

# Defining the core essential genome of *Pseudomonas aeruginosa*

## Supplementary information (datasets, methods and results, figures, tables)

### Datasets

Dataset S1 - sequencing depth and replicate concordance.xlsx  
Dataset S2 - genes without usable TA sites.xlsx  
Dataset S3 - gene probabilities of essentiality for all strains and media.xlsx  
Dataset S4 - 13 potential core essential genes.xlsx  
Dataset S5 - core and conditionally essential genes in core genome.xlsx  
Dataset S6 - core essential genes comparison with previous studies.xlsx  
Dataset S7 - primers used in this study.xlsx

### Raw data

Sequencing read count tally files and other relevant analysis files used in this study can be found here:

<https://data.broadinstitute.org/fitness/>

### Analysis pipeline

The pipeline to generate tally files from raw sequence data is UNIX-based; the *FiTnEss* analysis software is R-based. These analysis pipelines can be found here:

<https://github.com/broadinstitute/FiTnEss>

## Additional Methods and Results

**Plasmid construction.** pC9 was derived from pSAM-DGm (1) by digesting with ApaLI + AccI to remove the transposon, and pMAR was derived from pMAR2xT7 (2) by digesting with ApaLI + StuI to remove the transposase. Linearized vectors were each blunted, phosphorylated, ligated, and transformed into *E. coli* SM10 donor cells and selected on 100ug/ml carbenicillin (pC9) or 15ug/ml gentamicin (pMAR). Cloning reagents were obtained from New England Biolabs.

**Transposon library construction and sequencing.** Overnight cultures of *E. coli* SM10(pC9) and *E. coli* SM10(pMAR) donor cells were grown in LB medium with their respective antibiotics, sub-cultured 1:100, and grown at 37°C while shaking at 250 RPM for 3.5 hours until OD<sub>600nm</sub> reached ~0.5. Overnight cultures of recipient *P. aeruginosa* strains were grown in LB medium, sub-cultured 1:5, and grown at 42°C while shaking at 250 RPM for 3.5 hours. Cells were collected by centrifugation at 5000g for 10 minutes, washed once, and re-suspended in LB. A total of  $3 \times 10^{11}$  CFU were mixed in a 2:2:1 ratio of pC9:pMAR:recipient and collected by centrifugation. The cell mating mixture was re-suspended to an approximate concentration of  $10^{11}$  CFU/ml and 30  $\mu$ l spots were dispensed to a dry LB agar plate. Mating plates were incubated at 37°C for 1.5 hours before cells were scraped, resuspended in phosphate buffered saline (ThermoFisher), mixed with glycerol to a final concentration of 40%, aliquoted, and flash frozen in a dry ice/ethanol bath before storage at -80°C. A small aliquot of each mixture was thawed, diluted and plated to 5  $\mu$ g/ml irgasan + 30  $\mu$ g/ml gentamicin for CFU quantification of successful integrants. Matings were performed at least twice for each recipient strain. 250 mL of each medium containing 1.5% agar, 5  $\mu$ g/ml irgasan, and 30  $\mu$ g/ml gentamicin was prepared in a Biodish XL (Nunc). LB agar (US Biologicals), M9 minimal agar, synthetic cystic fibrosis medium (SCFM) (3) were prepared as previously described. Pooled, filter-sterilized urine, and fetal bovine serum (FBS) (ThermoFisher) were warmed to 55°C and mixed with a 5% agar solution (Teknova) to achieve a 1.5% final agar concentration. 500,000 CFU of each transposon-integrated strain were plated to each medium in duplicate and grown at 37°C for 24 hours (LB, FBS, SCFM) or 48 hours (urine, M9) before scraping and re-suspending cells in PBS. Genomic DNA was isolated using the DNeasy kit (Qiagen), and 5  $\mu$ g from each sample was sheared to 1.5 kb fragments by sonication (Covaris). End repair, dA-tailing, P5 adapter ligation, and PCR of the transposon-gDNA junction was performed using NEBNext enzymes (NEB) and custom primers from IDT (Fig. S1 and Dataset S7). Size selection was performed using Agencourt Ampure XP beads (Beckman Coulter) and ~500 bp libraries were quantified using D5000 ScreenTape System (Agilent). Sequencing was performed with an Illumina Nextseq platform to obtain 50 bp genomic DNA reads.

**Removal of confounding transposon insertion sites from analysis.** At these TA sites, the presence or absence of mapped insertions can be influenced by methodological artifacts unrelated to the essentiality of gene in which the TA is located. To avoid these confounding factors, we removed three classes of TA sites from analysis because of their potential to mislead:

(1) Non-permissive insertion sites – The sequence (GC)GNTANC(GC) was reported to be intolerant to Himar1 transposon insertions in *Mycobacterium tuberculosis* (4), which has a similar GC content to *P. aeruginosa*. This sequence occurs 6367 times in *P. aeruginosa* strain PA14 across 3389 genes. Indeed, we found that insertions mapped to these sites at a significantly reduced frequency compared to a random subsample of TA sites ( $p < 0.0001$ , Fig. S2) and thus excluded them from all subsequent analysis.

(2) Non-disruptive terminal insertions – Transposon insertions close to 5' and 3' gene termini can nevertheless result in the expression of a functional, albeit truncated version of the corresponding gene

product (5). Rather than selecting an arbitrary distance from the termini in which to exclude such potentially confounding TA sites, we empirically determined an optimal distance. Using the consensus 109 essential genes from previous transposon studies of strains PA14 and PAO1 as the truth set for essential genes (1, 2, 6-8), we found that 42 of these genes in our PA14-LB dataset contained >10 sequencing reads, all of which corresponded to TA site insertions located within 50 bp from the gene termini, regardless of gene size (Fig. S2). We thus eliminated from analysis all TA sites that fell within 50 nucleotides of either the 5' or 3' ends of each gene. Removal of these confounding TA sites resulted in the exclusion of 9829 TA sites.

(3) Homologous insertion sequences – Because insertions are assigned to a specific TA site in a specific gene based on the mapping of the genomic sequences flanking the ends of a transposon onto the entire genome, we removed from consideration TA sites whose flanking regions are not unique because of the possibility of mis-mapping reads.

### **FiTnEss: a statistical model to identify essential genes.**

**Motivation behind the model.** After removal of confounding TA sites, we used an exemplary dataset – PA14 LB – to first calculate the average number of reads per TA site in each gene,  $n_g/N_{TA}$ , where  $n_g$  is the total number of reads in the gene and  $N_{TA}$  is the number of TA sites. The distribution of  $n_g/N_{TA}$  across all genes in this dataset appears bimodal, with essential genes presumably on the far left and the bulk of the distribution (on the right) representing non-essential genes (Fig. S3A). However, when we separated this combined distribution based on gene-size (i.e.  $N_{TA} = 1, N_{TA} = 2, N_{TA} = 3$ , etc.), we observed that the distribution varies with  $N_{TA}$ : it is only bimodal in larger genes ( $N_{TA} \geq 5$ , Fig. S3B), and the non-essential peak becomes more narrow as  $N_{TA}$  further increases. Thus, selecting a single  $n_g/N_{TA}$  cutoff value for the whole distribution to define essential versus non-essential genes is not appropriate. Clearly, a given gene must be compared to the distribution calculated from genes of the same size. It turns out, as explained below, that a simple model with two parameters is sufficient for describing the behavior of all non-essential genes and can be used to derive thresholds for all  $N_{TA}$  values. Below, we describe in detail these observations about the data and the construction of the model that forms the basis for *FiTnEss*.

**Additional observations that support the model.** As noted above, calculating the average number of reads per TA site in a gene ( $n_g/N_{TA}$ ) of a given dataset results in a specific value that falls in a distribution. There is, however, variation between the numbers of reads at each TA site within that gene (e.g. a gene with 100 reads and 10 TA sites does not typically have 10 reads at each TA site). We refer to this as the “gene-specific distribution of reads”, whose mean is approximated by the average  $n_g/N_{TA}$ . As described above, Fig S3B demonstrated that the distribution of  $n_g/N_{TA}$  depends on  $N_{TA}$ . The simplest explanation for this dependence is that, like any average of independent variables, the variation of the average number of reads/TA at any gene is inversely proportional to the number of observations being averaged, namely  $N_{TA}$ . We demonstrate the  $N_{TA}$ -independence of the underlying gene-specific distributions by sampling once from each of them (instead of looking at the averages) (Fig. S3C,D). Indeed, the distributions for different  $N_{TA}$  categories look similar, which is also supported by the constant variance (compare Fig. S3E and S3F).

Given the bimodal distributions observed in the data (Fig. S3B), we based *FiTnEss* on modelling the collective behavior of gene-specific distributions for non-essential genes, and fitting the model parameters from the data, as explained in the next sections. In the final step, the data from each gene can be converted to the likelihood of that gene being non-essential by calculating p-values from the distributions.

**A model for non-essential genes.** We posited that the distribution of the number of reads at any TA site in a gene (the gene-specific distribution) is geometric with probability  $p_g$ , and that the expected number of reads ( $1/p_g$ ) is only a function of the fitness of the bacteria when the gene function is lost. Assuming a lognormal distribution of  $1/p_g$ , the model only requires two parameters, the mean and variance of this distribution, to be determined from the data. Since the model describes non-essential genes, it was important to avoid data from essential genes when fitting the model parameters. Thus, we used only genes with which we had high confidence in their non-essentiality ( $N_{TA}=10$ ; top 75% of the distribution). Given the fitted parameters of the model, a specific dependence of the distribution of  $n_g$  (and  $n_g/N_{TA}$ ) on the number of TA sites was predicted, which agrees well with the actual data (red curves in Fig. S3B).

We assume that each non-essential gene  $g$  is characterized by a parameter,  $p_g$ , the inverse of which comes from a log-normal distribution

$$p_g^{-1} \sim \text{Lognormal}(\mu, \sigma), \quad (1)$$

with parameters  $\mu, \sigma$ . We further assume that for any non-essential gene  $g$  with certain number of TA sites ( $N_{TA}$ ), the read counts at any of its TA sites,  $x_{g,i}$ , are iid (independent and identically distributed random variables), and are distributed according to

$$\begin{aligned} \text{For a specific gene } g: x_{g,i} &\sim \text{Geo}(p_g), \\ \text{for } i &= 1, \dots, N_{TA}. \end{aligned} \quad (2)$$

A possible interpretation of this model is that there is a distribution among non-essential genes of the small fitness costs of disabling them. Genes that are slightly more important would have a higher knockout cost  $p_g^{-1}$ , or a lower  $p_g$ , and thus a lower number of reads per TA site on average.

It follows that the distribution of  $n_g$ , the total number of reads in a given gene, follows a negative binomial distribution:

$$\begin{aligned} \text{For a specific gene } g: \\ n_g \equiv \sum_1^{N_{TA}} x_{g,i} &\sim \text{NB}(N_{TA}, p_g). \end{aligned} \quad (3)$$

(It should be noted that for the purpose of fitting the model and subsequent scoring of the essentiality of genes, using the distributions of  $n_g$  is completely equivalent to using those of  $n_g/N_{TA}$ .)

The distribution of  $n_g$  among all the genes for some value of  $N_{TA}$  is the convolution of the lognormal and the negative binomial:

$$\begin{aligned} F_{n_g}^*(n) \equiv \text{Prob}(n_g \leq n) = \\ \int_0^{+\infty} f_{LN}\left(\frac{1}{p}; \mu, \sigma\right) F_{NB}(n; N_{TA}, p) d\left(\frac{1}{p}\right), \end{aligned} \quad (4)$$

where  $f_{LN}$  is the probability density function of the lognormal distribution and  $F_{NB}$  the negative binomial cumulative distribution function.

**Fitting model parameters.** Model parameters are fitted by minimizing the distance between an expected distribution (given the parameters), and an empirical one. As a distance metric, we used a modification of the Cramér-von Mises criterion, which is a goodness of fit criterion, measuring the difference between the cumulative density functions of empirical and fitted distributions. We applied it here to the distributions of  $n_g$  in different  $N_{TA}$  categories (Equation 4). The empirical distribution

$F_{N_{TA}}$  is obtained directly from the data, and we use numerical sampling to approximate the theoretical  $F_{N_{TA}}^*$  (sampling 100,000 times for each pair of parameter values  $\mu, \sigma$ ). For any  $N_{TA}$  category, we have

$$\omega_{N_{TA}}^2 = \int_0^{+\infty} \left[ F_{n_g}(n; N_{TA}) - F_{n_g}^*(n; N_{TA}) \right]^2 dF_{n_g}^*(n; N_{TA}), \quad (5)$$

with  $\omega_{N_{TA}}^2$  denoting the integral of squared distance between two functions for all genes with  $N_{TA}$  TA sites. In order to fit model parameters that describe non-essential genes, we tried to avoid data that are potentially “contaminated” by essential ones in the parameter estimation phase. To address this, we use a modified version of the Cramér–von Mises criterion  $\omega^2$  as follows:

$$\omega_{N_{TA}}^2 = \int_{n_{[1/4]}}^{+\infty} \left[ F_{n_g}(n; N_{TA}) - F_{n_g}^*(n; N_{TA}) \right]^2 dF_{n_g}^*(n; N_{TA}), \quad (6)$$

where  $n_{[1/4]}$  is the low 25 percentile of read counts in this  $N_{TA}$  category. This practically means that minimizing this distance is only affected by the goodness of fit to the higher 75% of the empirical distribution, which is not expected to contain essential genes.

The model parameters can be determined by minimizing the sum of this modified  $\omega^2$  for any of the different  $N_{TA}$  categories, and the resulting parameters are not affected much by this choice. Yet we observed that for genes with a low number of TA sites there was not much separation between the essential and non-essential populations. Conversely, for the gene categories with large  $N_{TA}$ , where this separation is very pronounced, the number of genes in these categories is too small and leads to less robust fits. We have estimated the variability of the fitted parameters under perturbations of the data and concluded that using values of  $N_{TA}$  between 5 and 15 yield robust fits (Fig. S8). The parameters used for the results in this paper are based on fitting the distributions at  $N_{TA} = 10$ .

**Calling essential genes.** For each Tn-Seq dataset (= a replicate of strain x medium), after identifying parameters  $\mu, \sigma$  for non-essential genes, we construct the background distribution for each  $N_{TA}$  category by sampling  $10^5$  observations of  $(n_g^*)$  from the theoretical distribution (Equation 4). The actual number of reads for each gene is then compared to the background distribution for the corresponding  $N_{TA}$  category, and a p-value is calculated as the probability of obtaining this number reads ( $n_g$ ) or less “by chance”:

$$p\text{-value} = P(n_g^* < n_g). \quad (7)$$

In each medium and strain, we have more than 5000 genes being tested simultaneously. Accounting for multiple testing is required for obtaining true signals. Two-layer multiple comparison adjustments were conducted. First, in order to obtain a more conservative essential set, we adjusted for family-wise error rate. Family-wise error rate (FWER) is a conservative correction method for multiple hypothesis, by controlling type I error we allow low probability of making one or more false discoveries. In our analysis, we used Holm-Bonferroni method with type I error rate  $\alpha = 0.05$ , indicating that we have only 5% chance of obtaining even a single false positive call in the dataset. Second, to reduce the risk of losing important targets by being too conservative, we used Benjamini-Hochberg procedure, which is a less strict approach controlling for false-discovery rate. After either correction process, genes with

adjusted p-value smaller than 0.05 in both replicates are identified as “confident essential” (FWER) or “candidate essential” (FDR).

**Gene deletions.** Gene deletions were performed as previously described in strain PA14 (9). Briefly, 800-1200 bp regions flanking the target deletion gene of interest were PCR amplified, stitched and recombined into the pEXG2 (9) plasmid containing GentR and SacB markers using Gateway Cloning. Plasmids were conjugated into PA14 using the *E. coli* helper plasmid pRK2013 for 8 hours, followed by selection on LB agar containing 15 µg/ml irgasan + 30 µg/ml gentamicin. Individual colonies were grown in liquid LB for 4 hours, followed by streaking to LB agar supplemented with 10% sucrose and grown at 37°C for 16 hours. Colonies were confirmed to be GentS and successful gene deletions were confirmed by PCR amplification and sequencing.

**Comparison of *FiTnEss* to a HMM.** Analytical tools can also vary significantly, as they all have different strengths and weaknesses, often having been developed to answer different questions. One of the greatest challenges for the analytical tools is to translate measurements from Tn-Seq, which is really quantifying a continuum of fitness – from optimal growth in a particular condition to slow growth, from static for growth to cell death – to a binary classification of essential versus non-essential in the service of comprehensively defining candidate targets for antibiotic discovery. The different tools can vary dramatically both in the assumptions built into the analysis and how conservatively each model calls essentiality *i.e.*, whether one is more willing to tolerate false positives or false negatives.

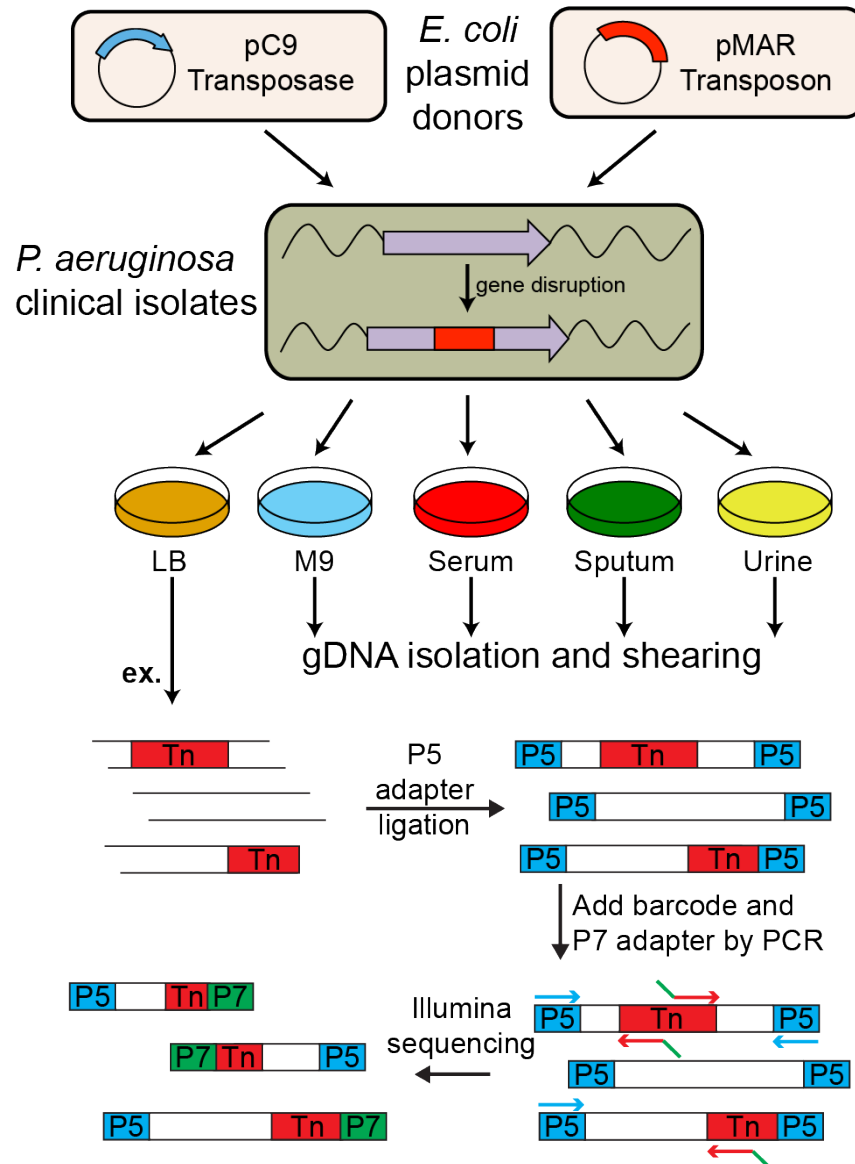
For example, a Hidden Markov Model (HMM) and sliding window approach rely on a stretch of TA sites that have zero to very low level insertions to denote an essential gene (10, 11). An advantage of these methods is that intergenic regions and essential domains within a larger gene can be queried. The disadvantage is that genes containing more insertions than the HMM and sliding window approaches tolerate in an essential gene, can in fact be essential with detectable insertions resulting because death of the corresponding mutant is slow or delayed. Because *FiTnEss* considers genes rather than individual TA sites as the basic unit for determining essentiality, it leverages all TA sites in a gene, allowing it to more easily distinguish whether low insertion numbers are indicative of low coverage in a non-essential gene or background noise in an essential gene.

Another set of genes that are often discrepant between analytical methods is short genes that may be flanked by genes of the opposite classification. Here, approaches that examine “windows” of adjacent TA sites (~5-10 adjacent sites) can misclassify the short gene of interest (<5 TA sites) by erroneously integrating in data from the flanking genes that are of the opposite classification; because *FiTnEss* examines the gene independent of flanking regions, it can avoid being misled by the behavior of the TA sites in the flanking genes. In the particular case of longer genes containing a mix of essential and nonessential domains, the power of *FiTnEss* to detect essentiality is reduced because of it cannot distinguish the even distribution of reads across the gene (resulting in a call of non-essentiality) with a bimodal distribution of reads among the essential and non-essential domains (which should result in a call of essentiality). Here, other methods such as HMM outperform *FiTnEss* (Fig. S7 for comparison of methods, Dataset S3 for complete *FiTnEss*/HMM gene calls).

While the methods are complementary, in the analysis of the datasets generated in this study, *FiTnEss* seemed to be generally more powerful for calling essential genes than HMM, with greater accuracy in calling the 115 conditionally essential gene-growth condition combinations that we validated using clean genetic deletions (Fig. 2 and Fig. S4). While the HMM method did not have any false positives (if combining essential and growth defective categories), it did miss calling many essential genes, *i.e.* tolerated a high false negative rate (Table S2). Meanwhile, *FiTnEss* attempted to balance false positive and false negative rates for this limited set of deletions resulting in greater overall accuracy. Of note, the genes selected for validation were skewed towards falling relatively clearly in the essential or non-essential distributions; thus, they may overestimate the positive predictive power of

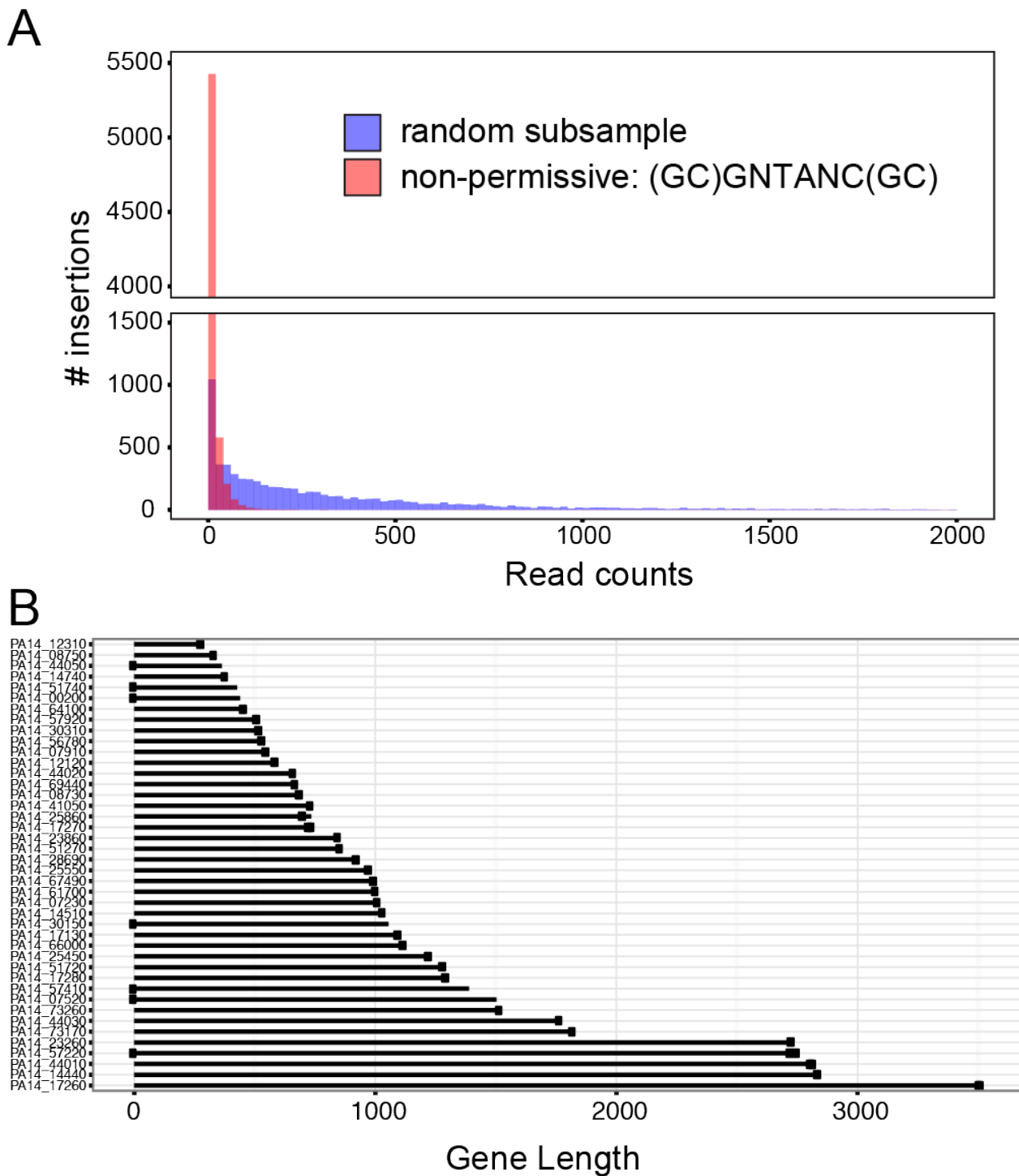
*FiTnEss*, particularly for genes that lay at the boundary of the bimodal distribution. Nevertheless, overall, *FiTnEss* appears to perform well in its binary classification of genes.

## Figures

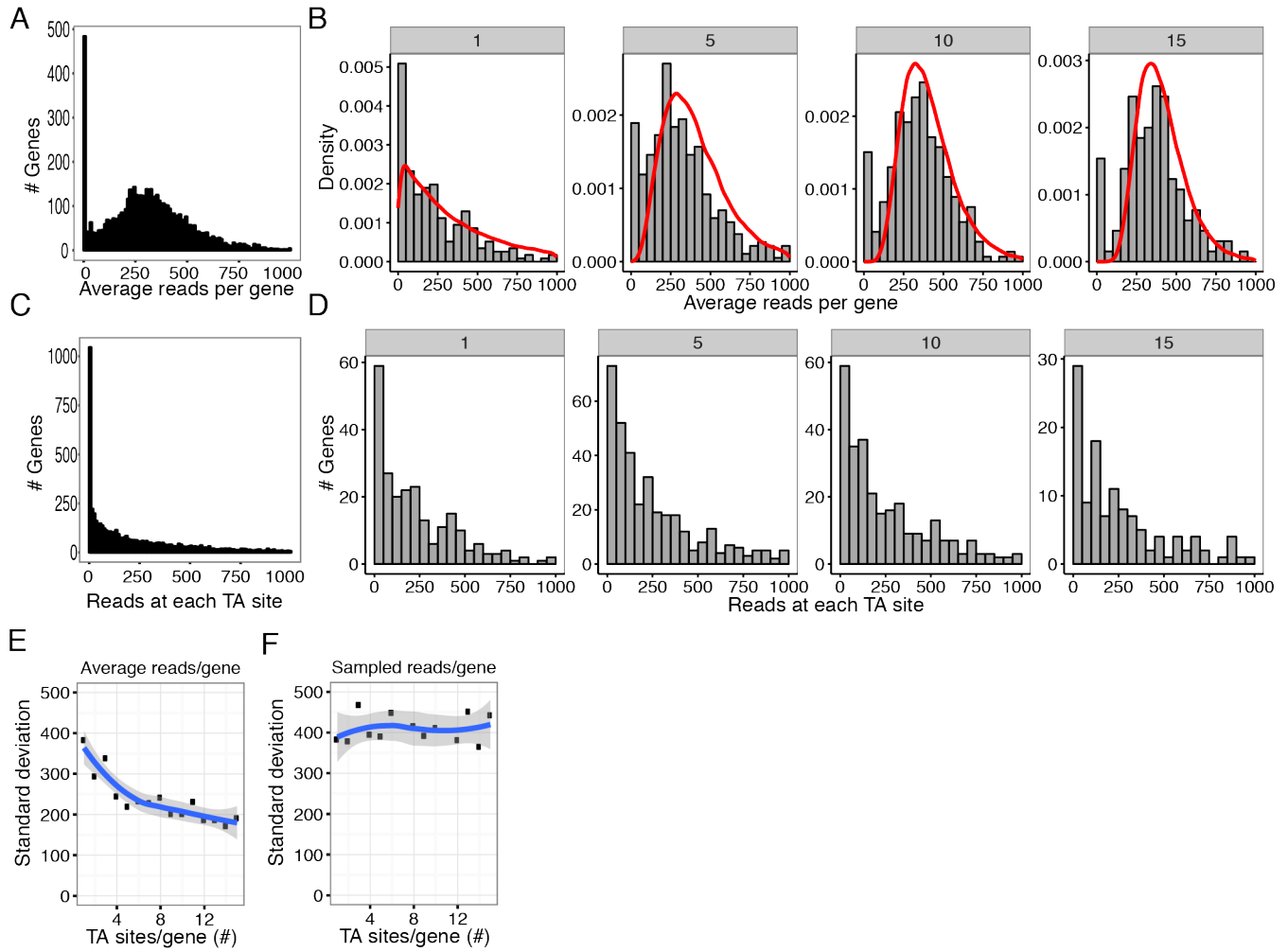


**Figure S1. Transposon mutagenesis and Illumina library construction.** *E. coli* SM10 donor cells containing either the pC9 transposase or pMAR transposon are mated with recipient *P. aeruginosa*. Transposon-integrated *P. aeruginosa* mutants are selected on solid medium: LB, M9 minimal, fetal bovine serum, synthetic cystic fibrosis sputum or urine followed by outgrowth, cell collection, and genomic DNA purification. Illumina libraries of isolated genomic DNA were constructed by: 1) end repair of ~1.5kb sheared DNA and ligation of Illumina P5 adapters; 2) PCR amplification using primers specific to the P5 ligated ends and transposon (Tn) while introducing the P7 Illumina flow cell binding site sequence; 3) size-selection of 400-500bp fragments containing the genomic-transposon junction.

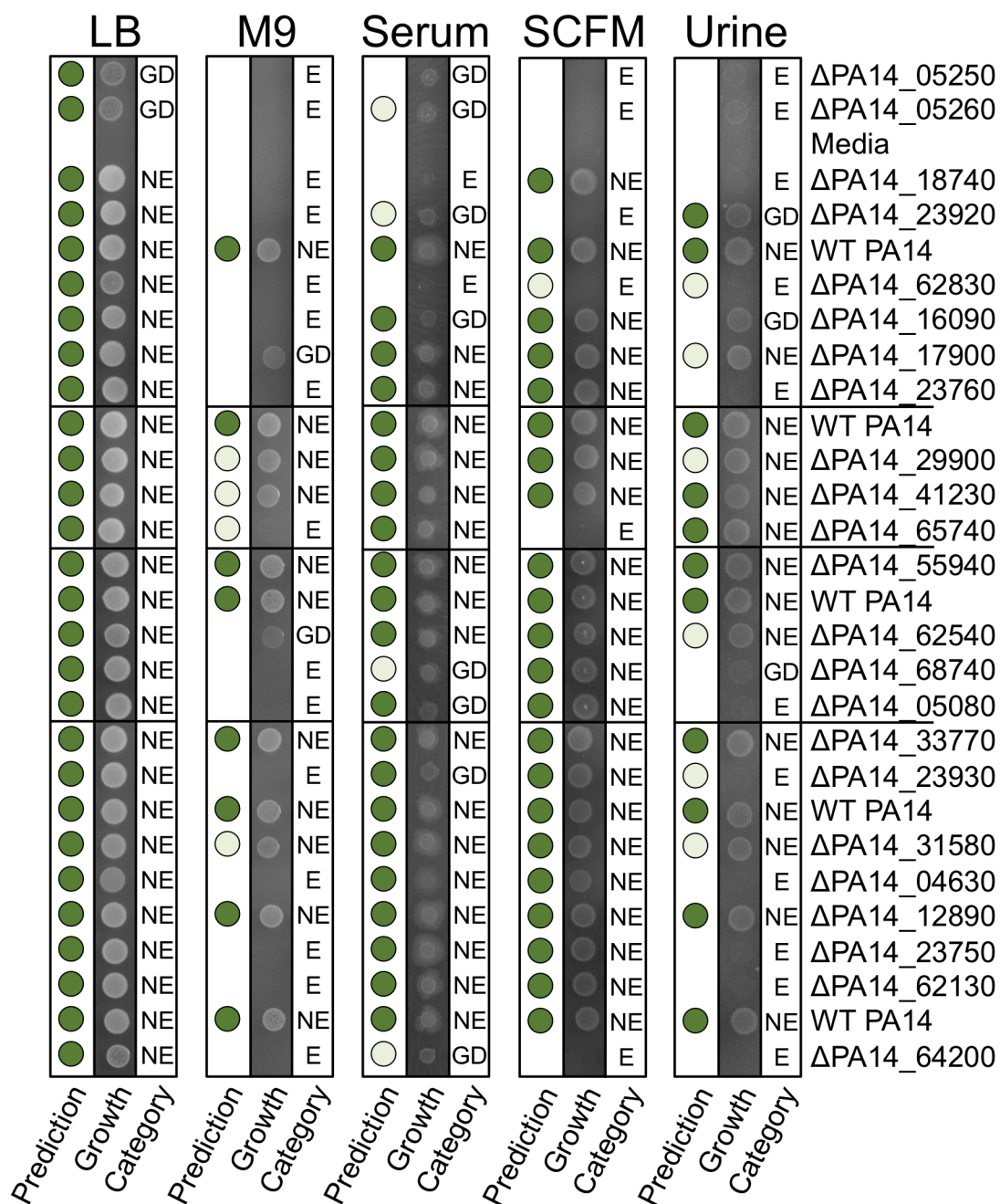




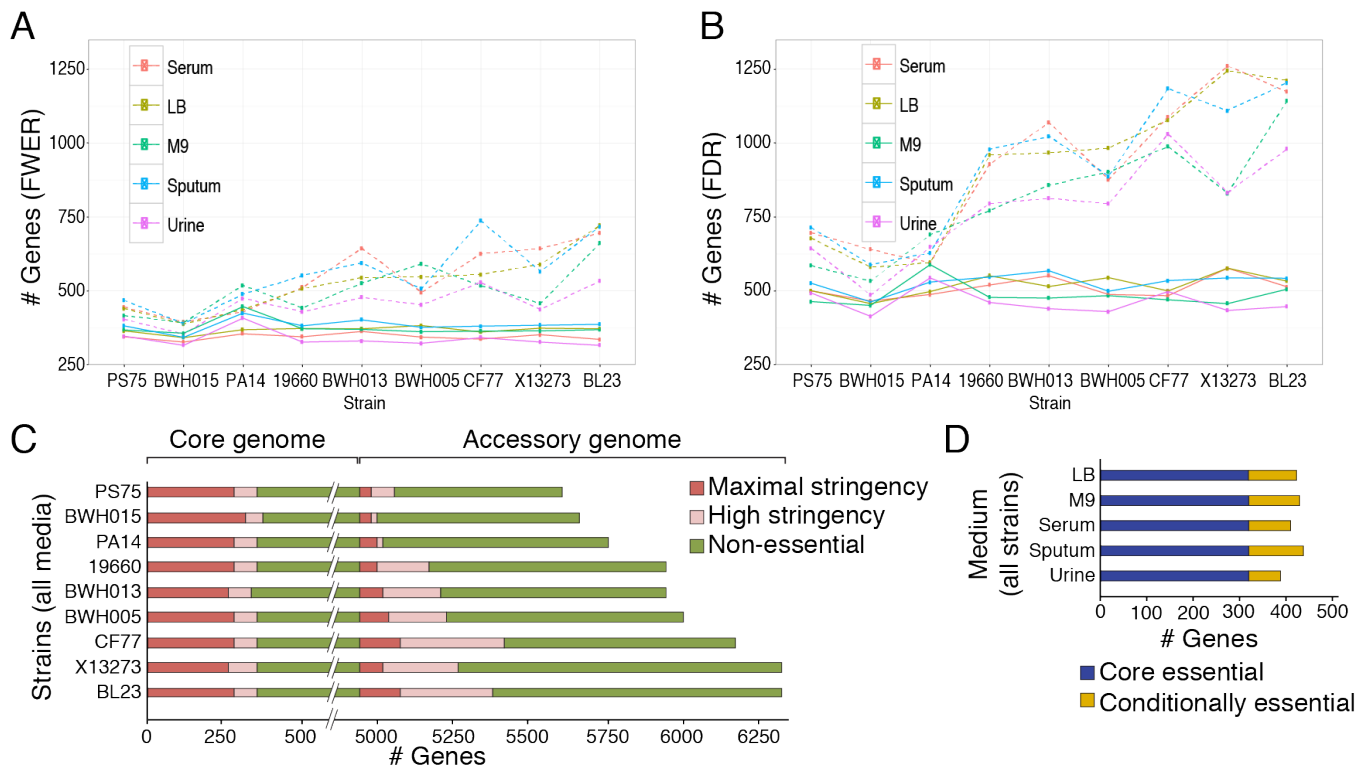
**Figure S2. Removal of non-permissive and gene-termini insertion sites from Tn-seq data analysis.**  
 A. Non-permissive TA sites. Reads at TA sites with the surrounding sequence (GC)GNTANC(GC) have significantly reduced reads compared to a random subsample of TA sites ( $p < 0.0001$ ,  $n = 6352$ , Kolmogorov-Smirnov test); PA14 LB dataset. B. Reads at termini of essential genes. 42 genes identified from our Tn-seq data that contain reads. These 42 genes are part of the consensus of 109 published essential genes (Dataset S6) and contain  $>10$  reads from the PA14-LB dataset, as indicated with a dot. Reads in these essential genes are found  $<50$  bp from either the 5' or 3' end regardless of gene size, thus distal TA sites were removed from analysis in all genes.



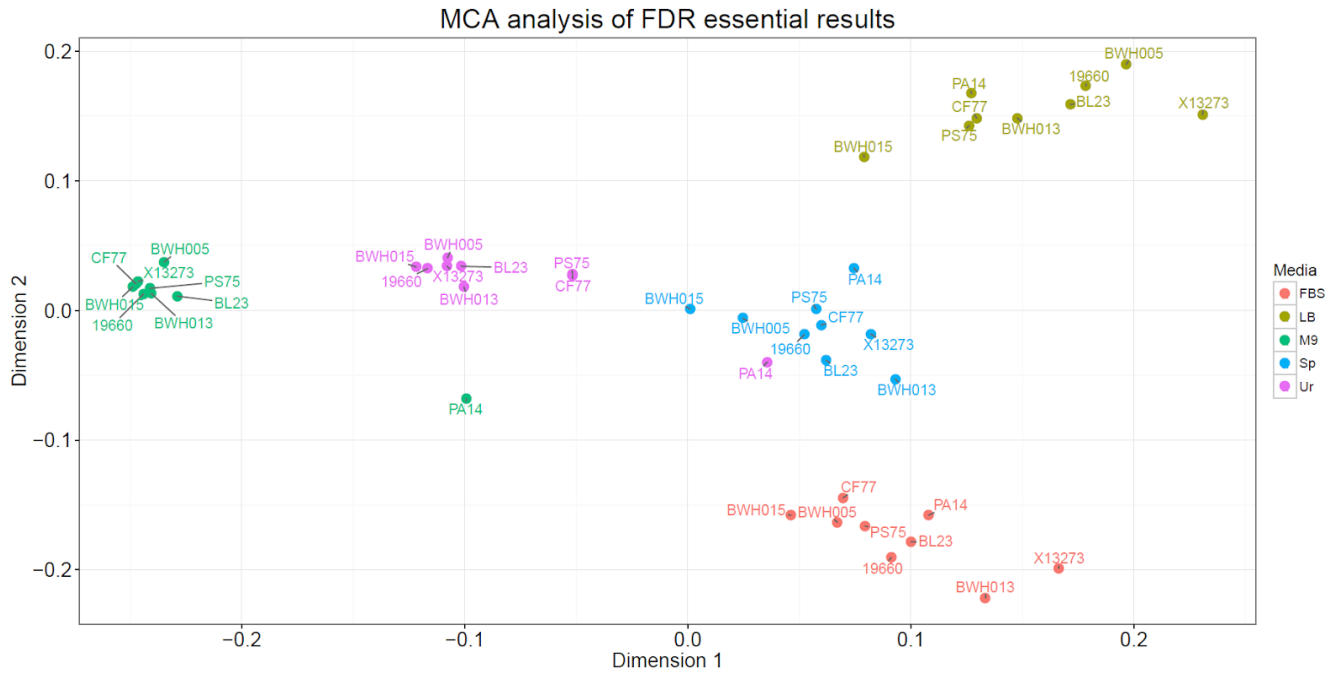
**Figure S3. Distributions of read numbers in Tn-Seq data in relation to gene size.** A. Distribution of average number of reads/TA site in a gene (ng/NTA) for all genes and B. in genes with only 1, 5, 10, 15 TA sites. The red curves are theoretical distributions for the non-essential genes simulated from our parameters; actual Tn-Seq data shown as histograms. C. Distribution of number of reads at one random sampled TA site in a gene for all genes and D. in genes with only 1, 5, 10, 15 TA sites. E. Standard deviation of average number of reads/TA site in a gene for these NTA categories is decreasing, as expected with increasing numbers of TA sites. F. Standard deviation of number of reads at one random TA site is relatively constant across different numbers of TA sites, thus showing that all TA sites are behaving similarly, regardless of gene length and numbers of TA sites in a gene.



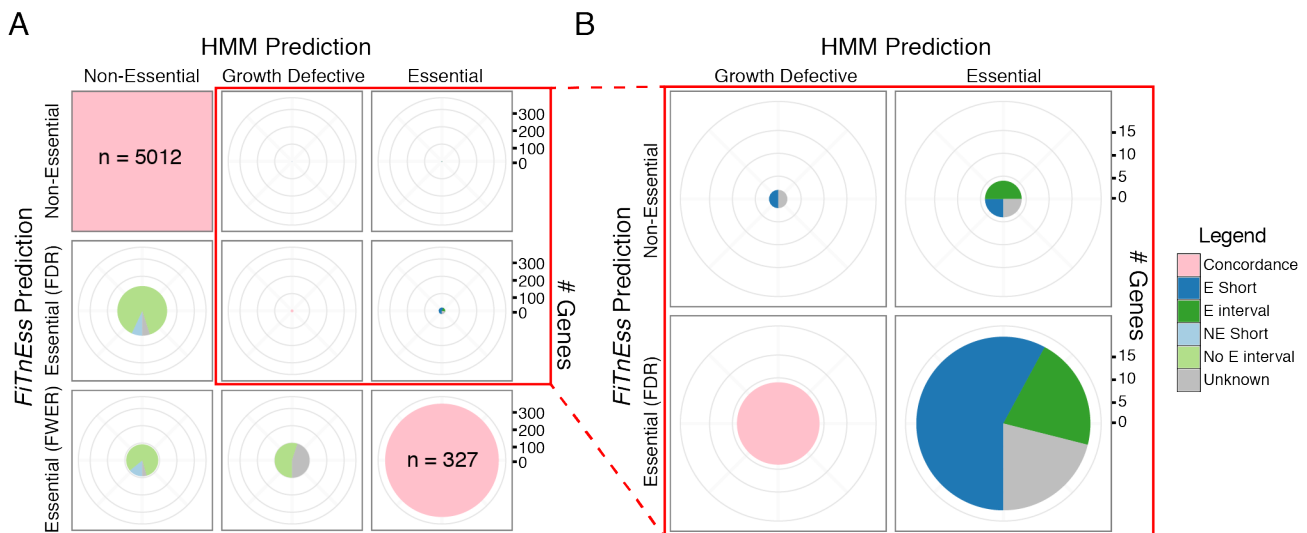
**Figure S4. *FiTnEss* validation using conditionally essential gene deletions.** 23 gene deletions in strain PA14 grown on 5 media (left to right: LB, M9, serum, SCFM sputum, urine). *FiTnEss* essentiality predictions (left columns) are depicted by: dark green circle, non-essential; no circle, maximal stringency essential; light green circle, high stringency essential. Images of growth (middle columns) are categorized using densitometry of 2 biological replicates relative to 10 replicates of WT (right columns): essential (E, 0-20% of WT), intermediate (GD, 21-50% of WT), non-essential (NE, > 50% of WT).



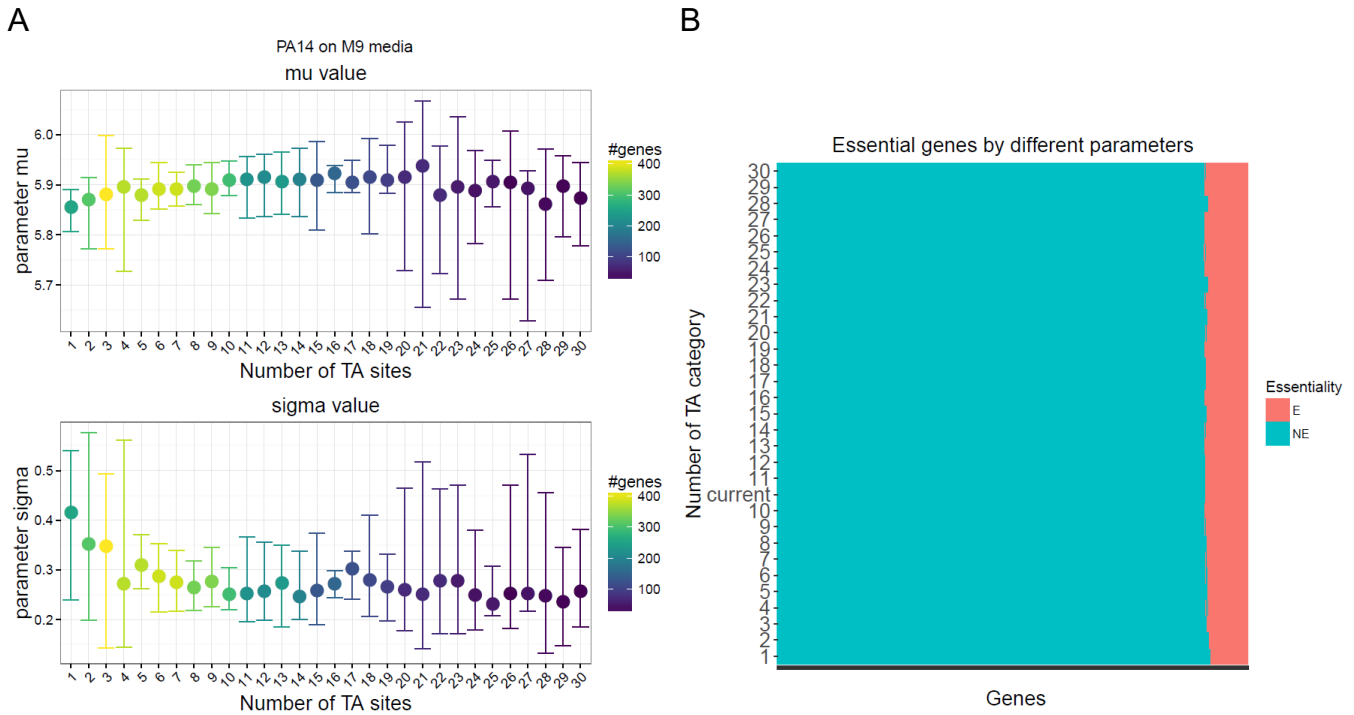
**Figure S5. Number of essential genes in genome as defined by *FitnEss*.** A. Maximal stringency (FWER) and B. high stringency (FDR) numbers of essential genes in whole genome for each strain and medium. Strains are ordered based on genome size, demonstrating the increase of accessory essential genes with genome size. Core genome essential genes are shown with solid lines, and total essential genes including the accessory genome essential genes are shown with dashed lines. C. Number of maximal stringency (red), high stringency (pink), and non-essential (green) genes common to each strain across all media, distributed between the 4903 core genes (left) and accessory genes (right). D. Number of essential genes common across all strains in each medium, highlighting the 321 core essential (blue) and conditionally essential (yellow) genes.



**Figure S6. MCA analysis of strain-media conditions.** MCA results of high-stringency essential genes identified by *FiTnEss* after FDR correction, projected on first two dimensions. Here we can see that conditions are grouped finely together based on their growth media, with PA14 being outlier in M9 minimal and urine media.



**Figure S7. Explanations of discrepancies between *FiTnEss* and TRANSIT (HMM method) predictions using the example PA14 on M9 media.** A. *FiTnEss* prediction categories (Essential (FWER), maximal stringency essential; Essential (FDR), high stringency essential; Non-Essential) versus HMM predictions (Essential; Growth Defective; Non-Essential). Pie sizes are proportional to number of genes that fall into each category (incremental circular scale shown), with pie components indicating explanations of discrepant calls. The vast majority of genes are concordant between the two methods and are colored in pink, with 5012 ‘Non-Essential’ genes and 327 essential genes. The ‘Growth Defective’ vs ‘Essential (FWER)’, ‘Non-Essential’ vs ‘Essential (FWER)’ and ‘Non-Essential’ vs ‘Essential (FDR)’ panels show genes that detected by *FiTnEss* as essential but are called less confidently (growth defective category) or completely non-essential by HMM. These genes for the most part fall unequivocally on the left mode of the ng distribution, and therefore justify *FiTnEss*’s calls. Closer examination of the data suggested that for most genes in this category, there are a moderately low number of reads that are distributed evenly across the genes. Since HMM focuses on consecutive runs of essential TA sites (that typically have near-zero reads), it has less power to detect the essentiality of such genes, whereas *FiTnEss* leverages all the TA sites in the gene and can thus more easily find a significant overall deficiency (**No E interval**). Another set of genes in this category are short genes with very few or no reads, flanked by non-essential genes with many reads. Here too, the HMM’s focus on runs of usually 5 or more essential TA sites reduces the method’s power to detect essentiality (**NE Short**). B. Scaled-up section from panel A, as indicated by the red boxes. ‘Essential’ vs ‘Essential (FDR)’, ‘Essential’ vs ‘Non-Essential’, and ‘Growth Defective’ vs ‘Non-Essential’ are genes that are called essential or growth defective by HMM but are called less confidently or not essential at all by *FiTnEss*. Many of the genes in this category do not fall convincingly in the left mode of ng distribution. These genes tend to be short and seem to “mislead” HMM by being flanked by essential genes (**E short**). Finally, there is a small set of genes that HMM seems justified in called essential as they exhibit a clear interval of essential TA sites, typically an essential region of a large gene, but that interval is being “diluted” over the entire gene and thus reduces the power of *FiTnEss* to detect them (**E interval**). In summary, while there is complementarity between the methods, *FiTnEss* seems to be more powerful for calling essential genes than HMM, with only a very small tradeoff for genes where the latter has an advantage.



**Figure S8. Robustness test of *FiTnEss* parameters.** A. Variation of fitted parameters using genes with different number of TA site categories. Genes with  $N_{TA}$  number of TA sites ( $N_{TA} \in [1, 30]$ ) were used to fit parameters respectively, and we repeated this process for 10 times in order to learn the variance. In this plot, dot showing the optimized parameter obtained from 10 runs for each  $N_{TA}$  category, while error bars showing 3 standard deviations around mean of those 10 runs. Color shows number of genes in each  $N_{TA}$  category. We found that parameters obtained using genes with 5-15 TA sites (median  $N_{TA} = 9$ ) were relatively consistent with small variations. In our analysis, we chose genes with 10 TA sites to fit parameters. B. Consistency of essential genes called by different parameters. We repeated essential gene calling process using optimized parameters obtained using genes of each  $N_{TA}$  category. Each row shows results using parameters obtained from corresponding  $N_{TA}$  categories shown on y-axis, with gene colored red to be essential, and blue to be non-essential. We could see that results are robust regardless of which  $N_{TA}$  category we used to fit parameters. Row with “current” label are our current results using genes with 10 TA sites.

## Tables

**Table S1. Summary of removed and usable TA insertion sites from analysis.**

<b>Strain</b>	<b>Total coding region sites</b>	<b>Non-permissive to insertion</b>	<b>Homologous surrounding sequences</b>	<b>Site at 5' or 3' gene end</b>	<b>Total sites removed<sup>a</sup></b>	<b>Total Usable TA sites</b>
19660	83542	6558	1942	10274	17684	65858
BL23	92065	6874	2720	11413	19732	72333
BWH005	84887	6574	1194	10400	17196	67691
BWH013	84102	6647	966	10120	16864	67238
BWH015	78683	6374	786	9448	15785	62898
CF77	89916	6936	2475	10826	19166	70750
PA14	81328	6367	1122	9829	16499	64829
PS75	75854	6199	671	9249	15309	60545
X13273	88610	6793	1416	10849	18085	70525

<sup>a</sup>TA sites may fall under multiple categorical reasons for removal from analysis



**Table S2. Summary of TRANSIT (HMM) performance based on gene deletion growth profiles from Fig. S4.**

Growth category	TRANSIT: HMM Prediction <sup>a</sup>		
	Essential	Growth-defective	Non-Essential
Essential	100% (15)	80% (12)	8% (7)
Intermediate	0% (0)	20% (3)	14% (12)
Non-essential	0% (0)	0% (0)	78% (66)

<sup>a</sup>Strain-medium instances are in parentheses

### References

1. Skurnik D, *et al.* (2013) A comprehensive analysis of in vitro and in vivo genetic fitness of *Pseudomonas aeruginosa* using high-throughput sequencing of transposon libraries. *PLoS Pathog* 9(9):e1003582.
2. Liberati NT, *et al.* (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc Natl Acad Sci U S A* 103(8):2833-2838.
3. Palmer KL, Aye LM, & Whiteley M (2007) Nutritional cues control *Pseudomonas aeruginosa* multicellular behavior in cystic fibrosis sputum. *J Bacteriol* 189(22):8079-8087.
4. DeJesus MA, *et al.* (2017) Comprehensive Essentiality Analysis of the *Mycobacterium tuberculosis* Genome via Saturating Transposon Mutagenesis. *MBio* 8(1).
5. Griffin JE, *et al.* (2011) High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog* 7(9):e1002251.
6. Jacobs MA, *et al.* (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A* 100(24):14339-14344.
7. Lee SA, *et al.* (2015) General and condition-specific essential functions of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A* 112(16):5189-5194.
8. Turner KH, Wessel AK, Palmer GC, Murray JL, & Whiteley M (2015) Essential genome of *Pseudomonas aeruginosa* in cystic fibrosis sputum. *Proc Natl Acad Sci U S A* 112(13):4110-4115.
9. Hmelo LR, *et al.* (2015) Precision-engineering the *Pseudomonas aeruginosa* genome with two-step allelic exchange. *Nat Protoc* 10(11):1820-1841.
10. Chao MC, Abel S, Davis BM, & Waldor MK (2016) The design and analysis of transposon insertion sequencing experiments. *Nat Rev Microbiol* 14(2):119-128.
11. DeJesus MA, Ambadipudi C, Baker R, Sassetti C, & Ioerger TR (2015) TRANSIT--A Software Tool for Himar1 TnSeq Analysis. *PLoS Comput Biol* 11(10):e1004401.