Title

Escherichia coli non-coding regulatory regions are highly conserved.

Permalink

https://escholarship.org/uc/item/45q726gs

Journal

NAR Genomics and Bioinformatics, 6(2)

Authors

Lamoureux, Cameron
Phaneuf, Patrick
Zielinski, Daniel
et al.

Publication Date

2024-06-01

DOI

10.1093/nargab/lqae041

Peer reviewed

NAR Genomics and Bioinformatics, 2024, **6**, lqae041
https://doi.org/10.1093/nargab/lqae041
Advance access publication date: 20 May 2024
Standard Article

OXFORD

Escherichia coli non-coding regulatory regions are highly conserved

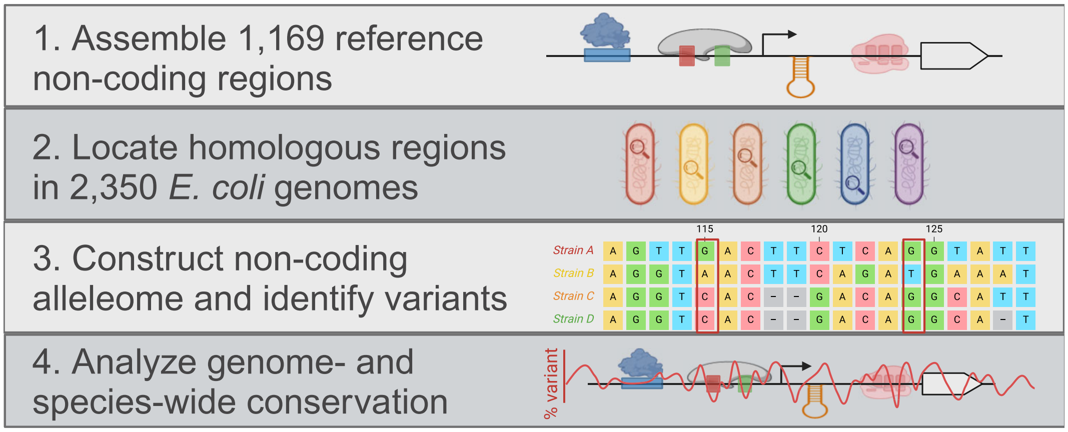
Cameron R. Lamoureux[1], **Patrick V. Phaneuf**[2], **Bernhard O. Palsson**[1,2] **and Daniel C. Zielinski**[1,*]

[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA
[2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Building 220, 2800 Kgs. Lyngby, Denmark
[*]To whom correspondence should be addressed. Tel: +1 858 246 1625; Fax: +1 858 822 3120; Email: dczielin@ucsd.edu

Abstract

Microbial genome sequences are rapidly accumulating, enabling large-scale studies of sequence variation. Existing studies primarily focus on coding regions to study amino acid substitution patterns in proteins. However, non-coding regulatory regions also play a distinct role in determining physiologic responses. To investigate intergenic sequence variation on a large-scale, we identified non-coding regulatory region alleles across 2350 *Escherichia coli* strains. This 'alleleome' consists of 117 781 unique alleles for 1169 reference regulatory regions (transcribing 1975 genes) at single base-pair resolution. We find that 64% of nucleotide positions are invariant, and variant positions vary in a median of just 0.6% of strains. Additionally, non-coding alleles are sufficient to recover *E. coli* phylogroups. We find that core promoter elements and transcription factor binding sites are significantly conserved, especially those located upstream of essential or highly-expressed genes. However, variability in conservation of transcription factor binding sites is significant both within and across regulons. Finally, we contrast mutations acquired during adaptive laboratory evolution with wild-type variation, finding that the former preferentially alter positions that the latter conserves. Overall, this analysis elucidates the wealth of information found in *E. coli* non-coding sequence variation and expands pangenomic studies to non-coding regulatory regions at single-nucleotide resolution.

Graphical abstract


1. Assemble 1,169 reference non-coding regions
2. Locate homologous regions in 2,350 *E. coli* genomes
3. Construct non-coding alleleome and identify variants
4. Analyze genome- and species-wide conservation
% variant

					115					120					125				
Strain A	A	G	T	T	G	A	C	T	T	C	T	C	A	G	G	T	A	T	T
Strain B	A	G	G	T	A	A	C	T	T	C	A	G	A	T	G	A	A	A	T
Strain C	A	G	G	T	C	A	C	–	–	G	A	C	A	G	G	C	A	T	T
Strain D	A	G	G	T	C	A	C	–	–	G	A	C	A	G	G	C	A	–	T

Introduction

Rapidly falling costs have yielded an explosion in complete genome sequences across organisms. Far from the first microbial genome assemblies almost 30 years ago, this wealth of sequence data necessitates genetic analyses that span thousands of genomes simultaneously (1). Pathogen genomes are particularly well-represented due to sequencing surveillance efforts (2,3); understanding genotype-phenotype relationships for these bacteria is critical. Pangenome analysis serves as a key tool towards this end by studying coding gene presence and absence across strains (4). Pangenome analyses have defined core (conserved) and accessory (variable) genomes for major microbial species (5) as well as identifying pangenome openness, or the continued discovery of unique genes as more genomes are sequenced(6). Advancements in understanding antimicrobial resistance (7), virulence (8) and metabolism (9) have all been empowered by pangenomics.

However, pangenome analysis has been primarily focused on coding regions. While coding sequences make up the majority of a typical microbial genome (10), non-coding regulatory regions play an outsize role in manifesting phenotype from genotype. In particular, promoters (11–13) and 5′ untranslated regions (5′ UTRs) (14,15) play a key role in executing the central dogma by modulating transcription and translation of operons. The core promoter features driving RNA polymerase binding via sigma factor recognition—the

Received: October 31, 2023. **Revised:** April 5, 2024. **Editorial Decision:** April 8, 2024. **Accepted:** May 15, 2024

−10 and −35 boxes—are well known(16–19). In turn, transcription start sites (TSS) have also been systematically identified (20) at base-pair resolution. Transcription factors (TFs) influence transcription in response to environmental stimuli via specific recognition of transcription factor binding sites (TFBS) within promoters and 5′ UTRs (21–24). Transcriptional attenuators also play an important role in regulating expression of certain genes(25).

All of these sequence features are encoded differently from genes. They are therefore primarily located in non-coding regulatory regions, rendering them invisible to pangenome analyses focused solely on coding genes. Quantifying variation and conservation within these regions would shed light on the evolutionary pressures affecting control of expression. Because of the fine-grained nature of these critical sequence features, a base-pair resolution view of non-coding variation amongst wild-type genomes is warranted. Coding sequence pangenome analyses typically focus on presence/absence of homologous gene clusters within each strain or organism studied. An existing pangenome tool (26) focused on intergenic regions identifies wholesale horizontal transfers but does not incorporate base-pair level sequence information. To analyze intergenic sequence variation, a non-coding alleleome must be established, where ‘alleleome’ refers to the aggregation of alleles for all sequence-based features of interest across a species (27–29). Such a construct would serve to more explicitly link pangenomics to the literature surrounding molecular evolution and variation, which also focuses primarily on coding regions (30–35).

We therefore built a non-coding alleleome for *Escherichia coli*, focusing on promoter and 5′ UTR regions. We amassed 2350 fully-sequenced *E. coli* strains from across the phylogenetic tree, isolated from a variety of hosts. From the reference strain K-12 MG1655, we extracted 1169 well-annotated non-coding regulatory regions that regulate transcription of 1975 genes. We then identified and aligned alleles for these intergenic regions across the 2350 strains. The resulting alleleome contains 117 781 unique alleles comprising over 400 000 base positions. Overall, we find the *E. coli* non-coding alleleome to be remarkably conserved. Furthermore, we: (i) cluster strains based solely on non-coding alleles, recovering phylogroups; (ii) quantify variation and conservation within key sequence features; (iii) identify essentiality and high expression as drivers of feature-specific conservation; (iv) characterize variation in conservation across transcription factor binding sites and (v) contrast adaptive laboratory evolution (ALE) mutations with wild-type variants. Taken together, the *E. coli* non-coding alleleome and analyses enabled by it represent an important expansion of large-scale genome sequence analysis to less-studied regions.

Materials and methods

Assembling complete *E. coli* genome sequences

Complete *E. coli* genome sequences and metadata were downloaded from BV-BRC (formerly known as PATRIC) (36). These genomes were subjected to the following quality control steps. Completeness and quality were verified by selecting genomes with ‘Contig L50’ of 1 (smallest number of contigs whose length sum makes up half of genome size; i.e. one contig should account for the majority of the genome) and ‘Contig N50’ >4 Mb (sequence length of the shortest contig at 50% of the total assembly length; i.e. the shortest contig should be at least 4 million bp, to capture minimum expected *E. coli* genome length). Furthermore, only genomes without ambiguous bases (i.e. only ACGT in sequence) were selected. Finally, genomes were selected only if they had coding sequences annotated (i.e. a GFF/FAA file was also downloaded). Phylogroups/clades were assigned for each genome sequence using the ClermonTyping *in silico* tool (37). Genomes annotated as ‘Non Escherichia’ or ‘Unknown’ were excluded. After these filtering steps, 2350 complete genome sequences remained (Supplemental Table 1).

Generating coding sequence pangenome

A coding sequence pangenome was generated as described previously (38). All FAA files for all amino acid sequences of all genes from all valid strains were combined into a single file and subjected to duplicate removal, yielding a listing of all 918 781 non-redundant protein sequences. This file was then provided to the CD-HIT protein sequence clustering program (v.4.8.1(39)) with the following non-default options: ‘-n 5 -c 0.8’. This processing yielded 80453 gene clusters. These clusters (and their constituent individual alleles) were then given unique identifiers and referenced back to the strain(s) from which they came.

Identifying reference non-coding regulatory regions and features

High-confidence transcription start sites (TSS) for the reference strain *Escherichia coli* K-12 MG1655 (genome accession number NC_000913.3, BV-BRC/PATRIC genome ID 511145.12) were accessed from RegulonDB (21). This resource has been extensively manually curated and comes with additional annotation of non-coding and regulatory features for these high-confidence TSSs. 2228 TSS were annotated as transcribing at least one coding gene. Each TSS was mapped to the first gene it transcribes. Then, a sequence region starting from 200 bp upstream of the TSS through 50 bp downstream of the first gene’s start codon was extracted for each of these TSS/first gene pairs. Thus, each reference non-coding regulatory contains (from upstream to downstream): the 200 bp upstream of the TSS—the TSS itself (1 bp)—all bases from the first base downstream of TSS to the base immediately upstream of the first transcribed gene’s start codon (the 5′ UTR; variable length)—the first 50 bp of the first transcribed gene (i.e. the first ∼7 codons, to aid in the next alignment step). At this stage, a separate region was extracted for alternate TSSs transcribing the same first gene, even if the regions partially overlapped. These nucleotide sequences were then written to a FASTA file.

Searching for reference non-coding regulatory regions across all strains

For each pangenome strain, coding genes that appeared in a cluster with a K-12 MG1655 gene were selected. Then, the largest distance from the reference strain gene start to a reference strain TSS transcribing that gene was determined: call this distance *reference_region_upstream_dist*. A local search region in the pangenome strain was extracted as follows: (100 + *reference_region_upstream_dist*) bp upstream from pangenome gene start to 100 bp downstream of pangenome gene start. Within each pangenome strain, all such search regions were combined into a single FASTA file and passed to create a BLAST search database with the BLAST+ (40) program *makeblastdb*. Then, *blastn* was used to search for all reference non-coding regulatory regions against this strain- and region-specific database. For each pangenome strain, only BLAST matches for a reference non-coding regulatory region in the local search region upstream of the pangenome strain gene corresponding to the appropriate reference strain gene were kept. If multiple alignments were found within the correct local search region, the alignment with the lowest *E*-value was selected. For each match, the corresponding nucleotide sequence of the non-coding regulatory region allele from each strain was extracted from the strain's genome. Finally, all sequence matches for a given reference non-coding regulatory region were grouped together.

Building the non-coding alleleome

For each set of non-coding sequences corresponding to a particular reference non-coding regulatory region (non-coding alleles), the nucleotide sequences were aligned using multiple sequence alignment tool MUSCLE (41) with all default arguments. Aligned sequences with greater than 20% gaps were filtered out. Then, only non-coding regulatory regions with an allele found in at least 75% of strains were kept; this step removed 13% (276/2074) of regions. This threshold ensured that the regions studied were actually well-represented within the set of strains, as our goal was to investigate base-pair level variation within these well-represented regions as opposed to presence/absence of less common regions between strains. At this point, due to alternate TSS for the same transcription unit, some non-coding regulatory regions could be subsets of others. Thus, for each set of alternate TSS, only the longest aligned set was selected for further analysis as the other regions would be subsets thereof. These steps led to the identification of 1169 final regions that—with all of their alleles—comprise the *E. coli* non-coding alleleome. These regions account for the transcription of 1975 genes–1970 protein-coding genes and 5 non-coding RNAs.

Annotating alleleome base pairs with variant and feature information

For each aligned base pair in the alleleome, variant percentage was calculated as the percentage of strains that have the non-dominant base at that position. Then, two *E. coli* reference strain regulatory element databases (Bitome(10) and RegulonDB v11.1(21)) were used to annotate each base pair in the alleleome for presence/absence of the following features: gene, TSS, core recognition element, −10 box, −35 box, −10/−35 spacer region, ribosome binding site (Shine-Dalgarno sequence), transcription factor binding site, and transcriptional attenuator. Because of the MUSCLE multiple sequence alignment, each alleleome base pair could be mapped to a reference strain base pair; then, any annotations corresponding to that reference strain base pair could be mapped to the aligned base pair(s) in the alleleome. All further analyses of these recognition elements were performed without any base-weighting based on the consensus motif (all bases considered equally). Furthermore, each non-coding regulatory region was annotated as essential or non-essential, with essential defined as any of the TSS in the non-coding regulatory region transcribing at least one gene annotated as essential in the Keio collection (42). Each non-coding regulatory region was also assigned a baseline expression level category of low, medium or high, based on the median of median expression levels across all genes transcribed from the region, using the PRECISE-1K definitions of the three categories (43). Finally, clusters of orthologous groups (COG) categories were assigned to each non-coding regulatory region based on the unique set of COGs assigned to genes transcribed from each region (a non-coding regulatory region could be assigned multiple COGs).

Clustering strains by non-coding alleles

The *linkage* function from the SciPy(44) hierarchical clustering package was used on a pairwise distance matrix between all 2350 strains, with non-default argument *method*=*'average'*. The pairwise distance was constructed by taking the complement of a similarity matrix, where the similarity between two strains was defined as the fraction of all 1169 non-coding regulatory regions for which the two strains had exactly the same allele. Then, flat clusters were computed using the Scipy *fcluster* function in *'maxclust'* mode. The optimal *'maxclust'* parameter was determined using a sensitivity analysis considering values from 2 through 25 inclusive and computing the mean silhouette score across all strains. This analysis selected 14 as the optimal number of clusters.

Assembling non-coding ALE mutations

Mutations acquired during ALE experiments were downloaded from ALEdb(45). This mutation data was cleaned as described in a previous publication(46). Unique mutations affecting the reference *E. coli* strain (K-12 MG1655, reference genome ID NC_000913.3) were selected. Additionally, only single nucleotide polymorphism (SNP) mutations were considered for this study. Then, only mutations annotated as occurring in non-coding regulatory regions were selected. These filtration steps yielded 1174 unique non-coding ALE-derived SNPs that could be mapped to reference strain positions. These mutations stem from multiple distinct ALE projects that are present in ALEdb.

Results

The *E. coli* non-coding alleleome captures variation across a broad range of strains and regulatory features

In order to construct the *E. coli* non-coding alleleome, we identified all nucleotide sequence variants (alleles) for each of 1169 well-annotated non-coding regulatory regions (from the reference strain K-12 MG1655) across 2350 complete-genome, wild-type (WT) *E. coli* strains (Figure 1A). We first amassed 2350 completely-sequenced *E. coli* strains from BV-BRC(36). The strains represented 14 distinct phylogroups as defined by ClermonTyping(37). The majority belonged to *E. coli sensu stricto* groups, while 109 strains come from more distantly related clades or the *fergusonii* and *albertii* groups (Figure 1B). The majority of strains with known hosts (67%) were isolated from humans, although other common domestic animals also provided strains (Figure 1C). Bodily excretions were the most common known sources of human-isolated strains.

These key non-coding regulatory regions capture 35.7% of the total non-coding positions in the reference strain and control the transcription of 1975 genes (Figure 1D). Importantly, these regions included majorities of key non-coding features,

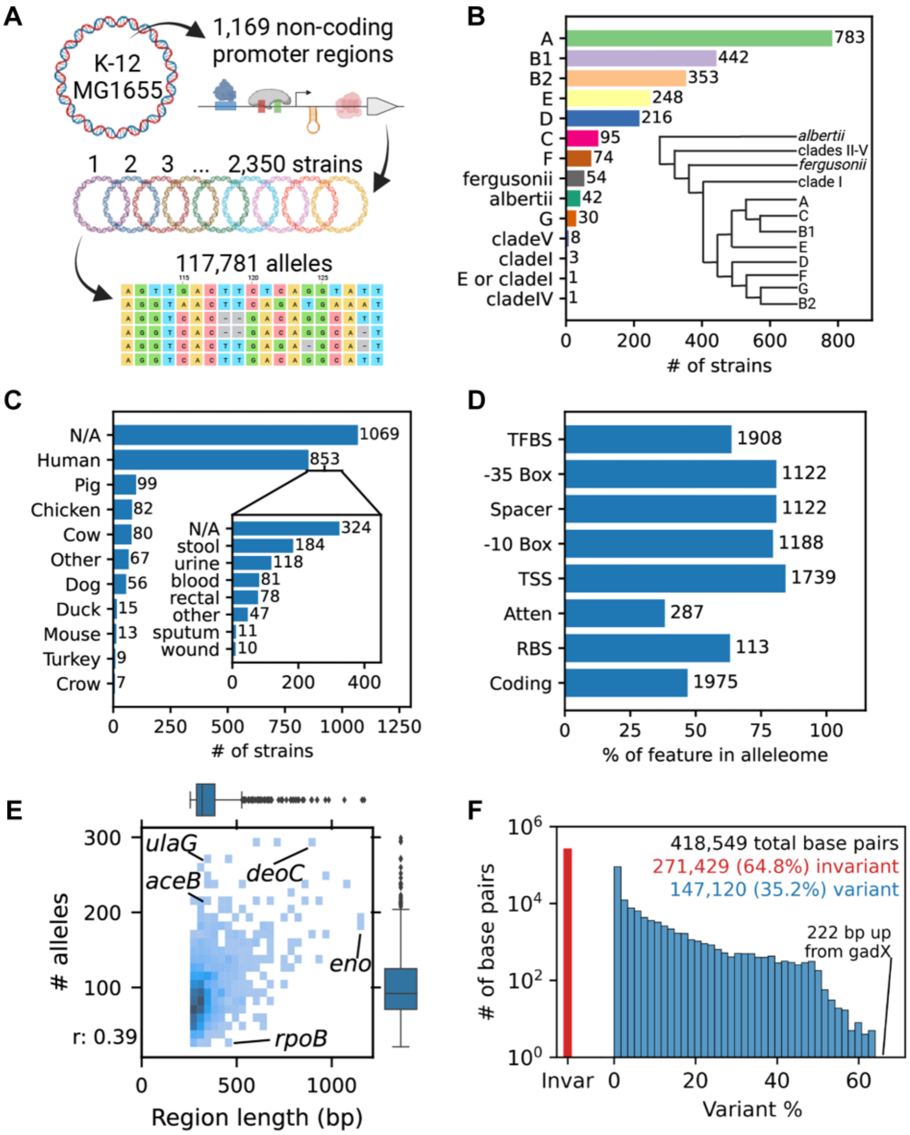


Figure 1. Constructing the *E. coli* non-coding alleleome. (**A**) Schematic representation of *E. coli* non-coding alleleome construction. 1169 non-coding promoter/5′ UTR regions from reference strain *E. coli* K-12 MG1655 were mapped across 2350 pangenome strains using BLAST (40). The resulting 117 781 alleles across all regions were aligned within each region using MUSCLE (41) to create the *E. coli* non-coding alleleome. (**B**) Strain counts for each of 14 *E. coli* categories assigned by ClermonTyping (37). Note: phylogenetic tree is not to scale. (**C**) Breakdown of *E. coli* strains by host common name (N/A indicates no host information). Inset indicates bodily fluid/tissue of origin for strains isolated from human hosts. (**D**) Counts and percentages of non-coding features from model strain K-12 MG1655 included in the non-coding alleleome. TFBS = transcription factor binding site, TSS = transcription start site, Atten = transcriptional attenuator, RBS = ribosome binding site. (**E**) 2-D histogram comparing length of aligned non-coding regulatory regions (n = 1169) to number of distinct alleles found for that region across the alleleome. r = Spearman's r. Note: the minimum possible allele length for a region is 250 bp; 50 bp downstream from gene start +200 bp upstream from TSS if TSS is at gene start (i.e. no 5′ UTR). (**F**) Histogram of variant percentages (i.e. percentage of non-dominant base pair) at each distinct aligned position in the *E. coli* alleleome. Blue histogram indicates variant % distribution for positions with non-zero variation. Red bar indicates the number of invariant base pairs.

including: 84% of TSS, 80% of core promoters, and 64% of TFBS. We then searched for these reference non-coding regulatory regions across the full set of *E. coli* strains, extracting homologous sequences from the expected local regions upstream of homologous genes in these other strains. Then, for each set of sequences corresponding to a reference region, we used multiple sequence alignment to determine the WT occurrence of every nucleotide (including indels) at every position.

In total, we identified 117781 distinct alleles across all regions and strains; these alleles comprise the *E. coli* non-coding alleleome. The median length of an aligned non-coding regulatory region was 319 base pairs, and the median region had 92 distinct alleles (Figure 1E). Region length and number of alleles were weakly correlated (0.39, Spearman). The promoter and 5′ UTR of *deoC*—encoding pyrimidine catabolism enzyme deoxyribose-phosphate aldolase—contained notable variation, with 294 distinct alleles in the 896-bp region. The upstream region of *eno* (encoding glycolysis and degradosome enzyme enolase), despite being the longest region considered at 1171 bp, had just 184 unique alleles. Much of *eno*'s upstream region overlaps with the upstream *pyrG* gene (transcribed in the same direction from distinct promoters), which may influence conservation in this multi-purpose region. A Cra-regulated promoter has also been reported upstream of *eno*(47). RNA polymerase core subunit gene *rpoB*'s region has just 32 distinct alleles despite a length of 442 bp, highlighting a level of conservation commensurate with this gene's essential role. Ascorbate degradation gene *ulaG* and glyoxylate cycle enzyme *aceB* featured particularly variable regulatory regions given their relatively short lengths. Overall, we assembled 418 549 aligned base pairs; of these, 65% are completely invariant (Figure 1F). The median variant percentage for variant positions was just 0.6%, highlighting an overall substantial level of sequence conservation across the non-coding alleleome. However, specific base pairs are particularly variant. The most variant base pair in the alleleome is found 222 base pairs upstream of the *gadX* gene start; an indel results in the dominant 'base' being a gap found in 34.4% of genomes, with A and G in 34.0% and 31.5% of cases, respectively.

Non-coding alleles recover phylogroups and highlight outliers

We next investigated the co-occurrence of non-coding alleles across strains. Hierarchical clustering of strains—with similarity defined as the fraction of shared alleles across the 1169 regions under consideration—yields 14 clusters, matching the number of groupings from ClermonTyping (Figure 2A). Overall, strains within the same group tend to be assigned to the same cluster (Figure 2B). Thus, as expected, non-coding alleles are more shared within strain groupings, and indeed are sufficient to discriminate between them in most cases. Interestingly, A, B1 and C—while most similar within their respective groups—are nonetheless similar enough with each other to be grouped together in Cluster 2. One strain identified by the phylogenetic method as ambiguous between phylogroup E and clade I clusters with all phylogroup E strains, again highlighting that non-coding alleles alone carry sufficient information to determine phylogroups.

Interestingly, clusters 4 and 5 contain single outlier strains from phylogroups B1 and A, respectively. A second B1 outlier strain appears in cluster 8 with most of the phylogroup D strains. Closer examination of the median pairwise distances of these two strains with all other B1 strains confirms that these strains indeed do share much fewer alleles than a typical pair of B1 strains (Figure 2C). We then further inspected B1 strain GF4-3 (the cluster 4 outlier) by identifying, for each of the non-coding regulatory regions, whether this strain's allele was completely unique within phylogroup B1 or shared with at least one other B1 strain. By analyzing the clusters of orthologous groups (COG) distributions of genes transcribed from the regions within these unique and shared groups, we identified the particular functional characteristics that contribute disproportionately to this strain's distinctiveness (Figure 2D). In particular, this strain has nearly four times more unique alleles related to secondary metabolite biosynthesis and nearly two times more related to energy metabolism. These unique characteristics may stem from this strain's host, the guineafowl (the only strain in this dataset isolated from this African bird).

aceB intergenic region provides a case study for analysis of sequence variation within functional sites

As a case study, we focused on a specific 331-bp non-coding regulatory region—the 5′ UTR and promoter region upstream of *aceB* (malate synthase A; a key enzyme in the glyoxylate cycle). We selected this region due to its relatively large number of alleles despite its short length. This region was identified in 2340 of 2350 strains, with the most common allele appearing in 20.2% (473/2340) of strains (Figure 3A). While some common alleles dominate, a variety of more niche alleles are also present. 90% of strains are accounted for by just 22% of the alleles; however, accounting for 99% of strains requires 90% of alleles. This region contains one transcription start site (TSS) with −10 and −35 elements, 11 TF binding sites of four distinct TFs, a transcriptional attenuator, and the very end of the next upstream gene, *metA* (homoserine *O*-succinyltransferase; catalyzes first step in methionine biosynthesis) (Figure 3B). Twenty significantly variant positions (those with variant base pairs present in at least 15% of strains) are mostly found upstream of the core promoter region, with a particular concentration in a specific IclR binding site. An additional 52% (173/331) of positions have minor variants, and 42% (141/331) are invariant.

Assessing variant presence in different genomic features provided a more detailed view of variation in this region. For example, while 33% of base pairs in this region are annotated as being part of at least one TF binding site, only 28% of all variant base pairs are found in TF binding sites (a factor of 0.15 fewer) (Figure 3C). Conversely, positions with no annotation accounted for 42% of the sequence but 49% of the variant base pairs. The core promoter elements—TSS, −10 and −35 elements—are particularly lacking in variants relative to their sequence exposures. No transcriptional terminators or ribosome binding sites are annotated for this region. While these observations hint at potential conservation patterns, one example non-coding regulatory region is insufficient to quantify systematic WT variation trends.

Aggregating non-coding alleles across the genome reveals conservation in functionally important regions

We repeated the *aceB* analysis across all 1169 non-coding regulatory regions and combined the results, revealing genome-wide trends in conservation within the non-coding alleleome (Figure 4A; Supplemental Table 2). On median, the top 16% most frequent alleles capture the sequence diversity of 90% of genomes, while a median of 76% of alleles are required to span 99% of genomes (Figure 4B). However, these distributions are quite broad - indeed, some regions are highly conserved, needing as few as 23% of alleles to cover 99% of genomes (region upstream of ribosomal protein *rpsM*). The most common allele covers a median of 33% of genomes; however, certain highly conserved regions—again including the region upstream of ribosomal protein *rpsM*—are covered almost entirely by a single dominant allele (Figure 4C). On the median, these regions are 65% invariant base pairs, 32% minor and 2% major variants (Figure 4D). There is notable variability across regions: for example, 31% of base pairs upstream of gluconeogenesis gene *pck* (encoding phosphoenolpyruvate carboxykinase) have major variation.

Most importantly, combining observations of variation across non-coding regulatory regions allows for assessment of conservation within annotated features (Supplemental Table 3). On median, non-coding base pairs without annotation vary just 3% more than expected based on their sequence

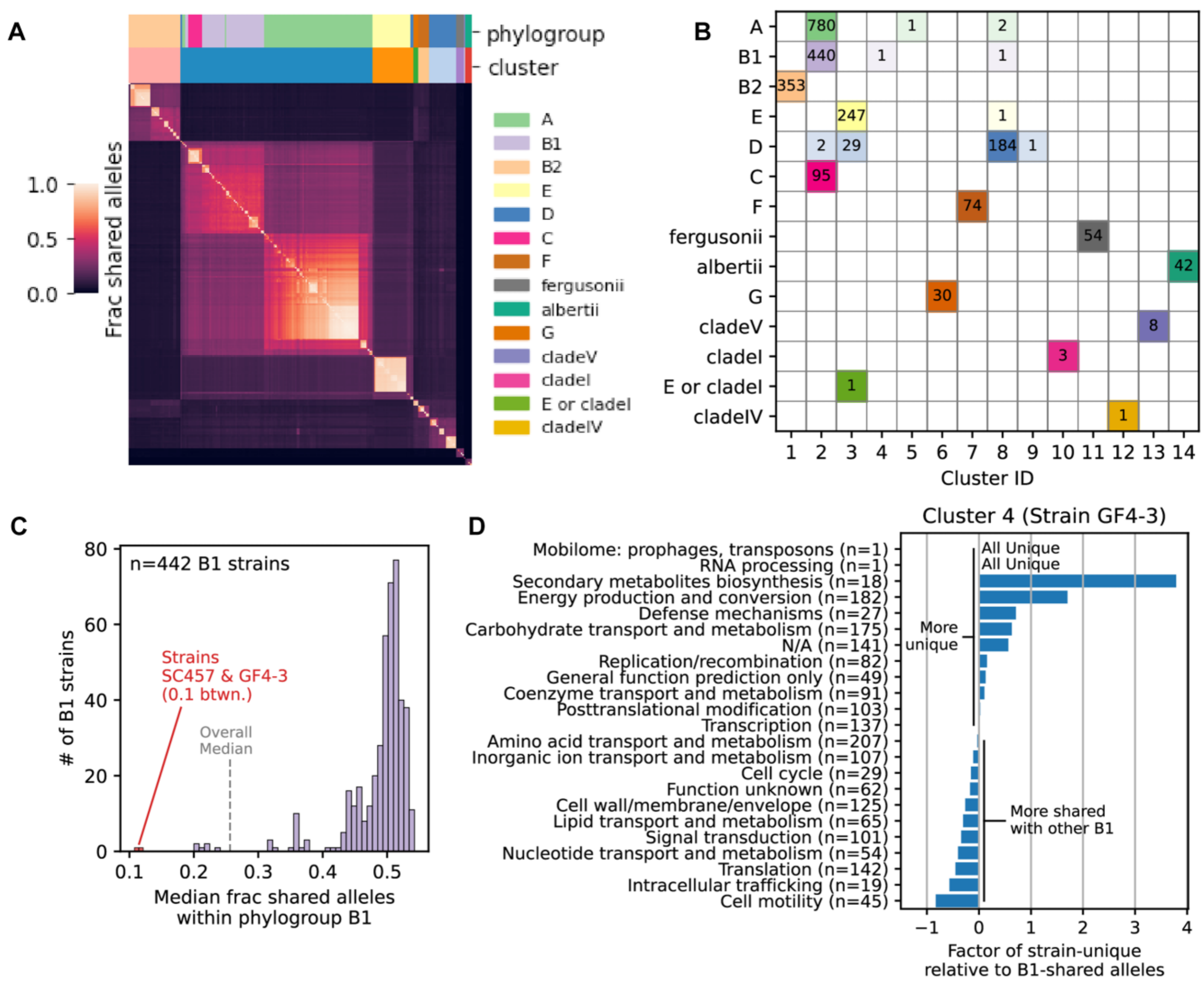


Figure 2. Non-coding allele clusters capture *E. coli* phylogroups. (**A**) Clustermap of 2350 *E. coli* strains, based on non-coding alleles. Heatmap displays a similarity matrix based on the fraction of shared alleles across 1169 distinct non-coding regulatory regions (i.e. a value of 1.0 indicates that strains have identical sequences at all 1169 non-coding regulatory regions. Upper colorbars indicate phylogroup/clade (determined separately using ClermonTyping (37); see Materials and Methods) and cluster assignment based on hierarchical clustering of distance matrix (1 − similarity matrix). Legend colors correspond to colors from the top colorbar row. (**B**) Cluster versus phylogroup/clade assignment matrix. Rows are phylogroups/clades, columns are clusters; e.g. all 30 strains in phylogroup G are in cluster 6, and cluster 6 contains no other strains. Phylogroup/clade color scheme same as panel A. $n = 2350$ total strains. (**C**) Histogram of pairwise similarity (defined as fraction of shared alleles) within phylogroup B1. Overall median = median pairwise similarity across all strain comparisons (i.e. median of all entries in heatmap from panel A. Strains GF4-3 and SC457 are outlier B1 strains, found in clusters 4 and 8 respectively. (**D**) Relative enrichment of fractions of clusters of orthologous groups (COGs) for genes transcribed by promoter alleles found uniquely in strain GF4-3 (cluster 4) versus by promoter alleles shared with at least one other B1 strain. For each COG, value is: (fraction of unique allele-transcribed gene COGs)/(fraction of shared allele-transcribed gene COGs) − 1; i.e. a value of 0 indicates that the COG comprises an equal fraction of the unique and shared sets. For example, the 'Defense mechanisms' COG is about 1.6× more represented in the unique set than the shared.

coverage, indicating minimal deviation from the background mutation rate. All non-coding features aside from attenuators vary significantly less than unannotated regions (Mann–Whitney U, FDR < 0.01). Ribosome binding sites and the core promoter elements (−10 and −35) are the most conserved sequences in non-coding regulatory regions. RBS, −10 elements, and −35 elements are on median 58%, 57% and 46% less variant than base pairs without functional annotation, respectively (Figure 4E). TF binding sites, the spacer between the core promoter elements, and the TSS and core recognition element (CRE) are all more conserved than unannotated regions (20%, 14%, 15% respectively). Coding regions included in this alleleome due to opposite strand overlap are just 6% less variant than unannotated non-coding base pairs, and only 3% less variant than expected based on sequence coverage. However, overlap with coding regions does significantly reduce variation in TF binding sites, spacers, CREs and attenuators (Supplemental Figure 1A). This effect is minimal when considering non-coding features that straddle coding region boundaries, suggesting that variation within any given feature is selected relatively uniformly (Supplemental Figure 1B).

Additionally, conservation in upstream non-coding regulatory regions relates to the functions of their gene products. Regions expressing genes in clusters of orthologous groups (COG) categories such as 'Translation, ribosomal structure and biogenesis' and 'Replication, recombination and repair' are amongst the most conserved (Supplemental Figure 2). Conversely, metabolic COGs appear to support more

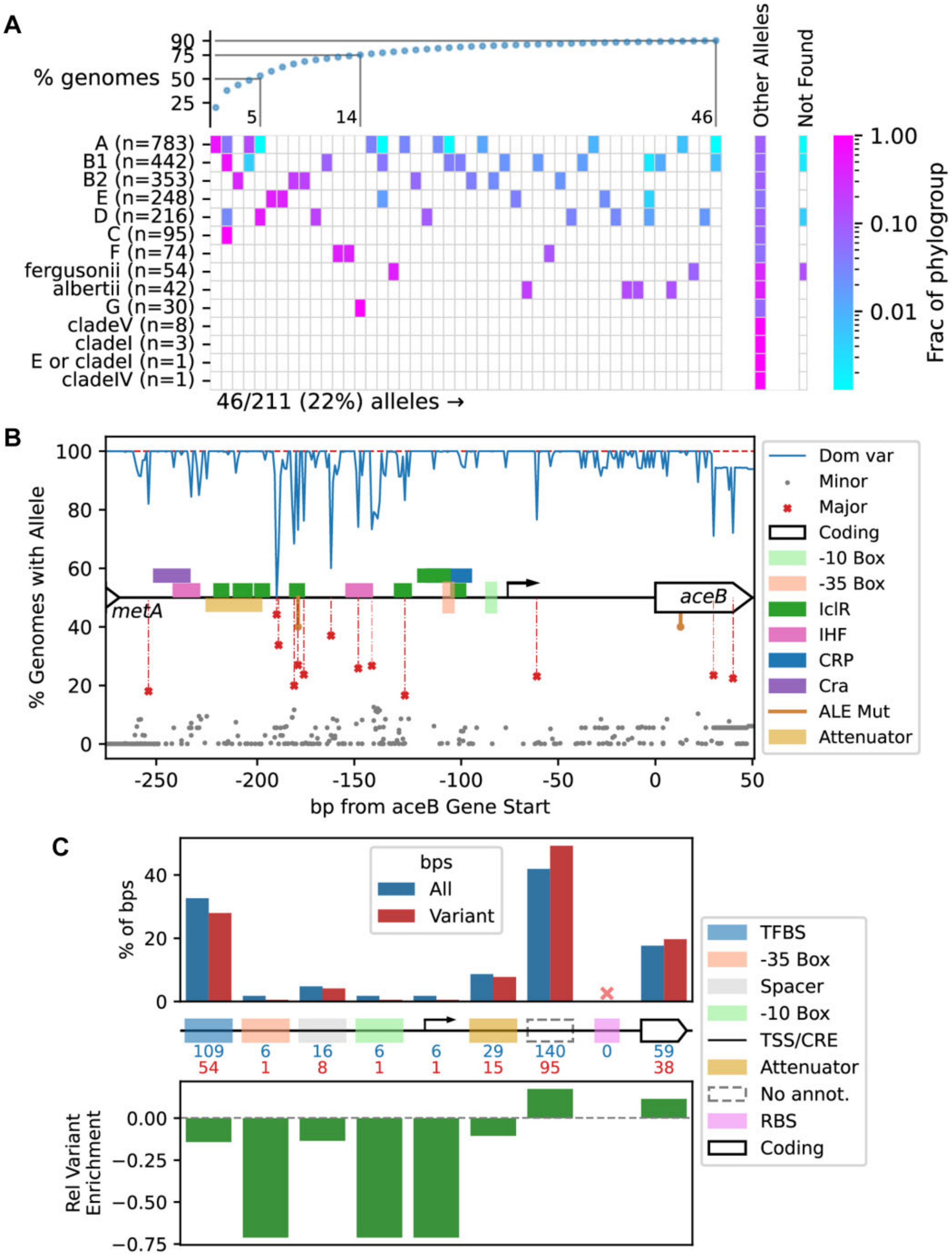


Figure 3. Alleleome for a single intergenic region; a case study. Statistics for a single region from the non-coding alleleome, comprising the 5′ UTR and promoter region for *aceB* (malate synthase A). (**A**) Heatmap of phylogroups/clades versus common alleles for this region. Colorbar is scaled per row (phylogroup); e.g. hot pink in phylogroup row G, allele column 14 indicates all G strains (fraction 1.0) have this allele. Alleles are sorted left-to-right in decreasing order of fraction of strains with allele. Scatterplot above heatmap indicates cumulative fraction of strains covered by corresponding number of alleles. Not found indicates fraction of each phylogroup/clade for which this region was not found at all (no allele). (**B**) Depiction of sequence features from reference strain (K-12 MG1655) in this 331-bp non-coding regulatory region (central track). Dom var (blue line): frequency of most common base; minor (gray dot): position and frequency of any $<$ 15% variants; major (red cross and line to help locate in region); position and frequency of $>$15% variants. (**C**) Breakdown of base pairs in this region based on presence in annotated promoter features. Blue bars indicate % of all 331 base pairs in the region that are annotated with each category. Red bars indicate % of 193 variant base pairs belonging to each category. Note that due to overlaps, percentages add up to larger than 100%. Blue and red numbers below the schematic indicate numbers of base pairs in each category. Red X in the bar plot indicates features with no presence in this region. Green bars indicate relative enrichment of variant base pairs in each category: (% variant bps in category / % all bps in category) / % all bps in category. For example, 109/331 (33%) of all bps are in a TF binding site, along with 54/193 (28%) of variant bps: (28% − 33%)/33% = −0.15.

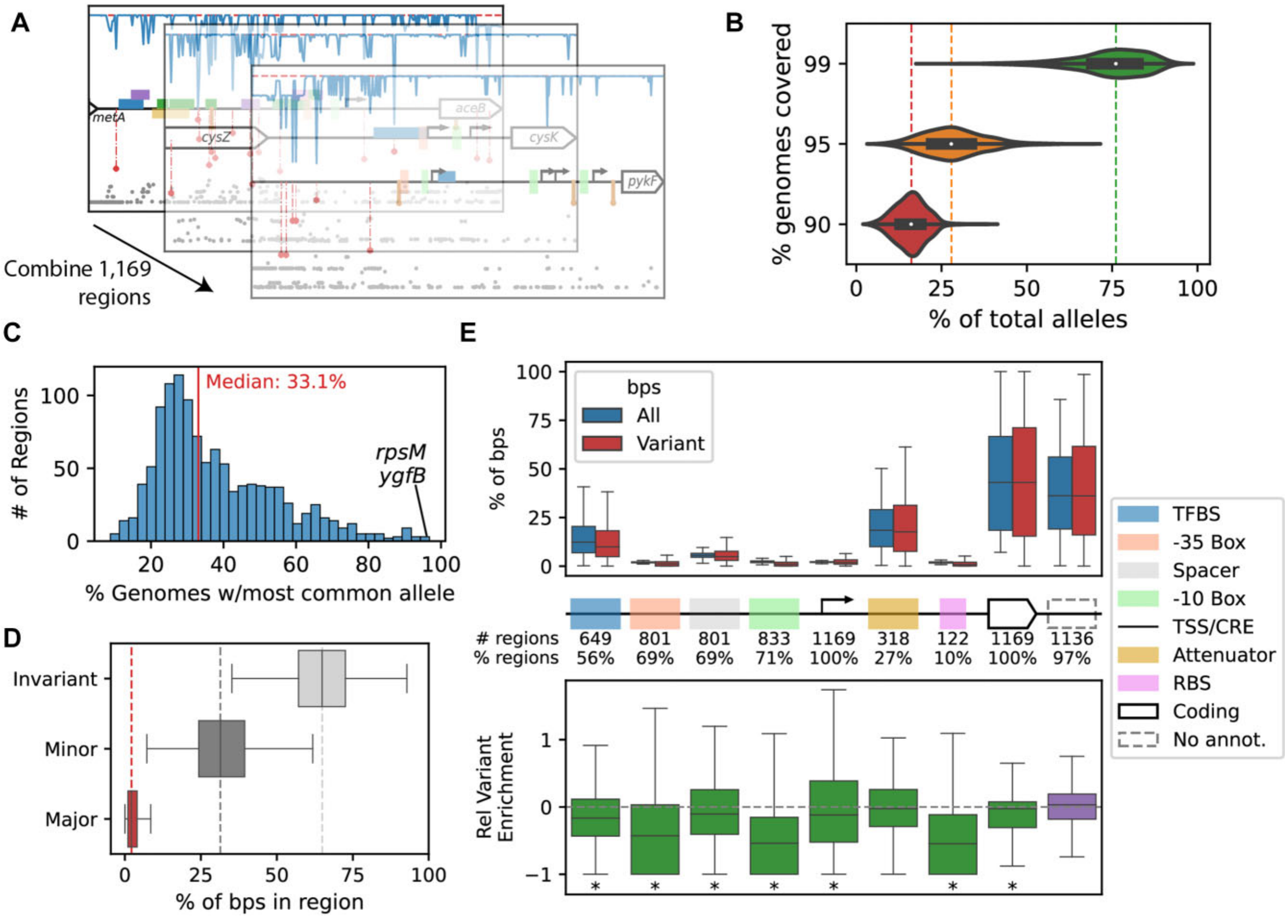


Figure 4. Summary statistics of sequence variation for all 1 169 non-coding regulatory regions' alleles. (**A**) All summary statistics for each of 1169 non-coding regulatory regions (represented by dashboards from Figure 3) were aggregated to investigate whole alleleome properties. (**B**) Violin plot showing distributions of allele percentages needed to cover 90/95/99% of genomes for a given non-coding regulatory region. (**C**) Histogram of % of strains covered by the most common allele; e.g. the non-coding regulatory regions upstream of *rpsM* and *ygfB* are identical in over 95% of strains in which the region was found. (**D**) Box plot showing distributions of variant types across non-coding regulatory regions. (**E**) Distributions of sequence and variant base pair percentages across annotation categories (see Figure 3C). #/% regions = number/percentage of regions that have at least one base pair annotated with the indicated category (e.g. 649/1169 (56%) of non-coding regulatory regions have at least one TF binding site). Asterisks indicate significant difference with purple (no annotation) regions (Mann–Whitney *U* test, FDR controlled at 0.01).

non-coding variation, such as amino acid metabolism and secondary metabolite biosynthesis. non-coding regulatory regions transcribing at least one essential gene are significantly more conserved than those that do not transcribe any essential genes (Supplemental Figure 3A). These essential-transcribing regions also have significantly more conserved transcription factor binding sites and promoters (Supplemental Figure 3B). The effect size is largest for the −10 and −35 elements of the promoter, highlighting the expected significance of these sigma factor binding regions. Similarly, non-coding regulatory regions also differ significantly in conservation depending on their baseline expression level (Supplemental Figure 3C). As with essential-transcribing regions, this conservation is also prevalent in the most critical promoter regions (Supplemental Figure 3D). Furthermore, non-coding regulatory regions have significantly different conservation depending on their locations relative to DNA replication (Supplemental Figure 4). In particular, regions located in the terminus region (opposite the origin of replication) are less conserved than regions on the leading or lagging strands. This effect may relate to the enrichment of highly-expressed, more conserved genes on the leading strand (48).

Transcription factor binding sites exhibit significant variation in conservation

The non-coding alleleome enables a detailed investigation of conservation within transcription factor binding sites. We identified 22 major transcription factors that have at least 10 binding sites and whose activity explains notable variation within the PRECISE-1K expression compendium. The median percentage of invariant base pairs within the binding sites of these TFs is not significantly correlated with the percentage of expression variation explained (Figure 5A). Most of these transcription factors have binding sites with a wide range of conservation (Figure 5B). Central carbon metabolism regulator Cra's binding sites are the most conserved, with a median of 84% invariance. Nucleotide metabolism regulator PurR's binding sites are consistently conserved; only one site falls <70% conservation. A subset of these TFs are further identifiable as dual regulators, with at least 10 binding sites annotated for both activation and repression roles. Mostly, this distinction does not result in a difference in binding site conservation (Figure 5C). However, the TF and nucleoid-associated protein IHF has significantly more conserved repressor sites than activator. IHF is known to be able to bind to DNA in

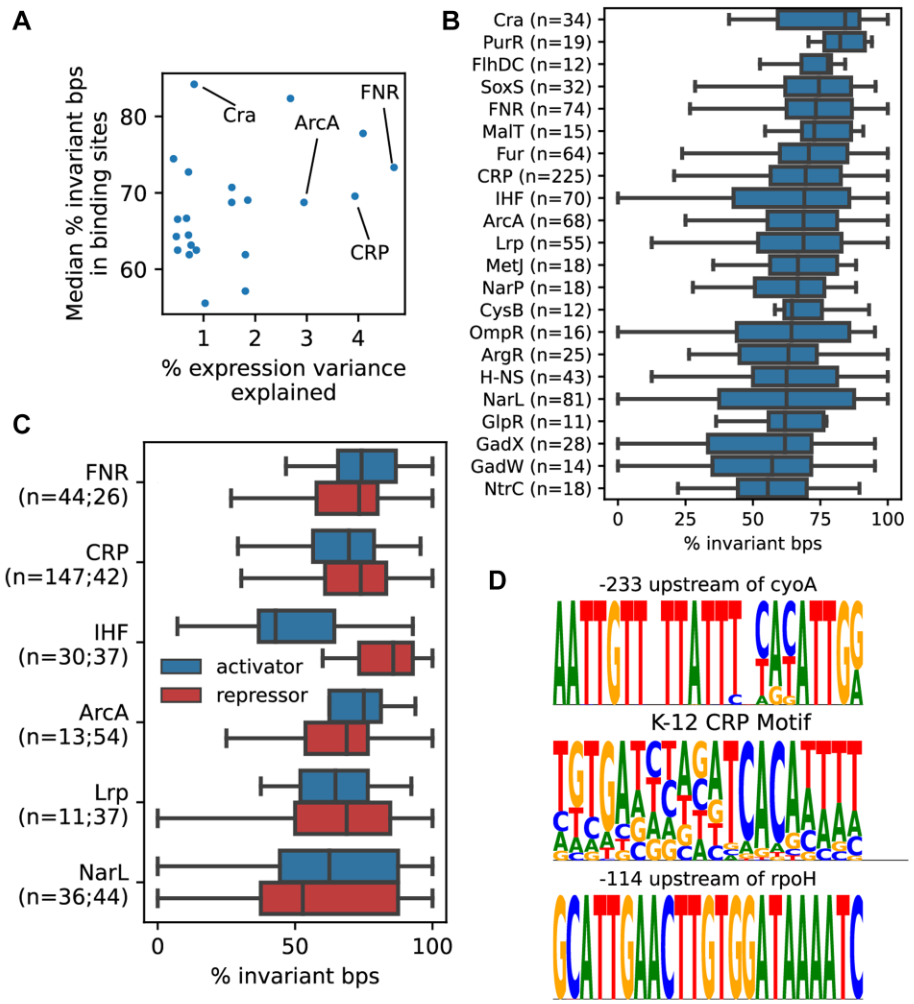


Figure 5. Transcription factor binding sites exhibit a wide range of conservation. (**A**) Comparison of expression variance explained to median percentage of invariant base pairs within binding sites for 22 major transcription factors (min 10 binding sites). Explained variance % is computed based on percentage of expression variance in PRECISE-1K expression compendium(43) explained by TF's iModulon (a gene grouping capturing the independent effect of the regulator). (**B**) Distributions of invariant base pair % for binding sites of 22 major transcription factors. (**C**) Distributions of invariant base pair % for major regulators with dual regulatory effect (min 10 annotated binding sites for each mode of regulation; individual sites annotated as 'dual' removed; effect data from RegulonDB(21)). (**D**) Example sequence logos for poorly conserved (top) and highly conserved (bottom) CRP binding sites. CRP motif from reference strain K-12 MG1655 (from RegulonDB) is shown in middle. Note: all base pairs in *cyoA* site have variants; some variants are too small to visualize. Note: position 1 in *rpoH* site has 7/2350 variants.

a non-specific manner and may even be redundant with AT-rich upstream regions in some cases(49). Example CRP binding sites highlight the range of conservations observed within binding site sequences (Figure 5D). A binding site upstream of *cyoA* has no completely conserved base pairs, while a CRP binding site regulating *rpoH* expression has only one position with any variation. Interestingly, the binding site upstream of *cyoA* exhibits particular variation across the alleleome in a relatively high-information region of the reference strain CRP motif, possibly indicating a functional impact of these variants on CRP activity.

ALE mutations are more likely than natural variants to impact functionally-relevant features

In contrast with natural variants, adaptive laboratory evolution (ALE) exerts selective pressure for cells to adapt to a particular stress or growth mode. A set of ALE mutations affecting non-coding regulatory regions was acquired from ALEdb (45), an online resource cataloging such experiments in *E. coli*. ALE's preference for high-impact mutations becomes clear when comparing the rates of ALE mutations in particular non-coding regulatory regions to wild-type variant rates. For example, ALE mutations are 75% more likely to occur in TF binding sites than these base pairs' sequence exposure would suggest (Figure 6A). Core promoter elements also exhibit this effect; −10 and −35 boxes are mutated 58% and 35% more often than expected. Not only are ALE mutations enriched in −10 boxes, but the mutations have a slight tendency to reduce the GC content of these regions (Figure 6B). The −10 box is typically the upstream location of DNA strand unwinding for transcriptional bubble formation upon RNA polymerase binding; thus, decreased GC content is likely to increase transcription at these sites. Wild-type variants at −10 boxes do not tend to alter GC content on median (Figure 6C).

Discussion

Here, we present a non-coding alleleome for *E. coli*, providing a deep look at variation in critical transcriptional

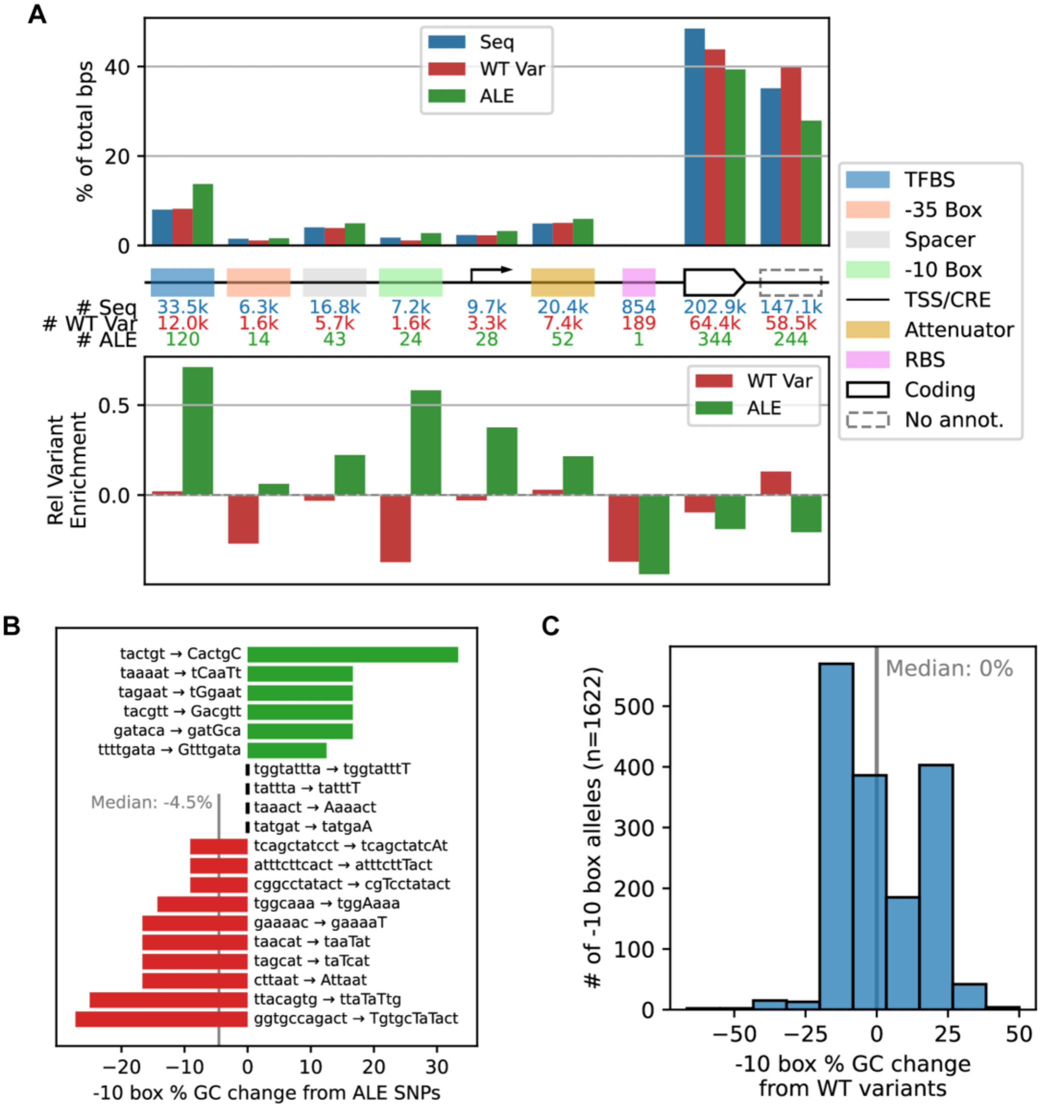


Figure 6. Adaptive laboratory evolution (ALE) mutations are over-represented in wild type-conserved non-coding regulatory regions. (**A**) Breakdown of percentages of all non-coding base pairs considered by annotation category. In upper panel, blue bars represent % of all base pairs annotated within each category; red bars represent % of all variant base pairs annotated within each category; green bars indicate percentage of all 1174 non-coding ALE mutations present in each category. (**B**) Effect of 20 ALE SNPs affecting −10 boxes on GC content. (**C**) Effect of 1622 distinct wild-type −10 box variant alleles on GC content relative to consensus −10 box sequence for that region.

and translational control regions. We assemble 2350 complete genomes across the *E. coli* phylogenetic tree and identify alleles for 1169 reference non-coding regulatory regions across this set of strains. We cluster strains based on their non-coding alleles, finding these to be largely sufficient for distinguishing phylogroups. Centrally, we find that 64% of positions are completely conserved, and variation at the remaining positions is just a median of 0.6%. As hypothesized, core promoter features and binding sites are more conserved than non-functional positions. We also show that essentiality and high expression drive significant conservation, again concentrated in functionally critical promoter features. The alleleome also provides a rich understanding of conservation across transcription factor binding sites, highlighting significant variation both between and within different regulators' binding sequences. Finally, we contrast wild-type variation with mutations acquired during ALE, determining that ALE mutations preferentially alter regions that natural variants conserve.

This analysis expands our understanding of natural sequence variation beyond coding regions. While 5′ UTR and promoter regions constitute a minor fraction of the genome sequence, they encode the critical control functions that enable *E. coli* to adapt its transcriptome in response to environmental signals. Our understanding of precisely how these sequences influence expression *in vivo* - as opposed to via synthetic promoter libraries - remains limited. This non-coding alleleome provides a new dimension with which the function of these regions may be further elucidated. For example, models that aim to predict expression level directly from promoter sequence may benefit from understanding how conserved each base pair in the promoter region is; a similar approach is important for the function of AlphaFold (50).

The identification of two unique strains whose non-coding alleles do not cluster notably with any other phylogroups highlights a potential bias in complete *E. coli* genome sequences currently available. The *E. coli* strain GF4-3, isolated from a guineafowl, harbors distinct non-coding alleles from any other strain observed in this study. *E. coli* sequence diversity may be significantly more rich than we currently realize due to over-representation of strains isolated from a handful of host organisms.

E. coli's genome is largely dominated by coding genes; in the reference strain K-12 MG1655, 87% of base pairs are part of a coding gene (10). As a result, many positions within the non-coding alleleome are technically also within coding regions. This situation may arise due to promoter regions found within operons, such that a promoter overlaps with the nearest upstream gene; or strand differences, where a gene encoded on one strand is directly opposite a promoter region on the other. Divergent promoters—where a relatively small promoter region is shared between two genes transcribed in opposite directions—are another overlap case. Indeed, these divergent promoters are a common regulatory motif across bacteria(51) that enable control of divergent operons, including for global regulators such as CRP (52). Our analysis indicates that, in general, coding positions vary at the expected rate based on their sequence coverage, implying primacy of coding functionality in these overlapping cases. However, further study is needed to determine whether the coding or non-coding functions encoded in these regions are driving conservation patterns.

The non-coding alleleome's quantification of variation in transcription factor binding sites provides an opportunity for expansion of binding motif definition. Motifs aim to summarize the specific sequence required for binding of a transcription factor to DNA by combining the sequences of experimentally-determined binding sites and indicating the probability of finding each base at each position. Motifs are typically generated by combining binding sites controlling different transcription units within the same strain. Frequently, real observed binding site sequences differ significantly from a canonical motif. Thus, alleleome variation within the same binding site may provide an alternative information source for assessment of binding site sequence importance by allowing comparison of alternative sequences within a more similar sequence and functional context. Any time a new experimental binding site is identified, alleleome variation within the proposed site can be assessed to provide context for the likely strength or importance of the site. Furthermore, this alleleome may provide an opportunity to suggest novel transcription factor binding sites. However, because we do also observe significant variation in conservation across binding sites, this approach may only provide one piece of information as part of a larger picture.

Overall, this *E. coli* non-coding alleleome quantifies base pair-level variation and conservation at genome- and species-scale. The data generated in this study provides a rich resource for analyzing non-coding regulatory regions in any *E. coli* genome. We believe that this type of analysis should be expanded to other organisms to enable comparative non-coding alleleomics. As sequence data continues to balloon, this study provides a blueprint for compiling, quantifying, and analyzing non-coding variation, revealing patterns of conservation and their relationship to phenotypic outcomes.

Data availability

All data and code for analysis and figures are available on GitHub at https://github.com/SBRG/noncoding_alleleome and Zenodo (https://zenodo.org/doi/10.5281/zenodo.10976515).

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

Conceptualization, B.O.P., C.R.L. and D.C.Z.; Methodology, C.R.L., D.C.Z., P.V.P.; Software, C.R.L., P.V.P.; Investigation, C.R.L.; Writing - Original Draft, C.R.L.; Writing - Review & Editing, C.R.L., D.C.Z. and B.O.P.; Visualization, C.R.L.; Supervision, D.C.V. and B.O.P.; Funding Acquisition, B.O.P.

Funding

Novo Nordisk Foundation [NNF20CC0035580]. Funding for open access charge: Novo Nordisk Fonden [NNF20CC0035580].

Conflict of Interest statement

None declared.

References

1. Giani,A.M., Gallo,G.R., Gianfranceschi,L. and Formenti,G. (2020) Long walk to genomics: history and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.*, **18**, 9–19.
2. Deng,X., den Bakker,H.C. and Hendriksen,R.S. (2016) Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu. Rev. Food Sci. Technol.*, **7**, 353–374.
3. Thomsen,M.C.F., Ahrenfeldt,J., Cisneros,J.L.B., Jurtz,V., Larsen,M.V., Hasman,H., Aarestrup,F.M. and Lund,O. (2016) A bacterial analysis platform: an integrated system for analysing bacterial whole genome sequencing data for clinical diagnostics and surveillance. *PLoS One*, **11**, e0157718.
4. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.
5. Rouli,L., Merhej,V., Fournier,P.-E. and Raoult,D. (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect*, **7**, 72–85.
6. Tettelin,H., Riley,D., Cattuto,C. and Medini,D. (2008) Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, **11**, 472–477.
7. Wood,S., Zhu,K., Surujon,D., Rosconi,F., Ortiz-Marquez,J.C. and van Opijnen,T. (2020) A pangenomic perspective on the emergence, maintenance, and predictability of antibiotic resistance. In: Tettelin,H. and Medini,D. (eds.) *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer, Cham.
8. Kim,Y., Gu,C., Kim,H.U. and Lee,S.Y. (2020) Current status of pan-genome analysis for pathogenic bacteria. *Curr. Opin. Biotechnol.*, **63**, 54–62.
9. Norsigian,C.J., Fang,X., Palsson,B.O. and Monk,J.M. (2020) Pangenome flux balance analysis toward panphenomes. In: Tettelin,H. and Medini,D. (eds.) *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer, Cham.
10. Lamoureux,C.R., Choudhary,K.S., King,Z.A., Sandberg,T.E., Gao,Y., Sastry,A.V., Phaneuf,P.V., Choe,D., Cho,B.-K. and Palsson,B.O. (2020) The Bitome: digitized genomic features reveal fundamental genome organization. *Nucleic Acids Res.*, **48**, 10157–10163.
11. Mulligan,M.E., Hawley,D.K., Entriken,R. and McClure,W.R. (1984) Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic Acids Res.*, **12**, 789–800.
12. Hawley,D.K. and McClure,W.R. (1983) Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res.*, **11**, 2237–2255.

13. Collado-Vides,J., Magasanik,B. and Gralla,J.D. (1991) Control site location and transcriptional regulation in Escherichia coli. *Microbiol. Rev.*, **55**, 371–394.
14. Chen,L.H., Emory,S.A., Bricker,A.L., Bouvet,P. and Belasco,J.G. (1991) Structure and function of a bacterial mRNA stabilizer: analysis of the 5' untranslated region of ompA mRNA. *J. Bacteriol.*, **173**, 4578–4586.
15. Yamanaka,K., Mitta,M. and Inouye,M. (1999) Mutation analysis of the 5' untranslated region of the cold shock cspA mRNA of Escherichia coli. *J. Bacteriol.*, **181**, 6284–6291.
16. Pribnow,D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 784–788.
17. Mejía-Almonte,C., Busby,S.J.W., Wade,J.T., van Helden,J., Arkin,A.P., Stormo,G.D., Eilbeck,K., Palsson,B.O., Galagan,J.E. and Collado-Vides,J. (2020) Redefining fundamental concepts of transcription initiation in bacteria. *Nat. Rev. Genet.*, **21**, 699–714.
18. Helmann,J.D. (2019) Where to begin? Sigma factors and the selectivity of transcription initiation in bacteria. *Mol. Microbiol.*, **112**, 335–347.
19. Browning,D.F. and Busby,S.J. (2004) The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.*, **2**, 57–65.
20. Mendoza-Vargas,A., Olvera,L., Olvera,M., Grande,R., Vega-Alvarado,L., Taboada,B., Jimenez-Jacinto,V., Salgado,H., Juárez,K., Contreras-Moreira,B., *et al.* (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One*, **4**, e7526.
21. Tierrafría,V.H., Rioualen,C., Salgado,H., Lara,P., Gama-Castro,S., Lally,P., Gómez-Romero,L., Peña-Loredo,P., López-Almazo,A.G., Alarcón-Carranza,G., *et al.* (2022) RegulonDB 11.0: Comprehensive High-throughput Datasets on Transcriptional Regulation in Escherichia coli K-12. *Microb. Genom.*, **8**, mgen000833.
22. Zheng,M. and Storz,G. (2000) Redox sensing by prokaryotic transcription factors. *Biochem. Pharmacol.*, **59**, 1–6.
23. Landis,L., Xu,J. and Johnson,R.C. (1999) The cAMP receptor protein CRP can function as an osmoregulator of transcription in Escherichia coli. *Genes Dev.*, **13**, 3081–3091.
24. Mukhopadhyay,P., Zheng,M., Bedzyk,L.A., LaRossa,R.A. and Storz,G. (2004) Prominent roles of the NorR and Fur regulators in the Escherichia coli transcriptional response to reactive nitrogen species. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 745–750.
25. Gollnick,P. and Babitzke,P. (2002) Transcription attenuation. *Biochim. Biophys. Acta*, **1577**, 240–250.
26. Thorpe,H.A., Bayliss,S.C., Sheppard,S.K. and Feil,E.J. (2018) Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience*, **7**, 1–11.
27. Catoiu,E.A., Phaneuf,P., Monk,J. and Palsson,B.O. (2023) Whole-genome sequences from wild-type and laboratory-evolved strains define the alleleome and establish its hallmarks. *Proc. Natl. Acad. Sci. U.S.A.*, **120**, e2218835120.
28. Phaneuf,P.V., Jarczynska,Z.D., Kandasamy,V., Chauhan,S.M., Feist,A.M. and Palsson,B.O. (2023) Using the E. coli alleleome in strain design. bioRxiv doi: https://doi.org/10.1101/2023.09.17.558058, 17 September 2023, preprint: not peer reviewed.
29. Harke,A.S., Josephs-Spauling,J., Mohite,O.S., Chauhan,S.M., Ardalani,O., Palsson,B. and Phaneuf,P.V. (2023) Genomic insights into lactobacillaceae: analyzing the 'alleleome' of core pangenomes for enhanced understanding of strain diversity and revealing phylogroup-specific unique variants. bioRxiv doi: https://doi.org/10.1101/2023.09.22.558971, 22 September 2023, preprint: not peer reviewed.
30. Kimura,M. (1983) In: *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
31. Drake,J.W. (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 7160–7164.
32. Wielgoss,S., Barrick,J.E., Tenaillon,O., Cruveiller,S., Chane-Woon-Ming,B., Médigue,C., Lenski,R.E. and Schneider,D. (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with Escherichia coli. *G3*, **1**, 183–186.
33. Nichols,B.P. and Yanofsky,C. (1979) Nucleotide sequences of trpA of Salmonella typhimurium and Escherichia coli: an evolutionary comparison. *Proc. Natl. Acad. Sci. U.S.A.*, **76**, 5244–5248.
34. Adelberg,E.A. and Burns,S.N. (1960) Genetic variation in the sex factor of Escherichia coli. *J. Bacteriol.*, **79**, 321–330.
35. Harshman,L. and Riley,M. (1980) Conservation and variation of nucleotide sequences in Escherichia coli strains isolated from nature. *J. Bacteriol.*, **144**, 560–568.
36. Olson,R.D., Assaf,R., Brettin,T., Conrad,N., Cucinell,C., Davis,J.J., Dempsey,D.M., Dickerman,A., Dietrich,E.M., Kenyon,R.W., *et al.* (2023) Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.*, **51**, D678–D689.
37. Beghain,J., Bridier-Nahmias,A., Le Nagard,H., Denamur,E. and Clermont,O. (2018) ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb. Genom.*, **4**, e000192.
38. Hyun,J.C., Monk,J.M. and Palsson,B.O. (2022) Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *Bmc Genomics [Electronic Resource]*, **23**, 7.
39. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
40. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf.*, **10**, 421.
41. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinf.*, **5**, 113.
42. Baba,T., Ara,T., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L. and Mori,H. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, **2**, 2006.0008.
43. Lamoureux,C.R., Decker,K.T., Sastry,A.V., Rychel,K., Gao,Y., McConn,J.L., Zielinski,D.C. and Palsson,B.O. (2023) A multi-scale expression and regulation knowledge base for Escherichia coli. *Nucleic Acids Res.*, **51**, 10176–10193.
44. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J., *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
45. Phaneuf,P.V., Gosting,D., Palsson,B.O. and Feist,A.M. (2019) ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. *Nucleic Acids Res.*, **47**, D1164–D1171.
46. Phaneuf,P.V., Zielinski,D.C., Yurkovich,J.T., Johnsen,J., Szubin,R., Yang,L., Kim,S.H., Schulz,S., Wu,M., Dalldorf,C., *et al.* (2021) Escherichia coli data-driven strain design using aggregated adaptive Laboratory evolution mutational data. *ACS Synth. Biol.*, **10**, 3379–3395.
47. Shimada,T., Fujita,N., Maeda,M. and Ishihama,A. (2005) Systematic search for the Cra-binding promoters using genomic SELEX system. *Genes Cells*, **10**, 907–918.
48. Price,M.N., Alm,E.J. and Arkin,A.P. (2005) Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. *Nucleic Acids Res.*, **33**, 3224–3234.
49. Yoshua,S.B., Watson,G.D., Howard,J.A.L., Velasco-Berrelleza,V., Leake,M.C. and Noy,A. (2021) Integration host factor bends and bridges DNA in a multiplicity of binding modes with varying specificity. *Nucleic Acids Res.*, **49**, 8684–8698.
50. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A.,

Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

51. Warman,E.A., Forrest,D., Guest,T., Haycocks,J.J.R.J., Wade,J.T. and Grainger,D.C. (2021) Widespread divergent transcription from bacterial and archaeal promoters is a consequence of DNA-sequence symmetry. *Nat. Microbiol.*, **6**, 746–756.
52. Shimada,T., Fujita,N., Yamamoto,K. and Ishihama,A. (2011) Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. *PLoS One*, **6**, e20081.

Received: October 31, 2023. **Revised:** April 5, 2024. **Editorial Decision:** April 8, 2024. **Accepted:** May 15, 2024