Defining the core essential genome of *Pseudomonas aeruginosa*

Bradley E. Poulsen[1,2,3] Rui Yang[3], Anne E. Clatworthy[1,3], Tiantian White[1], Sarah J. Osmulski[1], Li Li[1], Cristina Penaranda[1], Noam Shoresh[3], Deborah T. Hung[1,2,3*]

Pseudomonas aeruginosa is a clinically significant pathogen that has alarming antibiotic resistance rates and very few candidate drugs in development. The challenges of novel drug discovery are exacerbated by incomplete knowledge of the essential genes across the species required for survival under infection conditions. We thus sought to define the core essential genome of *P. aeruginosa* by performing transposon insertion sequencing (Tn-Seq) on nine strains of *P. aeruginosa* isolated from different infection sites including wound, eye, lung, urinary tract, and blood, with an environmentally isolated strain for comparison, in five different media conditions: three were infection relevant media (serum, sputum, urine) and two were lab-based media (LB and M9 minimal). We developed a novel statistical model, *FiTnEss*, to classify genes as essential versus non-essential across all strain-media combinations. *FiTnEss* required minimal assumptions and had good predictive power in a limited set of validation studies with mutant strains of PA14 containing clean gene deletions. A core set of 321 essential genes emerged that are the highest probability targets for successful novel drug discovery against this important pathogen.

Pseudomonas aeruginosa is a clinically significant pathogen that is a major cause of bacteremia, pulmonary, and urinary tract infections, with high mortality rates [1-3]. Due to its ability to evade current antibiotics or develop resistance, *P. aeruginosa* clinical strains are increasingly resistant to all current antibiotics [4, 5]. As such, *P. aeruginosa* has recently been classified as a priority pathogen in need of research investment and new drugs by the World Health Organization [6]. Alarmingly, only 1 in 5 antibacterial drugs succeed in clinical trials [7], and of the 48 potential antibacterials in development as of 2018, only 3 have activity against *P. aeruginosa* with only 1 of these having a new mechanism of action (www.pewtrusts.org/antibiotic-pipeline).

With the sequencing of the first bacterial genome in 1995 [8], the advent of the genomics era held the promise of revolutionizing the antibiotic discovery field by identifying a trove of potential new gene targets. However, the experience of two major pharmaceutical companies in the late 1990s to early 2000s suggests that this promise has failed to materialize [9, 10]. Among the factors contributing to failure is the existence of "genomic blind spots" that result in unforeseen gene redundancies that negate the value of a target within species subpopulations, as antibiotics with activity only against a subset of clinical isolates of a given species would be of little value [9]. A second factor contributing to failure is the concept of conditional essentiality, wherein the essentiality of a gene is dependent on its surrounding environment *i.e.,* its growth condition. It has become clear that being essential under *in vitro*, laboratory conditions does not guarantee essentiality *in vivo*, as illustrated by examples of targets or small molecules whose essentiality or efficacy, respectively, did not translate to *in vivo* conditions [11, 12].

Despite the arguably limited impact that genomics has had on identifying new, valuable targets for antibiotic discovery to date, advances in genomic technologies have significantly enabled the systematic studies of bacteria both by catalyzing the exponential increase in available bacterial genomes for comparative studies and the ability to functionally characterize bacteria on much larger-scales, including on numerous strains of a given species or in multiple conditions. Taking advantage of massively parallel sequencing, methods have emerged (Tn-Seq; also known as TIS, INseq, HITS, TraDIS, [13-17]) for performing genome-wide negative selections studies to quantitatively measure the relative fitness of a pool of mutants. Using transposon insertion sequencing, genes which are required for optimal growth under a specific growth condition can be identified by mapping the gene location of the disrupting transposon in fitness impaired mutants; at an extreme, mutants that cannot be detected at all in the pool correspond to genes that are essential under that condition. These methods have been applied in numerous bacterial studies, including two commonly studied reference strains of *P. aeruginosa*, PA14 and PAO1, with varying numbers of essential genes reported and varying essential gene identities (reviewed in [18]).

We sought to more definitively define the core essential genome – the complete set of essential genes that are

[1] Department of Molecular Biology and Center for Computational and Integrative Biology, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, United States
[2] Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, United States
[3] Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, Massachusetts 02142, United States
[*] To whom correspondence should be addressed. Email: hung@molbio.mgh.harvard.edu

common to all strains of *P. aeruginosa* and is as independent of growth condition as possible, with any bias being towards infection relevant conditions – in order to comprehensively elucidate the candidate drug targets in *P. aeruginosa.* To minimize the potential to be misled by genomic blindspots and false requirements in lab-based media, we performed Tn-Seq against a diverse set of nine *P. aeruginosa* strains from various isolation sources including pulmonary, urinary, blood, wound and ocular infections under five different growth conditions including media intended to simulate the conditions of human infection (sputum, serum, urine) and lab-based media (LB and M9). Because the value of a drug target is dependent on its binary classification as essential versus non-essential, we developed a novel, simple statistical method to map measurements of fitness to this binary classification with good predictive power. This simple model called *FiTnEss* (Finding Tn-Seq Essential genes) requires minimal assumptions and performed well, with a positive predictive value of 97% in a limited set of validation studies with mutant strains of PA14 containing clean gene deletions. We applied *FiTnEss* to the Tn-Seq data from all strain and media combinations in order to define the core set of essential genes under infection relevant conditions that are the highest probability targets for successful novel drug discovery against this important pathogen.

Results

Transposon mutagenesis, sequencing, and mapping of transposon insertions. In order to define the core essential genes across a diverse set of *P. aeruginosa* strains, we selected strains from a collection of 130 clinical *P. aeruginosa* isolates obtained from various sources (see Methods). After performing whole genome sequencing of the collection, mapping the isolates to the phylogenetic tree formed by 2560 *P. aeruginosa* genomes in NCBI, and testing a subset for their ability to be efficiently mutagenized by the Himar1-derived transposon MAR2xT7 [19-21] we focused on nine strains that represented five different infection types (blood, urine, respiratory, ocular and wound), with each strain representing a different branch of the dendrogram (NCBI ref; Fig. 1A). The genomes of these 9 strains varied from 6.34 to 7.15 Mbp.

We constructed transposon libraries by performing tripartite matings of these 9 *P. aeruginosa* strains with *E. coli* donor strain SM10 carrying an episomal MAR2xT7 transposon [20] and *E. coli* strain SM10 carrying an episomal hyperactive transposase that results in efficient integration at the dinucleotide sequence 'TA' [22]. Separating the transposase and transposon increased the efficiency of insertion sequencing and mapping, relative to the more common system of a single plasmid carrying both the transposase and the transposon, which resulted in a high percentage of reads that did not map to the recipient genome due to a second transposon in the donor plasmid which inserted itself with high frequency in the donor *E. coli* strain. Using tripartite matings, we obtained at least $5x10^6$ distinct mutants for each strain from at least two independent conjugations, and selected mutants on each of five different agar media directly to avoid a bottleneck from pre-selecting the libraries on a given medium. A total of $1x10^6$ mutants were selected on each medium in duplicate, yielding approximately a 10:1 mutant:TA insertion-site ratio, thus ensuring saturating mutagenesis. The media types included opposing rich (LB) and minimal (M9) laboratory media, and three media designed to simulate growth on infection site fluids: fetal bovine serum, synthetic cystic fibrosis sputum [23], and urine (Fig. 1B). We mapped the transposon insertions to the corresponding reference genomes for each of the 90 Tn-Seq datasets (9 strains grown on 5 media, performed in duplicate). Mapped read counts averaged approximately 10^7 for each of the 90 datasets, and reads at each TA site were highly concordant between replicates, with a mean $R^2 = 0.98$ (Dataset S1).

Qualitatively, when we examined the distribution of insertions across different strains in the different media, we could readily identify examples of genes that were variably essential under different growth conditions for a certain strain, illustrating the conditional essentiality of some genes (Fig. 1C). The relative absence of insertions in the thiamine synthesis genes *thiD* and *thiE* in M9 which does not contain thiamine, is in stark contrast to the abundance of insertions in these genes in rich LB media. Similar variability is seen for the *hemL* gene under different growth conditions. We also can clearly identify examples of genes that were variably essential in different strains under the same growth condition. The lack of insertions in *pilY1* of strain BWH013 compared to the corresponding abundance in the other 8 strains when grown on LB highlights the genomic plasticity of *P. aeruginosa* (Fig. 1D).

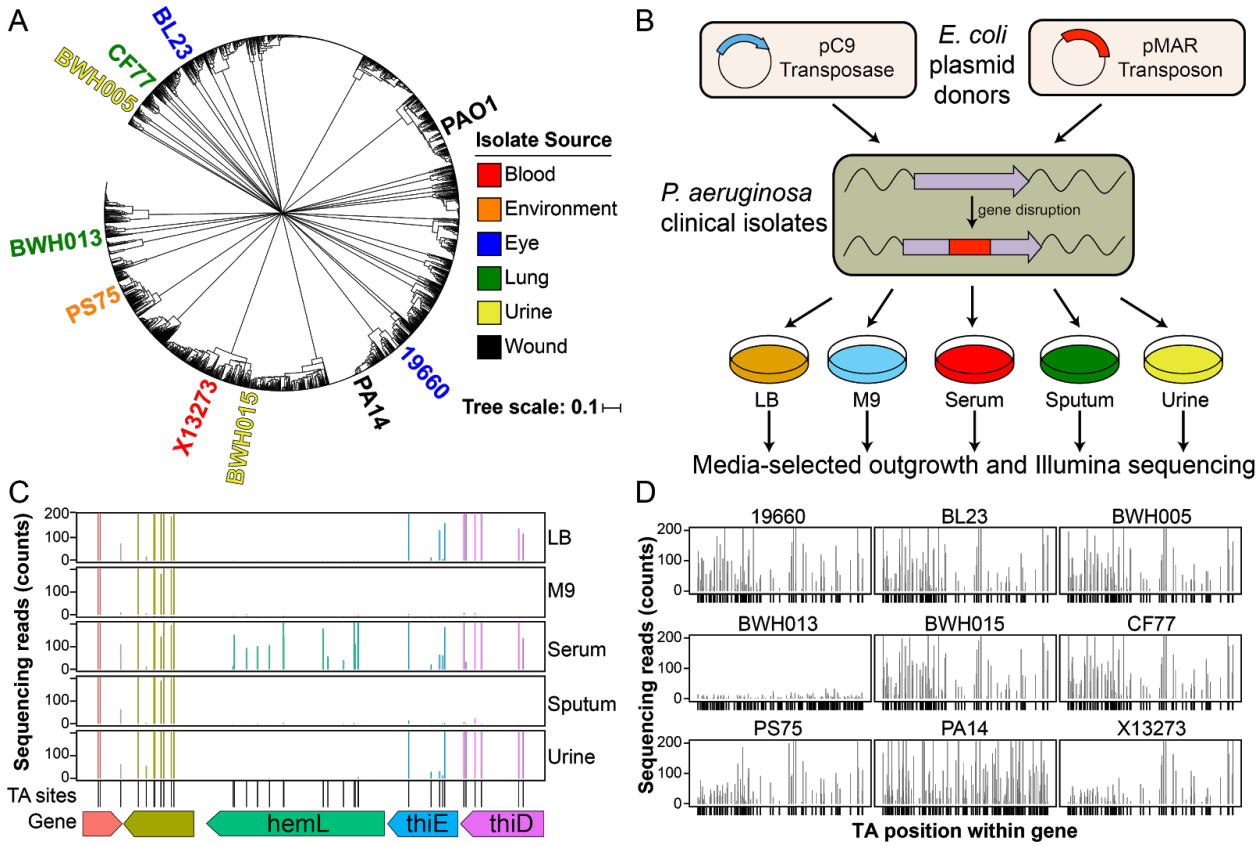

Figure 1. Tn-seq of *P. aeruginosa* clinical isolates. A. Strains selected for mutagenesis displayed on a dendrogram from the 2560 *P. aeruginosa* sequences available from NCBI. PAO1 is displayed for reference. B. *E. coli* SM10 donor cells containing either the pC9 transposase or pMAR transposon are 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 and Illumina sequencing of the transposon-genomic DNA junction. C. A highlighted region of five genes from strain PA14 showing sequencing reads mapped to TA integration sites demonstrates variable read counts mapping to *hemL, thiE* and *thiD* under different growth conditions, thereby highlighting the conditional essentiality of these genes. D. Normalized read counts for all strains in LB medium demonstrates variable read counts mapping to the *pilY1* gene, thereby highlighting the genomic heterogeneity of *P. aeruginosa* isolates.

To optimize our accuracy in calling genes essential or non-essential, we first removed confounding TA sites from the 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 recently reported to be intolerant to Himar1 transposon insertions in *Mycobacterium tuberculosis* [24], 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. S1) 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 [25]. 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 [20, 26-29], we found that 38 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. S1). 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

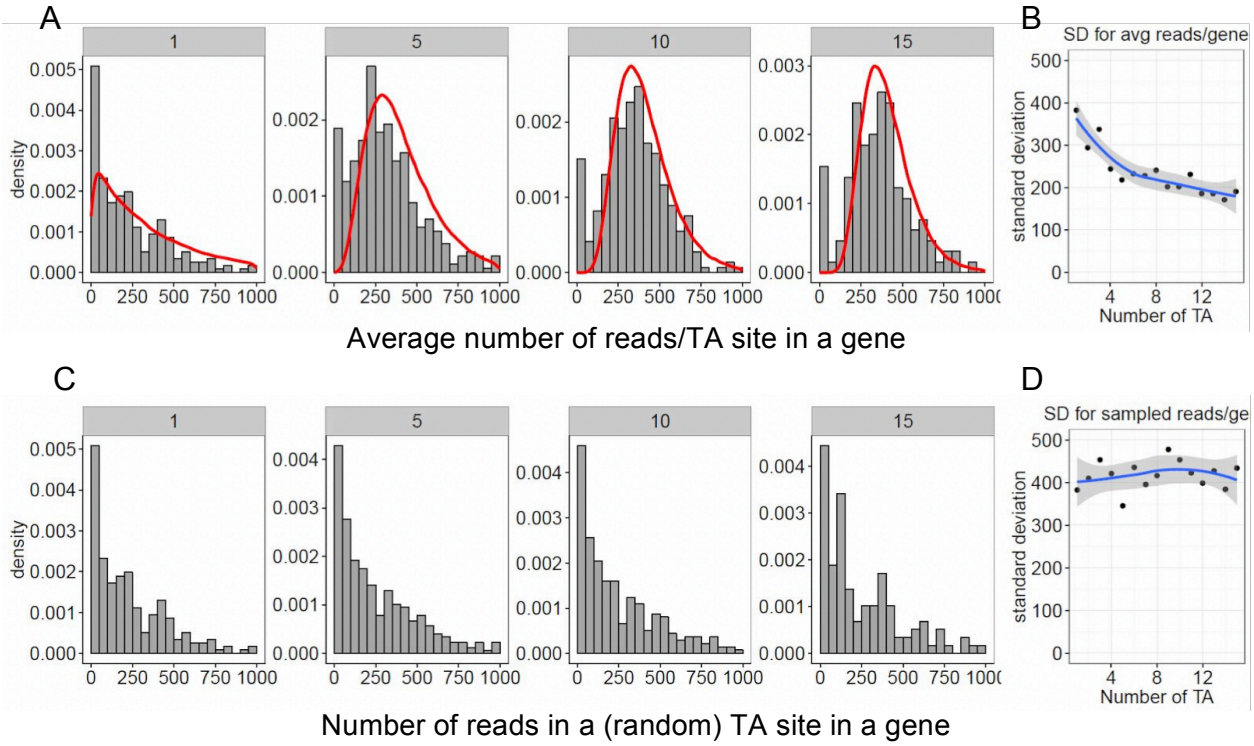

Figure 2. Distributions of numbers of reads in Tn-Seq data. A. Distribution of average number of reads/TA site in a gene (ng/NTA) for genes with 1,5,10,15 TA sites. The red curves are theoretical distributions for the non-essential genes simulated from our parameters; they matched well with the actual TnSeq data. B. 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. C. Distribution of number of reads at one random sampled TA site in a gene for genes with 1,5,10,15 TA sites. D. Standard deviation of number of reads at one random TA site is relatively constant across 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.

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. In PA14, 1122 such sites were found surrounding TA sites, 204 of which are from non-homologous genes. In total, by removing these three classes of TA sites from the exemplary genome of PA14, we omitted 16499 of 81328 TA sites (20%) from analysis, which resulted in our inability to assess 150 genes in PA14 (2.5%). Combining this with the inability to assess the essentiality of genes which contain no TA sites (35) for a total of 185 non-analyzable genes, we were able to assess the essentiality of 5708 out of the 5893 total genes in the PA14 genome (97%). We found similar trends for all strains analyzed and are summarized in Table S1.

***FiTnEss*: a statistical model to identify essential genes.** We next sought to perform a comprehensive and quantitative analysis of our 90 Tn-Seq datasets. However, while many methods exist for analyzing Tn-Seq data [13, 30-32], significant variation exists in the complexity of these methods and how conservative they are in calling a gene essential. Additionally, because of their complexity, many of them require implicit assumptions that may have contributed to their widely varying predictions of which genes are essential when applied to our datasets (data not shown). We thus developed a simple model and method (*FiTnEss*, Finding Tn-Seq Essentials) for identifying essential genes from Tn-Seq data that would require minimal assumptions with good predictive power.

FiTnEss identifies essential genes based on the numbers of sequencing reads at each TA site, absent the 3 classes of TA sites that we removed from analysis. Importantly, we evaluated essentiality based on consideration of the unit of the gene rather than the individual TA site. Thus, we began by calculating for each gene, the total number of reads **ng**, summing over the reads at all TA sites in the gene. Looking at the average number of reads per TA site for a gene, obtained by dividing the total number of reads per gene by the number of TA sites (**NTA)** in the gene (**ng/NTA**), we observed a clear bimodal distribution for genes containing more than a few (≥5 TA) sites (see NTA=5,10,15 in Fig. 2A). Presumably, in these distributions, essential genes are on the left with a small or zero ng/NTA, and non-essential ones are on the right (ng/NTA>0). Despite the fact that the distribution of the average number of reads per TA site in a gene (ng/NTA**)** is not the same for all NTA categories, with the standard deviation decreasing (unsurprisingly) with the number of TA sites (Fig. 2B), the distribution of reads at single TA sites was similar for all NTA categories (Fig. 2C,D). This suggested that all TA sites were behaving the same, independent of gene length. Given these bimodal distributions, we based *FiTnEss* on modelling of the read-number distribution for non-essential genes and fitting the model parameters from the data.

We posited that the distribution of the number of reads at any TA site in a gene is geometric with probability pg, and that the expected number of reads (1/pg) is only a function of the fitness of the bacteria when the gene function is lost. Assuming a lognormal distribution of 1/pg,

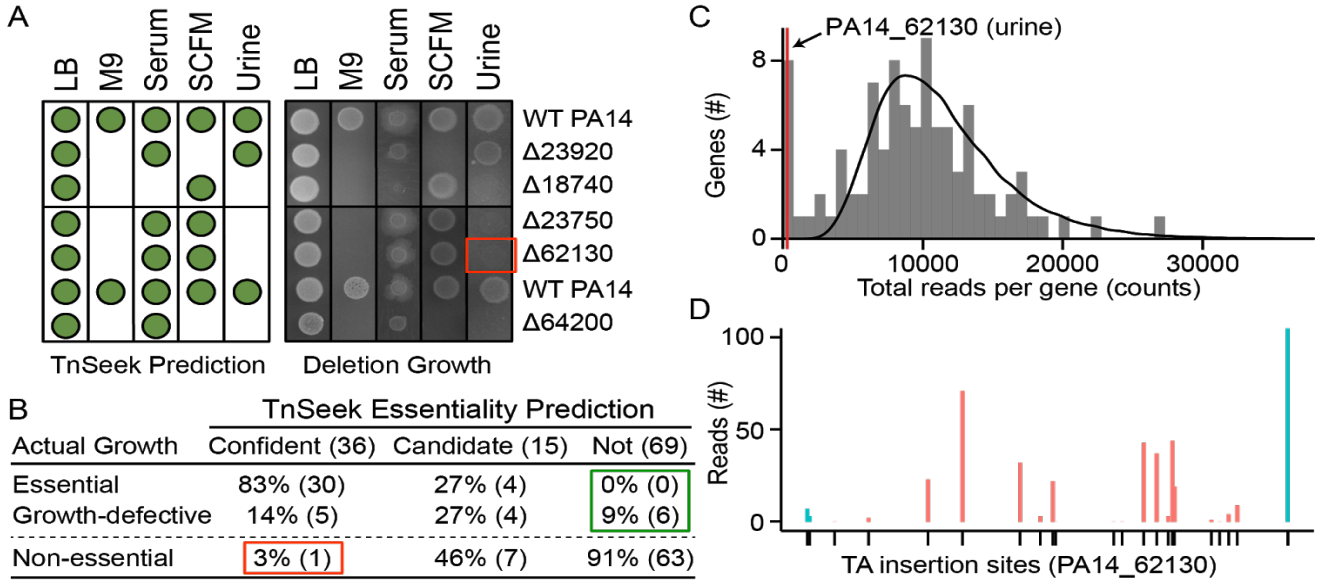

Figure 3. Validation of FiTnEss predictions on a set of conditionally essential gene deletions. A. FiTnEss essentiality predictions (left) of five representative gene deletions from strain PA14 mirror the actual growth on 5 media (right). The full growth profiles of 24 gene deletions can be found in the Supplementary material. The red box identifies the absence of growth of the PA1462130 (the deletion mutant for *ilvC* (see panel C and D below) thus experimentally confirming its essentiality on urine. B. A summary of FiTnEss performance based on actual gene deletion growth profiles. Confident and candidate essential gene categories are predicted based on family-wise error rate and false discovery rate corrections, respectively; gene/medium instances are indicated in parentheses; red and green boxes demonstrate false positive and negative rates, respectively. C and D. Example of FiTnEss prediction of essentiality for a gene containing multiple transposon insertions. Representative FiTnEss output for all conditionally essential genes in the validation set (C). The grey histogram shows the actual distribution of total reads for each gene, the black line shows the theoretical distribution calculated by FiTnEss, and the red line denotes the behavior of the gene *ilvC* (PA1462130) in urine, showing that it clearly falls in the far left (essential part) of the bimodal distribution. *ilvC* was demonstrated in panel A to be truly essential. Read numbers at each TA insertion site of *ilvC* (D). Blue bars indicate TA sites that were removed from analysis due to their proximity to the gene termini.

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 (NTA = 10; top 75% of the distribution). Given the fitted parameters of the model, a specific dependence of the distribution of ng on the number of TA sites was predicted, which agrees well with the actual data (red curves in Fig. 2A).

Applying this model to all 90 datasets, we determined p-values for the likelihood of any given gene to be drawn from the population of non-essential genes. Parameters for datasets from replicates of each condition (i.e. each combination of strain and medium) were fitted separately and applied to their corresponding dataset. All p-values were subsequently corrected in each dataset to account for multiple-hypothesis testing. Finally, we called a gene "essential" if its adjusted p-value is smaller than 0.05 in both replicates. We applied two methods of p-value adjustment: a very conservative family-wise error rate (FWER) correction offered a high confidence set of essential genes ("confidently essential"), and a more commonly used but less extreme false discovery rate (FDR) correction predicted a larger set of genes that not only included the "confidently essential" gene set but also an additional set of genes that are likely to be essential ("candidate essential"). Virtually all confident calls are expected to be true essential genes, while among the candidate essential set, a small number of false positives is expected.

Validating *FiTnEss* using strain PA14. In order to validate *FiTnEss*'s approach to predicting gene essentiality, we took advantage of the conditional essentiality of a subset of PA14 genes on the different growth media to compare *FiTnEss* predictions with actual viability and growth for a set of PA14 mutants in which we had disrupted particular genes of interest. We identified 24 genes that were identified as non-essential on LB, thus allowing us to create clean gene deletions of these genes: 18 of these genes were identified by *FiTnEss* as conditionally essential in at least one medium after both FWER and FDR corrections (confidently essential), 3 after FDR correction alone (candidate essential), and 3 were identified as non-essential by both corrections but had p-values approaching the cutoff for calling essential (Dataset S2). We determined the positive and negative predictive values of *FiTnEss* by growing the 24 mutants on the same 5 media as used in the original Tn-Seq experiments, for a total of 120 gene-medium combinations. Mutant strain viability was categorized as essential, growth-defective, and non-essential using densitometry (<20%, 20-50%, and >50% relative to WT, respectively). (Fig. XA-B and Fig. S2). Of the 36 combinations predicted to be confidently essential, 30 were correctly identified as essential and 5 were growth defective. Only a single false positive prediction was made, if growth defective strains are considered true positives, for the flagellar synthesis regulator gene *fleN* in the urine growth medium, Of the 69 combinations predicted to be non-essential, no instances were found as essential, but 6 instances were found to be growth defective. In this limited dataset, *FiTnEss* had a positive predictive value of 97% (if considering growth defective genes as essential) and 83% (if growth defective genes are considered non-essential), and a negative predictive value of 91% (if considering growth defective genes as essential) and 100% (if growth defective genes are considered non-essential) thus building confidence in its ability to accurately call essential and non-essential genes. Importantly, *FiTnEss* correctly predicted gene essentiality despite the presence of a small number of mapped insertions in the primary Tn-Seq data, as exemplified in the case of the *ilvC* gene encoding ketol-acid reductoisomerase (Fig. XC-D).

Predicting essential genes to define the core essential genome. We applied *FiTnEss* to all 90 datasets to identify confident or candidate essential genes based on FWER or FDR corrections, respectively (Table S2). We favored the extremely conservative FWER correction when considering a single strain-medium combination, and the FDR correction when considering multiple strains and/or multiple media conditions to avoid missing essential genes, as statistical power increases with sample size.
Before turning to examine essential genes, we first identified all genes (both essential and non-essential) that are common to all 9 strains, thus defining a 4903 gene common genome using the orthogroup clustering software Synerclust (https://www.broadinstitute.org/genome-sequencing-and-analysis/tool-development). Our common genome consisted of 4903 single-copy genes. The numbers of common genes across these 5 conditions is comparable to numbers which have been previously described for the common genome for *P. aeruginosa* using a much larger set of strains, after removal of genes that cannot be assessed by Tn-Seq (5001 genes, differing by 2%). The accessory

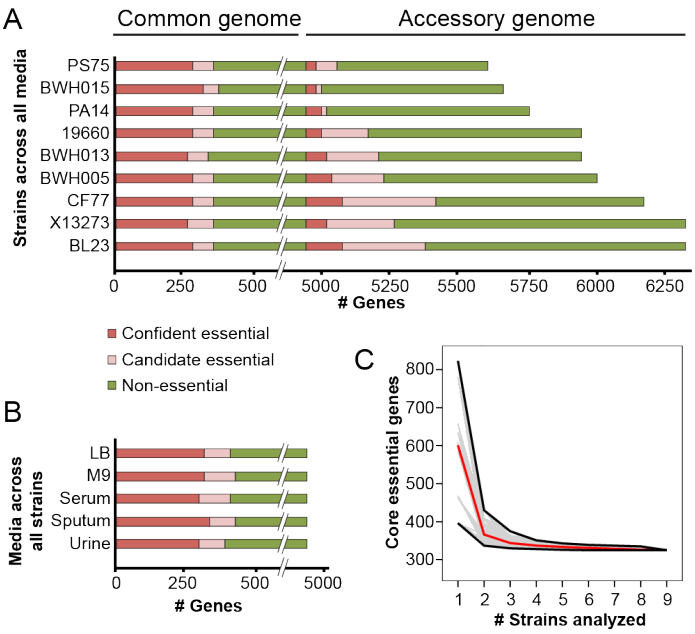

Figure 4. Assessing the essential genes in *P. aeruginosa* as determined by FiTnEss. A. The number of confident (red), candidate (pink), and non-essential (green) genes common to each strain across all media is shown, distributed between the 4903 common genes (left) and accessory genes (right). B. The number of confident, candidate, and non-essential genes common across all strains in each medium is shown. C. 10,000 random simulations (grey) of the trajectory of the number of core essential genes determined upon the sequential introduction of additional strains, up to a total of 9 strains. The largest and smallest simulated trajectories of core essential genome sizes are highlighted in black, and the mean in red.

genome (all genes that appear in one or more strains but not in all strains) consisted of 655-1369 genes.

Turning to the essential genes, we found that the number of essential genes in a single strain in any single medium varied between 354 and 727 genes (Table S2). The numbers of these genes belonging to the common versus the accessory genome revealed that most strain-growth condition combinations had approximately the same number of essential genes in the common genome (Fig. 4A). In contrast, there was significant variation in the numbers of essential genes in the accessory genome; interestingly, these numbers were proportional to genome size (Fig. 4A and Fig. S3).

Since candidate drug targets should be present and essential in all strains under relevant infection conditions, we sought to define the essential genes that were contained within the set of 4903 common genes, thus making up the set of common essential genes for each growth condition. Sputum and M9 had the highest number of common essential genes (439 and 431, respectively), consistent with these being the most nutritionally depleted media. LB had 424 common essential genes, while urine and serum had the fewest (400 and 412, respectively) (Fig. 4A and Table S2). While the numbers of common essential genes required in each growth condition did not vary significantly from condition to condition, the actual gene identities did vary such that the overlap of the 5 common essential gene sets derived from these 5 media was only 321 genes, now called the core essential genome. To ensure that we had analyzed sufficient strains to approach an asymptote of the core essential genome, we simulated the trajectory of numbers of essential genes upon the addition of an increasing number of strains in a random order (10,000 simulations) and found that indeed, an asymptote was reached after ~4 strains, regardless of which strains were selected (Fig. 4B).

We examined the identities and functions of the core essential genes. 263 of the 321 core essential genes correspond to cytosolic proteins, with 132 of the cytosolic proteins involved in metabolic pathways (50%) and 119 involved in macromolecular synthesis including DNA replication, transcription or translation (45%). 56 of the 321 genes correspond to cytoplasmic membrane, periplasmic and outer membrane proteins with the majority involved in cell structure and division, metabolite transport, or act as protein chaperones (14, 12 and 13 genes, respectively). 2 of the 321 genes are completely uncharacterized (Fig. 5B).

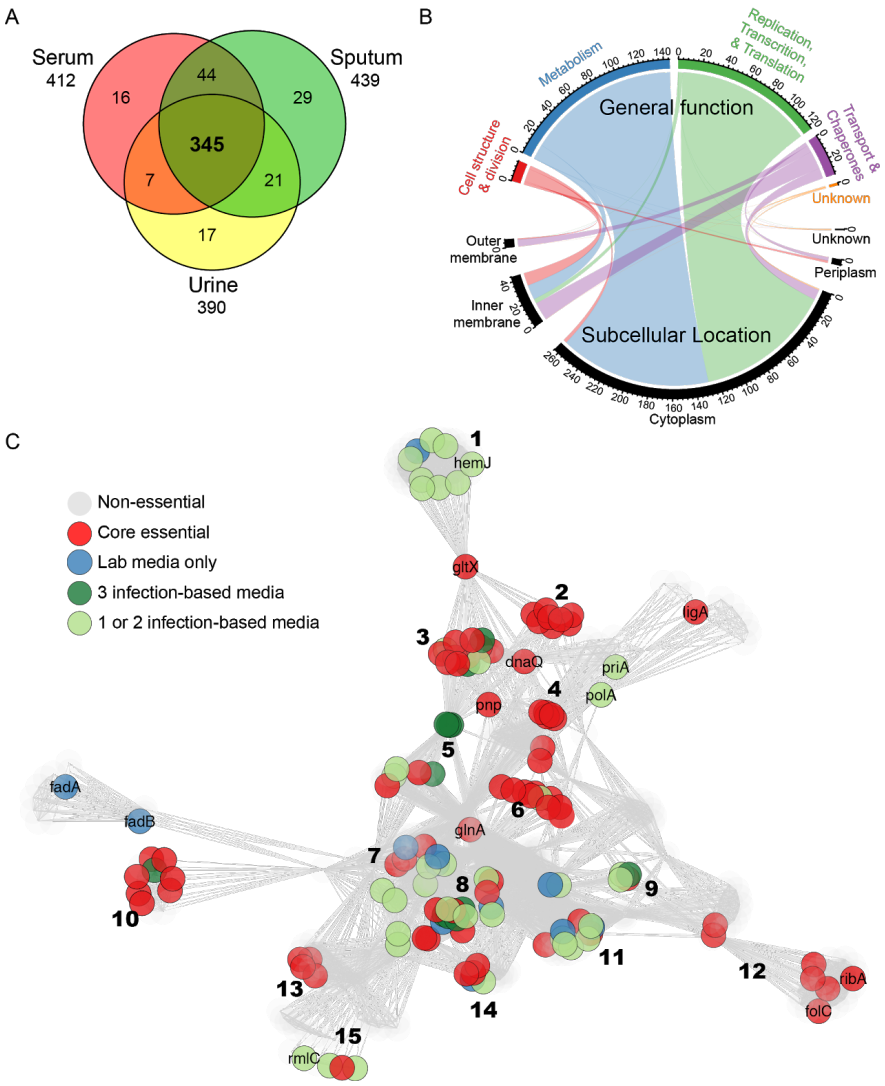

Figure 5. Core and conditional essential gene functions. A. Venn diagram representing the number of essential genes in all strains across three infection-relevant media. B. A chord diagram showing the relationship between subcellular location and general function of the 321 core essential genes. C. KEGG metabolic enrichment analysis of the core and conditionally essential genes. Interactions of genes defined by KEGG pathways are shown as gray lines and genes are shown in circles colored based on essentiality: core, red; lab-based media (LB and/or M9) only, blue, 3 infection-relevant media, dark green; 1 or 2 infection relevant media, light green; non-essential, light grey. Notable pathways are numbered: 1) Porphyrin synthesis; 2) Lipopolysaccharide biosynthesis; 3) Transcription; 4) DNA replication; 5) Purine and pyrimidine synthesis; 6) Terpenoid backbone biosynthesis; 7) Pyruvate metabolism; 8) Citrate cycle; 9) One carbon pool by folate; 10) Ubiquinone biosynthesis; 11) Glycine, serine and threonine metabolism; 12) Folate biosynthesis; 13) Amino sugar and nucleotide sugar metabolism; 14) Pentose phosphate pathway; 15) Polyketide sugar unit biosynthesis.

In addition to the core essential genes, the common genome also contains genes which are essential in one or more -- but not all -- media. These include 103 genes essential for the growth of all strains in LB, 110 genes required in M9, 91 genes required in serum, 118 genes required in sputum, and 69 genes required in urine (Fig. 4A and Dataset S4). In addition to the 321 core essential genes, there are an additional 24 essential genes required for growth in all three infection-relevant media (serum, sputum, & urine; Fig. 5A) but not in both of the lab-based media (LB, M9). Several of these genes are involved in pyrimidine and purine synthesis that are not essential in LB, suggesting that sufficient nucleotide intermediates may be present in this medium to sustain growth compared to the infection-based media.

Contained within the sets of conditionally essential genes are genes that are only essential in a single medium for all strains, termed unique conditionally essential genes. Considering only the three infection-relevant conditions while ignoring the laboratory conditions, sputum had 29, serum had 16, and urine had 17 unique conditionally

essential genes. These unique conditionally essential genes carry the intriguing potential of becoming infection site-specific targets for infection type specific antibiotics, i.e., a urine specific anti-pseudomonal antibiotic. The essential genes unique to sputum consist mainly of biosynthetic pathways such as thiamine, pyridoxine, and tryptophan synthesis. Even with tryptophan present in sputum, the quantities present may not be sufficient for growth [23]. Similarly, urine-specific essential genes almost exclusively consist of genes involved in amino acid biosynthesis, specifically targeting pathways for leucine, methionine, isoleucine, and arginine, suggesting that the quantities of these amino acids may not be sufficient to support growth of an auxotroph. In contrast, while genes required for amino acid biosynthesis are not required in serum, likely due to their relative abundance in serum [33], multiple cytochrome c-related proteins were uniquely essential in serum. Although porphyrin biosynthetic genes had previously been reported to be essential in *P. aeruginosa* [20, 26-29] (Dataset S3), this study found them to be dispensable in serum, with much higher levels of porphyrins in blood than urine [34], suggesting that *P. aeruginosa* can scavenge enough porphyrin from its environment, resulting in the dispensability of its synthesis in serum.

Comparison of the numbers of total essential genes gave the expected result: growth on nutrient rich LB requires fewer essential genes than growth on stringent M9. If we take into account all 5 media, it is also somewhat expected that the only condition that had conditionally essential genes that were unique to only itself (being essential in all 9 strains), but were not essential in any other strain in any other condition was M9 (20 genes). If we compare only LB and M9 conditionally essential genes, LB had 19 essential genes that are not required in M9 and M9, as expected, had more (61) essential genes required than in LB; genes essential in M9 that are not essential in LB are, not surprisingly, predominantly involved in metabolism. If we relaxed the stringency just a bit and required a gene to be essential in all 9 strains in LB, but allowed it to be essential in no more than a single strain in all other conditions, we were somewhat surprised to find a small number (5) of conditionally essential genes that are, for the most part, unique to LB, and are not required in any other condition. These genes include the *minC* and *minD* genes which play a role in determining the site of the septum during cell division [35]. However, *minC* and *minD* disrupted mutants have been reported [20], suggesting either that disruption of *minC* and *minD* may retard growth in LB more significantly than it does in other media or that the reported disrupted mutants may contain compensatory mutations or not truly disrupt the function of these genes; clearly, further studies are required to understand this interesting finding. Finally, when we applied Mulitple Correspondance Analysis (MCA) to all sets of essential genes for every strain-growth condition, we were reassured to find that all strains, for the most part, clustered together by growth condition (Fig. S4). Interestingly, one strain PA14 was an outlier under two conditions, M9 and urine. This behavior could be a result of the strain simply being a genetic outlier; alternatively, one might speculate that this might be a consequence of PA14 being the one laboratory strain which has adapted to laboratory conditions over a long period of time; further study is required to understand the anomalous behavior of this strain. Together, these datasets highlighted the tremendous differences required by *P. aeruginosa* in different microenvironments but also allowed the recognition of those genes that are conserved and may serve as good drug targets.

Discussion

Here we used Tn-Seq and a novel method of analysis, *FiTnEss*, to establish the core essential genome of *P. aeruginosa* within multiple clinical and environmental isolates and across five different lab and infection-relevant media in order to define essential targets for antibiotic discovery in this important pathogen. We determined that while a single strain has ~400-800 essential genes, the core essential genome across all strains analyzed is 321 genes which represent the most attractive candidates for discovery efforts. Further, there are an additional 24 essential genes required for growth in the three infection-relevant media examined, which are non-essential in LB and M9 media. Finally, we find that there are ~15-30 unique, conditionally essential genes for each of the infection-relevant medias examined, suggesting that the biological pathways to which they belong are important for survival only within a particular host tissue and that they may represent a unique set of targets for infection-type specific therapeutics.

Previous transposon mutagenesis studies of two common lab strains, PA14 and PAO1, have found varying numbers of essential genes, as reviewed in [18]. These studies have predominantly focused on identifying genes refractory to transposon mutagenesis when selected for growth on LB media [20, 26, 28], though more recent studies have also examined essentiality on minimal, sputum and BHI media [27, 29]. A comparison of all of these datasets combined, revealed an intersection of only 109

essential genes among these studies (Dataset S3). This low concordance may be due to methodological or analytical differences between the studies.

Experimental methods for identifying gene essentiality have varied greatly through the years. The advent of genomics made possible significant advancements in methods for defining fitness costs of gene disruptions. Nevertheless, several limitations to methods such as Tn-Seq continue to exist and must be kept in mind when applied to comprehensively defining candidate targets for antibiotic discovery. First, Tn-Seq studies, including this study, use pooled mutants when performing selection under a certain growth condition. However, some mutants may behave differently in a pool where there can be both competition as well *trans* complementation than when grown individually. Secondly, technically, transposon libraries are often constructed with an initial isolation on a rich medium and then subjected to a selection for growth on the condition of interest (often a more minimal media or media that models the host [13]. Because the isolation step is in fact a selection on rich media, genes that are essential in rich media cannot be evaluated in this way. To avoid this limitation, we omitted the initial isolation/selection step and plated the mating directly to the medium of interest. This approach allowed us to identify 103 conditionally essential genes which are required in LB, but not in at least one of the other 4 media. Finally, every transposon has limitations including the mariner transposon, which we used. We tried to minimize the confounding influences of this particular transposon by taking into account its sequence bias for insertion [24, 36], discounting from consideration TA sites within 50 bps of the 5' and 3' termini (a distance that we determined empirically based on the behavior of a truth set of essential genes), and omitting from analysis TA sites with surrounding homology that would result in inexact mapping of insertions. Removing these confounding sites from analysis, we were unable to query the essentiality of approximately 5% of the genome. Nevertheless, biases and limitations remain, including the TA site preference of the mariner transposon (which could be mollified by performing complementary studies with a different transposon with a different sequence preference) and the impact that polar effects can have on surrounding genes.

The 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 [13, 31]. 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 which 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. S5 for comparison of methods, Dataset S2 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 120 conditionally essential gene-growth condition combinations

that we validated using clean genetic deletions (Fig. X and Fig. S2). 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 S3). 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.

The great majority of core essential genes identified by *FiTnEss* can be broadly categorized as being involved in metabolic pathways, DNA replication, transcription or translation. Not surprisingly, these are already the categories of gene functions that are targeted by antibiotics. That the core essential genome is dominated by genes involved in macromolecular synthesis (*i.e.* protein and nucleic acid) may explain in part why most antibiotics seem to target a limited set of functions, *i.e.,* those involved in macromolecular synthesis. There has been greater reticence to target metabolic pathways because of concern over the ability of bacteria to scavenge nutrients from the host, thereby rendering their biosynthesis nonessential during infection. To do so requires validation of the target under all relevant infection conditions.

Importantly, we have identified several metabolic processes that are part of the core essential genome that have not previously and explicitly been defined as essential in *P. aeruginosa* by other studies (Dataset S3). For example, chorismate synthase (*aroC*), the last step of the shikimate pathway, is known to be essential in bacteria with its role in the biosynthesis of aromatic amino acids and folates [37]; previous studies however, have found that it is permissive to transposon insertion [20], albeit resulting in growth impairment. Here we find that it is indeed essential in all media; together with its absence in humans, this validation of its core essentiality demonstrates its value as a drug target. This example illustrates, *FiTnEss*'s assignment of the classification of essential to a gene, despite the fact that the corresponding mutant is growth impaired rather than non-viable. Such cases result from imposing a binary classification of genes as essential or non-essential on data that is really a relative fitness continuum of mutants within a pool. Additional examples include the genes *hfq rpoN*, and *gidA* where mutants containing disruptions of these genes are available [38-40]. *FiTnEss* classified them as confidently essential or candidate essential in most medias including LB, though mutants disrupted in these genes have been shown to be viable but with significant growth defects and/or reduced virulence in various models [38-40].

Genes that are conditionally essential in only infection-relevant media but not all lab media are of particular interest as they highlight the ability to be misled in understanding *P. aeruginosa* pathogenesis by limiting studies to and extrapolating behavior from artifical laboratory conditions. They also demonstrate the ability of these mutant libraries to probe and provide important insight into how bacteria cope with these respective growth conditions, as illustrated by the recognition of the dependence of amino acid biosynthesis and cytochrome C proteins on growth condition. With regards to these latter proteins, genes involved in oxidative phosphorylation including genes encoding the cytochrome bc_1 complex (PA14_57540, PA14_57560, PA14_57570) and cbb3-type cytochrome c oxidase proteins CcoI and CcoN (PA14_44440, PA14_44370), are conditionally essential in serum and CF sputum but not lab media nor urine, which suggests that alternative ubiquitinol pathways are insufficient to support growth in serum and CF sputum.

Importantly, despite the goal to define the core essential genome of *P. aeruginosa* as a means to comprehensively identify candidate drug targets, this study is inextricably linked to the concept of conditional essentiality by virtue of the fact that we had to select some limited set of conditions upon which to perform these negative genetic selection studies; the set of genes identified as the core essential gene set is thus conditionally dependent on the nature and intersection of these selected conditions. We tried to mitigate this conditionality by including two different lab medias, LB (rich) and M9 (minimal), to provide the boundaries (extremes of growth conditions) for essential gene identification. (Moreover, the inclusion of LB as a selection media was important as a benchmark for comparison of this study with previous studies.) Furthermore, given that the task of defining the common essential genome was tied to the goal of identifying relevant drug targets in infection, we wished any bias to be toward identifying genes relevant for growth in infection-relevant conditions. While selected *in vitro* conditions surely do not entirely replicate conditions seen by the bacteria in the human host, we propose that they provide a first step in validating genes potentially relevant to human infection.

Indeed, performing complementary *in vitro* and *in vivo* Tn-Seq studies may be the way forward toward defining and validating better targets. Performing selections on *in vitro* conditions simulating host physiologic conditions allowed us not only to perform Tn-Seq on many strains, but also to perform selections without first going through an LB growth bottleneck, which is typically required to define genes essential to *in vivo* animal infection models [28]. By plating Tn-Seq libraries directly onto the relevant media without pre-expanding and thus pre-selecting the transposon library in LB, we were able to query the essentiality of genes on the various medias, independent of its essentiality on LB, thus providing an important complement to *in vivo* studies.

The goal of this study was to understand the breadth of true core essential genes within a bacterial species that has great phylogenetic diversity on media that might most closely resemble the human host. We anticipated that these *in vitro* conditions would mirror the nutritional requirements available to *P. aeruginosa* within different host tissues at least to a first approximation. This study thus defines and refines the scope of potential targets for future drug-discovery efforts with the hope of potentially reducing the risk of false positive targets that may have caused prior drug discovery efforts to fail for reasons such as unforeseen gene redundancy within species subpopulations and false activities from lab-based media [11, 12, 41]. Further, we suggest that the prior apparent failure of genomics to transform antibiotic discovery is not due to an inherent failure in its ability to reveal valuable targets, but instead because the genomic experiment was performed too early, when only a limited number of bacterial genomes was available. The advancements in genomic technologies now make possible studies on a much greater scale, allowing us to define essential genes in an unprecedented way that will likely rectify previous shortcomings. To facilitate future large scale studies, we have developed a new, simple method for calling essential genes, *FiTnEss*. We anticipate that larger scale Tn-Seq studies -- across many strains and growth conditions and using *FiTnEss* to determine gene essentiality, will better define the core essential genomes for many other bacterial species, thereby comprehensively revealing targets for the discovery and development of new, much needed antimicrobial therapeutics.

Materials and Methods

Strain selection and plasmid construction. A genome tree report of 2560 sequenced *P. aerguinosa* strains was downloaded from NCBI (organism ID: 187) and visualized with iTOL [42]. Nine strains were selected for genetic diversity and graciously gifted from various sources: PA14, 19660, X13273 obtained from Frederick M. Ausubel [43]; BWH005, BWH013, BWH015 were collected through Brigham and Women's Hospital Specimen Bank per protocol previously described [44]; BL23 from Bausch & Lomb [45]; PS75 from Paula Suarez, Simon Bolivar University, Venezuela; and CF77 from Boston Children's Hospital [46]. pC9 was derived from pSAM-Bt [28] by digesting with ApaLI + AccI, and pMAR was derived from pMAR2xT7 [20] by digesting with ApaLI + StuI. 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 250RPM for 3.5 hours until OD600nm reached ~0.5. Overnight cultures of recipient *P. aeruginosa* strains were grown in LB medium, sub-cultured 1:3, and grown at 42°C while shaking at 250RPM for 3.5 hours. Cells were collected by centrifugation at 5000g for 10 minutes, washed once, and re-suspended in LB. Cells 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^{10} CFU/ml and 30ul 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 -80C. A small aliquot of each mixture was thawed, diluted and plated to 5ug/ml triclosan + 30ug/ml gentamicin for CFU quantification of successful integrants. Matings were performed at least twice for each recipient strain. 250mL of each medium containing 1.5% agar, 5ug/ml triclosan, and 30ug/ml gentamicin was prepared in a Biodish XL (Nunc). LB agar (US Biologicals), M9 minimal agar, synthetic cystic fibrosis medium (SCFM)[23] 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 5ug from each sample was sheared to 1.5kb 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. S7 and Dataset S6). Size selection was performed using Agencourt Ampure XP beads (Beckman Coulter) and ~500bp libraries were quantified using D5000 ScreenTape System (Agilent). Sequencing was performed with an Illumina Nextseq platform to obtain 50bp genomic DNA reads.

Transposon sequencing analysis. Genomes and annotations for each strain were obtained from www.pseudomonas.com [47]. Illumina reads were mapped to each respective genome using Bowtie [48] using the options for exact and unique read mapping. Reads potentially mapping to more than one location in a genome were discarded and homologous TA sites were removed from analysis by searching the genome using custom scripts. Reads mapped to each TA site were tallied using scripts from [32]. Non-permissive insertion sites containing the sequence (GC)GNTANC(GC) (ref) were removed using custom scripts. Gene clusters across strains were determined using Synerclust (https://www.broadinstitute.org/genome-sequencing-and-analysis/tool-development).

Model for non-essential genes.
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

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

with parameters μ, σ.

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, and are distributed according to

$$\text{For a specific gene g: } x_{g,i} \sim Geo(p_g), \quad for\ i = 1, \ldots, N_{TA}. \tag{2}$$

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:

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

The distribution of n_g among all the genes for some value of NTA is the convolution of the lognormal and the negative binomial:

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

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

Fitting model parameters.
Cramér-von Mises criterion is a goodness of fit criterion, measuring the difference between cumulative density functions of an empirical distribution and a fitted, theoretical one. We use it here for the distributions of the total number of reads in a gene with N_{TA} TA sites (Equation 4). The empirical distribution $F_{N_{T}}$ 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 μ, σ).

For any N_{TA} category, we have

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

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 which 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 ω^2 as follows:

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

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 ω^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 wasn't 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. S6). 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 μ, σ for non-essential genes, we construct the background distribution for each N_{TA} category by sampling 100000 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).

Method validation with clean gene deletions.
Gene deletions were performed as previously described in strain PA14 [49]. Briefly, 800-1200bp regions flanking the target deletion gene of interest were PCR amplified, stitched and recombined into the pEXG2 [49] 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 15ug/ml triclosan + 30ug/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. Successful gene deletion strains were grown in duplicate in LB at 37 °C for 16 hours before diluting 10^{-4} in PBS. 5ul diluted culture was spotted to the five solid media used in this study and grown at 37 °C for 24 hours. Images were captured and densitometry was performed using ImageJ and growth was categorized relative to 10 wild type replicates: essential (0-20%), growth-defective (21-50%), and non-essential (>50%).

Acknowledgments: This work was supported by a generous gift from Anita and Josh Bekenstein, NIH grant R33AI098705 (DTH), and a Cystic Fibrosis Canada Fellowship (BEP).

References

1. Folkesson, A., et al., *Adaptation of Pseudomonas aeruginosa to the cystic fibrosis airway: an evolutionary perspective.* Nat Rev Microbiol, 2012. **10**(12): p. 841-51.
2. Gellatly, S.L. and R.E. Hancock, *Pseudomonas aeruginosa: new insights into pathogenesis and host defenses.* Pathog Dis, 2013. **67**(3): p. 159-73.
3. Klevens, R.M., et al., *Estimating health care-associated infections and deaths in U.S. hospitals, 2002.* Public Health Rep, 2007. **122**(2): p. 160-6.

4. Del Barrio-Tofino, E., et al., *Genomics and Susceptibility Profiles of Extensively Drug-Resistant Pseudomonas aeruginosa Isolates from Spain.* Antimicrob Agents Chemother, 2017. **61**(11).
5. Strateva, T. and D. Yordanov, *Pseudomonas aeruginosa - a phenomenon of bacterial resistance.* J Med Microbiol, 2009. **58**(Pt 9): p. 1133-48.
6. Tacconelli, E., et al., *Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis.* Lancet Infect Dis, 2018. **18**(3): p. 318-327.
7. Hay, M., et al., *Clinical development success rates for investigational drugs.* Nat Biotechnol, 2014. **32**(1): p. 40-51.
8. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995. **269**(5223): p. 496-512.
9. Payne, D.J., et al., *Drugs for bad bugs: confronting the challenges of antibacterial discovery.* Nat Rev Drug Discov, 2007. **6**(1): p. 29-40.
10. Tommasi, R., et al., *ESKAPEing the labyrinth of antibacterial discovery.* Nat Rev Drug Discov, 2015. **14**(8): p. 529-42.
11. Brinster, S., et al., *Type II fatty acid synthesis is not a suitable antibiotic target for Gram-positive pathogens.* Nature, 2009. **458**(7234): p. 83-6.
12. Pethe, K., et al., *A chemical genetic screen in Mycobacterium tuberculosis identifies carbon-source-dependent growth inhibitors devoid of in vivo efficacy.* Nat Commun, 2010. **1**: p. 57.
13. Chao, M.C., et al., *The design and analysis of transposon insertion sequencing experiments.* Nat Rev Microbiol, 2016. **14**(2): p. 119-28.
14. Gawronski, J.D., et al., *Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung.* Proc Natl Acad Sci U S A, 2009. **106**(38): p. 16422-7.
15. Goodman, A.L., et al., *Identifying genetic determinants needed to establish a human gut symbiont in its habitat.* Cell Host Microbe, 2009. **6**(3): p. 279-89.
16. Langridge, G.C., et al., *Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants.* Genome Res, 2009. **19**(12): p. 2308-16.
17. van Opijnen, T., K.L. Bodi, and A. Camilli, *Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms.* Nat Methods, 2009. **6**(10): p. 767-72.
18. Juhas, M., *Pseudomonas aeruginosa essentials: an update on investigation of essential genes.* Microbiology, 2015. **161**(11): p. 2053-60.
19. Lampe, D.J., T.E. Grant, and H.M. Robertson, *Factors affecting transposition of the Himar1 mariner transposon in vitro.* Genetics, 1998. **149**(1): p. 179-87.
20. Liberati, N.T., et al., *An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants.* Proc Natl Acad Sci U S A, 2006. **103**(8): p. 2833-8.
21. Rubin, E.J., et al., *In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria.* Proc Natl Acad Sci U S A, 1999. **96**(4): p. 1645-50.
22. Lampe, D.J., et al., *Hyperactive transposase mutants of the Himar1 mariner transposon.* Proc Natl Acad Sci U S A, 1999. **96**(20): p. 11428-33.
23. Palmer, K.L., L.M. Aye, and M. Whiteley, *Nutritional cues control Pseudomonas aeruginosa multicellular behavior in cystic fibrosis sputum.* J Bacteriol, 2007. **189**(22): p. 8079-87.
24. DeJesus, M.A., et al., *Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis.* MBio, 2017. **8**(1).
25. Griffin, J.E., et al., *High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism.* PLoS Pathog, 2011. **7**(9): p. e1002251.
26. Jacobs, M.A., et al., *Comprehensive transposon mutant library of Pseudomonas aeruginosa.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 14339-44.
27. Lee, S.A., et al., *General and condition-specific essential functions of Pseudomonas aeruginosa.* Proc Natl Acad Sci U S A, 2015. **112**(16): p. 5189-94.
28. Skurnik, D., et al., *A comprehensive analysis of in vitro and in vivo genetic fitness of Pseudomonas aeruginosa using high-throughput sequencing of transposon libraries.* PLoS Pathog, 2013. **9**(9): p. e1003582.
29. Turner, K.H., et al., *Essential genome of Pseudomonas aeruginosa in cystic fibrosis sputum.* Proc Natl Acad Sci U S A, 2015. **112**(13): p. 4110-5.
30. Zomer, A., et al., *ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data.* PLoS One, 2012. **7**(8): p. e43012.
31. DeJesus, M.A., et al., *TRANSIT--A Software Tool for Himar1 TnSeq Analysis.* PLoS Comput Biol, 2015. **11**(10): p. e1004401.
32. Pritchard, J.R., et al., *ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing.* PLoS Genet, 2014. **10**(11): p. e1004782.
33. Stein, W.H. and S. Moore, *The free amino acids of human blood plasma.* J Biol Chem, 1954. **211**(2): p. 915-26.
34. *Porphyrins - urine test*. updated 2018 Aug 14, Bethesda (MD): National Library of Medicine (US): MedlinePlus [Internet]. p. 2
35. Bisicchia, P., et al., *MinC, MinD, and MinE drive counter-oscillation of early-cell-division proteins prior to Escherichia coli septum formation.* MBio, 2013. **4**(6): p. e00856-13.
36. Ason, B. and W.S. Reznikoff, *DNA sequence bias during Tn5 transposition.* J Mol Biol, 2004. **335**(5): p. 1213-25.

37. Dias, M.V., et al., *Chorismate synthase: an attractive target for drug development against orphan diseases.* Curr Drug Targets, 2007. **8**(3): p. 437-44.
38. Gupta, R., T.R. Gobble, and M. Schuster, *GidA posttranscriptionally regulates rhl quorum sensing in Pseudomonas aeruginosa.* J Bacteriol, 2009. **191**(18): p. 5785-92.
39. Hendrickson, E.L., et al., *Differential roles of the Pseudomonas aeruginosa PA14 rpoN gene in pathogenicity in plants, nematodes, insects, and mice.* J Bacteriol, 2001. **183**(24): p. 7126-34.
40. Sonnleitner, E., et al., *Reduced virulence of a hfq mutant of Pseudomonas aeruginosa O1.* Microb Pathog, 2003. **35**(5): p. 217-28.
41. Gentry, D.R., et al., *Variable sensitivity to bacterial methionyl-tRNA synthetase inhibitors reveals subpopulations of Streptococcus pneumoniae with two distinct methionyl-tRNA synthetase genes.* Antimicrob Agents Chemother, 2003. **47**(6): p. 1784-9.
42. Letunic, I. and P. Bork, *Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.* Nucleic Acids Res, 2016. **44**(W1): p. W242-5.
43. Lee, D.G., et al., *Genomic analysis reveals that Pseudomonas aeruginosa virulence is combinatorial.* Genome Biol, 2006. **7**(10): p. R90.
44. Pecora, N.D., et al., *Genomically Informed Surveillance for Carbapenem-Resistant Enterobacteriaceae in a Health Care System.* MBio, 2015. **6**(4): p. e01030.
45. Haas, W., et al., *Integrated analysis of three bacterial conjunctivitis trials of besifloxacin ophthalmic suspension, 0.6%: etiology of bacterial conjunctivitis and antibacterial susceptibility profile.* Clin Ophthalmol, 2011. **5**: p. 1369-79.
46. Ordonez, C.L., et al., *Inflammatory and microbiologic markers in induced sputum after intravenous antibiotics in cystic fibrosis.* Am J Respir Crit Care Med, 2003. **168**(12): p. 1471-5.
47. Winsor, G.L., et al., *Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database.* Nucleic Acids Res, 2016. **44**(D1): p. D646-53.
48. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
49. Hmelo, L.R., et al., *Precision-engineering the Pseudomonas aeruginosa genome with two-step allelic exchange.* Nat Protoc, 2015. **10**(11): p. 1820-41.