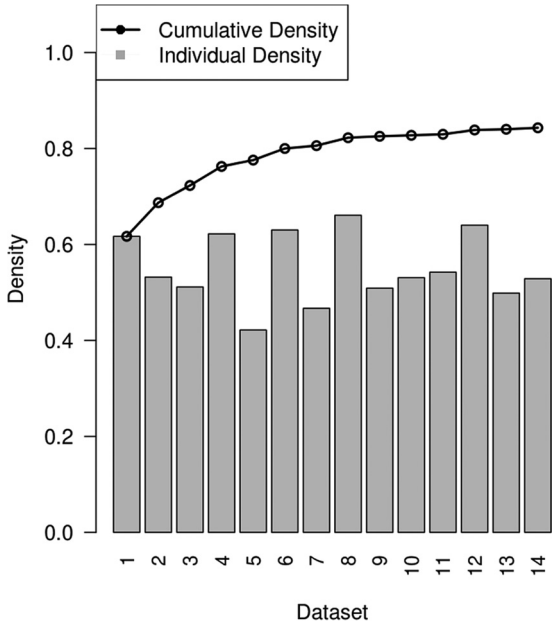

FIG 1 Cumulative fraction of TA sites represented as independent TnSeq data sets (black line). The gray bars show the saturation level of the individual data sets.

(sRNAs). The analysis of a saturated *Himar1* library provides a more reliable and comprehensive catalog of essential genomic features in *M. tuberculosis* and provides the basis for similar studies in other bacteria.

RESULTS

Generating a definitive map of essential genomic features in *M. tuberculosis* required the ability to interrogate small loci that contained few TA insertion sites. We reasoned that our ability to assess the essentiality of small features would be enhanced by producing a fully saturated library of mutants. To do this, we generated fourteen independent Tn insertion libraries in the H37Rv strain of *M. tuberculosis* by transfection with the ΦMycoMarT7 vector carrying the *Himar1* transposon (15, 17). In total, these libraries contained 35,314,576 independent insertion events, representing a level of coverage of 84.3% of the potential TA insertion sites. The transposon-chromosome junctions in each library were amplified and sequenced as previously described (15, 17). Sequencing the 14 libraries yielded an average of 2.5 million unique transposon-chromsosome junctions (termed "template counts"), which could be mapped to 42% to 64% of the TA dinucleotide sites in the chromosome in each individual library (see Table S1 for sequencing statistics).

Saturating mutagenesis of fitness-neutral sites can be achieved. To assess the level of saturation in the aggregated data set, we determined the cumulative density of insertions as each of the 14 libraries was included (in random order). Insertion density reached a plateau after about 12 of the 14 data sets were included (Fig. 1), indicating that virtually all available fitness-neutral sites contained an insertion in at least one library. In the aggregated data set, a *Himar1* insertion was detected in 84.3% of the TA sites, indicating that insertional mutants corresponding to 11,712 TA sites (15.7%) could not be recovered under these conditions. Many of the unoccupied sites clustered in regions of more than four consecutive TAs, and most of these gaps corresponded to ORFs that are essential for growth. When these operationally defined essential regions are excluded, the overall saturation increases to 92.5% (Table 1). The fact that 7.5% of the remaining TAs were unoccupied indicates the presence of many smaller essential genomic features that contained 1 to 4 TAs that did not sustain insertions in any of the libraries. These unoccupied sites were not generally correlated with sequencing failures

TABLE 1 Insertion count statistics of TA sites

TA site category	No. (%) of sites	% saturation	Mean read count (nonzero sites)
All	74,602	84.3	182.27
Those in putative nonessential regions[a]	67,992	92.5	182.27
Those in high-coverage regions[b]	57,452	96.5	189.54
Those in HC regions not matching NP sequence motif	52,672	98.8	192.18
All matching the NP motif	6,659 (9%)	59.7	50.00
All not matching the NP motif	67,943 (91%)	96.1	184.94

[a]Nonessential regions are defined as regions not containing a run of 4 or more unoccupied sites.
[b]High-coverage regions are based on labeling by the segmentation algorithm.

(i.e., did not correspond to poorly sequenced regions of the genome in other *M. tuberculosis* samples resequenced on the Illumina platform). However, it remained possible that some of these unoccupied sites corresponded to regions that were underrepresented as a result of either a moderate fitness defect in the corresponding mutant or previously uncharacterized insertion preferences of *Himar1* that are detectable only in a library as highly saturated as this. As described below, both of these effects contributed to the observed unoccupied sites and need to be taken into account.

Filtering out low-coverage regions reduces the fraction of unoccupied sites. We noted that many of the unoccupied sites were found in regions with lower-than-average insertion counts. Furthermore, the adjacent TA sites with low counts were often represented in only a small subset of the 14 independent libraries. If the representation of insertions in these libraries were stochastic, the number of independent libraries with an insertion at each biologically neutral site would be expected to be distributed binomially (sum of 14 Bernoulli trials). However, the observed distribution was significantly different from would be expected by chance (see Fig. S1a in the supplemental material). For example, sites with insertions in fewer than 3 of 14 replicates would be expected to be extremely rare if representation were completely random, and yet TA sites in putatively nonessential regions with insertions in just a few of the libraries (1–3) are frequently observed in the data. Template counts observed at each TA site exhibited a rough correlation with the number of replicates in which an insertion was detected (Fig. S1b), indicating that the abundance of a mutant is correlated with the probability that it would be detected. Note that the distributions shown in Fig. S1 exclude obvious "hit-free" essential regions and show that the insertions are still not completely random in the remaining nonessential regions but that there are various degrees of essentiality. These "low-coverage" (LC) regions likely correspond to genes that produce a quantitative growth defect upon disruption (15, 18), leading to underrepresentation in the libraries. Since detection of insertions in these low-coverage regions is less reliable, we sought to eliminate them from the analysis.

To systematically filter out TA sites in low-coverage regions, we implemented a probabilistic method for segmenting the genome into high-coverage (HC) and low-coverage (LC) regions (see Text S1 in the supplemental material). The segmentation algorithm effectively clusters TA sites together based on the local magnitude of read counts. The segmentation algorithm identified 833 LC regions, spanning 23% of the TA sites in the genome. This included all of the putative essential regions based on the operational definition given above (containing runs of 5 or more consecutive TA sites with no insertions), as well as isolated sites with 0 insertions adjacent to sites with depressed counts. Importantly, there were only 2,007 unoccupied TA sites remaining among the 57,452 TA sites in high-coverage regions (96.5% saturation). Of these remaining unoccupied sites, 86% were isolated TA sites surrounded by sites with insertions on both sides, and 12% were adjacent pairs of unoccupied sites.

Some TA sites appear to be nonpermissive for *Himar1* insertion. We then sought to determine if any of the remaining 2,007 unoccupied TA sites in HC regions might reflect a sequence preference for *Himar1* Tn insertion making them "nonpermissive" (NP) for insertion. To look for a sequence pattern associated with such sites, we selected

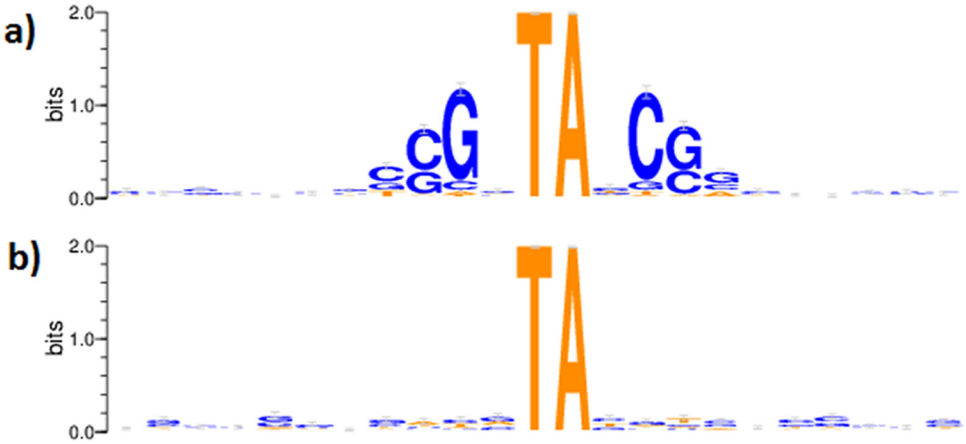

FIG 2 (a) Logo plot of $\log_2$ of nucleotide frequencies surrounding TA sites in the set of 1,746 unoccupied sites found in high-coverage regions (nonpermissive set). (b) Logo plot of $\log_2$ of nucleotide frequencies surrounding TA sites in the permissive set.

a putatively nonpermissive (NP) set of 1,746 unoccupied TA sites in high-coverage regions not associated with known genomic features other than ORFs (such as tRNAs, rRNAs, promoters, DNA methylation sites, etc., which could be devoid of insertions due to biological selection; see Materials and Methods). A corresponding set of 1,746 "permissive" sites (those with the highest mean template counts, >554) was chosen for comparison. The logo plot of the nucleotides surrounding TA sites in the nonpermissive set (Fig. 2a) reveals a strong sequence preference. Specifically, almost all of the nonpermissive sites have G at −2 and C at +2 around the TA dinucleotide, and sites at ±3 are either G or C. Forty-four percent of nonpermissive sites matching this pattern were nonpalindromic. Thus, the sequence pattern (GC)GNTANC(GC) is strongly associated with nonpermissiveness. In contrast, the permissive sites exhibit no local sequence motif that could be interpreted as favoring insertion (Fig. 2b).

Two-thirds of the TA sites in the nonpermissive set (1,223/1,746) match the sequence pattern (GC)GNTANC(GC), while none of the 1,746 sites in the permissive set match this pattern. The remaining sites not covered by this pattern could lack insertions because they are similar to the pattern but do not perfectly match the consensus, or for other reasons, including biological selection. Over the whole genome, 6,659 of 74,602 sites (9%) match the nonpermissive sequence pattern. Forty percent of these (2,682) have no insertions in any of the 14 replicates, and of those with insertions, 82% had insertions in 4 or fewer libraries (Fig. 3a). The insertion counts at these sites were also generally suppressed (Fig. 3b). Table S2 in the supplemental material indicates whether each TA site in the genome matches the nonpermissive sequence pattern, along with the observed insertion counts for each site.

After filtering out TA sites matching the nonpermissive (NP) sequence pattern, only 1.2% of the TA sites in high-coverage regions were devoid of insertions. Due to the saturation of the library, the odds that these sites are unoccupied due to chance are quite low. More likely, they are unoccupied because of unappreciated restrictions on Tn insertion or biological selection (i.e., disruption of a small essential genomic feature, the generation of a toxic truncated protein, etc.). For instance, Rv3397c contains an isolated but permissive site without insertions that coincides with a DNA methylation site, which could explain the apparent essentiality of this site. However, we cannot exclude the possibility that isolated sites in general are unoccupied because of biases against Tn insertion *per se*, making the interpretation of isolated unoccupied sites ambiguous. Thus, we are hesitant to conclude that isolated unoccupied TA sites necessarily correspond to a feature that is essential for growth. On the other hand, if we assume that unoccupied sites in nonessential regions are randomly distributed throughout the chromosome, the probability of two unoccupied (but permissive) sites being adjacent

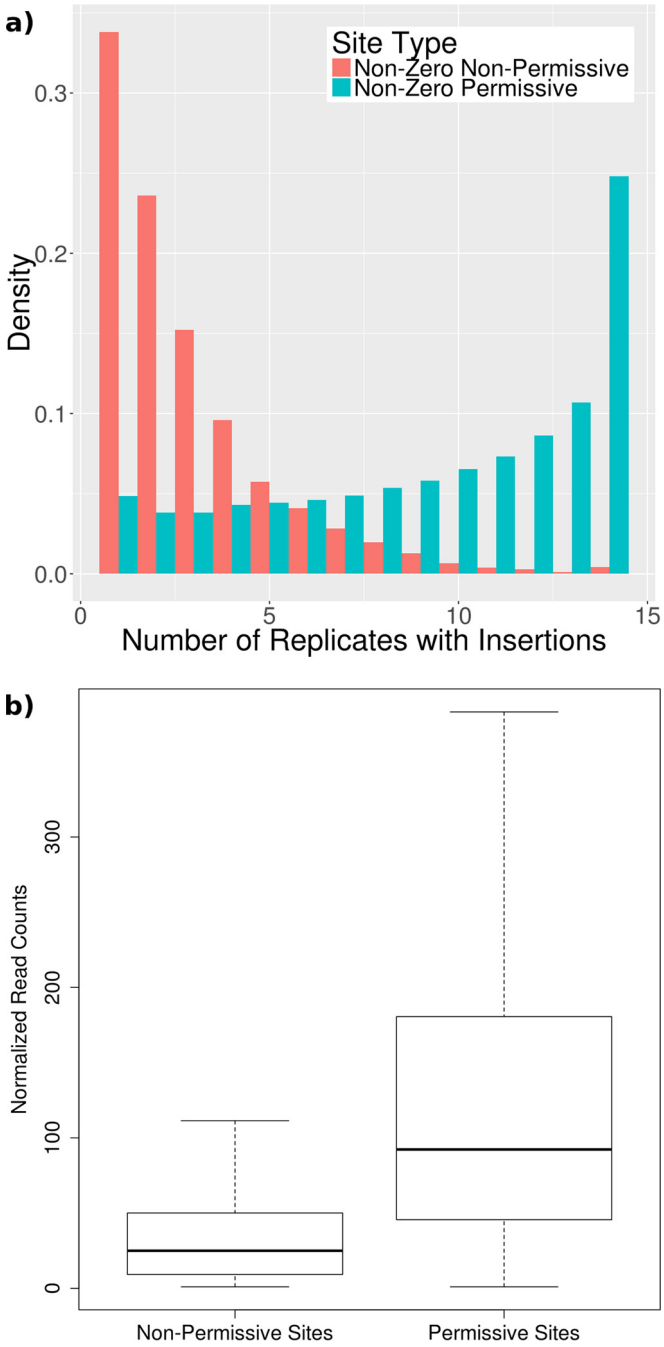

FIG 3 (a) Number of libraries representing sites matching (GC)GNTANC(GC) occupied by at least 1 insertion (red), compared to distribution over all TA sites (blue). (b) Box plot of nonzero insertion counts at sites matching the NP sequence motif versus sites not matching the motif. The boxes show the 25% to 75% interquartile range, while the whiskers show the majority of the range of insertion counts, except for the most extreme outliers.

is <0.0001. Hence, we restrict the analysis of essentiality to regions with two or more TA sites (separated by at least 15 bp, to ensure independence).

Nonpermissiveness for *Himar1* insertion is also observed in other prokaryotes. TA sites matching the NP sequence motif are underrepresented in other bacteria as well, suggesting that this pattern might reflect a general disinclination for insertion by the *Himar1* transposon (see Table 2). For example, in a Tn library of *Haemophilus influenzae* (1), the overall saturation was 53%, whereas only 6% of those sites matching

TABLE 2 Statistics for TnSeq data sets analyzed to determine transposon bias at sites matching NP motif

Organism	Tn	Study	%GC	No. of TA sites	No. of NP sites	% density	% density at NP sites
Mycobacterium tuberculosis	*Himar1*	This study	66	74,602	6,659	84	60
Caulobacter crescentus	*Himar1*	Murray et al. (19)	67	44,708	1,672	94	43
Haemophilus influenzae	*Himar1*	Gawronski et al. (1)	38	131,954	814	53	6
Vibrio cholerae	*Himar1*	Chao et al. (20)	47	192,681	4,439	56	7
Desulfovibrio vulgaris	Tn*5*	Fels et al. (69)	63	61,769	2,687	7[a]	8
Methanococcus maripaludis	Tn*5*	Sarmiento et al. (24)	33	133,503	514	5	6
Salmonella enterica serovar Typhi	Tn*5*	Langridge et al. (23)	53	233,259	9,289	3	4
Rhodopseudomonas palustris	Tn*5*	Pechter et al. (70)	65	72,385	3,844	3	3

[a]Although the Tn*5* transposon can insert at many different sites, the analysis was restricted to TA dinucleotides to investigate if Tn*5* had difficulty inserting in sites matching the NP motif as well.

the NP pattern were occupied by insertions. Similarly, in a high-saturation *Himar1* library of *Caulobacter crescentus* with 85.3% insertion density overall (19), 42.7% of TA sites matching the NP motif had insertions (and 20% of these were represented by only single reads). In a *Himar1* transposon-insertion library in *Vibrio cholerae* (20), insertions were observed in 56% of the TA sites, but only 7% of those matching the NP motif had insertions. The generality of this bias against insertions at such sites in organisms beyond mycobacteria suggests that it might reflect an intrinsic property of the *Himar1* transposase, which has been noted to require severe distortion (bending) of the DNA at the insertion site (21).

To discount the possibility that there might be a physical factor prohibiting transposon insertion at such sites, such as blocking by a widely distributed DNA-binding protein or a peculiar distortion of the DNA double helix by the adjacent G and C nucleotides, we examined several data sets from libraries generated with the Tn*5* transposon, which comes from a completely different family of transposases. Tn*5* can insert at a wide variety of genomic locations, due to a weaker sequence preference bias (22), and is not just restricted to TA sites, though some TA sites are candidate insertion sites for Tn*5*. If attention is restricted to just TA sites, then there does not appear to be a suppression of transposon insertions at the subset of TA sites matching the NP motif (Table 2). For example, in a Tn*5* library of *Salmonella enterica* serovar Typhi (23), the frequency of insertions at all TA sites was 3.0%, while the density among TA sites matching the NP motif was 3.9%. Similarly, 4.6% of all TA sites in a Tn*5* library of *Methanococcus maripaludis* (24) were occupied, whereas 5.6% of TA sites matching the NP motif were represented. This pattern is also observed for the magnitude of read counts. Sites matching the NP motif had significantly reduced read counts in the *Himar1* data sets compared to other sites (see Fig. 4). Together, these observations suggest that the suppression of insertions reported here represents a previously unidentified local sequence bias of the *Himar1* transposase.

Analysis of essential ORFs in H37Rv. The aggregated data set was analyzed using a HMM similar to one described in reference 25, extended to take the nonpermissive motif into account. The HMM is a statistical model for sequential data that uses likelihood functions and observed insertion counts to infer which essentiality state is most likely at each TA site, but also integrates this information with neighboring sites to produce a locally consistent (smoothed) interpretation of essentiality across the genome. The HMM parses the genome into distinct regions in a non-gene-centered way, unbiased by annotated ORF boundaries. Briefly, read counts are modeled as coming from geometric distributions conditioned on four different states of essentiality: essential (ES), growth defect (GD), nonessential (NE), and growth advantage (GA). Parameters for the expected read count distributions for each state were set relative to the mean read count (with ES being near 0, NE being near the mean, GD approximately 1/10 the mean, and GA 5 times the mean). The likelihood parameters for sites matching the nonpermissive motif were scaled down empirically to account for the suppression in read counts observed at these sites (see Materials and Methods).

The extended HMM was applied to the aggregated library and was used to determine the most likely sequence of essentiality states for TA sites across the entire H37Rv

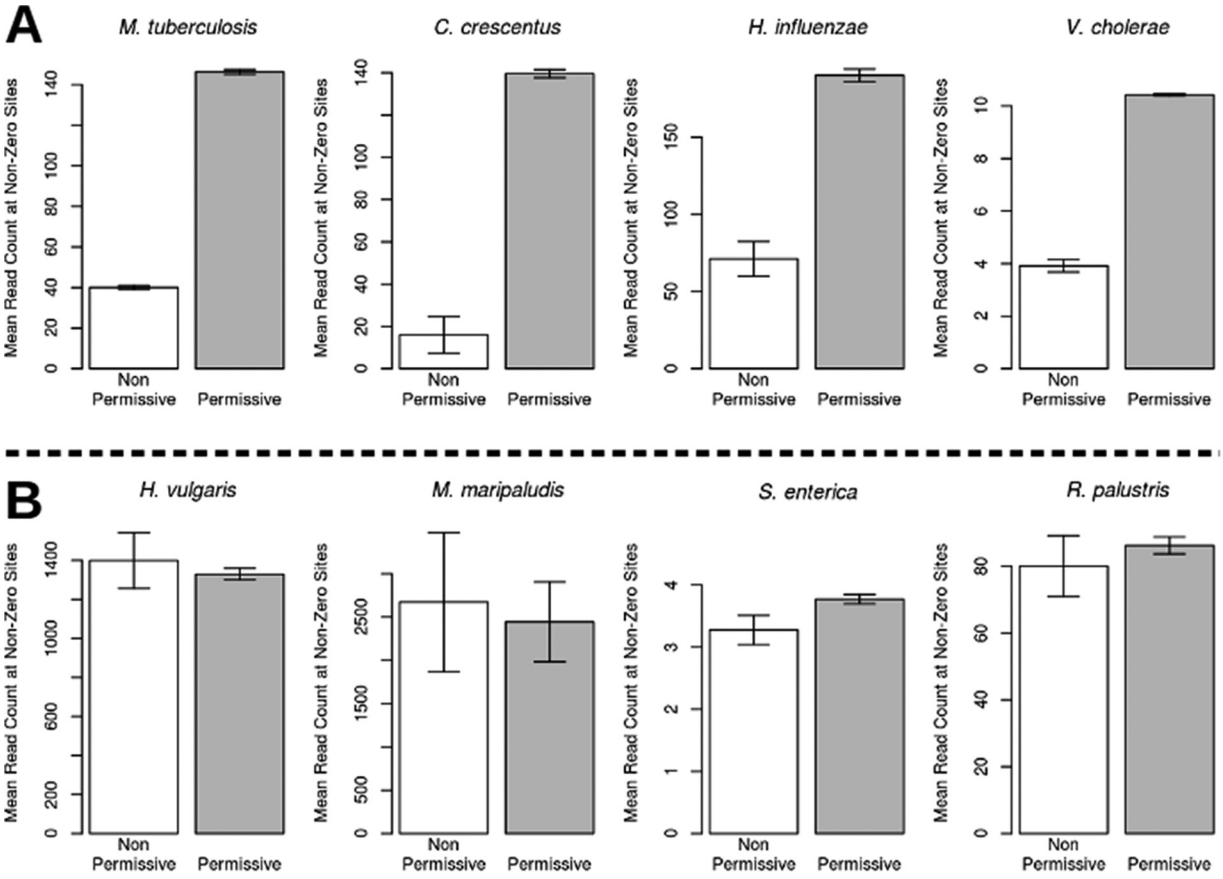

FIG 4 Mean read count at sites with at least one insertion for data sets made with the *Himar1* transposon (A) and the Tn*5* transposon (B). The nonpermissive sites (white bar), which match the nonpermissive motif identified in this study, significantly suppressed the read counts relative to sites that do not match the motif (permissive sites; grey bar). In contrast, limiting the analysis of the Tn*5* data sets to only insertions at TA dinucleotides, the mean read counts are similar for the permissive and nonpermissive sites. Error bars show the standard errors of the means.

genome, resulting in 11.6% of TA sites labeled ES, 77% NE, 3.5% GD, and 7.9% GA. The inferred state sequence was used to classify individual ORFs based on the most frequent label among TA sites in the ORF (the classification criteria are described in Materials and Methods). A total of 461 genes were identified as being essential and 135 genes with suppressed counts, whose disruption produces an apparent growth defect. As the HMM effectively recognizes large gaps (runs of consecutive TA sites lacking insertions) in ORFs, it is tolerant of a few insertions at the N and C termini of a gene. In addition, a subset ($n = 29$) are classified as "domain essentials," as they have both significant essential and nonessential regions. Combined, these represent 625 genes which are necessary for optimal growth *in vitro* (see Table S3). This number is similar to the overall number of essentials (614) in H37Rv determined by traditional hybridization-based methods (15) and includes many well-known genes known to be necessary for essential cellular functions. Unlike previous analyses performed with subsaturated libraries, which classified several members of the MmpL family of integral membrane proteins as essential (4), the current analysis predicts that only MmpL3 is essential whereas the other 12 MmpL family members are dispensable *in vitro*. This is consistent with genetic deletion studies (26, 27), as well as with recent reports of MmpL3 as a potential drug target (28–30). Similarly, none of the 61 Pro-Glu (PE) or Pro-Pro-Glu (PPE) (PE_PGRS) genes are classified as essential, and almost all of the 67 PPE and 33 PE genes are classified as nonessential, as expected, given that these highly duplicated gene families do not appear to play a critical biological role *in vitro* (31). Of these gene families, only Rv0285/PE5 and Rv0286/PPE4 (part of the ESX-3 locus, which plays an essential role in iron acquisition [32]) are categorized as essential in this analysis. Note

that PE_PGRS genes, which are GC rich and sometimes difficult to sequence, were enriched for nonpermissive sites; 35.6% of TA sites in PE_PGRS genes matched the NP pattern, compared to 9% overall. The inability to account for these insertional preferences is likely responsible for the inaccurate classification of some of these genes in previous studies. The gene encoding ICL (isocitrate lyase), which is a member of the glyoxylate shunt and is required for metabolizing fatty acids as a carbon source, was among the GD genes identified (33, 34). Because *M. tuberculosis* does not utilize carbon catabolite repression (34), cells with intact ICL grow better than ICL mutants on media containing fatty acids (e.g., oleate), but they can still survive without ICL, given the other carbon sources available in the rich medium used for culturing the libraries, explaining the growth defect phenotype.

Of the remaining genes, 3,008 were classified as NE by the HMM and as dispensable for growth *in vitro*, and 310 were classified as GA, indicating that disruption of these genes provides a growth advantage *in vitro*.

To compare these predictions to those of previous TnSeq studies, we reanalyzed data derived from a subsaturated library of H37Rv grown *in vitro* on glycerol as a carbon source (4) after applying a consistent multiple-testing correction and adjusted *P* value cutoff ($P_{adj} \leq 0.05$). We found that 79% (456) of the genes predicted to be essential in the prior study were classified as essential or growth defect genes using the saturated library. Many of the classifications of the genes that differ with respect to essentiality, such as *pckA* (encoding phosphoenolpyruvate carboxykinase), *glpK* (encoding glycerol kinase), and *bioABDF1* (biotin biosynthesis genes), can be explained by differences in growth medium. While Griffin used minimal media plus glycerol, our replicates were grown on rich media (7H9 or 7H10 [7H9/10] plus oleic acid-albumin-dextrose-catalase [OADC]) with a variety of potential carbon sources available (glycerol, dextrose, oleate, citrate, and glutamate). In addition, we suspect that a number of the genes were misclassified as essential in previous studies due to the incomplete saturation of the libraries.

Saturation enhances sensitivity for detecting small essential ORFs. The high saturation of our data set enabled us to identify small ORFs that are essential for growth. Typically, TnSeq analysis methods have been limited in their ability to detect genes with fewer than 9 TA sites (4, 6), even though these represent a larger number of ORFs in the *M. tuberculosis* genome (i.e., 1,274 ORFs with 2 to 9 TA sites). The difficulty in analyzing small ORFs is primarily due to the subsaturation of the TnSeq libraries, as well as limitations inherent in the choice of statistical method.

Using the motif-dependent HMM on the saturated libraries, we found 92 of 1,274 small ORFs (7.2%) to be essential *in vitro*, which is comparable to the fraction of larger genes that are classified as essential (Fig. S2). We identified several small essential genes with well-known functions important for survival, including *ssb* (encoding a single-stranded binding protein; 5 TA sites), which has no Tn insertions among the 14 independent libraries. Many of the ribosomal genes are categorized as essential or growth defect genes (15/23 *rpl* and 14/22 *rps* genes); most of those have sequences that were too short to analyze in previous studies. The essentiality of other small ORFs differs from conclusions of previous studies. For example, the *whiB1* transcription factor gene is classified as a growth defect gene, which is consistent with reports of the inability to delete it by homologous recombination (35–37), even though it was classified as nonessential in the initial TraSH study, likely due to the low resolution of that method (15).

To assess the relative impacts of both the high saturation of the libraries and the motif-dependent HMM on the detection of small ORFs, we compared our results to those produced by the Gumbel method, introduced by Griffin et al. (4), which uses the extreme-value distribution to identify unusually long stretches of sites without insertions (or "gaps"). To eliminate the growth condition (medium) as a variable, the Gumbel method was run on two of the libraries utilized in this study (grown in 7H9/10). The saturation of the combined data sets selected was 55.4%, which is similar to the saturation reported for previous TnSeq studies of H37Rv. The Gumbel method identi-

fied 0 small ORFs as essential, classifying all 1,274 as nonessential after correcting for multiple comparisons (the shortest gap that is significant at 55% saturation, based on the Gumbel distribution, is 10 TA sites long).

The inability of the Gumbel gap analysis method to identify small essential ORFs is primarily due to the saturation of the library. When all 14 libraries were analyzed instead, the method identified 93 small essential ORFs, and 84 of these (90.3%) are also identified as essential using the motif-dependent HMM. However, several of the remaining genes, which our motif-dependent HMM classifies differently, contain sites matching the nonpermissive motif. For instance, Rv3673c contains 3 empty sites that match the nonpermissive motif. While this can lead the method used by Griffin et al. (which focuses on identifying empty gaps) to identify the gap containing these sites as essential, the reduced probability of insertion at those sites makes it much likelier than they were missing by chance. Thus, taking this sequence-dependent bias into consideration can help better discriminate the essentiality for small ORFs.

Association of essential regions with small RNAs. We then assessed the association of unoccupied TA sites with sRNAs, which are short, highly structured transcripts that typically serve regulatory roles by modifying mRNA expression (38). sRNAs are often encoded as independent transcripts in intergenic regions but may also be derived from the 5′ untranslated region (5′UTR) or 3′UTR of an mRNA encoding a transcript (Storz et al. [39], Chao et al. [40], Loh et al. [41]). To date, these genomic features have not been included in TnSeq analyses because they contain only a few TA sites. In addition, there has been little consensus between studies describing the identification of sRNAs and their boundaries (42–45).

Determining the essentiality of the sRNAs first required definition of a set of sRNAs with accurate boundaries. We used a modified version of a bacterial small RNA sequencing (sRNA-Seq) protocol that utilizes size fractionation to enrich for small transcripts and ligate adapters to the natural ends of the transcripts (46). We next created a computational analysis, BS_finder (bacterial sRNA finder), to identify candidate sRNAs and distinguish them from other small transcript fragments such as those created through RNA degradation. BS_finder utilizes a sliding-window approach to identify small transcripts encoded completely or partially in an intergenic region with significantly greater read depth than the surrounding regions and sharp 5′ and 3′ boundaries. We employed stringent threshold criteria to identify 62 high-confidence sRNAs, ranging in size from 40 nucleotides (nt) to 268 nt, with an average size of 101 nt (Fig. S3a and Table S4). These sRNAs were named using the nomenclature proposed by Lamichhane et al. [47]. There was modest overlap between the sRNAs identified here and the sRNAs previously identified in studies enumerating intergenic transcripts or putative sRNAs (Fig. S3b) (Wang et al. [42], Arnvig et al. [48]). Many of the published putative sRNAs failed to reach our depth and boundary criteria designed to distinguish them from mRNA degradation products, but differences also likely reflect alterations in culture conditions, growth phases, and methods of library preparation. However, where the 5′ and 3′ boundaries of sRNAs have been experimentally determined by 5′ and 3′ rapid amplification of cDNA ends (RACE), there was excellent concordance with our analysis (45, 48–50). Of the 14 experimentally mapped *M. tuberculosis* sRNA ends, BS_finder mapped 12 within 3 bp of the experimentally defined end (Fig. S3c).

Using the HMM and our model (see Materials and Methods) to determine whether each sRNA is essential, 7 high-confidence sRNAs (ncRv0810c, ncRv0897, ncRv11315, ncRv1329, ncRv12783c, ncRv13418cA, and ncRv13418cB) and one previously identified sRNA, ncRNA3583A, were found to be essential or associated with a growth defect *in vitro*. Six of the essential sRNAs identified in this analysis appear to share transcriptional start sites (TSS) with essential or growth defect ORFs on the basis of TSS mapping data (51). It is possible that these regions lack insertions due to polar effects on expression of the neighboring essential gene. However, they could also represent distinct processed 5′UTRs of longer transcripts. Processing of 5′ and 3′ UTRs to generate independent *trans*-acting sRNAs has been previously described in several organisms (52, 53).

TABLE 3 Essentiality of non-ORF genomic features

Feature	Total no.	No. with ≥2 TA sites	No. essential
sRNA	62	48	7
tRNA	45	35	21
rRNA and other structural RNAs	5	5	5
DNA methylation site	362	55	0
Predicted rho-independent terminator	148	73	2
5′UTR	1,558	1,003	39
Promoter region	2,060	1,841	57

Thus, the association of these RNAs with essential ORFs does not eliminate the possibility that each represents an independently essential element. For example, ncRv12783c shares a TSS with the Rv2783c ORF. Additionally, one of the 2 TA sites between the annotated 3′ end of ncRv12783c and the translational start of Rv2783c tolerated insertions, supporting the possibility that Rv2783c and ncRv12783c are independently essential RNA species. It is possible that other small transcripts within this list are essential but cannot be confidently identified as such because they have only a single TA site. This is exemplified by the 4.5S RNA, which has a single TA site that tolerated no insertions. The 4.5S RNA is the RNA component of the signal recognition particle (SRP), which is essential in *Escherichia coli* (54). Importantly, the 4.5S RNA shares a TSS with Rv3722c in *M. tuberculosis* (Shell et al. [51]), exhibiting a pattern similar to that seen with the essential sRNAs identified here.

Association of unoccupied TA sites with other genomic features. In addition to sRNAs, we examined the essentiality of other known genomic features, including tRNAs, rRNAs, and other structural RNAs; promoters; 5′UTRs; terminator sequences; and DNA methylation sites (defined in Materials and Methods; see Table S5 for a full list). Table 3 contains a summary of the number of features that were found to be essential for each given type. A total of 21 of the 35 tRNAs with at least 2 TA sites were classified as essential, as expected (even though many tRNAs are small and contain only 3 TA sites on average). Eight of them, including Rvnt04/GlyU, were labeled as nonessential. The apparent dispensability of Rvnt04/GlyU (Gly tRNA) is likely due to redundancy with other Gly tRNAs annotated in the *M. tuberculosis* genome (Rvnt27/GlyV, Rvnt32/GlyT). The rRNAs (23S, 16S, 5S—*rrs*, *rrl*, *rrf*), as well as *rnpB* (nucleic acid component of RNase P), were all classified as essential, though *ssr* (which binds to and regulates the RNA polymerase) was nonessential. Conversely, almost all of the rho-independent terminators were observed to be nonessential, suggesting a lack of essential functionality under these conditions. A subset of promoter regions (~4.0%) was found to be essential. Several of these promoters were upstream of ORFs that are themselves essential, such as Rv0440/GroEL and Rv0667/RpoB. However, the correlation of essential promoters to essential genes was not strong, which could have been due to several factors. In some cases, promoters were classified essential not because the gene that they regulated was essential but due to overlap of the coding region of an adjacent essential gene. More generally, the choice of a broad window (−150 to +70 bp) around the transcription start site may have led to indiscriminate inclusion of parts of the upstream sequence that did not affect gene expression.

Identification of novel unannotated regions that are essential in the *M. tuberculosis* genome. Aside from ORFs, RNAs, and other annotated features, we wished to use our saturated library to discover novel genomic regions that are essential for *in vitro* growth. We segmented the remaining unannotated regions of the genome (i.e., excluding ORFs, RNAs, and other annotated features) into contiguous regions that were labeled as belonging to the same state (as determined by the HMM). A total of 17 unannotated regions were identified as essential, and 12 were identified as associated with a growth defect (Table S6). These regions could represent functionally important features whose function is not yet known. Among the largest of the unannotated growth defect segments was an intergenic region between Rv3616c (*espA*, encoding

ESX-1-associated protein) and Rv3617 (*ephA*, encoding epoxide hydrolase), spanning 15 TA sites. Another large segment spanning 13 sites occurred in the intergenic region between Rv1056 and Rv1057 (both hypothetical genes). Interestingly, while all four of these genes are nonessential, the intergenic regions contain multiple MprA binding sites (55, 56). MprAB is a two-component regulator implicated in stress response. Hence, we speculate that transposon insertion in the intergenic region might attenuate growth *in vitro* by disrupting the MprA regulon.

DISCUSSION

This work describes the first use of a *Himar1* library (compiled from 14 independent replicates) that has reached the practical limit of saturation for the definition of essential genomic regions. The high level of saturation reduces the ambiguity associated with TA sites lacking insertions, allowing us to determine essentiality with high confidence for even small genomic features. In this context, we define an "essential" feature as one whose disruption causes complete absence or significant suppression of read counts among the combined data sets. Using a hidden Markov model to segment the genome into different regions of essentiality, we identified 625 ORFs that are essential for optimal growth *in vitro* (including domain essentials and growth defect genes). This set was largely consistent with previous studies (4), though there were some differences. Some of these discrepancies could be explained by differences in growth medium, since the use of 7H9/10 plus OADC provides carbon sources and metabolites lacking in the minimal medium used by Griffin et al. (4). However, most differences were due to the identification of smaller ORFs (with as few as 2 TA sites) than was possible in previous studies. In contrast, the Griffin study (4) was unable to categorize ORFs with fewer than 9 TA sites as "uncertain" (after correcting for multiple tests), due largely to the lower level of saturation. The density of mutagenesis achieved in this study was due to a combination of factors—both the large number of random mutants generated and the relatively small number of TA sites in the GC-rich *M. tuberculosis* genome. Equivalent saturation of less-G+C-rich organisms with a much larger number of TA sites would require a correspondingly larger number of independent mutants.

If *Himar1* insertion into TA sites were random, the analysis of this data set would have been straightforward, as it would be highly unlikely that even a single TA site in a nonessential region would be unrepresented in all 14 libraries. Finding that the apparent saturation of our library plateaued at less than 100% in otherwise well-represented (i.e., high-coverage) regions prompted us to investigate the sequence specificity of transposition. We identified a local sequence pattern, (GC)GNTANC(GC), associated with TA sites with few to no insertions, suggesting that these sites are nonpermissive for *Himar1* insertion. This effect was observed in *Himar1* TnSeq data sets from other organisms as well. This appears to be specific to *Himar1*, since other transposons such as Tn*5* do not appear to be inhibited from integrating at such sites. This putative nonpermissive sequence pattern is similar to the complement of a preference bias reported for insertion of the Sleeping Beauty transposon, ANNTANNT, into the human genome, based on a consensus of 138 unique insertion sites (57). Like *Himar1*, Sleeping Beauty is in the *mariner* class of transposons (58). Unlike Sleeping Beauty, we observed a sequence pattern only in our set of nonpermissive TA sites, while permissive sites (those with high insertion counts) showed no apparent sequence bias. Mechanistically, the sequence preference for *Himar1* could be due to effects of the G and C at ±2 bp on local DNA structure (59) or possibly the hydrogen bonding between the transposase and edges of nucleotides in the major groove of the DNA (60). An alternative explanation could be that accessibility for transposon insertion at certain TA sites might be blocked by the presence of a DNA-binding protein recognizing a specific binding-site motif. For example, it has recently been shown that the insertion-site specificity pattern for *Himar1* in *Vibrio cholerae* is influenced by the binding of histone-like nuclear structuring protein, *h-ns* (61). However, blocking by DNA-binding proteins seems unlikely, given that the Tn*5* transposon appears to integrate at sites matching the NP pattern with same frequency as at other sites.

The *Himar1* transposon is widely reported to have little sequence specificity, other than the requirement for TA dinucleotides. No other local sequence dependence has been observed to date, although a weak preference for more bendable regions of DNA in the *E. coli* genome has been previously reported (14). As a result, most previous gene essentiality studies using *Himar1* have been based on the assumption that insertion is equally probable at all TA sites. Our studies indicate that this assumption is not completely accurate and that insertions at TA sites matching the NP pattern are substantially less frequent than at other TA sites. In fact, the insertion bias of the *Himar1* transposon does not appear to be limited to mycobacteria, as it could be observed in TnSeq data sets of other organisms as well. This insertion bias could be a consequence of the transposase-DNA interaction, which involves significant distortion of the DNA, as observed in the recently described crystal structure of the transposase-DNA strand transfer complex (21), hence biasing against sequences that cannot accommodate this distortion. Because nearly 10% of all TA sites in the *M. tuberculosis* genome match the identified nonpermissive sequence pattern, indiscriminate inclusion of these nonpermissive sites in TnSeq analyses could artificially inflate the number of predicted essential regions determined by the use of a statistical framework that assumes random integration. Indeed, this insertional bias of *Himar1* may have contributed to the previous misclassification of genes in certain families, such as the MmpL and PE_PGRS genes. However, there is general agreement between our analysis of a saturated library and previous studies using subsaturated libraries (analyzed with statistical methods that assume unbiased random insertion). Thus, in subsaturated libraries, where nonpermissive sites represent a smaller subset of all sites lacking insertions, the assumption of random integration appears to have a less-significant impact on essentiality analyses. It is primarily in the context of nearly complete saturation that the insertional preferences of *Himar1* need to be taken into account.

The statistical analysis that we employed in this study was a hidden Markov model that takes advantage of both the identified sequence preference for insertion and the high level of saturation. Specifically, the HMM uses geometric distributions as likelihood functions to evaluate the read counts observed at TA sites in the entire genome, conditioning parameters based on whether the sites match the motif identified in this study. A significant advantage of the HMM over other analysis methods is that it is capable of determining different degrees of essentiality, representing increased or decreased levels of fitness. This includes growth defect genes, which exhibit an intermediate level of transposon insertions possibly reflecting a reduction in fitness, and growth advantage genes which result in an improvement in fitness when disrupted. Another advantage of the HMM is that the analysis is not limited to predefined gene boundaries but is instead capable of assessing essentiality across the entire genome in an unbiased (non-gene-centered) way. The high-resolution analysis afforded by the nearly complete saturation of the library enabled us, in an unprecedented way, to analyze the essentiality of smaller features of the *M. tuberculosis* genome, such as RNAs and regulatory regions such as promoters and terminators. Other than RNAs, the most frequent essential feature was the presence of 5′UTRs and promoter regions, while only a few of the predicted terminators and DNA methylation sites appeared to be essential. This could imply that adenosine methylation and rho-independent termination are not functionally relevant under these conditions, or possibly that the sequence models used to predict these sites are imperfect. We identified some essential regulatory regions (promoters and 5′ UTRs), though they were not well correlated with essential genes. We also developed a reliable list of 62 sRNAs based on RNA-Seq analysis, which took advantage of a customized sequencing protocol using direct adapter ligation to precisely characterize both 5′ and 3′ ends of the sRNA transcripts. Seven sRNAs were found to be essential for optimal growth *in vitro*, six of which occurred in 5′UTRs of essential genes and likely represent processed transcripts. Taking the data together, the results of high-resolution statistical analysis of essential regions of the genome, in conjunction with the high-resolution annotation of sRNA, will provide a useful reference for the *M. tubercu-*

losis research community. More broadly, these studies represent an experimental and statistical framework to extend saturation mutagenesis to other organisms.

MATERIALS AND METHODS

Construction of Tn libraries. *M. tuberculosis* H37Rv transposon libraries were constructed by *Himar1* mutagenesis as previously described (17). Briefly, 100 ml of mid-log-phase *M. tuberculosis* culture (optical density at 600 nm [OD_{600}], ~0.7 to 1.0) was incubated with 1×10^{11} to 2×10^{11} PFU of ΦMycoMarT7 phage (62) at 37°C for 4 to 18 h; cultures were then washed and plated on the media indicated in Table S1 in the supplemental material and incubated for 18 to 21 days at 37°C. Libraries with titers of greater than 100,000 CFU were used in the analysis to ensure comprehensive coverage of the possible TA dinucleotide insertion sites in the *M. tuberculosis* genome.

Sequencing of Tn libraries. Genomic DNA was extracted from the transposon libraries, and mutant composition was determined by sequencing amplicons of the transposon-genome junctions as previously described (17). Briefly, genomic DNA was sheared into ~500-bp fragments by ultrasonication by the use of a Covaris E220 system, and fragments were subjected to end repair and A-tailing with *Taq* polymerase and ligated to T-tailed adapters bearing random 7-nucleotide barcodes to distinguish between unique fragments before subsequent PCR amplification. Fragments containing transposon-genome junctions were selectively enriched in a first PCR amplification, size selected in the 400-to-600-bp range, and amplified in a second heminested PCR to add adapter sequences for Illumina sequencing. PCR amplicons were subjected to 75-to-100-bp paired-end sequencing on an Illumina HiSeq platform, and raw sequence data were exported to fastq files for further analysis.

Data processing. The sequence data were processed using the TPP tool included with TRANSIT (12). Reads were mapped to the genome using the Burroughs-Wheeler Aligner (bwa [63]). The set of reads in "read1" with a prefix matching the end of the *Himar1* transposon were mapped to the corresponding TA site in the genome (stripping off the transposon prefix). The read counts were reduced to template counts by discarding duplicates with the same barcode in "read2." The final template counts were normalized across all the data sets using Trimmed Total Reads (TTR) normalization. TTR normalizes data sets so that they have the same mean template count, while ignoring ("trimming") the top and bottom 5% of read counts to reduce the influence of outliers.

RNA extraction, library construction, and sequencing. *Mycobacterium tuberculosis* H37Rv was cultured in 7H9 media (Difco Laboratories) with 10% oleic acid-albumin-dextrose-catalase, and cells were harvested at an OD of 1.0. Cell pellets were resuspended in TRIzol reagent (Ambion, Life Technologies, Inc.) and lysed using a FastPrep 24 instrument. RNA was purified using a Direct-Zol MiniPrep kit (Zymo Research), and DNA was digested using Turbo DNase (Ambion, Life Technologies, Inc.). Size selection to achieve an approximately 300-nt cutoff was performed using RNA Clean and Concentrator-25 columns (Zymo Research) according to the manufacturer's instructions with the following modification: preparation of the adjusted RNA binding buffer was done by adding an equal volume of RNA binding buffer and 70% ethanol. The small RNA fraction was then depleted of rRNA using a RiboZero-Bacteria kit (Illumina). The small RNA library was then prepared as described previously using a modified TruSeq small-RNA-sequencing kit (Illumina). Briefly, RNA was dephosphorylated using RppH enzyme (New England BioLabs) followed by an ethanol precipitation step. 3′ and 5′ adapters were then ligated to the RNA followed by reverse transcription and PCR amplification. The libraries were then purified using Agencourt AMPure XP beads (Beckman Coulter, Inc.). Paired-end sequencing (150 bp) was then performed on a MiSeq sequencer. Sequenced reads were aligned to the *Mycobacterium tuberculosis* NC_000962.3 genome (National Center for Biotechnology Information database) using the Burrows-Wheeler alignment tool (63).

sRNA identification. Per-base coverage of the genome was obtained using the *genomecov* tool of the BEDtools suite (64). sRNAs were then identified using BS_finder (bacterial sRNA finder), which uses a sliding-window approach to sRNA identification. Briefly, a sliding window scans the per-base read depth and searches for positive slope (the intensity of which can be modified to adjust stringency) across the window, which demarcates a 5′ end (searching the plus strand) or a 3′ end (searching the minus strand) if the position passes a modifiable read depth threshold. A nested sliding window then searches for decreases in slope across the window, which demarcates either a 3′ or 5′ end. If the discovered feature is not entirely within an ORF and has a 5′ end at least 50 bp away from the nearest ORF translational start site, then it is output as a candidate sRNA. BS_finder was run on the data set using default parameters: a sliding-window size of 2 bp with a slope threshold of 3 and a read depth threshold of 500.

Statistical analysis of essentiality. The hidden Markov model described in reference 25 was extended to take into consideration the suppression of insertions observed at sites matching the NP motif. Read counts were modeled as coming from four different states of essentiality: the essential, growth defect, nonessential, and growth advantage states. Geometric distributions, conditioned on the local sequence (i.e., matching the NP motif or not), were used to assess the likelihood of observing a read count from a given state as follows:

$$P\left(O_t|Q_t = s,\ np = 0\right) = \text{geometric}\left(O_t|P = \frac{1}{\mu_s}\right)$$

$$P\left(O_t|Q_t = s,\ np = 1\right) = \text{geometric}\left(O_t|P = \frac{1}{\mu_s \times \omega}\right)$$

where O_t values represent the insertion counts observed at each site (normalized and summed over the 14 replicates), Q_t represents the state labels ($Q_t \in$ <ES, GD, NE, GA>), μ_s represents the expected read

count for state *s* (which is chosen to represent different levels of fitness [25]), *np* indicates whether the site matches the NP motif (1) or not (0), and ω represents the ratio between the mean read count at sites matching the NP motif and the mean read count at other sites (accounting for the reduction in read counts at nonpermissive sites) as follows:

$$\omega = \frac{\mu_{np}}{\mu_P}$$

The transition probabilities were set such that the states would have a high probability of remaining in the current state, thus requiring a significant change in local read counts to prompt a change of the essentiality state (25). The Viterbi algorithm (65) was used to identify the most likely state sequence for the entire genome.

To obtain essentiality classifications for individual ORFs or features, we employed the following criteria. Most of the ORFs were classified by the plurality of state labels (the most frequent) among its TA sites. ORFs which have an essential domain (i.e., a significantly large ES segment, as well as a nonessential region spanning at least 200 bp) were identified using the extreme-value distribution by calculating the probability of observing the corresponding sequence of ES states in a row using the formulas given in reference 4. Small ORFs (≤3 TA sites), which could be influenced strongly by the essentiality of surrounding sites, were classified as essential only if they were significantly devoid of insertions as determined by a binomial distribution [P value < 0.05; $P(k; n, P) = \sum_{i=0}^{k} \binom{n}{i} P^i (1 - P)^{n-i}$], provided that they span at least 15 bp. Small ORFs which did not pass this threshold but which were completely devoid of insertions nonetheless (mostly ORFs with just 1 TA site) were considered to be "uncertain."

Analysis of non-ORF genomic features. Promoter regions were defined based on a set of 2,060 transcriptional start sites (TSSs) defined in reference 51 that are within 500 bp of a translational start site (on the same strand). A region around each TSS (−150 to +70 bp) was used to define the promoter region, since regulatory signals such as transcription factor binding sites are often found in this broad region (not just the −10 and −35 regions of the sigma factor binding site) (66). 5′UTRs were defined as the region between the transcriptional and the translational start sites (for transcripts with leaders), which can contain regulatory features such as ribosomal binding sites or riboswitches. The motif "GATN$_4$RTAC" was used to define DNA methylation sites, since it is the only 1 of 3 motifs identified to date (27, 67) that has the possibility of containing a TA site and thus of being capable of being disrupted by a *Himar1* transposon insertion. Finally, a predictive model for rho-independent termination signals (~65-bp pseudopalindromic sequence) was applied to identify putative terminators in the 3′ ends of transcripts (68).

The essentiality of these features, together with the remaining unannotated regions in the genome, was analyzed using the method outlined above.

Accession number(s). All TnSeq data sets are publicly available on the NCBI Sequence Read Archive with accession number SRP083947 and BioProject accession number PRJNA341349.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.02133-16.

TEXT S1, PDF file, 0.1 MB.
FIG S1, TIF file, 8.1 MB.
FIG S2, TIF file, 3.8 MB.
FIG S3, TIF file, 2.7 MB.
TABLE S1, DOCX file, 0.01 MB.
TABLE S2, XLSX file, 3.1 MB.
TABLE S3, XLSX file, 0.3 MB.
TABLE S4, XLSX file, 0.01 MB.
TABLE S5, XLSX file, 0.3 MB.
TABLE S6, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

This work was supported by NIH grant U19 AI107774 (T.R.I., E.J.R., D.S., S.E., S.M.F., and C.M.S.) and NSF grant DGE1144152 (E.R.G.).

REFERENCES

1. Gawronski JD, Wong SM, Giannoukos G, Ward DV, Akerley BJ. 2009. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for haemophilus genes required in the lung. Proc Natl Acad Sci U S A 106:16422–16427. https://doi.org/10.1073/pnas.0906627106.
2. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC. 1999. Global transposon mutagenesis and a minimal mycoplasma genome. Science 286:2165–2169. https://doi.org/10.1126/science.286.5447.2165.
3. Jacobs MA, Alwood A, Thaipisuttikul I, Spencer D, Haugen E, Ernst S, Will O, Kaul R, Raymond C, Levy R, Chun-Rong L, Guenthner D, Bovee D, Olson MV, Manoil C. 2003. Comprehensive transposon mutant library of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 100:14339–14344. https://doi.org/10.1073/pnas.2036282100.
4. Griffin JE, Gawronski JD, DeJesus MA, Ioerger TR, Akerley BJ, Sassetti CM. 2011. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog 7:e1002251. https://doi.org/10.1371/journal.ppat.1002251.

5. Joshi SM, Pandey AK, Capite N, Fortune SM, Rubin EJ, Sassetti CM. 2006. Characterization of mycobacterial virulence genes through genetic interaction mapping. Proc Natl Acad Sci U S A 103:11760–11765. https://doi.org/10.1073/pnas.0603179103.
6. Zhang YJ, Ioerger TR, Huttenhower C, Long JE, Sassetti CM, Sacchettini JC, Rubin EJ. 2012. Global assessment of genomic regions required for growth in Mycobacterium tuberculosis. PLoS Pathog 8:e1002946. https://doi.org/10.1371/journal.ppat.1002946.
7. Kieser KJ, Baranowski C, Chao MC, Long JE, Sassetti CM, Waldor MK, Sacchettini JC, Ioerger TR, Rubin EJ. 2015. Peptidoglycan synthesis in Mycobacterium tuberculosis is organized into networks with varying drug susceptibility. Proc Natl Acad Sci U S A 112:13087–13092. https://doi.org/10.1073/pnas.1514135112.
8. Nambi S, Long JE, Mishra BB, Baker R, Murphy KC, Olive AJ, Nguyen HP, Shaffer SA, Sassetti CM. 2015. The oxidative stress network of Mycobacterium tuberculosis reveals coordination between radical detoxification systems. Cell Host Microbe 17:829–837. https://doi.org/10.1016/j.chom.2015.05.008.
9. van Opijnen T, Camilli A. 2013. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol 11:435–442. https://doi.org/10.1038/nrmicro3033.
10. Zomer A, Burghout P, Bootsma HJ, Hermans PW, van Hijum SA. 2012. Essentials: software for rapid analysis of high throughput transposon insertion sequencing data. PLoS One 7:e43012. https://doi.org/10.1371/journal.pone.0043012.
11. Pritchard JR, Chao MC, Abel S, Davis BM, Baranowski C, Zhang YJ, Rubin EJ, Waldor MK. 2014. ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing. PLoS Genet 10:e1004782. https://doi.org/10.1371/journal.pgen.1004782.
12. DeJesus MA, Ambadipudi C, Baker R, Sassetti C, Ioerger TR. 2015. TRANSIT—a software tool for Himar1 TnSeq analysis. PLoS Comput Biol 11:e1004401. https://doi.org/10.1371/journal.pcbi.1004401.
13. DeJesus MA, Sacchettini JC, Ioerger TR. 2013. Reannotation of translational start sites in the genome of Mycobacterium tuberculosis. Tuberculosis (Edinb) 93:18–25. https://doi.org/10.1016/j.tube.2012.11.012.
14. Lampe DJ, Grant TE, Robertson HM. 1998. Factors affecting transposition of the Himar1 mariner transposon in vitro. Genetics 149:179–187.
15. Sassetti CM, Boyd DH, Rubin EJ. 2003. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 48:77–84. https://doi.org/10.1046/j.1365-2958.2003.03425.x.
16. Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabási AL, Oltvai ZN, Osterman AL. 2003. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol 185:5673–5684. https://doi.org/10.1128/JB.185.19.5673-5684.2003.
17. Long JE, DeJesus M, Ward D, Baker RE, Ioerger T, Sassetti CM. 2015. Identifying essential genes in Mycobacterium tuberculosis by global phenotypic profiling. Methods Mol Biol 1279:79–95. https://doi.org/10.1007/978-1-4939-2398-4_6.
18. van Opijnen T, Bodi KL, Camilli A. 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 6:767–772. https://doi.org/10.1038/nmeth.1377.
19. Murray SM, Panis G, Fumeaux C, Viollier PH, Howard M. 2013. Computational and genetic reduction of a cell cycle to its simplest, primordial components. PLoS Biol 11:e1001749. https://doi.org/10.1371/journal.pbio.1001749.
20. Chao MC, Pritchard JR, Zhang YJ, Rubin EJ, Livny J, Davis BM, Waldor MK. 2013. High-resolution definition of the Vibrio cholerae essential gene set with hidden Markov model-based analyses of transposon-insertion sequencing data. Nucleic Acids Res 41:9033–9048. https://doi.org/10.1093/nar/gkt654.
21. Morris ER, Grey H, McKenzie G, Jones AC, Richardson JM. 2016. A bend, flip and trap mechanism for transposon integration. Elife 5. https://doi.org/10.7554/eLife.15537.
22. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS. 1998. Tn*5*/IS*50* target recognition. Proc Natl Acad Sci U S A 95:10716–10721. https://doi.org/10.1073/pnas.95.18.10716.
23. Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK. 2009. Simultaneous assay of every Salmonella typhi gene using one million transposon mutants. Genome Res 19:2308–2316. https://doi.org/10.1101/gr.097097.109.
24. Sarmiento F, Mrázek J, Whitman WB. 2013. Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon Methanococcus maripaludis. Proc Natl Acad Sci U S A 110:4726–4731. https://doi.org/10.1073/pnas.1220225110.
25. DeJesus MA, Ioerger TR. 2013. A hidden Markov model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. BMC Bioinformatics 14:303. https://doi.org/10.1186/1471-2105-14-303.
26. Domenech P, Reed MB, Barry CE, III. 2005. Contribution of the Mycobacterium tuberculosis MmpL protein family to virulence and drug resistance. Infect Immun 73:3492–3501. https://doi.org/10.1128/IAI.73.6.3492-3501.2005.
27. Zhu L, Zhong J, Jia X, Liu G, Kang Y, Dong M, Zhang X, Li Q, Yue L, Li C, Fu J, Xiao J, Yan J, Zhang B, Lei M, Chen S, Lv L, Zhu B, Huang H, Chen F. 2016. Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. Nucleic Acids Res 44:730–743. https://doi.org/10.1093/nar/gkv1498.
28. La Rosa V, Poce G, Canseco JO, Buroni S, Pasca MR, Biava M, Raju RM, Porretta GC, Alfonso S, Battilocchio C, Javid B, Sorrentino F, Ioerger TR, Sacchettini JC, Manetti F, Botta M, De Logu A, Rubin EJ, De Rossi E. 2012. MmpL3 is the cellular target of the antitubercular pyrrole derivative BM212. Antimicrob Agents Chemother 56:324–331. https://doi.org/10.1128/AAC.05270-11.
29. Li W, Upadhyay A, Fontes FL, North EJ, Wang Y, Crans DC, Grzegorzewicz AE, Jones V, Franzblau SG, Lee RE, Crick DC, Jackson M. 2014. Novel insights into the mechanism of inhibition of MmpL3, a target of multiple pharmacophores in Mycobacterium tuberculosis. Antimicrob Agents Chemother 58:6413–6423. https://doi.org/10.1128/AAC.03229-14.
30. Tahlan K, Wilson R, Kastrinsky DB, Arora K, Nair V, Fischer E, Barnes SW, Walker JR, Alland D, Barry CE, III, Boshoff HI. 2012. SQ109 targets MmpL3, a membrane transporter of trehalose monomycolate involved in mycolic acid donation to the cell wall core of Mycobacterium tuberculosis. Antimicrob Agents Chemother 56:1797–1809. https://doi.org/10.1128/AAC.05708-11.
31. McEvoy CR, Cloete R, Müller B, Schürch AC, van Helden PD, Gagneux S, Warren RM, Gey van Pittius NC. 2012. Comparative analysis of Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. PLoS One 7:e30593. https://doi.org/10.1371/journal.pone.0030593.
32. Tufariello JM, Chapman JR, Kerantzas CA, Wong KW, Vilchèze C, Jones CM, Cole LE, Tinaztepe E, Thompson V, Fenyö D, Niederweis M, Ueberheide B, Philips JA, Jacobs WR, Jr. 2016. Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence. Proc Natl Acad Sci U S A 113:E348–E357. https://doi.org/10.1073/pnas.1523321113.
33. Muñoz-Elías EJ, McKinney JD. 2005. Mycobacterium tuberculosis isocitrate lyases 1 and 2 are jointly required for in vivo growth and virulence. Nat Med 11:638–644. https://doi.org/10.1038/nm1252.
34. de Carvalho LP, Fischer SM, Marrero J, Nathan C, Ehrt S, Rhee KY. 2010. Metabolomics of Mycobacterium tuberculosis reveals compartmentalized co-catabolism of carbon substrates. Chem Biol 17:1122–1131. https://doi.org/10.1016/j.chembiol.2010.08.009.
35. Smith LJ, Stapleton MR, Fullstone GJ, Crack JC, Thomson AJ, Le Brun NE, Hunt DM, Harvey E, Adinolfi S, Buxton RS, Green J. 2010. Mycobacterium tuberculosis WhiB1 is an essential DNA-binding protein with a nitric oxide-sensitive iron-sulfur cluster. Biochem J 432:417–427. https://doi.org/10.1042/BJ20101440.
36. Gomez JE, Bishai WR. 2000. whmD is an essential mycobacterial gene required for proper septation and cell division. Proc Natl Acad Sci U S A 97:8554–8559. https://doi.org/10.1073/pnas.140225297.
37. Rybniker J, Nowag A, van Gumpel E, Nissen N, Robinson N, Plum G, Hartmann P. 2010. Insights into the function of the WhiB-like protein of mycobacteriophage TM4—a transcriptional inhibitor of WhiB2. Mol Microbiol 77:642–657. https://doi.org/10.1111/j.1365-2958.2010.07235.x.
38. Arnvig K, Young D. 2012. Non-coding RNA and its potential role in Mycobacterium tuberculosis pathogenesis. RNA Biol 9:427–436. https://doi.org/10.4161/rna.20105.
39. Storz G, Vogel J, Wassarman KM. 2011. Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell 43:880–891. https://doi.org/10.1016/j.molcel.2011.08.022.
40. Chao Y, Papenfort K, Reinhardt R, Sharma CM, Vogel J. 2012. An atlas of Hfq-bound transcripts reveals 3′ UTRs as a genomic reservoir of regulatory small RNAs. EMBO J 31:4005–4019. https://doi.org/10.1038/emboj.2012.229.

41. Loh E, Dussurget O, Gripenland J, Vaitkevicius K, Tiensuu T, Mandin P, Repoila F, Buchrieser C, Cossart P, Johansson J. 2009. A trans-acting riboswitch controls expression of the virulence regulator PrfA in Listeria monocytogenes. Cell 139:770–779. https://doi.org/10.1016/j.cell.2009.08.046.
42. Wang M, Fleming J, Li Z, Li C, Zhang H, Xue Y, Chen M, Zhang Z, Zhang XE, Bi L. 2016. An automated approach for global identification of sRNA-encoding regions in RNA-Seq data from Mycobacterium tuberculosis. Acta Biochim Biophys Sin (Shanghai) 48:544–553. https://doi.org/10.1093/abbs/gmw037.
43. Haning K, Cho SH, Contreras LM. 2014. Small RNAs in mycobacteria: an unfolding story. Front Cell Infect Microbiol 4:96. https://doi.org/10.3389/fcimb.2014.00096.
44. Pellin D, Miotto P, Ambrosi A, Cirillo DM, Di Serio C. 2012. A genome-wide identification analysis of small regulatory RNAs in Mycobacterium tuberculosis by RNA-Seq and conservation analysis. PLoS One 7:e32723. https://doi.org/10.1371/journal.pone.0032723.
45. Arnvig KB, Young DB. 2009. Identification of small RNAs in Mycobacterium tuberculosis. Mol Microbiol 73:397–408. https://doi.org/10.1111/j.1365-2958.2009.06777.x.
46. Gómez-Lozano M, Marvig RL, Molin S, Long KS. 2014. Identification of bacterial small RNAs by RNA sequencing. Methods Mol Biol 1149:433–456. https://doi.org/10.1007/978-1-4939-0473-0_34.
47. Lamichhane G, Arnvig KB, McDonough KA. 2013. Definition and annotation of (myco)bacterial non-coding RNA. Tuberculosis (Edinb) 93:26–29. https://doi.org/10.1016/j.tube.2012.11.010.
48. Arnvig KB, Comas I, Thomson NR, Houghton J, Boshoff HI, Croucher NJ, Rose G, Perkins TT, Parkhill J, Dougan G, Young DB. 2011. Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. PLoS Pathog 7:e1002342. https://doi.org/10.1371/journal.ppat.1002342.
49. DiChiara JM, Contreras-Martinez LM, Livny J, Smith D, McDonough KA, Belfort M. 2010. Multiple small RNAs identified in Mycobacterium bovis BCG are also expressed in Mycobacterium tuberculosis and Mycobacterium smegmatis. Nucleic Acids Res 38:4067–4078. https://doi.org/10.1093/nar/gkq101.
50. Miotto P, Forti F, Ambrosi A, Pellin D, Veiga DF, Balazsi G, Gennaro ML, Di Serio C, Ghisotti D, Cirillo DM. 2012. Genome-wide discovery of small RNAs in Mycobacterium tuberculosis. PLoS One 7:e51950. https://doi.org/10.1371/journal.pone.0051950.
51. Shell SS, Wang J, Lapierre P, Mir M, Chase MR, Pyle MM, Gawande R, Ahmad R, Sarracino DA, Ioerger TR, Fortune SM, Derbyshire KM, Wade JT, Gray TA. 2015. Leaderless transcripts and small proteins are common features of the mycobacterial translational landscape. PLoS Genet 11:e1005641. https://doi.org/10.1371/journal.pgen.1005641.
52. Chao Y, Vogel J. 2016. A 3′ UTR-derived small RNA provides the regulatory noncoding arm of the inner membrane stress response. Mol Cell 61:352–363. https://doi.org/10.1016/j.molcel.2015.12.023.
53. Lalaouna D, Carrier MC, Semsey S, Brouard JS, Wang J, Wade JT, Massé E. 2015. A 3′ external transcribed spacer in a tRNA transcript acts as a sponge for small RNAs to prevent transcriptional noise. Mol Cell 58:393–405. https://doi.org/10.1016/j.molcel.2015.03.013.
54. Brown S, Fournier MJ. 1984. The 4.5 S RNA gene of Escherichia coli is essential for cell growth. J Mol Biol 178:533–550. https://doi.org/10.1016/0022-2836(84)90237-7.
55. Pang X, Cao G, Neuenschwander PF, Haydel SE, Hou G, Howard ST. 2011. The beta-propeller gene Rv1057 of Mycobacterium tuberculosis has a complex promoter directly regulated by both the MprAB and TrcRS two-component systems. Tuberculosis 91(Suppl 1):S142–S149. https://doi.org/10.1016/j.tube.2011.10.024.
56. Pang X, Samten B, Cao G, Wang X, Tvinnereim AR, Chen XL, Howard ST. 2013. MprAB regulates the espA operon in Mycobacterium tuberculosis and modulates ESX-1 function and host cytokine response. J Bacteriol 195:66–75. https://doi.org/10.1128/JB.01067-12.
57. Vigdal TJ, Kaufman CD, Izsvák Z, Voytas DF, Ivics Z. 2002. Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. J Mol Biol 323:441–452. https://doi.org/10.1016/S0022-2836(02)00991-9.
58. Ivics Z, Hackett PB, Plasterk RH, Izsvák Z. 1997. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91:501–510. https://doi.org/10.1016/S0092-8674(00)80436-5.
59. Liu G, Geurts AM, Yae K, Srinivasan AR, Fahrenkrug SC, Largaespada DA, Takeda J, Horie K, Olson WK, Hackett PB. 2005. Target-site preferences of Sleeping Beauty transposons. J Mol Biol 346:161–173. https://doi.org/10.1016/j.jmb.2004.09.086.
60. Voigt F, Wiedemann L, Zuliani C, Querques I, Sebe A, Mátés L, Izsvák Z, Ivics Z, Barabas O. 2016. Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering. Nat Commun 7:11126. https://doi.org/10.1038/ncomms11126.
61. Kimura S, Hubbard TP, Davis BM, Waldor MK. 2016. The nucleoid binding protein H-NS biases genome-wide transposon insertion landscapes. mBio 7:e01351-16. https://doi.org/10.1128/mBio.01351-16.
62. Sassetti CM, Boyd DH, Rubin EJ. 2001. Comprehensive identification of conditionally essential genes in mycobacteria. Proc Natl Acad Sci U S A 98:12712–12717. https://doi.org/10.1073/pnas.231275498.
63. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
64. Quinlan AR. 2014. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics 47:11.12.1–11.12.34.
65. Rabiner LR. 1989. A tutorial on Hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286. https://doi.org/10.1109/5.18626.
66. Minch KJ, Rustad TR, Peterson EJ, Winkler J, Reiss DJ, Ma S, Hickey M, Brabant W, Morrison B, Turkarslan S, Mawhinney C, Galagan JE, Price ND, Baliga NS, Sherman DR. 2015. The DNA-binding network of Mycobacterium tuberculosis. Nat Commun 6:5829. https://doi.org/10.1038/ncomms6829.
67. Shell SS, Prestwich EG, Baek SH, Shah RR, Sassetti CM, Dedon PC, Fortune SM. 2013. DNA methylation impacts gene expression and ensures hypoxic survival of Mycobacterium tuberculosis. PLoS Pathog 9:e1003419. https://doi.org/10.1371/journal.ppat.1003419.
68. Gardner PP, Barquist L, Bateman A, Nawrocki EP, Weinberg Z. 2011. RNIE: genome-wide prediction of bacterial intrinsic terminators. Nucleic Acids Res 39:5845–5852. https://doi.org/10.1093/nar/gkr168.
69. Fels SR, Zane GM, Blake SM, Wall JD. 2013. Rapid transposon liquid enrichment sequencing (TnLE-seq) for gene fitness evaluation in underdeveloped bacterial systems. Appl Environ Microbiol 79:7510–7517. https://doi.org/10.1128/AEM.02051-13.
70. Pechter KB, Gallagher L, Pyles H, Manoil CS, Harwood CS. 2015. Essential genome of the metabolically versatile alphaproteobacterium Rhodopseudomonas palustris. J Bacteriol 198:867–876. https://doi.org/10.1128/JB.00771-15.