

High-throughput transposon mutagenesis in the family Enterobacteriaceae reveals core essential genes and rapid turnover of essentiality

Fatemeh A. Ghomi,¹ Jakob J. Jung,² Gemma C. Langridge,³ Amy K. Cain,⁴ Christine J. Boinett,⁵ Moataz Abd El Ghany,^{6,7,8,9} Derek J. Pickard,^{5,10} Robert A. Kingsley,^{3,11} Nicholas R. Thomson,^{5,12} Julian Parkhill,¹³ Paul P. Gardner,^{1,14} Lars Barquist^{2,15,16}

AUTHOR AFFILIATIONS See affiliation list on p. 19.

ABSTRACT The Enterobacteriaceae are a scientifically and medically important clade of bacteria, containing the model organism *Escherichia coli*, as well as major human pathogens including *Salmonella enterica* and *Klebsiella pneumoniae*. Essential gene sets have been determined for several members of the Enterobacteriaceae, with the Keio *E. coli* single-gene deletion library often regarded as a gold standard. However, it remains unclear how gene essentiality varies between related strains and species. To investigate this, we have assembled a collection of 13 sequenced high-density transposon mutant libraries from five genera within the Enterobacteriaceae. We first assess several gene essentiality prediction approaches, investigate the effects of transposon density on essentiality prediction, and identify biases in transposon insertion sequencing data. Based on these investigations, we develop a new classifier for gene essentiality. Using this new classifier, we define a core essential genome in the Enterobacteriaceae of 201 universally essential genes. Despite the presence of a large cohort of variably essential genes, we find an absence of evidence for genus-specific essential genes. A clear example of this sporadic essentiality is given by the set of genes regulating the σE extracytoplasmic stress response, which appears to have independently acquired essentiality multiple times in the Enterobacteriaceae. Finally, we compare our essential gene sets to the natural experiment of gene loss in obligate insect endosymbionts that have emerged from within the Enterobacteriaceae. This isolates a remarkably small set of genes absolutely required for survival and identifies several instances of essential stress responses masked by redundancy in free-living bacteria.

IMPORTANCE The essential genome, that is the set of genes absolutely required to sustain life, is a core concept in genetics. Essential genes in bacteria serve as drug targets, put constraints on the engineering of biological chassis for technological or industrial purposes, and are key to constructing synthetic life. Despite decades of study, relatively little is known about how gene essentiality varies across related bacteria. In this study, we have collected gene essentiality data for 13 bacteria related to the model organism *Escherichia coli*, including several human pathogens, and investigated the conservation of essentiality. We find that approximately a third of the genes essential in any particular strain are non-essential in another related strain. Surprisingly, we do not find evidence for essential genes unique to specific genera; rather it appears a substantial fraction of the essential genome rapidly gains or loses essentiality during evolution. This suggests that essentiality is not an immutable characteristic but depends crucially on the genomic context. We illustrate this through a comparison of our essential genes in free-living bacteria to genes conserved in 34 insect endosymbionts with naturally reduced genomes, finding several cases where genes generally regarded as being

Invited Editor David S. Guttman, University of Toronto, Toronto, Ontario, Canada

Editor Samuel I. Miller, University of Washington, Seattle, Washington, USA

Address correspondence to Lars Barquist, lars.barquist@helmholtz-hiri.de.

Fatemeh A. Ghomi and Jakob J. Jung contributed equally to this article. Author order was determined by seniority.

The authors declare no conflict of interest.

See the funding table on p. 19.

Received 2 July 2024

Accepted 5 August 2024

Published 29 August 2024

Copyright © 2024 A. Ghomi et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

important for specific stress responses appear to have become essential in endosymbionts due to a loss of functional redundancy in the genome.

KEYWORDS gene essentiality, transposon mutagenesis, Enterobacteriaceae, *Escherichia coli*, *Salmonella*, *Klebsiella*, *Citrobacter*

Gene essentiality is a core concept with importance in sub-fields spanning the breadth of microbiology and genetics (1). The motivation for the earliest attempts at determining the complete set of essential genes in bacteria, through comparisons of the first two bacterial genomes sequenced, was to determine a minimal gene set sufficient to support life (2, 3). While a somewhat esoteric object of study, the minimal gene set has important practical implications across a range of applications (4). For instance, antibiotics operate by interfering with essential cellular functions, though target-based screens for new antibiotic leads based on essential gene lists have so far met with limited success for a variety of technical reasons (5, 6). Essential gene sets are also relevant for biotechnology applications, where the essential genome puts constraints on our ability to engineer bacterial chassis (7). Finally, the construction of the first fully synthetic genomes has been deeply reliant on investigations of the essential genome (8, 9) and has been important in accelerating technological progress in synthetic biology (10).

High-throughput sequencing has accelerated the study of gene essentiality. Relatively inexpensive genome sequencing has enabled the exploration of diverse reduced genomes from obligate pathogens or endosymbionts, finding bacteria that can survive with less than 200 genes, albeit under controlled conditions within host cells (11). These organisms provide natural experiments ablating the bacterial genome to its core essential components, providing information on the basic essential functions necessary to sustain life. Sequencing has also led to new experimental approaches for exploring gene essentiality, notably transposon insertion sequencing (TIS) (12) and more recently pooled CRISPR interference (CRISPRi) screens using catalytically inactive Cas proteins to silence gene transcription (13, 14). In contrast to traditional approaches that relied on labor-intensive mapping of single transposon insertion mutants (15) or targeted gene deletion (16), these approaches enable the simultaneous assay of hundreds to millions of pooled mutants in a single experiment. This has enabled the first comparisons of experimentally determined gene essentiality between strains within bacterial species (17–21), finding evidence for strain-level variation in the essential gene complement.

The Enterobacteriaceae are perhaps the single most well-studied family of bacteria. They include the human gut commensal *Escherichia coli*, which has served as a model organism for 135 years (22). The *E. coli* species also includes a wide range of pathovars adapted to various intestinal and extraintestinal niches, including the polyphyletic *Shigella* lineages. Several other major human pathogens have also arisen within this family, including causative agents of typhoidal and nontyphoidal salmonellosis (*Salmonella*), dysentery (*Shigella*), yersiniosis and plague (*Yersinia*), and pneumonia and nosocomial infections (*Klebsiella*). This has led to intense interest in members of the family for medical as well as purely scientific reasons. Indeed, one of the first arrayed single gene deletion libraries (the Keio collection) was constructed in *E. coli* (16) and has served as a major reference point for the community studying essential genes since.

Despite the intense interest, there have been few comparisons of gene essentiality between members of the Enterobacteriaceae. Using TIS, we previously showed evidence for a conserved core of 281 genes essential in both the salmonellosis-associated *Salmonella enterica* serovar Typhimurium and the typhoidal *Salmonella enterica* serovar Typhi, with a smaller set of 228 also essential in *E. coli* (18). A similar study comparing *Shigella flexneri* with *E. coli* found a largely overlapping essential gene set, with most differences in gene essentiality being due to a loss of redundant genes or pathways in *S. flexneri* (23), which is expected as the shigellae are simply specialized pathotypes of *E.*

coli (24). How these findings extend to differences in gene essentiality across the family Enterobacteriaceae is currently unclear.

To investigate gene essentiality in the family Enterobacteriaceae, we gathered a collection of previously reported and newly generated Tn5 TIS data sets from the genera *Escherichia*, *Salmonella*, *Citrobacter*, and *Klebsiella* generated using the Transposon Directed Insertion-site Sequencing (TraDIS) technique (Fig. 1A) (25, 26). We benchmark methods for identifying essential genes and developing a new classifier based on empirical clustering using DBSCAN (27). We examine how biases in TraDIS data affect gene essentiality predictions and correct for these. We compare essential gene sets between the strains in our collection, identifying a core conserved essential gene set of 201 genes and reconstructing an ancestral gene set of 296. Despite substantial variation in the essential gene content across strains, we find little evidence for private essential

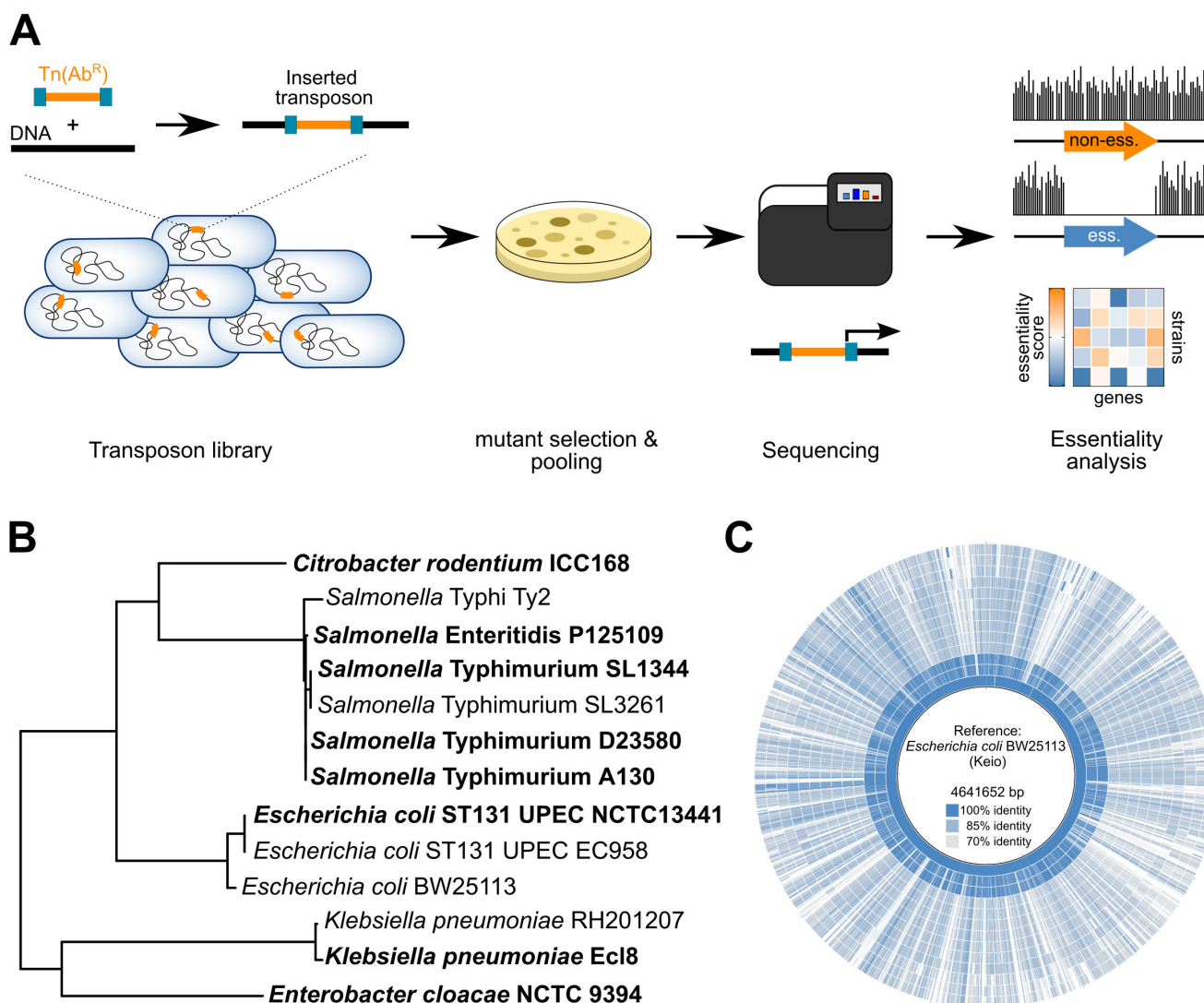


FIG 1 Overview of a collection of TraDIS libraries for five genera within the Enterobacteriaceae. (A) Overview of TraDIS experiments performed to identify essential genes. A random Tn5 library is generated and selected on solid antibiotic medium, before being sequenced using transposon-specific primers. Essential genes are identified by depletion of identified insertion sites (see text and methods for details). (B) An estimated tree showing phylogenetic relationships between the strains used in this study. The tree was constructed using RAxML (28) on a concatenated set of single-copy core genes (29) (see Methods). The TraDIS libraries first reported here are in bold font. (C) Genome alignment for all the genomes in this study compared to *E. coli* BW25113 from Keio collection (16), generated with BRIG (30). The genomes from the inner circle to the outer circle are *E. coli* BW25113, *E. coli* UPEC ST131 EC958, *E. coli* UPEC ST131, *S. Typhimurium* A130, *S. Typhimurium* D23580, *S. Typhimurium* SL3261, *S. Typhimurium* SL1344, *S. Enteritidis* P125109, *S. Typhi* Ty2, *C. rodentium* ICC168, *E. cloacae* NCTC 9394, *K. pneumoniae* Ecl8, and *K. pneumoniae* RH201207, respectively.

genes within genera. Finally, we compare our essential genes in free-living Enterobacteriaceae to related endosymbionts characterized by extreme genome reduction. We identify both a conserved core of genes absolutely required for survival, as well as a number of genes typically considered to play roles in stress responses that appear to become essential in the context of a minimal genome.

RESULTS

A TraDIS gene essentiality compendium encompassing five genera of the Enterobacteriaceae

To investigate gene essentiality in the Enterobacteriaceae, we assembled a collection of new and previously published data that cover five medically and scientifically important genera of the Enterobacteriaceae: *Escherichia*, *Salmonella*, *Citrobacter*, *Klebsiella*, and *Enterobacter* (Fig. 1B and C). This includes eight TraDIS libraries that have not been previously reported (see Methods), and all have been constructed in rich media with growth at 37°C. The selected *Escherichia* strains include the parent strain of the Keio collection (16), *Escherichia coli* BW25113 (31), as well as two sequence type 131 (ST131) strains associated with urinary tract infections, NCTC 13441 (32) and EC958 (33). Our *Salmonella* data include data from the CVD908-*htrA* vaccine derivative of *Salmonella enterica* serovar Typhi Ty2 (25), the common *S. Typhimurium* lab strain SL1344 and its vaccine derivative SL3261 (18), the *S. Enteritidis* PT4 strain P125109 (34) representative of the global gastroenteritis epidemic, and *S. Typhimurium* D23580 (35) and A130, representatives of the two major clades involved in the HIV-associated invasive non-typhoidal salmonellosis epidemic in sub-Saharan Africa (36). We also have included *Citrobacter rodentium* ICC168 (37), a model mouse pathogen that carries the type III secretion system encoding locus of enterocyte effacement (LEE) pathogenicity island in common with the enteropathogenic (EPEC) and enterohemorrhagic (EHEC) pathovars of *E. coli* (38). We have also included two strains of *Klebsiella pneumoniae*, a major cause of multi-drug-resistant nosocomial infections. These are Ecl8, a genetically manipulable model strain (39), and RH201207, a multi-drug-resistant clinical isolate (40). Finally, we included a representative of the opportunistic pathogen *Enterobacter cloacae*, strain NCTC 9394.

Assessing essentiality classification

Several methods have been used for evaluating the essentiality of genes using transposon insertion data. These use different features that report on gene essentiality, which can include features of the sequenced libraries themselves, such as the number of insertion sites divided by gene length [the so-called insertion index (25)], or comparisons of experimentally determined transposon read counts against synthetic reference libraries (41). Freed and colleagues assessed 11 of these features, finding the insertion index, the mean distance between insertions, and the largest uninterrupted fraction of gene length were most predictive of essentiality (23).

Here we evaluated four of these measures: the average distance between insertion sites in a gene, the largest uninterrupted fraction, the insertion index, and the Monte Carlo method combined with DESeq proposed by Turner and colleagues (41) (Fig. 2A). In addition, we constructed a predictor using all four measures of gene essentiality using principal component analysis (PCA), taking the weighted sum given by the first principal component as our combined measure. To evaluate the accuracy, we compared essential genes predicted from an *E. coli* K-12 BW25113 transposon library (31) with each method to curate essential genes in the EcoGene database (42). EcoGene is a refinement of essential genes determined by directed single-gene knockouts in the Keio collection (16). The four measures performed similarly, as evaluated by the area under the receiver operating characteristic curve (AUC). Our PCA-based predictors lead to a modest increase in performance, indicating that these measures are not completely redundant. However, on average, this improvement only led to an average increase of approximately 1.2

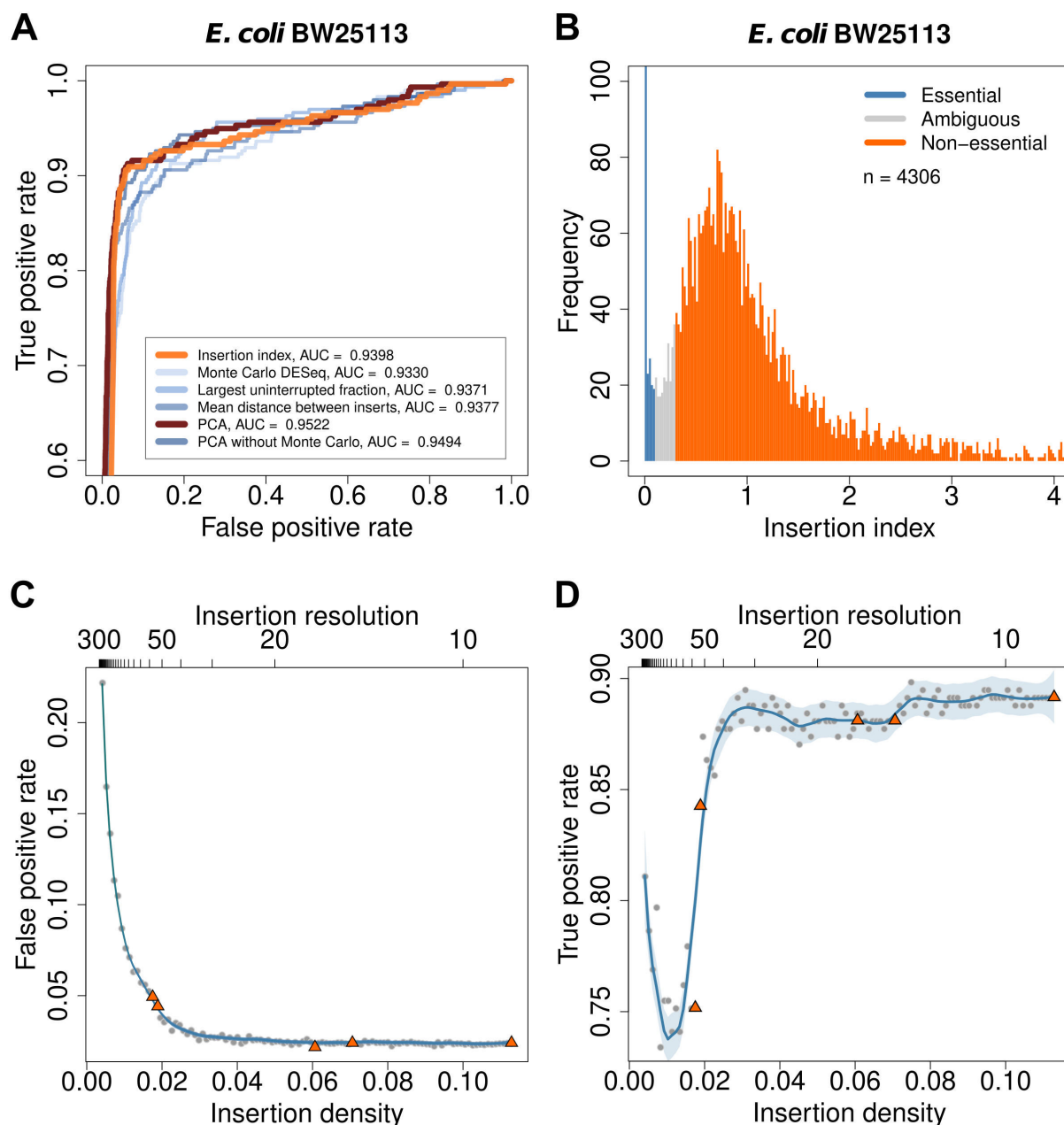


FIG 2 Assessing gene essentiality predictions. (A) Receiver operating characteristic (ROC) curves show the accuracy of six methods for predicting essential genes. True positives are genes that are predicted as essential for *E. coli* K-12 BW25113 by TraDIS and classified as essential in EcoGene. False positives are genes that are predicted as essential but are classified as non-essential in EcoGene. Genes for which the Monte Carlo method returned NA were omitted from this comparison. (B) The bimodal distribution of insertion indices illustrates essentiality classification using DBSCAN. Essential genes have the lowest insertion indices, non-essential genes have higher insertion indices, and ambiguous genes are located between these groups. (C and D) Simulation of insertion density effects: The orange triangles and gray dots are obtained from real and down-sampled data from *S. Typhimurium* SL1344, respectively. The blue lines show loess curves with 0.2 span, and the light blue regions show 95% confidence intervals. Insertion resolution is calculated by dividing genome length by the number of unique insertion sites. The false-positive rate decreases with increasing insertion density (number of insertions divided by genome length) and remains constant after it reaches 0.04, or approximately one insertion every 25 bases (C). The true-positive rate converges around an insertion density of ~0.03, or approximately one insertion every 30 bases (D). The false-positive and true-positive rates are calculated by comparing predicted essential genes with the EcoGene database.

correctly determined essential genes across our strain collection at a false-positive rate of 0.05 (Table S1). Given this modest improvement and the relative ease of calculation, and

particularly the straightforward interpretability of the insertion index compared to PCA, we have used the insertion index for further work.

A second issue in building an essential gene classifier is choosing a threshold for essentiality. In previous work, gamma distributions have been fit to the bimodal distribution observed in the insertion index (Fig. 2B), and a log-likelihood ratio was used to determine the essentiality cut-off (25, 26). This method is heuristic, based on the separation of the two-component distributions and maximum likelihood fits. We have observed that this, on occasion, can lead to poor fits and instability in the classification, which can require manual intervention to correct. To make this procedure more robust, we have adopted a non-parametric clustering approach using DBSCAN (27). This leads to consistent improvements in classification accuracy compared to automated gamma fitting (26) as measured by the Matthews correlation coefficient across all our data sets using EcoGene essential genes as a gold standard (Fig. S1).

The final component in determining gene essentiality is the transposon library itself; insufficient density will lead to genes being devoid of insertion by chance. Often, density is reported in terms of the number of insertions divided by genome length. However, this can be misleading, as insertion density is clearly non-uniform. To address this, we have constructed a series of five increasingly dense *Tn5* libraries in *S. Typhimurium* SL1344, with an estimated 50K, 100K, 250K, 500K, and 1,000K mutants based on transformation efficiency control plates. After sequencing and read-mapping, we recovered 85,477, 92,145, 295,854, 344,380, and 550,657 unique insertion sites, respectively. We then computationally subsampled these insertion libraries and predicted gene essentiality using the insertion index and DBSCAN, to plot both false-positive and true-positive rates for a full range of library complexities (Fig. 2C and D). These curves show that classifier performance stabilizes at around one transposon mutant per 35 nucleotides of genomic sequence, providing a clear minimum density for accurate prediction of gene essentiality with *Tn5* mutagenesis. All of the data sets included in our study meet this threshold.

Evaluating and correcting for biases in *Tn5* insertion

A number of studies have reported genomic biases in *Tn5* insertion (17, 18, 25, 43–47). However, the findings of different studies are contradictory, and it remains unclear whether these biases impact the quality of gene essentiality predictions. Using our combined data set, we examined the evidence and impact of four potential biases: positional bias within genes and genomes, and compositional bias in the form of preferred sequence motifs and G + C content.

Within genes, it has been noted that the effect of transposon insertions at the start and end of the sequence can diverge from that observed across the entire gene (15). To investigate this in our own data, we calculated local averaged insertion indices across percentiles of essential genes (Fig. 3A). There was a clear enrichment in transposon insertions in the first 5% and last 20% of the essential gene sequences. Insertions early in essential genes may be due to either biological alternative start codons, or possibly more likely, the bias in gene prediction algorithms toward longer coding sequences resulting in incorrect start-codon annotations (48). The enrichment of insertions at the 3' end of essential genes may indicate that functional proteins can frequently be produced from truncated transcripts, as can be seen in the specific case of RNase E in *K. pneumoniae* RH201207 (Fig. 3B).

Within genomes, positional effects in transposon density have been observed with differences in density between the origin and terminus of replication (44). These differences are presumably due to the dosage effect of multiple copies of DNA near the origin of replication during exponential growth (18, 50). We observed this phenomenon to varying degrees in our data (Fig. 3C; Fig. S2); this variation is likely due in part to individual optimization of the transformation protocol during library creation, leading to different bacterial libraries being created in slightly different growth phases. In some genomes (e.g., *E. cloacae* NCTC9394), positional bias in transposon density did not clearly

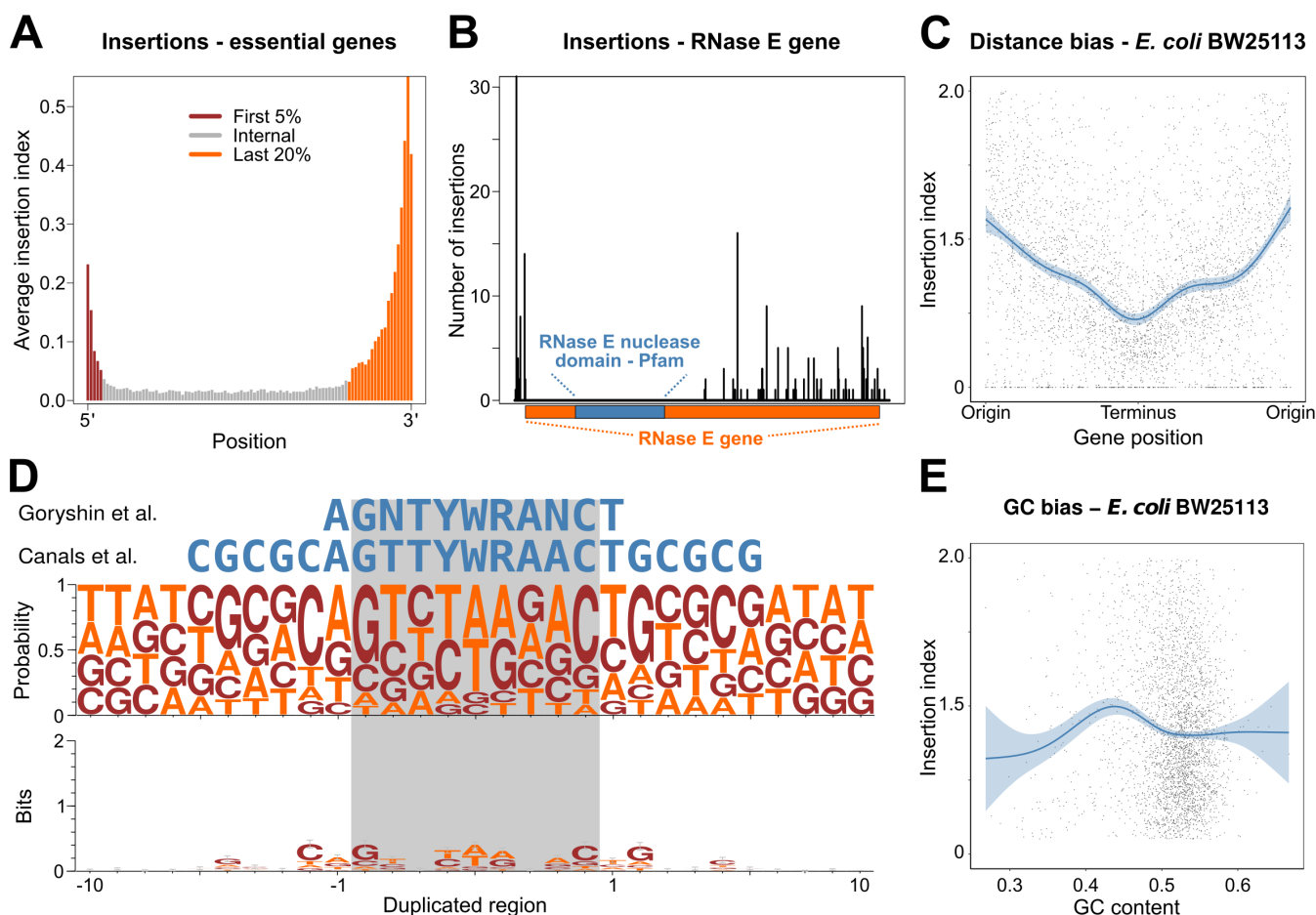


FIG 3 An analysis of putative sources of bias in TraDIS data. (A) The average insertion index for each length percentile of all essential genes from TraDIS data. The genes are divided into three segments: 5% of the gene length on the 5' end (dark red), 20% of the gene length on the 3' end (orange), and the rest in the middle (gray). (B) Number of insertions and their location in the RNase E gene in *Klebsiella pneumoniae* RH201207. The 5' end of the gene is located on the left-hand side. There are no insertions in the nuclease domain predicted by Pfam (49). (C) The insertion index versus the distance from the *dnaA* gene (usually found near the origin of replication). The blue curve shows a fitted GAM (Generalized Additive Model) curve, and the shading shows the 95% confidence interval. (D) Sequence logo plots generated using sequences from the 10 nucleotides flanking the 100 topmost frequent insertion sites from each genome. Character height represents frequency (top) or information content per position (bottom), calculated by multiplying the frequency of each base by the total information content of the position in bits. (E) Insertion index versus G-C content of genes. The blue curve shows the fitted GAM curve, and the shading shows the 95% confidence interval.

correlate with distance to the origin (Fig. S2). This may be indicative of errors in genome scaffolding and contiguation, as the genome sequence remains unclosed.

We next examined the existence of preferred sequence motifs for Tn5. When the Tn5 transposon inserts into DNA, a region of 9 nucleotides is duplicated on each side of the transposon (43). An early study of Tn5 insertion based on circa 100 example insertions suggested a preferred A-GNTYWRANC-T motif (43), and this was later supported and extended by transposon sequencing in *Salmonella* (17). Other higher-throughput studies have failed to find a clear motif but have suggested a preference for elevated G + C content regions (45, 47). To address the presence of a preferred motif, we generated a sequence logo from the 10 nucleotides flanking each insertion site for the 100 most frequent insertion sites in each genome (Fig. 3D). The nucleotide frequency observed is broadly consistent with the sequence motif proposed by Goryshin and Canals; however, the information content across the duplicated region was consistently low, suggesting this motif is not a major constraint on transposon insertion. Similarly, examining the relationship between G + C content and insertion density failed to give consistent results across genomes. Lower G + C genes tended to have slightly elevated insertion densities

(Fig. 3E), but this may be an artifact of the small number of data points below the ~50% to 60% G + C range and the association of low G + C content with horizontally acquired DNA in enterobacterial species (27).

Correcting for the three biases most easily adjusted for (positional bias within genes, positional bias within the genome, and G + C content bias) each led to modest improvements in classifier performance. Combining the three corrections improved performance as measured by the AUC by ~1%. However, this did lead to a substantial increase in the true-positive rate in the critical low false positive range of the ROC curve (Fig. S3).

Core and ancestral gene essentiality in the Enterobacteriaceae

Having investigated the factors affecting the accuracy of gene essentiality calls, we turned to the investigation of essential genes in our TraDIS collection. We called essential genes in all 13 strains using the insertion index after correcting for positional bias within genes, positional bias within genomes, and G + C content bias using our DBSCAN-based classification. We then defined an “essentiality score” as the $\log_2$ transformation of the ratio of the insertion index to the DBSCAN cut-off. Intuitively, this results in a score where values less than or equal to 0 indicate essentiality and the values indicate twofold changes from this threshold. Genes with an insertion index of 0 were arbitrarily recoded to -6.5 , as there were no observed scores lower than this number, and $\log_2(0)$ is undefined.

This resulted in between ~300 and ~440 genes classified as essential in each strain (Fig. 4A). This is comparable to previous results in the literature applying transposon insertion sequencing to the study of gene essentiality (51), with larger essential gene sets likely reflecting the difficulty in segregating truly essential genes from those with strong growth effects in high-throughput studies. To compare essential genes between strains, we defined groups of orthologous proteins using Hieranoid (52) and a phylogenetic tree constructed using RaXML (28) from concatenated core gene alignments produced by PhyloSift (29) (see Methods). On manual examination, we found that a number of essential genes were absent in the genome sequence of *Enterobacter cloacae* NCTC 9394 due to an incomplete assembly, and this strain was excluded from further analyses.

As a gold standard comparator, we included a curated collection of *E. coli* K-12 essential genes from the EcoGene database (42). Unfortunately, this resource is no longer maintained, but an archival copy is available in the Internet Archive [<https://web.archive.org/web/20170228010715/http://www.ecogene.org/?q=topic/5>], and we additionally provide the gene list in Table S2. This list curates the Keio collection of directed gene deletions (16) with observations from later studies that have investigated technical artifacts in the Keio collection and contains 299 essential genes. Major differences from the Keio collection are the inclusion of 15 genes not considered essential by Keio which were subsequently shown to be artifacts of gene duplication, namely *glyS*, *ileS*, *alaS*, *coaA*, *coaE*, *dnaG*, *glmM*, *groL*, *parC*, *prfB*, *polA*, *rho*, *rpoD*, *rpsU*, and *lptB*. Genes considered essential in the original Keio paper but later shown to be non-essential include the ribosomal proteins *rpsL*, *rpsM*, and *rpsQ*, mutants of which grow extremely slowly (53), and *rnc* encoding RNase III, whose deletion is thought to have a polar effect on the downstream essential gene *era* with which it is co-transcribed (54).

Our initial expectation was that differences in gene essentiality would largely be concordant with the phylogeny, that is, we expected to find groups of genus-specific essential genes. However, this was not the case (Fig. 4A). To investigate further, we took multiple approaches to exploring the contents of our essential gene collection. As a first approach, we defined the core essential genes in our genomes as the intersection of essential genes across all TraDIS libraries and the EcoGene list. This resulted in a set of 201 genes essential across all strains in our collection. As this gene set excluded a number of essential functions, we also applied Fitch’s algorithm (55) as a less conservative approach to reconstruct the set of genes essential in the ancestor of all strains included in our collection resulting in 296 ancestrally essential

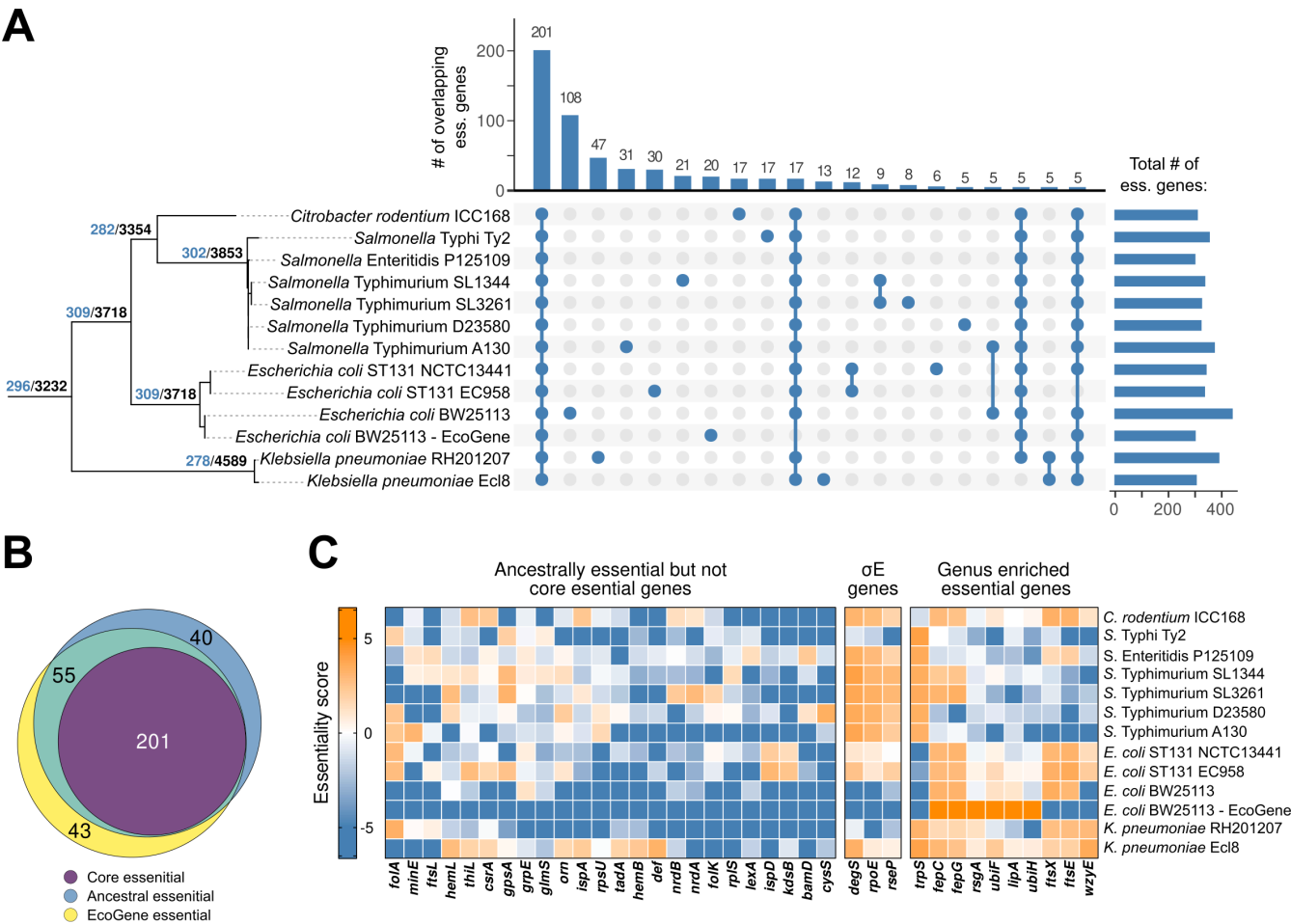


FIG 4 The ancestral and core essential genomes of Enterobacteriaceae. (A) Left, reconstruction of the ancestral essential genome. Numbers on branches indicate the number of reconstructed essential genes (blue) compared to the total number of genes reconstructed to be present (black) on each ancestral branch. Right, upset plot illustrating the overlap of essential genes between strains. Overlaps with <5 essential genes are not shown. (B) Venn diagram comparing the ancestral and core essential gene sets to genes designated essential in EcoGene. (C) Heatmaps illustrating variable gene essentiality across the Enterobacteriaceae. Essentiality scores are indicated by color; blue indicates essentiality while orange indicates non-essentiality. Left, 24 genes essential in EcoGene and ancestrally essential, but not core essential, with clear evidence for non-essentiality (essentiality score >1) in at least one strain in the collection. Center, genes involved in the σ^E response. Right, genes with genus-associated essentiality as determined by a permutation test ($P < 0.05$).

genes (Fig. 4A and B). Fitch's algorithm is based on parsimony and attempts to reconstruct the ancestral essentiality state by minimizing the number of transitions between essentiality and non-essentiality given the observed data. To check that our choice of reconstruction method did not majorly bias our results, we also used PastML, a likelihood-based method, to infer ancestral essentiality (56). This yielded similar but slightly more conservative results with fewer genes classified as ancestrally essential. All genes except for one (272 out of 273) that were classified as ancestrally essential using PastML were also considered ancestrally essential using Fitch's reconstruction. Conversely, 24 genes that were not classified as ancestral using PastML were considered ancestral using Fitch's method (272 out of 296, Fig. S4 and S5). Core and ancestrally essential genes are summarized in Table 1.

We found it somewhat surprising that despite the range in the number of genes predicted to be essential across our collection, the ancestral essential gene set was of similar size to that in the EcoGene gold standard. We examined the intersection of the core, ancestral, and EcoGene gene sets in an attempt to understand the differences between them (Fig. 4B). By design, the core essential genes are a proper subset of the

TABLE 1 Ancestral essential gene functions in the Enterobacteriaceae^a

Biological process	Subprocess	Genes
Cell division		<i>ftsABH*ILQW*YZ</i> , <i>minE</i> , <i>mukB</i> , <i>zipA</i>
DNA replication	Polymerases I, II, and III	<i>dnaENQX</i> , <i>holABD</i>
	Supercoiling	<i>gyrAB</i> , <i>parCE</i>
	Primosome associated	<i>dnaBCGT</i> , <i>priA</i> , <i>ssb*</i>
	Other	<i>dnaA</i> , <i>ligA</i> , <i>topA*</i> , <i>mukEF</i> , <i>nrdAB</i> , <i>rpoABC</i>
Transcription	RNA polymerase	<i>nusA*BG</i> , <i>rpoDH</i> , <i>rho</i>
	Sigma, elongation, anti- and termination factors	
Translation	tRNA synthetases	<i>alaS</i> , <i>argS</i> , <i>asnS</i> , <i>aspS</i> , <i>cysS</i> , <i>glnS</i> , <i>gltX</i> , <i>glyQS</i> , <i>hisS</i> , <i>ileS</i> , <i>leuS</i> , <i>lysS</i> , <i>metG</i> , <i>pheST</i> , <i>proS</i> , <i>serS</i> , <i>thrS</i> , <i>tilS</i> , <i>tyrS</i> , <i>valS</i>
	Ribosome components	<i>rplABC*DEFJKLMNQP*QRST*UVWXY</i> , <i>rpmABC**D*HI</i> , <i>rpsABCDEFGHIJKLMNQP*QRSTU</i>
	Initiation, elongation, and peptide chain release factors	<i>fusA</i> , <i>infA*BC</i> , <i>prfAB</i> , <i>tsf</i> , <i>yrdC</i>
	Other	<i>cca*</i> , <i>def</i> , <i>der</i> , <i>era</i> , <i>frr</i> , <i>fmr*</i> , <i>rimM</i> , <i>rnpA</i> , <i>mnmA</i> , <i>pth</i> , <i>rnc</i> , <i>tadA</i> , <i>trmD</i> , <i>tsaBC*DE</i> , <i>ybeY</i> , <i>ygfZ</i> , <i>yqgF</i>
Biosynthetic pathways		
Peptidoglycan		<i>mraY</i> , <i>mrdA*B*</i> , <i>murABCDEFGHIJ</i>
Fatty acids		<i>accAB**CD</i> , <i>acpP*S</i> , <i>fabA*BDGHIZ</i>
Phospholipids		<i>cdsA</i> , <i>pgsA</i> , <i>plsBC</i> , <i>psd</i> , <i>pssA</i>
Amino acids		<i>asd</i> , <i>dapABDE</i> , <i>metK</i>
Nucleotide		<i>adk</i> , <i>dut</i> , <i>nadDEK</i> , <i>prs</i> , <i>pyrGH</i> , <i>thyA</i> , <i>tmk</i>
Coenzyme A		<i>coaADE</i> , <i>dfp</i>
Isoprenoids, heme, ubiquinone		<i>dxr</i> , <i>dxs</i> , <i>hemABCDEFGHIJL</i> , <i>ispABDEFGHU</i> , <i>ubiABDEGHX</i>
B Vitamin		<i>folABCD*EK</i> , <i>ribABCDEFGH</i> , <i>thiL</i>
Miscellaneous processes		
Regulation		<i>birA*</i> , <i>cohE</i> , <i>csrA</i> , <i>lexA</i>
Protein localization and transport		<i>bamAD</i> , <i>ffh</i> , <i>lepB</i> , <i>secADEFY</i> , <i>yidC</i>
Outer membrane		<i>kdsAB</i> , <i>lpxABCDHK</i> , <i>lptABDEFG</i> , <i>msbA</i>
Lipoprotein synthesis and transport		<i>lgt</i> , <i>Int</i> , <i>lolAB*CDE</i> , <i>lspA</i> , <i>waaA</i>
Glycolysis		<i>eno*</i> , <i>fbaA*</i> , <i>gapA</i> , <i>lpd</i> , <i>pgk</i>
Other ancestrally essential genes		<i>dnaK</i> , <i>erpA</i> , <i>fldA</i> , <i>glmMSU</i> , <i>glyA</i> , <i>gmk</i> , <i>gpsA</i> , <i>groLS</i> , <i>grpE</i> , <i>iscS</i> , <i>mreB*C*D*</i> , <i>obgE</i> , <i>orn</i> , <i>ppa</i> , <i>rodZ</i> , <i>suhB*</i> , <i>yihA</i>

^aAncestral essential genes are determined by Fitch's algorithm, and core essential genes are highlighted in bold. Genes that miss the essentiality threshold (<2-fold) in up to two strains are highlighted with an asterisk. Genes that are nonessential in only one strain, likely due to misannotation, are marked with a double asterisk.

ancestral and EcoGene sets, and contain ~2/3rds of the genes essential in either of these sets. Fifty-five genes occur in both the ancestral and EcoGene sets but are not core. Manual inspection led us to conclude that two of these were excluded from the core gene set due to annotation errors: strains D23580 and A130 lacked annotations for *rpmC*, though the gene sequence appears to exist at the same locus as other *Salmonella* strains. Similarly, the A130 annotation appears to contain a truncated annotation for *accB*, which was not properly clustered by Hieranoid. The gene *cohE*, encoding a phage repressor protein, was essential and present in *E. coli* and *K. pneumoniae* strains but not others, suggesting independent horizontal acquisition of similar prophage in the two lineages and misclassification as ancestral. Twenty-seven of the remaining genes were predicted as essential in most strains (Table 1), with only one or two strains scoring slightly above (<twofold higher) the essentiality threshold, indicating that most of these genes are likely either core essential or at least cause severe fitness defects in all strains surveyed.

This analysis left 24 genes that were ancestrally essential and essential in EcoGene, but not classified as core essential, where at least one strain had a score greater than twofold higher than the essentiality threshold (Fig. 4C). Some of these had clear-cut explanations. For instance, *folA*, encoding dihydrofolate reductase, is the target of

the antibiotic trimethoprim commonly used in the treatment of urinary tract infections. *folA* non-essentiality could be explained in all cases by the acquisition of alternative trimethoprim-resistant dihydrofolate reductase enzymes either in the plasmid or chromosome. Another explained example is *cysS*, encoding a cysteine-tRNA ligase uniquely non-essential in *S. Typhimurium* D23580, which was previously shown to be displaced by a plasmid-borne copy in this strain (57). Most others were less easily explained. For instance, *ispD*, *ispA*, and *hemL* in the isoprenoid and heme biosynthesis pathways were called as clearly non-essential in multiple strains. The gene *ispD* in particular, encoding a key cytidyltransferase in isoprenoid biosynthesis and a target of fosmidomycin (58), was non-essential in both ST131 UPEC strains. These are particularly interesting as the methyl-D-erythritol (MEP) pathway is not present in metazoans and hence has served as a target for therapeutic intervention (59).

Evidence for differences in essentiality among the Enterobacteriaceae

Forty-three genes were essential in the EcoGene list, but are neither core nor ancestral (Fig. 4B), constituting almost 15% of essential genes in *E. coli*. This is of particular importance, as the essential genes of *E. coli* are often taken as representative of related species. Fourteen were variably present in our strain collection, and a third of these encoded the antitoxin components of toxin-antitoxin systems (*chpS*, *mazE*, *mqsA*, *yafN*, *yefM*). Phage repressors were also in this set of essential, variably present genes as seen in previous comparisons of gene essentiality between strains (18), notably including *dicA* encoding the c2 repressor in the cryptic Qin prophage (60) that controls expression of the small protein DicB and small RNA DicF that affect cell division and phage susceptibility (61). Others include the *trpS* tryptophan tRNA-ligase, of which an ancestral second copy appears to have been lost in the lineage leading to *E. coli* and *C. rodentium*.

Approximately 60% (27/43) of the genes essential in the EcoGene list but not core or ancestrally essential were also non-essential in our *E. coli* BW25113 data set. We cross-compared our results to an independently generated and analyzed *E. coli* BW25113 TraDIS library (62) and found that with the exception of two genes very close to the essentiality decision boundary (*yrfF* and *can*), these results agreed with ours in all cases. Some of these disagreements with EcoGene could be explained by domain essentiality within a gene otherwise tolerant of insertions on visual examination. For instance, *rne*, *polA*, *yefM*, and *spoT* all appeared to contain essential domains across multiple strains and were additionally found to be essential across a collection of *E. coli* strains in a recent CRISPRi screen (21). However, the reasons behind the majority of these disagreements between EcoGene and TraDIS data remain unclear, and many of these genes remain poorly characterized.

One example of a set of genes essential in EcoGene and not core or ancestrally essential is associated with *rpoE*, which encodes the extracytoplasmic stress-responsive sigma factor σ^E (Fig. 4C). While σ^E has long been known to be essential in *E. coli* (63), and suppressor mutants are easily obtained (64), the exact reason why it is essential remains unclear, though may have to do with the accumulation of misfolded outer membrane proteins (OMPs) in the deletion strain (65). For activation, σ^E requires release from the anti-sigma factor RseA via the activity of the DegS and RseP proteases. All three genes *rpoE*, *degS*, and *rseP* were essential in our BW25113 data in agreement with EcoGene; however, none of these genes were essential in either of the ST131 *E. coli* strains included in our collection. These genes have also previously been reported to be essential in *S. Typhi* CVD908-htrA (18), confirmed both here and by an independent study in *S. Typhi* Ty2 (17). We also find all three to be essential in our clinical isolate of *K. pneumoniae* RH201207. Why this system is essential in three phylogenetically separated strains and not others is unclear. It has been speculated that σ^E essentiality may depend on the presence of *ycdQ/hicB* (64), though this was later shown to be the result of HicB toxin activation (66) and neither *S. Typhi* CVD908-htrA nor *K. pneumoniae* RH201207 encode the HicAB toxin-antitoxin system. The OMP content of Enterobacterial membranes can vary widely between strains, so this could possibly be a driving factor

in this differential essentiality if the accumulation of misfolded OMPs plays a role in the differential requirement for σ^E . Whatever the ultimate reasons for σ^E essentiality in these strains, it is clear that this trait has clearly arisen independently multiple times within the Enterobacteriaceae and can occur in strains of clinical relevance.

To further test our observation that changes in essentiality frequently happen between related strains in a genus, we developed a permutation test to identify genes more likely to be essential within a genus than without (see Methods). We found only 10 genes that were significantly specific to either *Salmonella* or *Escherichia*, the two genera where we had the most samples ($P < 0.05$, Fig. 4C, right panel). Of these, only one (*trpS*, discussed previously) was fully concordant with the phylogeny, being essential only in *Citrobacter* and *Escherichia*. All others were either sporadically non-essential within the genus (e.g., *fepC*, *fepG*, *ftsX*, *ftsE* in *Salmonella*), or also sporadically essential in other strains outside the genus (e.g., *rsgA*, *ubiF*, *lipA*, *ubiH*, *wzyE* in *Salmonella*). Our inability to identify genes whose essentiality is concordant with the phylogeny suggests that changes in essentiality status for non-core essential genes are rapid, occurring frequently within genera.

Analysis of essential genes conserved in reduced genomes identifies essential stress responses

Two major approaches have been taken to determining gene essentiality in the literature: experimental approaches, like the TraDIS technique used here, and comparative genomics based on the conservation of individual genes across bacterial phylogeny. Bacteria with reduced genomes, such as obligate pathogens and endosymbionts, are particularly informative for this purpose as they can survive with just hundreds of genes (11). The family Enterobacteriaceae contains a number of obligate insect endosymbionts, including the model symbionts *Buchnera aphidicola* in aphids and *Wigglesworthia glossinidia* in the tsetse fly (67). These bacteria typically reside in specialized fat cells termed bacteriocytes, which provide nutrients to the endosymbiont in exchange for essential amino acids or vitamins absent from the insect diet. While phylogenetic inference on highly reduced genomes is challenging, it does appear the endosymbiont lifestyle has emerged independently from free-living or facultative intracellular ancestors several times within the Enterobacteriaceae (68). Investigating gene conservation in these “natural experiments” provides an orthogonal approach to our TraDIS screens for determining core gene essentiality.

We collected a set of 34 endosymbiont genomes from within the Enterobacteriaceae on which to conduct conservation analysis. Half (17) were *Buchnera aphidicola* strains but also included representatives of the *Sodalis*, *Wigglesworthia*, *Blochmannia*, *Baumannia* genera; two secondary symbionts of psyllids; and *Moranella endobia*, a symbiont which lives within the betaproteobacterium *Tremblaya princeps*, itself a symbiont of mealybugs (69). Using the same Hieranoid-based method as we used for our essentiality analysis, we identified orthologous genes between these strains and our collection of strains used for TraDIS analysis. This allowed us to identify genes universally conserved among endosymbionts, as well as compare them to our essential gene sets for free-living Enterobacteriaceae on rich media.

We identified a remarkably small set of 120 genes universally conserved in the endosymbionts (Fig. 5A), and 73% (88 of 120) overlapped with either our core or ancestral essential gene sets determined by TraDIS. Nearly half of these (41 genes) are ribosomal proteins, comprising all but eight of the essential ribosomal proteins in our ancestral essential gene set (Table 1; Table S2). Other large classes of genes essential in free-living Enterobacteriaceae and conserved in endosymbionts included a subset of aminoacyl-tRNA biogenesis (nine genes, including *trpS*), genes involved in translation (*der*, *frr*, *fnt*, *infA*, *infB*, *mnmA*, *tadA*, *tsaD*, *pth*, *rnc*, *ybeY*, *ygfZ*), DNA replication (*dnaB*, *dnaN*, *dnaQ*, *dnaX*, *gyrB*, *holB*, *ssb*), and fatty acid biosynthesis (*acpS*, *fabB*, *fabG*, *fabI*).

While endosymbiont lineages largely conserved a core of genes involved in information transfer, few essential biosynthetic pathways were universally conserved

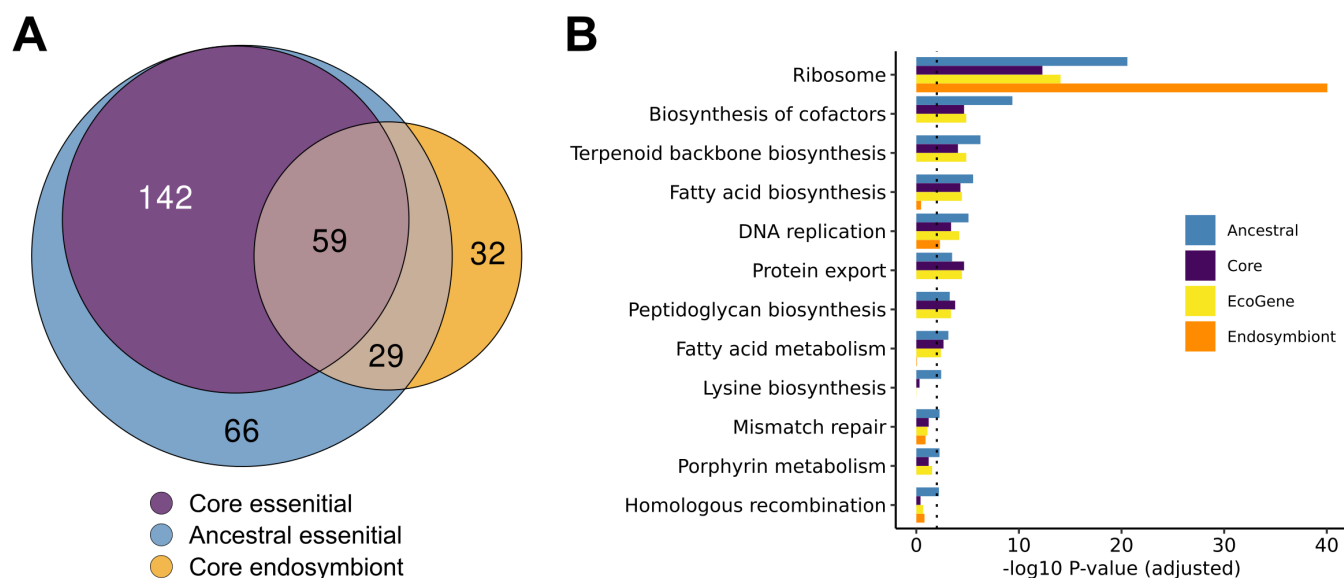


FIG 5 Comparing gene essentiality in free-living Enterobacteriaceae to conservation in reduced genomes. (A) Venn diagram comparing ancestral and core essential gene sets to genes classified as core endosymbiont genes. Core endosymbiont core genes are defined as those universally conserved across a set of 34 endosymbiont genomes from within the family Enterobacteriaceae. (B) Barplot for gene set overrepresentation analysis of KEGG pathways in the four essential gene classes. Bars depict FDR-adjusted hypergeometric P -values of significantly enriched pathways ($P(\text{adj}) < 0.01$) per gene class. All pathways with at least one significantly overrepresented class are shown. The dotted gray line shows the significance threshold ($P(\text{adj}) = 0.01$).

(Fig. 5B). Loss of genes involved in ubiquinone synthesis has been noted since the sequencing of the first *Buchnera* genome (70); we find that these and other associated isoprenoid and heme biosynthesis genes involved in electron transport have been lost independently multiple times. Similarly, essential genes associated with the synthesis of the outer membrane, peptidoglycan, and lipopolysaccharides have been lost multiple times, including the genes responsible for the synthesis of phospholipids. It has been speculated that the host may provide phospholipids for the symbiont membrane (71), though the details of how this may occur remain unclear. In the case of *Moranella*, it has been shown that host-encoded peptidoglycan synthesis genes of diverse bacterial origin can complement the absence of these genes in the endosymbiont (72, 73).

Thirty-two genes were universally conserved in endosymbionts despite not being essential in free-living species (Fig. 5A; Table 2). Some of these could be explained by their relationship to core information transfer processes: for instance, six encoding non-essential ribosomal proteins (*rplI*, *rpmEFGJ*, *rpsI*), and four encoding ribosomal- and tRNA-modifying enzymes (*mnmE*, *gidA/mnmG*, *rluD*, and *rsmH*). Others appeared to be associated with stress responses in free-living bacteria, which was unexpected as the host is generally thought to maintain a stable environment for its symbionts (71). These included, for instance, *lepA*, encoding a non-essential translation factor that contributes to survival of stress conditions (74) and whose deletion has previously been shown to negatively interact with the deletion of ubiquinone biosynthesis genes in *E. coli* (75), and *ahpC*, encoding alkyl hydroperoxide reductase required to scavenge hydrogen peroxide in the absence of catalase (76).

The requirement for stress response genes extends to entire systems, including for instance the trans-translation system comprising SsrA (the transfer-messenger RNA), SmpB, and the ClpPX protease that collectively rescue stalled ribosomes (77). While generally thought to be important for the survival of stress conditions in free-living bacteria (78), tmRNA has been found in some rare mitochondria (79) suggesting it may also play an important role in reduced genomes. A final striking example is given by the *cspE*, encoding the cold shock protein CspE. As the name suggests, cold shock proteins were discovered as regulators of the cold shock response (80), and the best

TABLE 2 Genes universally conserved in endosymbiont genomes and not ancestrally essential in the Enterobacteriaceae, excluding those encoding ribosomal proteins

Gene	Process	Annotated protein function
<i>aceF</i>	Glycolysis	Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex
<i>ahpC</i>	Oxidative stress response	Alkyl hydroperoxide reductase C
<i>aroK</i>	Biosynthesis of amino acids	Shikimate kinase 1
<i>clpP</i>	Protein degradation	ATP-dependent Clp protease proteolytic subunit
<i>clpX</i>	Protein degradation	ATP-dependent Clp protease ATP-binding subunit
<i>cspE</i>	RNA chaperone	Cold shock-like protein CspE
<i>deaD</i>	RNA degradation	ATP-dependent RNA helicase DeaD
<i>efp</i>	Translation	Elongation factor P
<i>gpmA</i>	Glycolysis	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase
<i>greA</i>	Transcription	Transcription elongation factor
<i>grxD</i>	Iron-sulfur cluster biogenesis	Glutaredoxin 4
<i>lepA</i>	Translation	Elongation factor 4
<i>lon</i>	Protein degradation	Lon protease
<i>map</i>	Protein maturation	Methionine aminopeptidase
<i>mnmE</i>	Translation	tRNA modification GTPase
<i>mnmG</i>	Translation	tRNA uridine 5-carboxymethylaminomethyl modification enzyme
<i>nfuA</i>	Iron-sulfur cluster biogenesis	Fe/S biogenesis protein
<i>prmC</i>	Translation	Release factor glutamine methyltransferase
<i>rluD</i>	Translation	Ribosomal large subunit pseudouridine synthase D
<i>rsmH</i>	Translation	Ribosomal RNA small subunit methyltransferase H
<i>smpB</i>	Translation	SsrA-binding protein
<i>sufA</i>	Iron-sulfur cluster biogenesis	Fe/S cluster assembly protein
<i>tkt</i>	Pentose phosphate pathway	Transketolase 1
<i>tpiA</i>	Glycolysis	Triosephosphate isomerase
<i>trpS</i>	Translation	Tryptophan--tRNA ligase
<i>ycfH</i>	Unknown	Uncharacterized metal-dependent hydrolase

characterized, CspA, acts to melt RNA structures to ease translation at low temperatures (81). Free-living Enterobacteriaceae genomes encode multiple cold shock protein homologs, with *E. coli* carrying nine, for example. CspE in particular has recently been characterized as playing a key role in the virulence of *S. enterica* (82), possibly through stabilizing mRNA transcripts (83), though the nature of this role is unclear and appears to be largely redundant with the paralogous protein CspC. A recent study of *Buchnera* genomes found that *cspC* has been repeatedly independently lost while *cspE* has been maintained in different lineages (84), a finding in keeping with our expanded analysis of other endosymbiont genomes. This suggests some level of non-redundancy in CspC and E function and further suggests cold shock proteins may collectively provide some essential function during normal growth masked by their high degree of redundancy in free-living Enterobacteriaceae.

DISCUSSION

An understanding of gene essentiality is critical to a wide variety of fields, from industrial biotechnology to medicine. Here we have investigated the essential genomes of thirteen strains of the Enterobacteriaceae, including the archetypal model organism *E. coli*. Our analysis found approximately one-third of essential genes in any given strain changing in their essentiality status over evolution across the genera *Escherichia*, *Salmonella*, *Klebsiella*, and *Citrobacter*. Rather than being an immutable gene set, this implies a remarkable flexibility in the essential genome. Beyond this study, mounting evidence supports this view. CRISPRi screens in diverse *E. coli* strains have shown that gene

essentiality varies widely between strains and that phylogenetic distance is a poor predictor of the fitness effect of any particular gene (21); similar observations have been made using transposon mutagenesis in a collection of *Pseudomonas aeruginosa* strains (19). While horizontal gene transfer appears to play a major role in essentiality changes, and we have highlighted instances where phage or plasmid acquisition appears to drive acquisition or loss of essentiality, horizontal transfer cannot fully explain these differences. Even during a relatively short period of adaptation to a static environment, recent work investigating gene essentiality in the Long-Term Evolution Experiment (LTEE) found nearly 200 genes that change in essentiality status in clonal populations (85). Here, we had initially expected to identify genus-specific essential genes but were unable to support this with our TraDIS data set. While it is possible that confounding factors in TraDIS screens such as difficulties in detecting domain essentiality or protection of DNA from transposition by high molecular weight DNA-binding proteins (86, 87) may complicate the identification of genus-specific essential genes, it is difficult to explain our complete inability to find them. This suggests that the variation in gene essentiality observed between closely related strains is similar to that seen across much larger phylogenetic distances.

Despite this variation, we also found a core essential gene set comprising 201 genes whose essentiality status remained unchanged across all strains in our collection. Remarkably, similarly sized sets of conserved essential genes have been previously reported between *E. coli* and the alphaproteobacterium *Caulobacter crescentus* (88), and even the Gram-positive *Bacillus subtilis* (89). How can we reconcile the deep conservation of this core essential gene set with the turnover observed across the other third of the essential gene set? Recent work in *Streptococcus pneumoniae* investigating core and strain-dependent essential genes has suggested that the fitness context is a driving feature discriminating the two (90): while disruption of core essential genes was only ever successful through the generation of merodiploids carrying a wild-type copy, disruption of strain-dependent essential genes frequently gave rise to suppressor mutations at distal loci. This suggests the existence of more paths to non-essentiality for some essential genes than others. The variable essentiality of the σ^E response within our TraDIS collection is a canonical example of a system with many paths to non-essentiality, as its essentiality can be suppressed by a number of distinct spontaneous mutations (64), which may partially explain how it has transitioned between essential and non-essential states multiple times over evolution within the Enterobacteriaceae.

If fitness context defines core and accessory essential genes, then obligate endosymbionts represent an almost pathological case study. The isolation and extreme bottlenecks these organisms have been subjected to have pruned the genome of redundancy and routes for compensation, uncovering essential processes masked in the larger genomes of free-living organisms. A prime example of this is the trans-translation system for rescuing stalled ribosomes, centered on tmRNA. In *E. coli*, trans-translation is important for the survival of a wide variety of stresses, including heat, acid stress, and starvation (78). Genetic interaction analysis has shown that tmRNA is only non-essential because of redundancy with other ribosome release systems (91, 92), which have been lost in the majority of the endosymbiont genomes we investigated here. The association of many genes universally conserved in endosymbionts with multiple stress responses in free-living bacteria appears to be a general trend, and suggests that at least some of what is generally considered a stress response is actually essential for cellular maintenance even in relatively static conditions, but masked by the redundancy of most bacterial genomes.

Given that essentiality is contextual, how might we best quantify this in the future? Our observations suggest several future directions. The first is investigating the relationship between conditional essentiality in response to stress and the core essentiality of a process. Many non-essential core conserved genes in endosymbionts are known to have pleiotropic phenotypes in free-living bacteria. It seems plausible that the more unrelated phenotypic traits a gene affects, the more difficult it will be

to compensate for during evolution. Indeed, genes whose deletion leads to sensitivity to a wide range of conditions tend to be involved in similar processes to essential genes (93). A second direction may be to directly reduce redundancy through genetic interaction screening. Methods for this have traditionally been extremely labor intensive, involving individually mating thousands of deletion strains (94, 95). Transposon insertion sequencing has made this analysis more straightforward for single query genes (12, 96), but would still require the construction of thousands of mutant libraries for true genome-wide analysis. CRISPR array methods, which allow multiplex expression of several guides within a single cell simultaneously (97, 98), could provide a viable approach to combinatorial silencing of genes allowing for truly comprehensive screening of genetic interactions. This would allow us to peel back the layers of redundancy and come to a better understanding of the minimal requirements for cellular viability.

MATERIALS AND METHODS

All data analysis scripts are freely available at <https://github.com/Gardner-BinLab/Enterobacteriaceae-TraDIS>. Raw sequencing reads have been deposited in GEO with accession GSE216013.

Construction and sequencing of TraDIS libraries

All TraDIS libraries first described here were constructed using a Tn5-derived transposon as described previously (25). Briefly, transposons carrying either a Kanamycin (*Citrobacter rodentium* ICC168, *Enterobacter cloacae* NCTC 9394, *Klebsiella pneumoniae* Ecl8, *Salmonella* Typhimurium D23580) or Chloramphenicol (*Salmonella* Typhimurium SL1344, *Salmonella* Typhimurium A130, *Salmonella* Enteritidis P125109, *Escherichia coli* ST131) resistance cassette were incubated with EZ-Tn5 transposase (Epicenter, Madison, USA) and electroporated into the target bacterium. Transformants were selected by plating on LB agar containing appropriate antibiotics and harvested directly from the plates following overnight incubation at 37°C. Multiple electroporation batches were combined to produce transposon mutant libraries of high complexity.

Genomic DNA from the pooled library samples was extracted using a Qiagen Genomic-tip 100/G kit (Qiagen) and used to prepare TraDIS sequencing libraries as previously described (26). Briefly, ~2 µg of genomic DNA was fragmented to a size of ~300 base pairs using a Covaris Focused ultrasonicator (Covaris), followed by end repair using the NEBNext End Repair module (NEB), A-tailing using the NEBNext dA-Tailing module (NEB), and adapter ligation using the NEBNext Quick Ligation Module with a custom splinkerette adaptor (SplA5). PCR enrichment of fragments containing transposon sequence was then carried out using transposon- and adaptor-specific primers using Kapa HiFi HotStart ReadyMix (Kapa Biosystems) for 10 to 20 cycles. PCR products were analyzed on an Agilent Bioanalyzer using a High Sensitivity DNA kit (Agilent Technologies) and by SYBR green qPCR (Kapa Biosystems). Sequencing was then performed on an Illumina HiSeq 2000, HiSeq 2500, or MiSeq using transposon-specific primers and a custom dark-cycle sequencing protocol at the Wellcome Sanger Institute or the King Abdullah University of Science and Technology Bioscience Core Laboratory (*Salmonella* Typhimurium A130 and D23580). All custom adaptors, primers, and sequencing protocols are described in reference (26) and publicly available at <https://github.com/sanger-pathogens/Bio-TraDIS/tree/master/recipes>.

TraDIS read mapping and quantification

Read mapping and quantification were performed using the Bio-TraDIS pipeline (26). FASTQ files for previously published TraDIS libraries were retrieved from the NCBI SRA (99), except for the *Escherichia coli* BW25113 data (31) which was retrieved from <https://genomics.lbl.gov/supplemental/rbarseq/>. The FASTQ files for non-standard multiplexed or barcoded samples [*Escherichia coli* EC958 (33), *Escherichia coli* BW25113 (31)] were modified to remove barcode sequences and leave a short transposon tag compatible

with Bio-TraDIS. Both new and previously published libraries were processed in a uniform manner. Reads were trimmed and quality filtered on their right (genomic sequence) ends using PRINSEQ (100) with the command line options:

```
prinseq-lite.pl -fastq stdin -out_good stdout -out_bad null -min_len 30
-trim_tail_right 2 -trim_ns_right 1 -trim_qual_right 20 -trim_qual_window 3
```

Reads were filtered for 10 bases matching the expected transposon tag sequence, and filtered reads were then mapped against the appropriate reference genome and plasmid sequences, available at <https://github.com/Gardner-BinfLab/EnTrl> using the Bio-TraDIS pipeline with the following command line options:

```
bacteria_tradis -v --smalt_r 0 --smalt_k 8 --smalt_s 1 --smalt_y 0.9 -m 0 -f
fastqs.txt -t TAAGAGACAG -r ref.fa
```

Insertion sites and associated read counts were tabulated in plot files appropriate for visualization with Artemis (101) and for further downstream analyses.

Correcting for biases in the insertion index

For each gene in each organism, we calculated an insertion index as a measure of gene essentiality. The insertion index is calculated as $\frac{(ng/lg)}{(nG/lG)}$ where ng is the number of unique insertion sites in the gene, lg is the length of the gene, nG is the number of insertion sites in the genome, and lG is the length of the genome. We calculated the insertion index for each gene after trimming 5% of the sequence from the 5' end and 20% of the sequence from the 3' end, following our analysis of insertion sites in essential genes (Fig. 3A).

To correct for biases in the insertion index, we used fitted values from a generalized additive model (GAM). For each genome, we fitted a GAM curve (using the MGCV package in R) with the formula $y \sim s(x)$ to the distance vs insertion index plot for distance normalization. The normalized insertion indices were calculated by dividing the insertion indices by the predicted insertion index from the GAM, and then multiplying them by the average insertion index. The same procedure was repeated for G-C normalization.

Classifying essential genes using DBSCAN

We used the DBSCAN R package (25) to cluster the normalized insertion indices using the DBSCAN function with parameters $\text{minPts} = 200$ and $\text{eps} = 0.05$. This density-based clustering finds two clusters for essential and non-essential genes, which are the two peaks shown by blue and orange in Fig. 2B, respectively. A region of noise data was scattered in between these two regions shown in light grey. The long tail of data points on the right side of Fig. 2B were also separated as noise by the DBSCAN algorithm, but as they have a high number of insertions, we grouped them with other non-essential genes.

Phylogenetic tree construction

We constructed a single-copy marker gene-based phylogenetic tree for the bacteria in this study using PhyloSift (29) and RaXML (28). We first used PhyloSift-search on all genome sequences used in this study to find marker genes in the PhyloSift database (29) by running `phylosift search --isolate --besthit genome_file`. Subsequently, we used PhyloSift-align to extract marker gene sequences by running the command `phylosift align --isolate --besthit genome_file`. Finally, we concatenated the alignments and used RaXMLHPC to generate a final tree by running `raxmlHPC -s protein_alignment_file -n output_tree -m PROTGAMMALG4M -p 1234 -f a -x 1234 -# 100`.

Clustering orthologous genes

We used Hieranoid (52) with default parameters to identify orthologous genes using BLAST (102) as a similarity search tool. The RaXML species tree described above was used for resolving orthology relationships.

Permutation test

To investigate the genus specificity of essential genes, we developed a permutation test for the genera *Salmonella* and *Escherichia*. The test evaluated genes that were called essential in at least one strain of the data set. Each gene was then evaluated to determine its essentiality within and outside its respective genus by comparing the fraction F_{within} , the number of strains within the genus where the gene is essentially divided by the total number of tested strains in the respective genus, to the fraction F_{outside} , the number of strains outside the genus where the gene is essentially divided by the total number of tested strains outside the respective genus. We then assigned genus specificity to genes where $F_{\text{within}} > F_{\text{outside}}$. To determine the statistical significance of genus specificity for these genes, we randomly shuffled genus assignments 10,000 times across strains and recalculated genus specificity similarly to generate a null distribution. We then calculated the P -value by dividing the number of permutations with an F_{within} greater than or equal to the actual result by the total amount of shuffles. We defined genes with a P -value < 0.05 as genus-enriched essential genes and show their essentiality scores across all strains in Fig. 4C.

Defining ancestrally essential and ancestrally conserved genes

To reconstruct gene essentiality along the phylogenetic tree, we used the parsimony-based Fitch's algorithm (55) with a binary alphabet for essentiality (0 for non-essential and 1 for essential). The parent of every pair of nodes in the tree was labeled with the intersection of its children's assigned sets if the intersection was not empty. Otherwise, it was labeled with the union of its children's assigned sets. We continued this process until we reached the root of the tree. At the root, if the assigned set only contained 1, the gene was considered ancestrally essential. Otherwise, it was considered ancestrally non-essential. The same procedure was performed, using gene presence rather than essentiality, to predict ancestral conservation in Fig. 4A. In addition, we applied the PastML algorithm, which uses a likelihood-based method to infer ancestral essentiality (56). Briefly, we performed the same preparation steps as for Fitch's algorithm but then ran the PastML algorithm using the command-line tool with default parameters. We then re-generated Fig. 4 and 5 using the PastML-inferred ancestral essential genes to compare the two approaches (Fig. S5 and S6).

Pathway analysis

We performed a KEGG pathway analysis using hypergeometric tests to determine pathway overrepresentation in the different classes of essential genes, that is, core essential genes, ancestral essential genes, EcoGene essential genes, and endosymbiont essential genes. We used the R package KEGGREST (v1.28.0) to assign genes to KEGG groups. Then, we generated hypergeometric P -values for each pathway, comparing the number of essential genes per category to the expectation for randomly drawn gene sets. Finally, we adjusted the P -values for multiple testing using the Benjamini–Hochberg method (103) and plotted the results in Fig. 5B.

ACKNOWLEDGMENTS

L.B. is supported in part by a Bayresq.net project grant (Rbiotics) from the Bavarian State Ministry for Science and Art and an NSERC Discovery Grant (RGPIN-2024-04305). G.L. and R.K. were supported by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/X011011/1 and BB/R012504/1 and its constituent projects BBS/E/QU/230002B, BBS/E/F/000PR10348, and BBS/E/F/000PR10349. A.K.C. is supported by an Australian Research Council Future Fellowship (FT220100152). P.P.G. is supported by the Marsden Fund (19-UOO-040 & 17-UOO-050), a Rutherford Discovery Fellowship (10-UOC-013), and a Ministry of Business, Innovation and Employment Grants (UOOX1709 & UOAX1932). Sequencing was supported by the Wellcome Sanger Institute and the KAUST Bioscience Core Laboratory.

AUTHOR AFFILIATIONS

- ¹Biomolecular Interactions Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand
- ²Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Center for Infection Research (HZI), Würzburg, Germany
- ³Microbes in the Food Chain, Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom
- ⁴ARC Centre of Excellence in Synthetic Biology, School of Natural Sciences, Macquarie University, Sydney, Australia
- ⁵Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom
- ⁶The Westmead Institute for Medical Research, University of Sydney, Sydney, Australia
- ⁷Sydney Institute for Infectious Diseases, University of Sydney, Sydney, Australia
- ⁸School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia
- ⁹King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
- ¹⁰Department of Medicine, University of Cambridge, Cambridge, United Kingdom
- ¹¹Department of Biological Sciences, University of East Anglia, Norwich, United Kingdom
- ¹²London School of Hygiene and Tropical Medicine, London, United Kingdom
- ¹³Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom
- ¹⁴Department of Biochemistry, Otago University, Dunedin, New Zealand
- ¹⁵Faculty of Medicine, University of Würzburg, Würzburg, Germany
- ¹⁶Department of Biology, University of Toronto, Mississauga, Ontario, Canada

AUTHOR ORCID*s*

Moataz Abd El Ghany  <http://orcid.org/0000-0002-0020-261X>

Robert A. Kingsley  <http://orcid.org/0000-0002-0194-6485>

Julian Parkhill  <http://orcid.org/0000-0002-7069-5958>

Paul P. Gardner  <http://orcid.org/0000-0002-7808-1213>

Lars Barquist  <http://orcid.org/0000-0003-4732-2667>

FUNDING

Funder	Grant(s)	Author(s)
Bayerisches Staatsministerium für Wissenschaft und Kunst (Bavarian State Ministry of Science and the Arts)	bayresq.net	Lars Barquist
Canadian Government Natural Sciences and Engineering Research Council of Canada (NSERC)	RGPIN-2024-04305	Lars Barquist
UKRI Biotechnology and Biological Sciences Research Council (BBSRC)	BB/X011011/1,BB/R012504/1,BBS/E/QU/230002B,BBS/E/F/000PR10348,BBS/E/F/000PR10349	Gemma C. Langridge Robert A. Kingsley
Australian Research Council	FT220100152	Amy K. Cain
Marsden Fund (Royal Society of New Zealand Marsden Fund)	19-UOO-040,17-UOO-050,10-UOC-013	Paul P. Gardner
Ministry of Business, Innovation and Employment (MBIE)	UOOX1709,UOAX1932	Paul P. Gardner

AUTHOR CONTRIBUTIONS

Fatemeh A. Ghomi, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft | Jakob J. Jung, Data curation, Formal analysis,

Investigation, Methodology, Software, Visualization, Writing – review and editing | Gemma C. Langridge, Resources, Writing – review and editing | Amy K. Cain, Resources, Writing – review and editing | Christine J. Boinett, Resources, Writing – review and editing | Moataz Abd El Ghany, Resources, Writing – review and editing | Derek J. Pickard, Resources, Writing – review and editing | Robert A. Kingsley, Resources, Writing – review and editing | Julian Parkhill, Resources, Writing – review and editing | Paul P. Gardner, Conceptualization, Funding acquisition, Investigation, Resources, Supervision, Writing – original draft, Writing – review and editing | Lars Barquist, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review and editing.

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

Supplemental material (mBio01798-24-s0001.pdf). Figures S1-S6 and Table S1.
Table S2 (mBio01798-24-s0002.xlsx). Complete study data.

REFERENCES

- Rancati G, Moffat J, Typas A, Pavelka N. 2018. Emerging and evolving concepts in gene essentiality. *Nat Rev Genet* 19:34–49. <https://doi.org/10.1038/nrg.2017.74>
- Mushegian AR, Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* 93:10268–10273. <https://doi.org/10.1073/pnas.93.19.10268>
- Koonin EV. 2000. How many genes can make a cell: the minimal-gene-set concept. *Annu Rev Genomics Hum Genet* 1:99–116. <https://doi.org/10.1146/annurev.genom.1.1.99>
- Juhas M, Eberl L, Church GM. 2012. Essential genes as antimicrobial targets and cornerstones of synthetic biology. *Trends Biotechnol* 30:601–607. <https://doi.org/10.1016/j.tibtech.2012.08.002>
- Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. 2007. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov* 6:29–40. <https://doi.org/10.1038/nrd2201>
- Tommasi R, Brown DG, Walkup GK, Manchester JJ, Miller AA. 2015. ESKAPEing the labyrinth of antibacterial discovery. *Nat Rev Drug Discov* 14:529–542. <https://doi.org/10.1038/nrd4572>
- Calero P, Nikel PI. 2019. Chasing bacterial chassis for metabolic engineering: a perspective review from classical to non-traditional microorganisms. *Microb Biotechnol* 12:98–124. <https://doi.org/10.1111/1751-7915.13292>
- Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang R-Y, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkuch C, Denisova EA, Young L, Qi Z-Q, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA, Smith HO, Venter JC. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329:52–56. <https://doi.org/10.1126/science.1190719>
- Hutchison CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinc TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L, Pelletier JF, Qi Z-Q, Richter RA, Strychalski EA, Sun L, Suzuki Y, Tsvetanova B, Wise KS, Smith HO, Glass JI, Merryman C, Gibson DG, Venter JC. 2016. Design and synthesis of a minimal bacterial genome. *Science* 351:aad6253. <https://doi.org/10.1126/science.aad6253>
- Boeke JD, Church G, Hessel A, Kelley NJ, Arkin A, Cai Y, Carlson R, Chakravarti A, Cornish VW, Holt L, Isaacs FJ, Kuiken T, Lajoie M, Lessor T, Lunshof J, Maurano MT, Mitchell LA, Rine J, Rosser S, Sanjana NE, Silver PA, Valle D, Wang H, Way JC, Yang L. 2016. GENOME ENGINEERING. The genome project-write. *Science* 353:126–127. <https://doi.org/10.1126/science.aaf6850>
- Moran NA, Bennett GM. 2014. The tiniest tiny genomes. *Annu Rev Microbiol* 68:195–215. <https://doi.org/10.1146/annurev-micro-091213-112901>
- Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, van Opijnen T. 2020. A decade of advances in transposon-insertion sequencing. *Nat Rev Genet* 21:526–540. <https://doi.org/10.1038/s41576-020-0244-x>
- Todor H, Silvis MR, Osadnik H, Gross CA. 2021. Bacterial CRISPR screens for gene function. *Curr Opin Microbiol* 59:102–109. <https://doi.org/10.1016/j.mib.2020.11.005>
- Yu Y, Gawlitt S, de Andrade e Sousa LB, Merdivan E, Piraud M, Beisel CL, Barquist L. 2024. Improved prediction of bacterial CRISPRi guide efficiency from depletion screens through mixed-effect machine learning and data integration. *Genome Biol* 25:13. <https://doi.org/10.1186/s13059-023-03153-y>
- Jacobs MA, Alwood A, Thaipisuttikul I, Spencer D, Haugen E, Ernst S, Will O, Kaul R, Raymond C, Levy R, Chun-Rong L, Guenther D, Bovee D, Olson MV, Manoil C. 2003. Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A* 100:14339–14344. <https://doi.org/10.1073/pnas.2036282100>
- Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2:2006. <https://doi.org/10.1038/msb4100050>
- Canals R, Xia X-Q, Fronick C, Clifton SW, Ahmer BMM, Andrews-Polymeris HL, Porwollik S, McClelland M. 2012. High-throughput comparison of gene fitness among related bacteria. *BMC Genomics* 13:212. <https://doi.org/10.1186/1471-2164-13-212>
- Barquist L, Langridge GC, Turner DJ, Phan M-D, Turner AK, Bateman A, Parkhill J, Wain J, Gardner PP. 2013. A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. *Nucleic Acids Res* 41:4549–4564. <https://doi.org/10.1093/nar/gkt148>
- Poulsen BE, Yang R, Clatworthy AE, White T, Osmulski SJ, Li L, Penaranda C, Lander ES, Shores N, Hung DT. 2019. Defining the core essential genome of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A* 116:10072–10080. <https://doi.org/10.1073/pnas.1900570116>
- Abd El Ghany M, Barquist L, Clare S, Brandt C, Mayho M, Joffe E, Sjöling Å, Turner AK, Klena JD, Kingsley RA, Hill-Cawthorne GA, Dougan G, Pickard D. 2021. Functional analysis of colonization factor antigen I positive enterotoxigenic *Escherichia coli* identifies genes implicated in survival in water and host colonization. *Microb Genom* 7:000554. <https://doi.org/10.1099/mgen.0.000554>
- Rousset F, Cabezas-Caballero J, Piastra-Facon F, Fernández-Rodríguez J, Clermont O, Denamur E, Rocha EPC, Bikard D. 2021. The impact of genetic diversity on gene essentiality within the *Escherichia coli* species. *Nat Microbiol* 6:301–312. <https://doi.org/10.1038/s41564-020-00839-y>
- Blount ZD. 2015. The unexhausted potential of *E. coli*. *Elife* 4:e05826. <https://doi.org/10.7554/eLife.05826>

23. Freed NE, Bumann D, Silander OK. 2016. Combining *Shigella* Tn-seq data with gold-standard *E. coli* gene deletion data suggests rare transitions between essential and non-essential gene functionality. *BMC Microbiol* 16:203. <https://doi.org/10.1186/s12866-016-0818-0>
24. Belotserkovsky I, Sansonetti PJ. 2018. *Shigella* and enteroinvasive *Escherichia coli*, p 1–26. In Frankel G, Ron EZ (ed), *Escherichia coli*, a versatile pathogen. Springer International Publishing, Cham.
25. Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK. 2009. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res* 19:2308–2316. <https://doi.org/10.1101/gr.097097.109>
26. Barquist L, Mayho M, Cummins C, Cain AK, Boinett CJ, Page AJ, Langridge GC, Quail MA, Keane JA, Parkhill J. 2016. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics* 32:1109–1111. <https://doi.org/10.1093/bioinformatics/btw022>
27. Ester M, Kriegel HP, Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD.
28. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
29. Darling AE, Jospin G, Lowe E, Matsen FA IV, Bik HM, Eisen JA. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243. <https://doi.org/10.7717/peerj.243>
30. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. <https://doi.org/10.1186/1471-2164-12-402>
31. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, Blow MJ, Bristow J, Butland G, Arkin AP, Deutschbauer A. 2015. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *mBio* 6:e00306-15. <https://doi.org/10.1128/mBio.00306-15>
32. Sharp C, Boinett C, Cain A, Housden NG, Kumar S, Turner K, Parkhill J, Kleanthous C. 2019. O-antigen-dependent colicin insensitivity of uropathogenic *Escherichia coli*. *J Bacteriol* 201:e00545-18. <https://doi.org/10.1128/JB.00545-18>
33. Phan M-D, Peters KM, Sarkar S, Lukowski SW, Allsopp LP, Gomes Moriel D, Achard MES, Totsika M, Marshall VM, Upton M, Beatson SA, Schembri MA. 2013. The serum resistome of a globally disseminated multidrug resistant uropathogenic *Escherichia coli* clone. *PLoS Genet* 9:e1003834. <https://doi.org/10.1371/journal.pgen.1003834>
34. Thomson NR, Clayton DJ, Windhorst D, Vernikov G, Davidson S, Churcher C, Quail MA, Stevens M, Jones MA, Watson M, et al. 2008. Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res* 18:1624–1637. <https://doi.org/10.1101/gr.077404.108>
35. Ondari EM, Klemm EJ, Msefula CL, El Ghany MA, Heath JN, Pickard DJ, Barquist L, Dougan G, Kingsley RA, MacLennan CA. 2019. Rapid transcriptional responses to serum exposure are associated with sensitivity and resistance to antibody-mediated complement killing in invasive *Salmonella* Typhimurium ST313. *Wellcome Open Res* 4:74. <https://doi.org/10.12688/wellcomeopenres.15059.1>
36. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, Kariuki S, Msefula CL, Gordon MA, de Pinna E, Wain J, Heyderman RS, Obaro S, Alonso PL, Mandomando I, MacLennan CA, Tapia MD, Levine MM, Tennant SM, Parkhill J, Dougan G. 2012. Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. *Nat Genet* 44:1215–1221. <https://doi.org/10.1038/ng.2423>
37. Petty NK, Feltwell T, Pickard D, Clare S, Toribio AL, Fookes M, Roberts K, Monson R, Nair S, Kingsley RA, Bulgin R, Wiles S id, et al. 2011. *Citrobacter rodentium* is an unstable pathogen showing evidence of significant genomic flux. *PLoS Pathog* 7:e1002018. <https://doi.org/10.1371/journal.ppat.1002018>
38. Collins JW, Keeney KM, Crepin VF, Rathinam VAK, Fitzgerald KA, Finlay BB, Frankel G. 2014. *Citrobacter rodentium*: infection, inflammation and the microbiota. *Nat Rev Microbiol* 12:612–623. <https://doi.org/10.1038/nrmicro3315>
39. Fookes M, Yu J, De Majumdar S, Thomson N, Schneiders T. 2013. Genome sequence of *Klebsiella pneumoniae* Ecl8, a reference strain for targeted genetic manipulation. *Genome Announc* 1:e00027-12. <https://doi.org/10.1128/genomeA.00027-12>
40. Jana B, Cain AK, Doerrler WT, Boinett CJ, Fookes MC, Parkhill J, Guardabassi L. 2017. The secondary resistome of multidrug-resistant *Klebsiella pneumoniae*. *Sci Rep* 7:42483. <https://doi.org/10.1038/srep42483>
41. 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* 112:4110–4115. <https://doi.org/10.1073/pnas.1419677112>
42. Zhou J, Rudd KE. 2013. EcoGene 3.0. *Nucleic Acids Res* 41:D613–24. <https://doi.org/10.1093/nar/gks1235>
43. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS. 1998. Tn5/IS50 target recognition. *Proc Natl Acad Sci U S A* 95:10716–10721. <https://doi.org/10.1073/pnas.95.18.10716>
44. Gallagher LA, Shendure J, Manoil C. 2011. Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. *mBio* 2:e00315-10. <https://doi.org/10.1128/mBio.00315-10>
45. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. 2012. Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mob DNA* 3:3. <https://doi.org/10.1186/1759-8753-3-3>
46. Zomer A, Burghout P, Bootsma HJ, Hermans PWM, van Hijum S. 2012. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS One* 7:e43012. <https://doi.org/10.1371/journal.pone.0043012>
47. Rubin BE, Wetmore KM, Price MN, Diamond S, Shultzaberger RK, Lowe LC, Curtin G, Arkin AP, Deutschbauer A, Golden SS. 2015. The essential gene set of a photosynthetic organism. *Proc Natl Acad Sci U S A* 112:E6634–E6643. <https://doi.org/10.1073/pnas.1519220112>
48. Korandla DR, Wozniak JM, Campeau A, Gonzalez DJ, Wright ES. 2020. AssessORF: combining evolutionary conservation and proteomics to assess prokaryotic gene predictions. *Bioinformatics* 36:1022–1029. <https://doi.org/10.1093/bioinformatics/btz714>
49. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res* 49:D412–D419. <https://doi.org/10.1093/nar/gkaa913>
50. Sullivan GJ, Barquist L, Cain AK. 2024. A method to correct for local alterations in DNA copy number that bias functional genomics assays applied to antibiotic-treated bacteria. *mSystems* 9:e0066523. <https://doi.org/10.1128/mSystems.00665-23>
51. Barquist L, Boinett CJ, Cain AK. 2013. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biol* 10:1161–1169. <https://doi.org/10.4161/rna.24765>
52. Schreiber F, Sonnhammer ELL. 2013. Hieranoid: hierarchical orthology inference. *J Mol Biol* 425:2072–2081. <https://doi.org/10.1016/j.jmb.2013.02.018>
53. Bubunenko M, Baker T, Court DL. 2007. Essentiality of ribosomal and transcription antitermination proteins analyzed by systematic gene replacement in *Escherichia coli*. *J Bacteriol* 189:2844–2853. <https://doi.org/10.1128/JB.01713-06>
54. Takiff HE, Chen SM, Court DL. 1989. Genetic analysis of the rnc operon of *Escherichia coli*. *J Bacteriol* 171:2581–2590. <https://doi.org/10.1128/jb.171.5.2581-2590.1989>
55. Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Biol* 20:406–416. <https://doi.org/10.1093/sysbio/20.4.406>
56. Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. 2019. A fast likelihood method to reconstruct and visualize ancestral scenarios. *Mol Biol Evol* 36:2069–2085. <https://doi.org/10.1093/molbev/msz131>
57. Canals R, Hammarlöf DL, Kröger C, Owen SV, Fong WY, Lacharme-Lora L, Zhu X, Wenner N, Carden SE, Honeycutt J, Monack DM, Kingsley RA, Brownridge P, Chaudhuri RR, Rowe WPM, Predeus AV, Hokamp K, Gordon MA, Hinton JCD. 2019. Adding function to the genome of African *Salmonella* Typhimurium ST313 strain D23580. *PLoS Biol* 17:e3000059. <https://doi.org/10.1371/journal.pbio.3000059>
58. Zhang B, Watts KM, Hodge D, Kemp LM, Hunstad DA, Hicks LM, Odom AR. 2011. A second target of the antimalarial and antibacterial agent

- fosmidomycin revealed by cellular metabolic profiling. *Biochemistry* 50:3570–3577. <https://doi.org/10.1021/bi200113y>
59. Singh KS, Sharma R, Reddy PAN, Vonteddu P, Good M, Sundararajan A, Choi H, Muthumani K, Kossenkov A, Goldman AR, Tang H-Y, Totrov M, Cassel J, Murphy ME, Somasundaram R, Herlyn M, Salvino JM, Ditiwala F. 2021. IspH inhibitors kill Gram-negative bacteria and mobilize immune clearance. *Nat New Biol* 589:597–602. <https://doi.org/10.1038/s41586-020-03074-x>
 60. B  jar S, Bouch   F, Bouch   JP. 1988. Cell division inhibition gene *dicB* is regulated by a locus similar to lambdoid bacteriophage immunity loci. *Mol Gen Genet* 212:11–19. <https://doi.org/10.1007/BF00322439>
 61. Ragunathan PT, Vanderpool CK. 2019. Cryptic-prophage-encoded small protein DicB protects *Escherichia coli* from phage infection by inhibiting inner membrane receptor proteins. *J Bacteriol* 201:e00475–19. <https://doi.org/10.1128/JB.00475-19>
 62. Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, Cunningham AF, Lund PA, Cole JA, Henderson IR. 2018. The essential genome of *Escherichia coli* K-12. *mBio* 9:e02096–17. <https://doi.org/10.1128/mBio.02096-17>
 63. De Las Pe  as A, Connolly L, Gross CA. 1997. SigmaE is an essential sigma factor in *Escherichia coli*. *J Bacteriol* 179:6862–6864. <https://doi.org/10.1128/jb.179.21.6862-6864.1997>
 64. Button JE, Silhavy TJ, Ruiz N. 2007. A suppressor of cell death caused by the loss of sigmaE downregulates extracytoplasmic stress responses and outer membrane vesicle production in *Escherichia coli*. *J Bacteriol* 189:1523–1530. <https://doi.org/10.1128/JB.01534-06>
 65. Konovalova A, Grabowicz M, Balibar CJ, Malinverni JC, Painter RE, Riley D, Mann PA, Wang H, Garlisi CG, Sherborne B, Rigel NW, Ricci DP, Black TA, Roemer T, Silhavy TJ, Walker SS. 2018. Inhibitor of intramembrane protease RseP blocks the σ^E response causing lethal accumulation of unfolded outer membrane proteins. *Proc Natl Acad Sci U S A* 115:E6614–E6621. <https://doi.org/10.1073/pnas.1806107115>
 66. Daimon Y, Narita S, Akiyama Y. 2015. Activation of toxin-antitoxin system toxins suppresses lethality caused by the loss of σ^E in *Escherichia coli*. *J Bacteriol* 197:2316–2324. <https://doi.org/10.1128/JB.00079-15>
 67. Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet* 42:165–190. <https://doi.org/10.1146/annurev.genet.41.110306.130119>
 68. Husnik F, Chrudimsk   T, Hyp  a V. 2011. Multiple origins of endosymbiosis within the Enterobacteriaceae (γ -Proteobacteria): convergence of complex phylogenetic approaches. *BMC Biol* 9:87. <https://doi.org/10.1186/1741-7007-9-87>
 69. McCutcheon JP, von Dohlen CD. 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Curr Biol* 21:1366–1372. <https://doi.org/10.1016/j.cub.2011.06.051>
 70. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature New Biol* 407:81–86. <https://doi.org/10.1038/35024074>
 71. Baumann P. 2005. Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annu Rev Microbiol* 59:155–189. <https://doi.org/10.1146/annurev.micro.59.030804.121041>
 72. Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, von Dohlen CD, Fukatsu T, McCutcheon JP. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* 153:1567–1578. <https://doi.org/10.1016/j.cell.2013.05.040>
 73. Bublitz DC, Chadwick GL, Magyar JS, Sandoz KM, Brooks DM, Mesnage S, Ladinsky MS, Garber AI, Bjorkman PJ, Orphan VJ, McCutcheon JP. 2019. Peptidoglycan production by an insect-bacterial mosaic. *Cell* 179:703–712. <https://doi.org/10.1016/j.cell.2019.08.054>
 74. Starosta AL, Lassak J, Jung K, Wilson DN. 2014. The bacterial translation stress response. *FEMS Microbiol Rev* 38:1172–1201. <https://doi.org/10.1111/1574-6976.12083>
 75. Balakrishnan R, Oman K, Shoji S, Bundschuh R, Fredrick K. 2014. The conserved GTPase LepA contributes mainly to translation initiation in *Escherichia coli*. *Nucleic Acids Res* 42:13370–13383. <https://doi.org/10.1093/nar/gku1098>
 76. Seaver LC, Imlay JA. 2001. Alkyl hydroperoxide reductase is the primary scavenger of endogenous hydrogen peroxide in *Escherichia coli*. *J Bacteriol* 183:7173–7181. <https://doi.org/10.1128/JB.183.24.7173-7181.2001>
 77. Gottesman S, Roche E, Zhou Y, Sauer RT. 1998. The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes Dev* 12:1338–1347. <https://doi.org/10.1101/gad.12.9.1338>
 78. Li J, Ji L, Shi W, Xie J, Zhang Y. 2013. Trans-translation mediates tolerance to multiple antibiotics and stresses in *Escherichia coli*. *J Antimicrob Chemother* 68:2477–2481. <https://doi.org/10.1093/jac/dkt231>
 79. Burger G, Gray MW, Forget L, Lang BF. 2013. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol Evol* 5:418–438. <https://doi.org/10.1093/gbe/evt008>
 80. Phadtare S, Alsina J, Inouye M. 1999. Cold-shock response and cold-shock proteins. *Curr Opin Microbiol* 2:175–180. [https://doi.org/10.1016/S1369-5274\(99\)80031-9](https://doi.org/10.1016/S1369-5274(99)80031-9)
 81. Zhang Y, Burkhardt DH, Rouskin S, Li G-W, Weissman JS, Gross CA. 2018. A stress response that monitors and regulates mRNA structure is central to cold shock adaptation. *Mol Cell* 70:274–286. <https://doi.org/10.1016/j.molcel.2018.02.035>
 82. Michaux C, Holmqvist E, Vasicek E, Sharan M, Barquist L, Westermann AJ, Gunn JS, Vogel J. 2017. RNA target profiles direct the discovery of virulence functions for the cold-shock proteins CspC and CspE. *Proc Natl Acad Sci U S A* 114:6824–6829. <https://doi.org/10.1073/pnas.1620772114>
 83. Jenniches L, Michaux C, Popella L, Reichardt S, Vogel J, Westermann AJ, Barquist L. 2024. Improved RNA stability estimation through Bayesian modeling reveals most *Salmonella* transcripts have subminute half-lives. *Proc Natl Acad Sci U S A* 121:e2308814121. <https://doi.org/10.1073/pnas.2308814121>
 84. Chong RA, Park H, Moran NA. 2019. Genome evolution of the obligate endosymbiont *Buchnera aphidicola*. *Mol Biol Evol* 36:1481–1489. <https://doi.org/10.1093/molbev/msz082>
 85. Couce A, Limdi A, Magnan M, Owen SV, Herren CM, Lenski RE, Tenaillon O, Baym M. 2024. Changing fitness effects of mutations through long-term bacterial evolution. *Science* 383:eadd1417. <https://doi.org/10.1126/science.add1417>
 86. Manna D, Porwollik S, McClelland M, Tan R, Higgins NP. 2007. Microarray analysis of Mu transposition in *Salmonella enterica*, serovar Typhimurium: transposon exclusion by high-density DNA binding proteins. *Mol Microbiol* 66:315–328. <https://doi.org/10.1111/j.1365-2958.2007.05915.x>
 87. Kimura S, Hubbard TP, Davis BM, Waldor MK. 2016. The nucleoid binding protein H-NS biases genome-wide transposon insertion landscapes. *mBio* 7:e01351–16. <https://doi.org/10.1128/mBio.01351-16>
 88. Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Collier JA, Fero MJ, McAdams HH, Shapiro L. 2011. The essential genome of a bacterium. *Mol Syst Biol* 7:528. <https://doi.org/10.1038/msb.2011.58>
 89. Koo B-M, Kritikos G, Farelli JD, Todor H, Tong K, Kimsey H, Wapinski I, Galardini M, Cabal A, Peters JM, Hachmann A-B, Rudner DZ, Allen KN, Typas A, Gross CA. 2017. Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. *Cell Syst* 4:291–305. <https://doi.org/10.1016/j.cels.2016.12.013>
 90. Rosconi F, Rudmann E, Li J, Surujon D, Anthony J, Frank M, Jones DS, Rock C, Rosch JW, Johnston CD, van Opijnen T. 2022. A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nat Microbiol* 7:1580–1592. <https://doi.org/10.1038/s41564-022-01208-7>
 91. Chadani Y, Ono K, Ozawa S, Takahashi Y, Takai K, Nanamiya H, Tozawa Y, Kutsukake K, Abo T. 2010. Ribosome rescue by *Escherichia coli* ArfA (YhdL) in the absence of trans-translation system. *Mol Microbiol* 78:796–808. <https://doi.org/10.1111/j.1365-2958.2010.07375.x>
 92. Feaga HA, Viollier PH, Keiler KC. 2014. Release of nonstop ribosomes is essential. *mBio* 5:e01916. <https://doi.org/10.1128/mBio.01916-14>
 93. Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA. 2011. Phenotypic landscape of a bacterial cell. *Cell* 144:143–156. <https://doi.org/10.1016/j.cell.2010.11.052>
 94. Typas A, Nichols RJ, Siegle DA, Shales M, Collins SR, Lim B, Braberg H, Yamamoto N, Takeuchi R, Wanner BL, Mori H, Weissman JS, Krogan NJ, Gross CA. 2008. High-throughput, quantitative analyses of genetic

- interactions in *E. coli*. *Nat Methods* 5:781–787. <https://doi.org/10.1038/nmeth.1240>
95. Butland G, Babu M, Díaz-Mejía JJ, Bohdana F, Phanse S, Gold B, Yang W, Li J, Gagarinova AG, Pogoutse O, et al. 2008. eSGA: *E. coli* synthetic genetic array analysis. *Nat Methods* 5:789–795. <https://doi.org/10.1038/nmeth.1239>
96. van Opijnen T, Bodi KL, Camilli A. 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* 6:767–772. <https://doi.org/10.1038/nmeth.1377>
97. Reis AC, Halper SM, Vezeau GE, Cetnar DP, Hossain A, Clauer PR, Salis HM. 2019. Simultaneous repression of multiple bacterial genes using nonrepetitive extra-long sgRNA arrays. *Nat Biotechnol* 37:1294–1301. <https://doi.org/10.1038/s41587-019-0286-9>
98. Liao C, Ttofali F, Slotkowski RA, Denny SR, Cecil TD, Leenay RT, Keung AJ, Beisel CL. 2019. Modular one-pot assembly of CRISPR arrays enables library generation and reveals factors influencing crRNA biogenesis. *Nat Commun* 10:2948. <https://doi.org/10.1038/s41467-019-10747-3>
99. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. 2022. The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res* 50:D387–D390. <https://doi.org/10.1093/nar/gkab1053>
100. Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27:863–864. <https://doi.org/10.1093/bioinformatics/btr026>
101. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 28:464–469. <https://doi.org/10.1093/bioinformatics/btr703>
102. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
103. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57:289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>