Title

Whole-genome sequences from wild-type and laboratory-evolved strains define the alleleome and establish its hallmarks

Permalink

https://escholarship.org/uc/item/5jw6q1x4

Journal

Proceedings of the National Academy of Sciences of the United States of America, 120(15)

ISSN

0027-8424

Authors

Catoiu, Edward Alexander
Phaneuf, Patrick
Monk, Jonathan
et al.

Publication Date

2023-04-11

DOI

10.1073/pnas.2218835120

Peer reviewed

PNAS RESEARCH ARTICLE | MICROBIOLOGY OPEN ACCESS

Whole-genome sequences from wild-type and laboratory-evolved strains define the alleleome and establish its hallmarks

Edward Alexander Catoiu[a], Patrick Phaneuf[b], Jonathan Monk[c], and Bernhard O. Palsson[a,b,1]

Edited by Jens Nielsen, BioInnovation Institute, Copenhagen, Denmark; received November 4, 2022; accepted March 6, 2023

The genomic diversity across strains of a species forms the genetic basis for differences in their behavior. A large-scale assessment of sequence variation has been made possible by the growing availability of strain-specific whole-genome sequences (WGS) and with the advent of large-scale databases of laboratory-acquired mutations. We define the *Escherichia coli* "alleleome" through a genome-scale assessment of amino acid (AA) sequence diversity in open reading frames across 2,661 WGS from wild-type strains. We observe a highly conserved alleleome enriched in mutations unlikely to affect protein function. In contrast, 33,000 mutations acquired in laboratory evolution experiments result in more severe AA substitutions that are rarely achieved by natural selection. Large-scale assessment of the alleleome establishes a method for the quantification of bacterial allelic diversity, reveals opportunities for synthetic biology to explore novel sequence space, and offers insights into the constraints governing evolution.

allele | mutation | laboratory evolution | sequence | alignment

In the late 2000s, DNA sequencing costs dramatically decreased. Throughout the 2010s, inexpensive sequencing led to a steady increase in the number of publicly available sequenced genomes in strains of a bacterial species (1–5). Thus, sequence variation among bacterial strains can now be studied at an unprecedented scale. The first forays into the study of bacterial sequence variation led to the development of phylotyping, representing the grouping of bacterial strains into "clades" based on the presence of a select set of "housekeeping" genes (6–9). Such phylotyping can identify the environmental origin, the evolutionary lineage, and the potential pathogenicity of a given strain (10–12). More recently, sequence variants within the same gene ("alleles") have been shown to affect niche phenotypes such as bacterial cell–host adhesion and interaction (13–19). Given the recent availability of whole-genome sequences (WGS), a full definition and characterization of the open reading frame ("ORF alleleome")—the collection of every allele for all the genes in an organism—is now possible.

Concurrent with the increase of publicly available sequenced genomes of wild-type (WT) strains, laboratory evolution has emerged as a new approach to address biological questions and develop new phenotypic traits (20–24). Bacterial strains have been evolved in a variety of different laboratory environments, and a large number of laboratory-acquired mutations are now found in databases (25, 26). Ongoing since 1988, the *Escherichia coli* long-term evolution experiment (LTEE) has produced more than 10,000 unique mutations in over 70,000 generations grown in a consistent medium (27–30). Conversely, adaptive laboratory evolution (ALE)—numerous short-term evolutions in response to different selection pressures—has produced more than 45,000 unique mutations in *E. coli* strains (31–37). Thus, the availability of thousands of fully sequenced genomes and laboratory-acquired mutations allows for the detailed genome-scale comparison between the sequence variation in WT strains of a species and the mutations fixed in laboratory strains of the same species.

Using a collection of 2,661 fully sequenced WT strains belonging to various phylogroups, isolated from various hosts and geographic regions (*SI Appendix*, Fig. S1 and Dataset S1), this work establishes a unique method to quantify the intragenic natural sequence variation of the *E. coli* "alleleome" at the genome scale. We find a surprisingly limited diversity to be the hallmark characteristic of the ORF alleleome: variation is found in relatively few codon positions in the *E. coli* genome and is limited to a few alternate amino acid (AA) substitutions unlikely to affect protein function. Against this limited diversity, we show that laboratory-acquired mutations in ALE and LTEE evolution experiments reveal a novel sequence space that falls outside of the natural sequence diversity of *E. coli* that is much more likely to yield AA substitutions predicted to have an impact on protein function. Taken together, this work defines the genome-scale characterization of the *E. coli* alleleome and finds that natural and laboratory evolution produces largely

Significance

The wide range of behaviors exhibited by *Escherichia coli* isolated from diverse environments is expected to be reflected in the sequence variation of its genome. Large-scale multi-strain assessment of the *E. coli* genome finds that the coding-region is highly conserved and that its scant variation is enriched in benign mutations. Contrastingly, mutations acquired through laboratory evolutions are more severe and are rarely found in nature. The antagonistic roles of general evolutionary pressures between wild-type and laboratory-evolved strains may explain these differences. Our study suggests that natural evolution produces intraspecies phenotypic diversity primarily by modulating protein abundances—rather than by altering protein properties. In comparing natural and synthetic *E. coli* mutations, we identify "sequence space" that may guide future experimental design.

Author contributions: E.A.C. and B.O.P. designed research; E.A.C. and B.O.P. performed research; E.A.C. contributed new reagents/analytic tools; E.A.C., P.P., J.M., and B.O.P. analyzed data; P.P. data acquisition of ALE mutations; J.M. data acquisition of WT genome sequences and identification of ORFs; and E.A.C. and B.O.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: palsson@ucsd.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2218835120/-/DCSupplemental.

Published April 3, 2023.

nonoverlapping sets of mutations with significantly disjoint preferences for codon selection.

Results

Establishing a Methodology for Quantifying Natural Sequence Variation. To determine natural AA sequence variation, we identified all sequence variants (alleles) for every gene present in a collection of 2,661 fully sequenced WT *E. coli* strains (*SI Appendix*, Dataset S2). Alignment of each gene's alleles (*SI Appendix*, Dataset S3, QCQA described in *SI Appendix*, Fig. S2) allowed us to determine the WT occurrence of every distinct AA residue at every AA position (Fig. 1*A*). Thus, we were able to determine the "dominant" (of highest occurrence) AA residue for each codon and thus characterize the full set of AA substitutions (nondominant, "variant" AAs) at every position across a gene (ORF).

The occurrence of dominant and variant AA residues in a given ORF can be displayed as a 3D histogram (Fig. 1*B*). The scarcity of prominent AA substitutions allowed us to describe the AA variation on a 3D structure of the protein (Fig. 1*C*). The dominant AA at each position was used to define a WT "consensus sequence" for the ORF (Fig. 1*D*, black). The consensus sequence shows the AA positions in the ORF that are fully conserved, the positions that are the most variable, and the occurrence and location of significant variants (Fig. 1*D*, cyan). To achieve a position-independent view of sequence variation in each ORF, we prepare a histogram of the dominant and variant AA frequencies (c_{dom} and c_{var}) by normalizing to the total number of strains carrying the gene ($0 < c \leq 1$) (Fig. 1*E*). Thus, we can quickly quantify the conservation of all AA positions, and the frequency and extent with which AA substitutions are found for any given ORF. The alleleome is illustrated for a single ORF in Fig. 1.

The *E. coli* Alleleome Is Highly Conserved, Is "Narrow," and Is Enriched in Inconsequential Mutations. We can combine single ORF histograms (Fig. 1*E*) for all ORFs (Fig. 2*A*) to generate

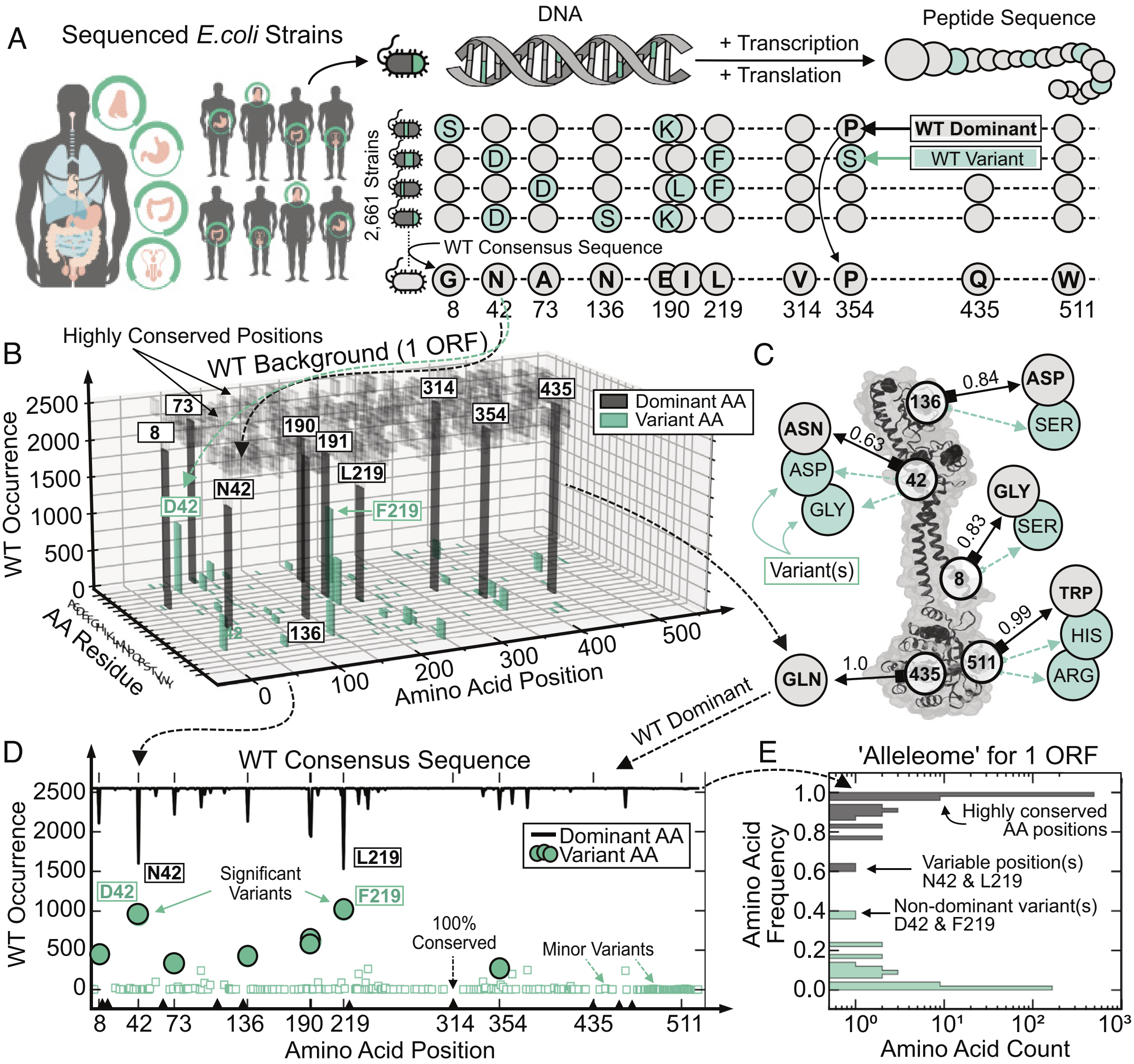

Fig. 1. The natural sequence variation in one gene. (*A*) Alignment of all unique AA sequences (alleles) of a single gene (an ORF – in this case *pdeB*) present in the WGS of up to (*SI Appendix*, Fig. S2) 2,661 WT *E. coli* strains is used to calculate (*B*) the WT occurrence of every AA residue at every AA position across the ORF. (*C*) Dominant (of highest occurrence) and variant (nondominant) AAs and their respective normalized WT occurrences for select AA positions are shown on a protein structure. (*D*) The dominant AA residues are used to define the "WT consensus sequence" while deviations from this sequence describe the full set of AA substitutions in the ORF. (*E*) A position-independent and normalized view of the WT consensus sequence and AA substitutions found in an ORF are used to describe the alleleome of a single gene.

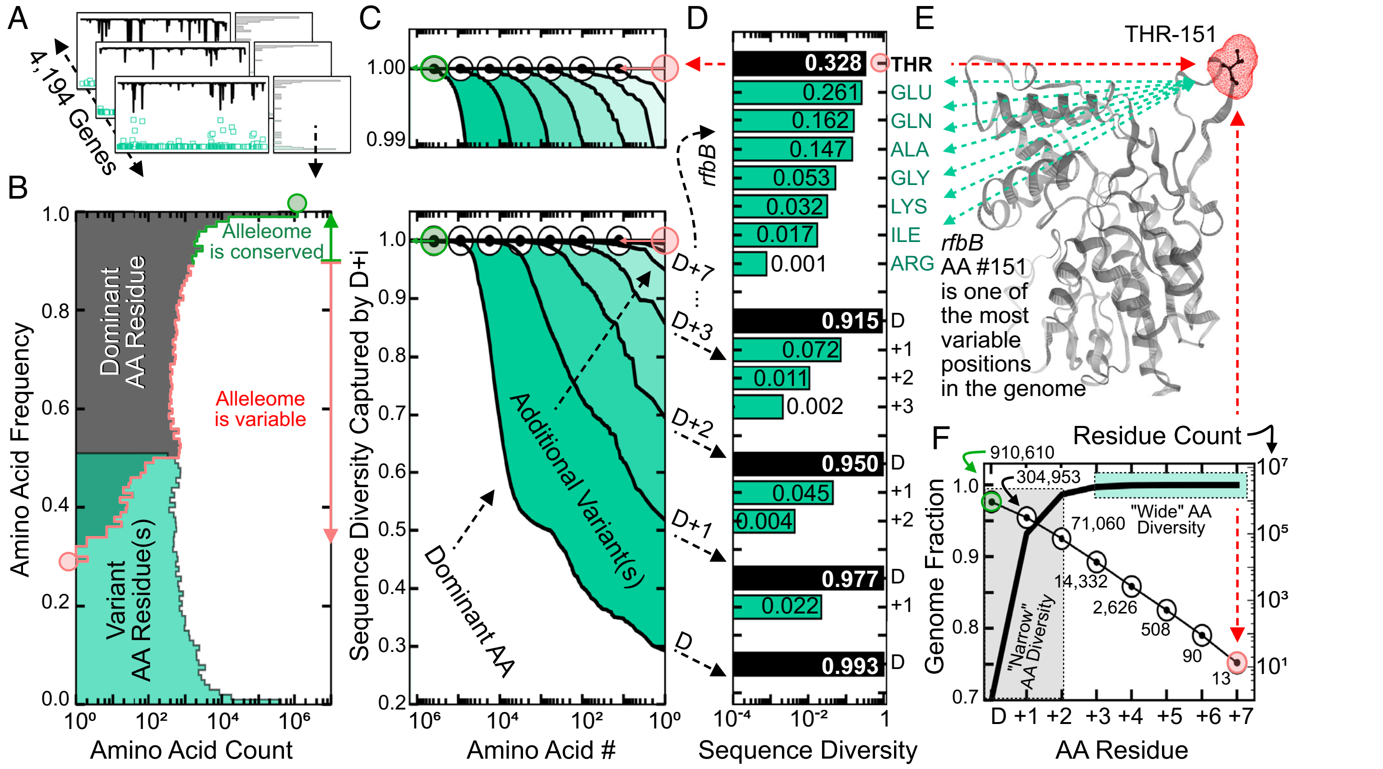

Fig. 2. The *E. coli* alleleome is "highly conserved" and "narrow". (*A*) The individual ORF histogram in Fig. 1*E* can be generated for all genes and combined to show (*B*) the frequencies of dominant AA and AA substitutions in all 1.3 million codon positions that make up the "*E. coli* ORF alleleome." In-frame deletions and insertions are also counted as substitutions, but are rarely observed (*SI Appendix*, Fig. S3 *C* and *D*). Conserved regions of the alleleome are defined by positions where the dominant AA frequency is at least 90% (c ≥ 90). (*C*) The sequence diversity at each position is defined by the sum of the AA frequencies, c, of the dominant AA (*D*), and any additional variants (+1,+2,+3,...+i) present. By sequentially considering additional variants at each position, the sequence diversity in an increasing number of positions of the alleleome can be fully described ($\Sigma c_{D+i} = 1$) (circles). (*D*) The average sequence diversity captured at positions with at least D+i AAs is shown. The majority of the D+6 and D+7 codons are found in unstructured regions of the protein. (*E*) One of the 13 D+7 codons is shown on its protein structure and is highlighted in red throughout the figure. (*F*) The total number of positions in the alleleome whose sequence diversity can be fully described ($\Sigma c_{D+i} = 1$) by exactly D+i AAs decreases logarithmically as more variants are considered (circles). The sequence diversity in 99% of the genome can be fully described by a dominant AA and narrow selection of, at most, two variant AAs (D+2) (grey) (*SI Appendix*, Fig. S3*B*).

a consolidated histogram reflecting the *E. coli* ORF alleleome (Fig. 2*B*). This alleleome represents the full natural sequence variation background of the 2,661 WT *E. coli* genome sequences isolated from diverse environments, which corresponds to a collection of 365,021 AA alleles (*SI Appendix*, Dataset S3) across 4,194 ORFs containing 1.3 million AA positions. It shows the occurrence of the dominant AA for each codon (Fig. 2*B*, gray) and also provides a global assessment of all AA substitutions found in the *E. coli* proteome (Fig. 2*B*, cyan). This ORF alleleome is a global representation of DNA and AA sequence variation in a species based on the available WGS for strains in the species (*SI Appendix*, Dataset S4).

We define a "conserved" region of the alleleome that consists of AA positions where the dominant AA frequency is at least 90% (c ≥ 0.90) (Fig. 2*B*, green). In this conserved region of the alleleome, there are 910,610 AA positions (70%) for which there is absolutely no sequence variation (c = 1.0) among the WT strains (Fig. 2, green circle). There are an additional 328,034 positions (25%) for which the dominant AA is found at a rate of 99% or greater (0.99 ≤ c < 1). Finally, there are 39,358 positions (3.0%) where the dominant AA frequency is greater than 90% (0.90 ≤ c < 0.99), showing that 1.28 million AA positions (98.0%) of the alleleome are ≥90% conserved (*SI Appendix*, Fig. S3*A*). These results show that the WT *E. coli* alleleome is highly conserved.

Among the 1.3 million dominant AAs, the sequence diversity of the *E. coli* alleleome is defined by 503,744 unique AA variants (Fig. 2*B* and *SI Appendix*, Fig. S3*A*, cyan) distributed across 393,580 codon positions (Fig. 2*C*). By calculating the sequence diversity captured by alternate AAs found in these positions (Fig. 2*C*), we determine that the *E. coli* alleleome diversity is extremely narrow—99% of all AA positions are characterized by three or fewer AAs (i.e., a dominant AA and up to two less common variants) (Fig. 2*F* and *SI Appendix*, Fig. S3*B*). These results show that the WT *E. coli* alleleome exhibits a narrow range of alternate residues in AA substitutions.

Since the alleleome provides a global assessment of all ORF sequence variation for *E. coli*, we next characterize the likely effects of the observed mutations on protein function. When we order these mutations by their global occurrence in the alleleome, we find that the vast majority of sequence variation occurs at the codon level (*SI Appendix*, Fig. S3 *C* and *D*), and the resulting mutations are synonymous (Fig. 3*A*, dark blue, *SI Appendix*, Dataset S6).

The Grantham score (GS)—a measure of physiochemical differences between two AA residues—can be used to assess the severity of each AA substitution (38). A GS below 50 reflects a "conservative" mutation, while a GS above 150 describes a "radical" mutation.

For 17.9 million observed nonsynonymous mutations, we find that the resulting AA substitutions are enriched in lower numerical values of the GS (μ = 62, a "moderately conservative" mutation) (Fig. 3*A*). We find only a small minority (2.7%) of AA substitutions with radical (>150) GSs (Fig. 3*B*, red). A summary of the WT mutations and AA substitutions analyzed is found in Table 1.

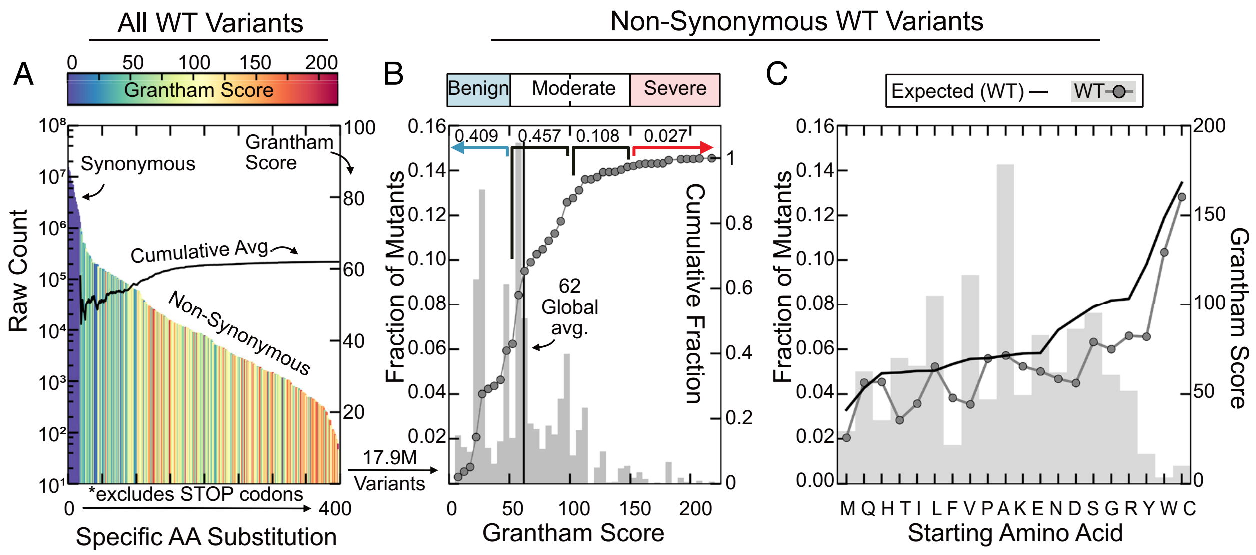

Fig. 3. The set of observed alleleomic AA substitutions is enriched in low Grantham scores (GS). (*A*) The alleleome is enriched in synonymous mutations and (*B*) in nonsynonymous AA substitutions with low GSs. The average AA substitution is moderately conservative. (*C*) The average GS of all AA substitutions originating from the same "WT-dominant" AA is lower than expected (*SI Appendix*, Fig. S4) for the majority of AAs in the alleleome. This value represents the average severity of all mutations stemming from the same AA across all positions of the alleleome.

To determine whether the enrichment of predictably "benign" AA substitutions holds true for all AAs, we develop a simple mathematical model (*SI Appendix*, Fig. S4) to calculate the expected GS from a given starting AA. We find that the observed GSs fall below their expected values for 85% of starting AAs. Only mutations in three AAs (alanine, threonine, and phenylalanine) were marginally (+0.2%, +3.7%, and +6.0%, respectively) above their expected severity (Fig. 3*C*).

These results show that the AA substitutions observed in the alleleome (Fig. 2, cyan) are enriched in substitutions that are not likely to affect protein function. This characteristic makes the ORF alleleome effectively narrower than suggested by considering only DNA sequence variation.

Laboratory Evolutions Produce *E. coli* Mutants Rarely Found in Nature. ALE is an experimental approach for biological inquiry and a method for developing phenotypic traits (20). Cultures of bacteria are serially passaged in a defined environment until their growth rate does not notably change with subsequent passages, and one or more strains from the end point population are selected for genome sequencing. A collection of the mutations acquired during ALE has been assembled in a publicly available database (ALEdb.org), which has grown exponentially since its inception in 2019 (25). Presently, ALEdb contains 22,045 publicly available mutations obtained from 1,864 bacterial isolates from 108 ALE experiments under a wide range of environments (i.e., nonglucose medium, oxidative stress, temperature stress, etc.). Our analysis is based on 45,413 unique ALE mutations (QCQA described in *SI Appendix*, Fig. S5), many of which have not yet appeared in peer-reviewed publications or in ALEdb. The ALE mutations analyzed in this study, including those previously unpublished, can be found in Table 1 and in *SI Appendix*, Dataset S5*A*.

We can assess the genetic differences between laboratory ("synthetic") evolution and natural evolution by comparing the WT alleleome and mutations found in ALEdb. We begin by looking at the mutations fixed during ALE in the *pdeB* gene (Fig. 4*A*) and display them on the WT AA occurrence diagram (i.e., Fig. 1*D*). The result, graphed in Fig. 4*B*, reveals three types of nonsynonymous mutations: first, there are seven distinct mutations occurring in conserved positions where the AA substitution falls outside of the WT alleleome (shown in red); second, an AA substitution resulting in a switch from an AA of lower occurrence (a nondominant WT variant) to the AA of dominant occurrence (D42N, shown in green) that can be thought of as being a "revertant" to consensus; and third, there is an AA substitution from a dominant occurrence to one of a lower occurrence (N136S, shown in orange). Seven of the nine nonsynonymous ALE mutations found in *pdeB* fall outside of

Table 1. Summary of the mutations and sequence variation analyzed in this study

Data source	Mutation type	Unique mutations	Genome positions (%)	Figure(s) analyzing data directly	Data
WT	Invariant (DNA)	N/A	443,357 (34%)	*SI Appendix*, Fig. S3	*SI Appendix*, Dataset S4
WT	Invariant (AA)		910,610 (70%)	Fig. 2 and *SI Appendix*, Fig. S3	
WT	Nonsynonymous*	478,749*	358,324* (27%)	Figs. 2$^{*}_{\text{cyan}}$, 3, 6 *C*–*G*, and 7 and *SI Appendix*, Fig. S3	*SI Appendix*, Datasets S4* and S6
WT	Synonymous	847,951	671,872 (52%)	Figs. 3*A* and 7 and *SI Appendix*, Fig. S3	
ALE	Nonsynonymous*	25,470*	9,459* (0.7%)	Figs. 5*–7	*SI Appendix*, Datasets S5* and S6
ALE	Synonymous	19,943	8,041 (0.6%)	Figs. 5, 6*A*, and 7	
LTEE	Nonsynonymous*	7,788*	7,753* (0.6%)	Figs. 6 and 7 and *SI Appendix*, Fig. S6*	*SI Appendix*, Datasets S5* and S6
LTEE	Synonymous	2,785	2,783 (0.2%)	Figs. 6*B* and 7 and *SI Appendix*, Fig. S6	

*A minor fraction of indels is included in the calculations of some figures and datasets (*SI Appendix*, Fig. S3).

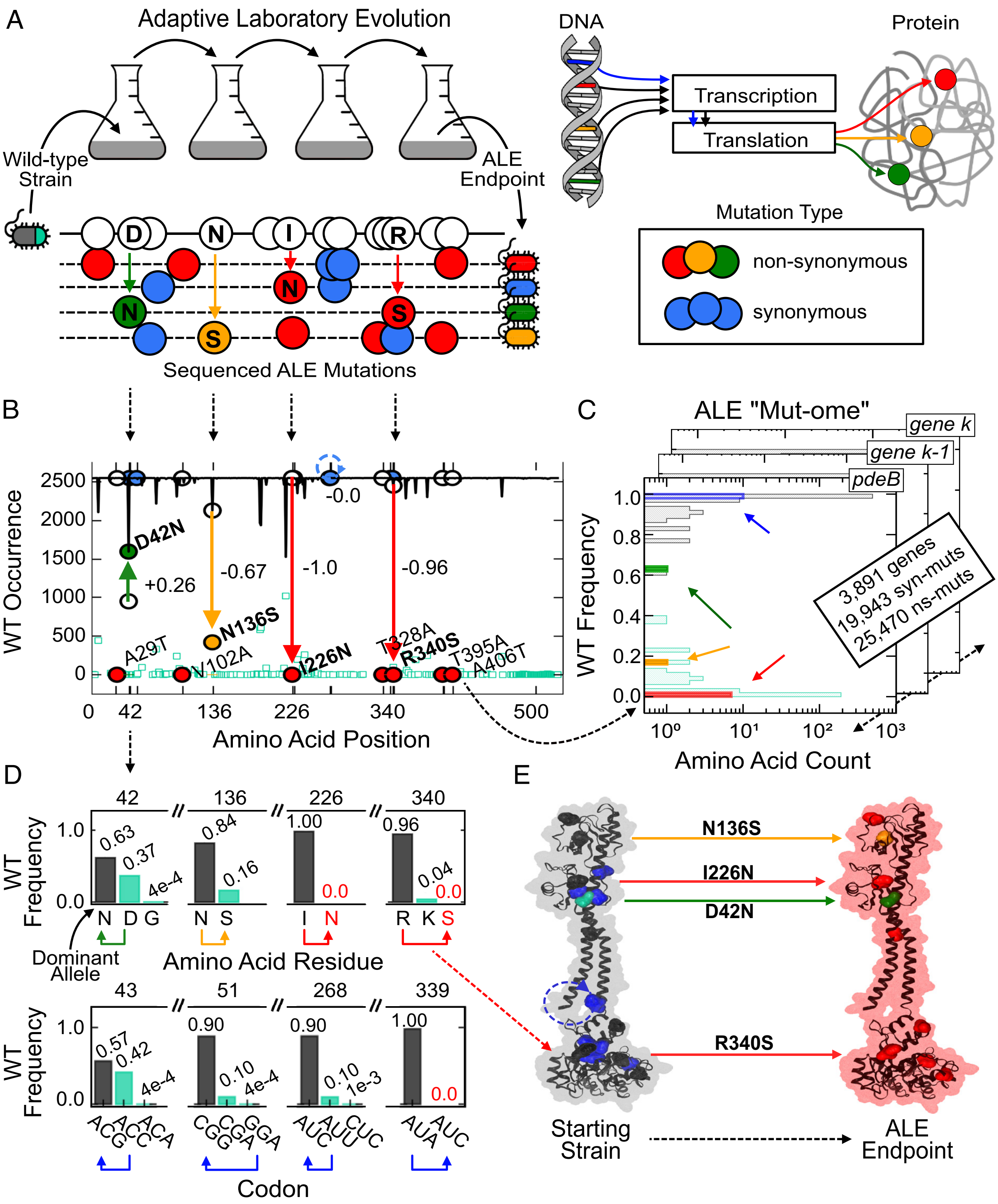

Fig. 4. Laboratory-acquired mutations can be contextualized against the natural sequence variation of *E. coli*. (*A*) Mutations are identified in the sequenced genomes of ALE *E. coli* strains. Nonsynonymous mutations resulting in AA substitutions i) that are not found (c = 0) or are rarely found (c ≤ 0.01) in the *E. coli* alleleome (red); ii) that are WT variants in the alleleome (0.01 < c < $c_{\text{WT consensus AA}}$) (orange); and iii) that are the WT-dominant AA (green) are shown. (*B*) ALE mutations in phosphodiesterase gene, *pdeB*, are mapped onto the WT consensus AA sequence. Arrows connecting the AAs in the starting strain (white circles) to the observed AA substitutions in the ALE end point (colored circles) show the change in WT occurrence for each mutation. (*C*) Position-independent normalized WT occurrences (WT frequency) of ALE mutants and natural variants of *pdeB* (*D*) are displayed for eight specific ALE mutations. Arrows indicate the AA or codon substitutions at each position (e.g., D42N reflects an aspartate—in the starting strain—to asparagine—in the ALE end point—AA substitution at position 42). (*E*) ALE mutations are shown on a rendering of *pdeB* 3D protein structure. WT-dominant AAs (black) and variants (cyan) as well as synonymous mutations (blue) are shown in the starting strain.

the WT alleleome, while the remaining two are found in two variable positions in the protein (Fig. 4*C*). Thus, ALE can provide a selection pressure that selects for a nondominant WT AA. The frequency and type of AAs where substitution takes place are detailed in Fig. 4*D*, and with such few mutations in *pdeB*, they can be viewed on the 3D protein structure (Fig. 4*E*).

As for the WT alleleome (Fig. 2 *A* and *B*), the representation of mutational data from ALE can be scaled up from one ORF

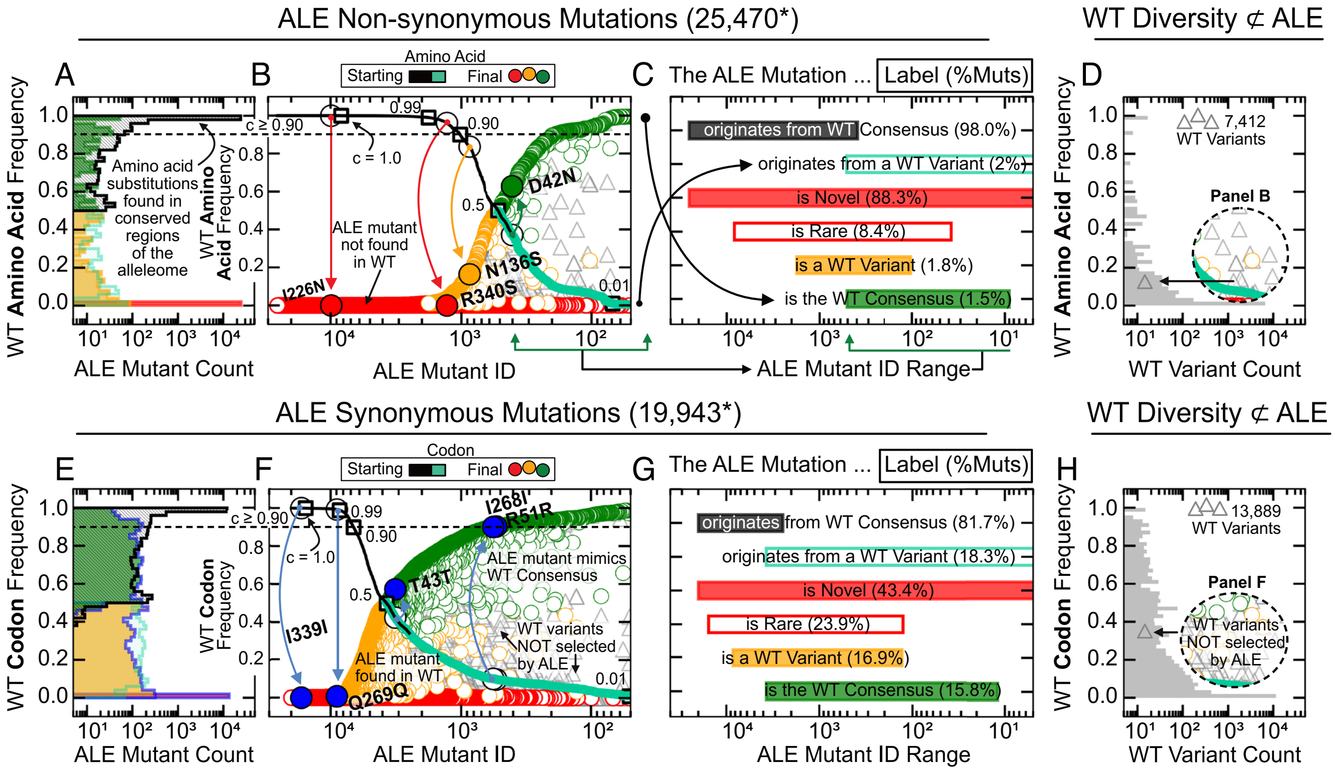

Fig. 5. ALE mutations explore a novel sequence space outside that of the *E. coli* alleleome. (*A*) The WT (WT) frequency of AAs involved in 25,470* unique nonsynonymous ALE mutations is (*B*) rank ordered by the WT frequency of the AA in the starting strain (premutation). *In-frame deletions are also counted, but are rarely observed (*SI Appendix*, Dataset S5*A*). The AA in the starting strain can reflect the WT-dominant (black) or a WT variant (cyan) AA. The WT frequency of the AA substitution is plotted (circles) according to the color scheme in Fig. 4*A*. ALE mutations in *pdeB* are shown. An inverse log scale (*x* axis) is used to highlight the ALE mutations that are found in WT strains. (*C*) The WT frequency of the premutation AA can classify mutations as "originating from the WT consensus (or from a WT variant)". The WT frequency of the postmutation AA is used to classify an ALE mutation as "novel," "rare," "a WT variant," or "the WT consensus." The fraction of ALE mutants represented by each group is shown. (*D*) The unexplored natural sequence space is represented by WT variants that exist in the same positions as ALE mutations but that are NOT selected for by ALE (gray triangles). (*E*–*H*) The analysis in Panels *A*–*D* is replicated using the WT frequency of the codons involved in ALE-acquired synonymous mutations. The ALE mutations and associated WT frequencies are provided in *SI Appendix*, Dataset S5*A*.

(Fig. 4*C*) to all the 3,891 ORFs with ALE-acquired mutations. The results give us a global view of all the acquired ALE mutations relative to the WT alleleome (Fig. 5*A*). To assess the mutations in ALEdb, we identify two features of each mutation: first, the frequency with which the AA in the starting ALE strain (i.e., MG1655) is found in the WT alleleome (i.e., is the original residue the WT consensus residue, or a WT variant residue), and second, the frequency with which the AA substitution in the ALE end point is found in the WT alleleome (i.e., is the mutant novel, rare, a WT variant, or a reversion to the WT consensus). The change in WT frequency of the starting and final AA residue for each mutant is shown in Fig. 5*B*. Of the observed nonsynonymous mutations in ALEdb, 98.0% occur in consensus positions of the WT alleleome and the majority of the resulting AA substitutions are novel (c = 0, 88.3%) or rarely found (0 < c ≤ 0.01, 8.4%) in the WT sequence variation (Fig. 5*C*).

Since the majority of ALE mutations occur in highly conserved consensus regions of the WT alleleome and result in AA residues that are not found in the natural sequence variation, it is likely that laboratory and natural selection pressures are largely disjoint. This divergence thus suggests that serial passaging may subject laboratory strains to significantly different evolutionary pressures than those experienced by WT strains. In fact, we find that on a per-gene basis, mutations found in WT strains are predominantly driven by a purifying selection pressure whereas the broad range of laboratory conditions used in ALE experiments to drive evolution often creates a diversifying selection pressure on laboratory-evolved strains (see Fig. 7).

In positions where ALE-acquired AA substitutions occur, the 7,412 variants found only in the WT sequence variation reveal an additional sequence space yet to be explored by laboratory evolutions (Fig. 5*D*). We find similar results in the analyses of synonymous mutations found in ALEdb (Fig. 5 *E*–*H*) and for 10,574 unique mutations across 211 isolates acquired in the LTEE (of multidecade duration) pioneered by Lenski (27–30) (*SI Appendix*, Fig. S6). The LTEE mutations analyzed in this study can be found in Table 1 and in *SI Appendix*, Dataset S5B.

Consequential Mutations Are More Frequently Acquired in Laboratory-Evolved *E. coli* Strains. A more detailed analysis of the mutations found in ALE and the LTEE (of multidecade duration) reveals that laboratory mutations are more likely to produce mutations that result in changes in protein properties. As for the WT variants (Fig. 3), the Granthan score (GS) is used to predict possible consequences of each laboratory-acquired mutation. We find the global average GS for nonsynonymous mutations found in laboratory evolutions to be 77 (in ALE and 78 in LTEE)—a 15-point increase in expected severity observed in WT mutants (Fig. 6 *A* and *B* and *SI Appendix*, Dataset S6). Compared to WT variants, we observe a more than threefold increase in AA substitutions characterized by severe changes in chemical properties (GS > 150) and a 1.5-fold increase in moderately severe AA substitutions (150 > GS > 100) in laboratory-evolved strains (Fig. 6 *C* and *D*).

The Grantham "space"—the set of all GSs that can be achieved—at a given genomic position depends on the original (premutation)

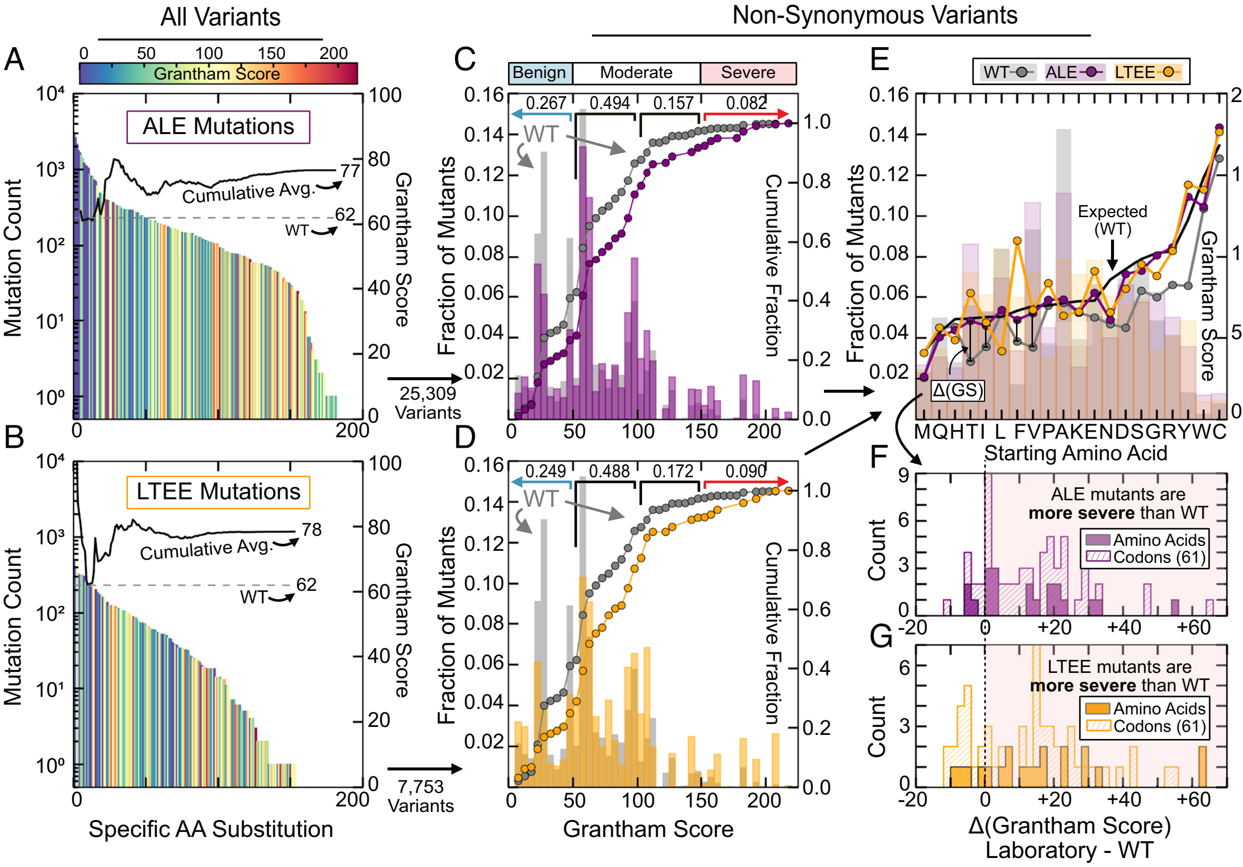

Fig. 6. Laboratory evolutions produce more severe mutations than found in WT strains. (*A*) All distinct mutations ($AA_{original}$→AA_{final}) found in ALE and (*B*) LTEE strains are rank ordered by prevalence and colored by their calculated GSs (*SI Appendix*, Dataset S6). The cumulative average GS (black) is calculated only for nonsynonymous mutations and is greater than that of WT mutations. (*C*) The histogram shows the normalized occurrence of 25,309 and 7,753 nonsynonymous mutations with varying GSs in ALE and (*D*) LTEE strains, respectively. WT occurrences are shown (gray). The cumulative fraction of mutations with predictably benign, moderate, and severe impacts on protein properties is shown. We show that the observed distribution of GSs in the WT and laboratory strains is distinct (P < 0.01, Kolmogorov–Smirnov two-sample test, *SI Appendix*, Fig. S7 and Dataset S7). (*E*) The normalized occurrence of mutations with the same starting AA is shown (histogram) for WT, ALE, and LTEE strains. We show that the observed rates of specific mutations in WT and laboratory strains are distinct for a majority of nonsynonymous AA substitutions (P < 0.01, Chi-squared test, *SI Appendix*, Figs. S8 and S9 and Dataset S7). The average GS for mutations stemming from the same AA was calculated. (*F*) The nonsynonymous AA substitutions observed in ALE and (*G*) in LTEE strains are more severe than those observed in the WT sequence variation for a majority of starting codons and AAs.

AA. For example, only three AA substitutions result in GSs greater than 200: Cys ↔ Lys (GS 202), Cys ↔ Phe (GS 205), and Cys ↔ Trp (GS = 215). Thus, a mutation in genomic positions originally containing a cysteine residue is three times more likely to yield a GS greater than 200 than in those containing lysine, phenylalanine, or tryptophan. The Grantham space is further constrained by codon selection. For example, the change of a glycine (small and aliphatic) AA for a tryptophan (large aromatic) AA is quite severe (GS 184). Four codons (GGU, GGC, GGA, and GGG) encode glycine, while only one (UGG) encodes tryptophan. One point mutation in the first position is sufficient to change the GGG-glycine codon into the UGG-tryptophan codon, whereas a minimum of two sequential point mutations are required for GGU-Gly, GGC-Gly, and GGA-Gly codons to become UGG-Tyr. Thus, to determine differences between the predicted severity of mutations in WT and laboratory-evolved strains, the original codon and AA for each mutation must be taken into account.

We find differences in the original residues involved in AA substitutions; notably, an increased aversion of WT strains to mutate AAs (Gly, Arg, Tyr, Trp, and Cys) with propensities for severe changes in chemical properties (Fig. 6*E*, bars). Normalizing across all AA substitutions derived from the same starting AA, we find an increase in predicted severity for mutations acquired through laboratory evolution for the majority of AAs and a notable increase in predicted severity in substitutions in positions containing tyrosine, aspartate, arginine, and leucine (WTGS +59, +33, +30, and +28, respectively) (Fig. 6 *E*–*G*).

General Evolutionary Pressures Play Disparate Roles in WT and Laboratory-Evolved Strains. Mutational differences in codon selection can be influenced by general evolutionary selection pressures. Purifying (negative) selection is responsible for removing severe mutations out of a population. Diversifying (positive) selection is responsible for selecting mutants resulting in cellular phenotypes with improved fitness. The large-scale assessment of natural sequence variation and laboratory-acquired mutations allows for the quantification of general selection pressures acting at the gene-level in WT and laboratory strains. We use the ratio of nonsynonymous to synonymous codon substitutions (*dN*/*dS*) in each ORF to determine the strength and mode of selection pressures acting upon each gene. Genes with *dN*/*dS* ratios greater than 1 are influenced by diversifying (positive) selection pressure, while *dN*/*dS* ratios less than 1 indicate purifying (negative) selection pressure (Fig. 7*A*). In WT strains, we find that purifying selection plays a dominant role in the natural sequence variation of a majority (4033/4335) of genes (Fig. 7*B*). In contrast, we find that the majority (2628/3891 and 2248/3058) of genes are

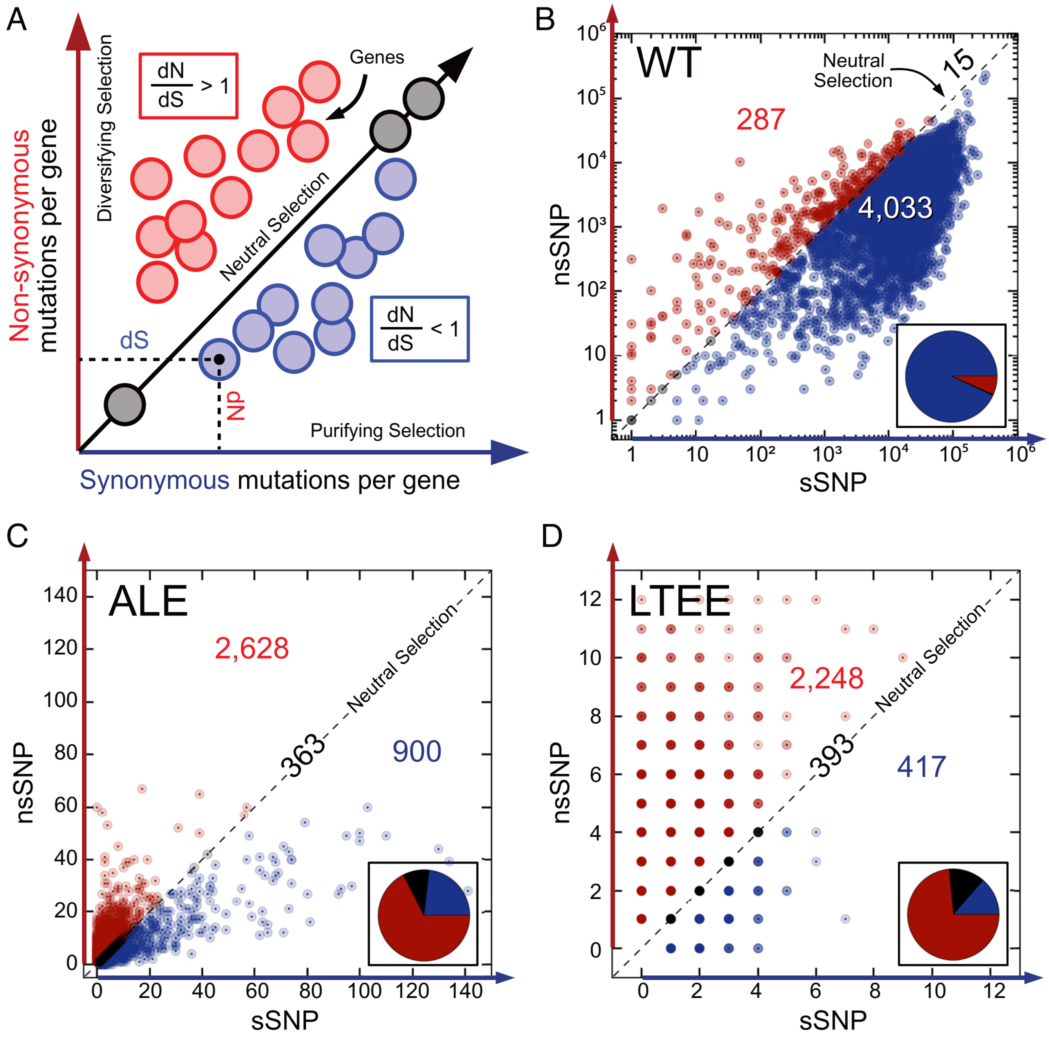

Fig. 7. Purifying selection and diversifying selection have opposite influences on WT and laboratory-evolved strains of *E. coli*. (*A*) The number of synonymous and nonsynonymous mutations in each gene. Genes are colored by their dominant selection pressure: purifying selection (blue), diversifying selection (red), and neutral selection (gray). Where mutations were found, purifying vs. diversifying selection was determined for (*B*) 4,335 genes in the WT (*C*) 3,891 genes mutated in ALE experiments (25) and (*D*) 3,058 genes mutated in the LTEE (26). The pie chart *Insets* reflect the fraction of genes influenced by the selection pressures and show that WT and laboratory strains are predominantly influenced by opposite selection pressures. The mutations used in this figure can be calculated from *SI Appendix*, Dataset S4 (WT) and *SI Appendix*, Dataset S5 (ALE and LTEE).

predominantly influenced by diversifying selection pressure in laboratory-evolved strains (ALE and LTEE, respectively) (Fig. 7 *C* and *D*).

Discussion

The deluge of recently available bacterial WGS has allowed us to develop a unique method to analyze the natural sequence diversity among WT *E. coli* strains in the form of the ORF alleleome—a collection of the sequence variations in the 1.3 million codon positions in the coding region of the *E. coli* genome. We find that the alleleome is highly conserved: 98% of AAs positions are conserved in more than 90% strains. Alleleomic variation is further limited by a narrow range (typically one or two) of alternate AAs.

In view of the surprisingly conserved and narrow alleleome, we point out that a number of pangenome analyses have attributed differences in phenotypic traits between strains to differential gene counts in multiple strains of a species (39). Furthermore, ALE experiments suggest that regulatory mutations play a bigger role in adaptation than structural mutations (40) and that proteome allocation may be more important than changes in protein properties (41). Future studies of the intragenic diversity may thus reveal greater consequential sequence diversity than that found here in the ORF alleleome. Allele variation, however, is important in special cases such as antimicrobial resistance studies where strong selection is based on changes in a target protein (42).

Concurrently, inexpensive DNA sequencing has enabled the sequencing of a large number of clones from laboratory evolutions and finds causal mutations relative to a specific selection pressure. This large-scale comparison of the natural sequence variation in *E. coli* and laboratory-evolved mutants finds fundamentally divergent sets of mutations, suggesting that the selection pressures that have been used in experimental evolutions to date may not reflect those found in nature.

This finding is consistent with many observations of mutations in experimental populations. In WT strains, nonsynonymous mutations that may improve a strain's fitness in one environment are likely deleterious for a strain exposed to changing environments over an extended period of time. Thus, nonsynonymous variants are purified out of the WT population, selecting only ("generalist") strains with improved fitness across multiple environments. In contrast, when exposed to a well-defined environment, laboratory strains may acquire nonsynonymous mutations that improve the strain's fitness in a specific environment ("specialist") at the cost of decreasing the strain's fitness when exposed to changing environments.

Differing adaptive fitness strategies have been observed in various ALE experiments. When *E. coli* was grown in alternating sugar substrates, some populations developed a generalist strain that persisted across multiple substrates, while other populations developed two specialist subpopulations that alternated dominance between environmental conditions (34). This multi-scale adaptive behavior of *E. coli* strains may be explained by the antagonistic pleiotropic effects of specific mutations. For example, ALE-acquired single-mutation variants of *rpoB* were shown to grow faster in glucose but required a longer diauxic shift to support growth on acetate when compared to the WT strain (43), suggesting an inability to quickly shift proteome allocation in response to a changing environment. Antagonistic pleiotropy and adaptation trade-offs have been studied in great detail for *E. coli* in the LTEE (44–48), and these effects have been observed in other organisms (49, 50).

In this study, we confirm opposing general selection pressures between WT and laboratory strains that are consistent with the differences in environmental constancy and evolutionary time scales expected between natural and laboratory evolution. As ALE has been shown to be useful in generating desirable bioprocess phenotypes—such as those exhibiting increased tolerance for end product toxicity and improved substrate readiness (31, 34)—the uniqueness of the vast majority of experimental evolution mutations is advantageous from a synthetic biology point of view.

Remarkably, a subset of mutations in ALE were observed to switch between the WT dominant and secondary AAs (Fig. 5 green & yellow, respectively), suggesting that well-defined selection pressures can identify naturally occurring "toggle-switches" in the *E. coli* sequence space. This observation may guide the generation of future ALE experimental design. For instance, ALE mutants that revert to the WT consensus in one experiment can be subjected to a range of selection pressures until the mutant returns to the original AA in the first experiment. Identifying the selection pressures that can wobble specific AAs between WT consensus residues and variants may help identify specific positions across the genome that reflect generalist or specialist adaptation strategies and may offer insights into divergence in evolution between natural "WT" strains and laboratory strains.

Genome-scale comparison of WT variants and mutations acquired through laboratory evolution also offers insights into the differential propensity of certain AAs to become mutated. Compared against the WT alleleome, we show that both ALE and the LTEE yield mutations which are more likely to impact protein function. Among WT strains, we also find an increased avoidance of fixing mutations in positions with AAs whose Grantham space has a propensity for severe mutations, supporting the causality of these severe mutations when they occur in laboratory-evolved strains. Future studies may benefit from a gene-by-gene level determination if these mutations fall within key regions of the protein (e.g., active sites, protein–protein interfaces, etc.). This cumulative experience shows that there is useful sequence space to be explored for biological design purposes which may be nonobvious and thus advantageous from a discovery standpoint.

Materials and Methods

Procurement of a Diverse Set of *E. coli* Genome Sequences and Identification of ORFs. WGS of 2,661 *E. coli* strains were downloaded from the Pathosystems Resource Integration Center (PATRIC) (51). The PATRIC metadata included the host and geographic information for each strain. To demonstrate the diversity of our strains, phylogrouping the strain collection was completed using EzClermont (52). The genomic distances were calculated with Mashtree (53) and visualized using Interactive Tree of Life (54) (*SI Appendix*, Fig. S1 and Dataset S1). The bidirectional best blast hit tool from Rapid Annotation using Subsytems Technology (RAST) (55) (https://rast.nmpdr.org) was used to match annotated loci from *E. coli* K-12 MG1655 (Blattner Numbers) (in PATRIC) to orthologs in the genomes of WT strains. For 4,349 genes, each unique nucleic acid (NA) sequence match from the WT strains was assigned a NA allele ID (*SI Appendix*, Dataset S2). The WT occurrence and NA sequence for each allele were recorded (*SI Appendix*, Dataset S3).

Quality Control and Analysis of Nucleic Acid Allele Sequences. Each NA allele was read three nucleobases at a time until the first stop codon to determine the codon sequence of the gene. NA alleles with early truncations resulting in a gene loss greater than 20% of the gene length were removed. The codon sequence was translated into an AA (AA) sequence. The distribution of AA allele sequence lengths for each gene was calculated. AA alleles found to be more than two SDs shorter than the mean AA sequence length were removed. The genes with remaining alleles represented in less than 133 (5% of) WT strains were removed. This QCQA analysis yielded 729,212 NA alleles and 365,021 AA alleles distributed across 4,194 genes (*SI Appendix*, Fig. S2 and Dataset S3).

Defining the WT Alleleome for 4,194 ORFs. For each gene, Multiple Sequence Comparison by Log-Expectation (MUSCLE) (56) was used to align the AA sequences of its alleles. For each AA position in the alignment, the occurrence (number of strains) of all AA variants was calculated. The WT dominant AA–the AA of highest occurrence–at each position was used to define a WT consensus sequence. In the case of the dominant AA being a "deletion" (e.g., an insertion is found in a nonmajority of strains), this residue position was removed from the overall WT consensus sequence. The occurrence of nondominant AA residues (variants) was also determined. The WT occurrence of nondominant in-frame deletions is counted as AA variants in the WT alleleome (Figs. 2 and 5) but not in the GS analyses (Figs. 3 and 6).

Once the AA sequence (and in-frame deletions) of all alleles was determined, the codon sequence for each allele was recreated to include any gaps detected by the MUSCLE alignment of AA sequences, allowing standardized position and codon information across all alleles. The WT alleleome was defined for 1.3 million codon positions distributed across 4,194 genes where dominant and variant AA occurrences could be determined. The complete WT alleleome, codon-level variants and AA substitutions can be found in *SI Appendix*, Dataset S4. All nondominant variants (including in-frame deletions and insertions) are described in *SI Appendix*, Fig. S3.

Mapping Laboratory-Acquired Mutations to the WT Alleleome. We use a series of quality control assessments (*SI Appendix*, Fig. S5) to ensure the proper mapping of laboratory mutations in ALEdb (and additional unpublished mutants) and in the LTEE dataset (https://barricklab.org/shiny/LTEE-Ecoli/) to the WT alleleome (*SI Appendix*, Dataset S4). Each mutation mapped to the WT alleleome i) must be acquired in an *E. coli* strain (applicable only for ALEdb); ii) must be found within an ORF; iii) must be annotated to a gene name that can be mapped to a Blattner number [either directly or through a gene synonym confirmed by EcoCyc (57)]; iv) must have agreement between the codon found in the reference genome sequence (REL606/7 for LTEE, multiple for ALEdb) at the genome position given with the codon found at the gene location given in the mutation annotation; v) must NOT result in a truncation larger than 20% of the gene; vi) must map to a gene with sufficient alleles (i.e., is present in >5% of WT strains, see *SI Appendix*, Fig. S2); and vii) is only used once in the dataset (repeat mutations are removed).

After QCQA, we were able to identify 45,413 unique mutations found across 4,181 ALE isolates from 284 ALE experiments (*SI Appendix*, Dataset S5*A*) and 10,574 unique mutations isolated from 211 strains in the LTEE (*SI Appendix*, Dataset S5*B*). For each nonsynonymous mutation (including a small number of early truncations and in-frame deletions) in ALE strains, the WT occurrence of the premutation and postmutation AA (or termination) was determined (Fig. 5 *A*–*D*). The WT occurrence of codons involved in synonymous ALE mutations was also determined (Fig. 5 *E*–*H*). The analyses were repeated for mutations found in LTEE strains (*SI Appendix*, Fig. S6).

Mathematical "Null" Model for Computing Rates of Specific Mutations Expected in WT Strains. We calculate the expected distribution of codon variants (codon2s) found in all mutations originating from the same initial codon

(codon1) and compare this to the observed WT distribution. We can calculate the GS for each specific codon change and use the expected distribution of specific codon changes to calculate the expected GS of a mutation originating from a specific codon.

A codon can be converted into any other codon by acquiring up to three sequential and distinct point mutations ("hops") (e.g., a final codon that requires two point mutations in the original codon is referred to as a 2-hop mutant) (*SI Appendix*, Fig. S4*A*). Given that a mutation occurs and the initial codon (codon1) is known, the expected distribution of all specific codon changes from the original codon depends on i) the number of sequential point mutations needed to reach the final codon (codon2) and the probability of reaching suitable intermediates in the mutation pathway (*SI Appendix*, Fig. S4*B*). Since the mutation occurs, the sum of the expected rates of all final codons must be equal to 1. Using these assumptions, we can write an equation for the sum of all expected rates of codon changes:

$$\sum_{k=1}^{3} N_k * \left(P_{k-1}\right) * \left(f_{i,k-1}\right) * \mu = 9 * \mu + 27 * (\mu) * \left(\frac{2}{9}\right) * \mu + 27\left(\frac{2\mu^2}{9}\right)\left(\frac{3}{27}\right)\mu = 1,$$

where *k* represents the number of sequential point mutations (definition of all terms can be found in *SI Appendix*, Fig. S4*B*). This equation allows us to solve for μ, the probability that a mutation results in a specific 1-hop mutant. Likewise, we solve for (2u^2/9) and (2u^3/81), the probability of finding a specific 2-hop and 3-hop mutant, respectively. For each initial codon, the expected distribution of mutants across the codon table is calculated (*SI Appendix*, Fig. S4*D*). Using this distribution, we calculate the expected average GS for all mutants derived from an initial codon. The deviation between the model-predicted average GS and the GSs observed in the alleleome is calculated for each initial codon (*SI Appendix*, Fig. S4*E*). These codon-level calculations can be grouped by their resulting AA residues to yield the severity of specific AA substitutions (as in Fig. 3).

We analyze only the GSs of nonsynonymous AA substitutions. As such, each initial codon will have a constrained number of codon variants counted in the observed data. To account for this, we constrain our model to only compute mutation rates for this reduced set of codon changes (*SI Appendix*, Fig. S4*C*).

Determining the Observed Mutation Rates of Specific Codon Changes. In the WT alleleome, the dominant codon (codon1) at each position is determined. For each nondominant codon variant (codon2) at that position, the total number of strains represented at the position by the nondominant codon is counted as the number of mutations from the WT consensus sequence. The mutation counts of mutations that share the same specific codon change (codon1→codon2 pair) are combined to determine the rates of specific codon changes across the alleleome (*SI Appendix*, Dataset S6, "WT"). Codon changes are further separated by their effect on the AA sequence (synonymous vs. nonsynonymous AA substitution). The rates of specific codon changes in the ALE and LTEE strains were calculated using the metadata provided in *SI Appendix*, Datasets S5*A* and S5*B*, respectively. These mutation rates are recorded in *SI Appendix*, Dataset S6.

Identification of General Selection Pressures. The rates of synonymous and nonsynonymous mutations were determined (as described above) on a per-gene basis. The ratio of nonsynonymous mutations to synonymous mutations (*dN/dS*) was used to infer general selection pressure acting on each gene. These values were plotted for genes mutated in WT, ALE, and LTEE strains (Fig. 7).

Calculating the Average GS (Per Codon & AA). Each specific premutation codon to postmutation codon pair (Codon1→Codon2) is counted across WT and laboratory strains and recorded in *SI Appendix*, Dataset S6. The distribution of all mutations originating from the same codon is determined. Each mutation's GS and weighted occurrence in the distribution is used to find the average GS of all nonsynonymous AA substitutions originating from the same codon. This GS represents the average severity of all mutations stemming from the same initial codon across all positions in the alleleome. This analysis is repeated for all (61 of 64) premutation nonterminating codons. The per-codon results are used to analyze differences in GS severity between WT and laboratory strains (Fig. 6 *F* and *G*).

The observed codon changes are also grouped by their premutation AA (Fig. 6*E*, histogram). For all 20 AAs, the distribution of all nonsynonymous AA substitutions that originate from the same premutation AA is used to calculate the average GS observed in a particular AA (Fig. 6*E*, lines). This GS represents the average severity of all mutations stemming from the same initial AA across all positions in the alleleome. Differences between GSs in WT and laboratory strains are calculated per each AA (Fig. 6 *F* and *G*).

Statistical Analysis of GS Distributions Observed in WT and Laboratory Strains. The distribution of GSs is determined using the mutation rates in *SI Appendix*, Dataset S6 for all nonsynonymous mutations in WT (Fig. 3*B*) and laboratory strains (Fig. 6 *C* and *D*). The mutations are grouped by premutation codon or by premutation AA (see above). For each nonterminating premutation codon (codon1), the distribution of GSs for all mutations originating from codon1 is calculated. A Kolmogorov–Smirnov two-sample test is used to compare the per-codon GS distributions in WT and laboratory mutations (*SI Appendix*, Fig. S7 *A* and *B*). This test is repeated at the AA level (*SI Appendix*, Fig. S7 *C* and *D*). Statistical metrics are provided in *SI Appendix*, Dataset S7.

Statistical Analysis of Observed Rates of Specific Mutations in WT and Laboratory Strains. The observed rates of all specific mutations are recorded in *SI Appendix*, Dataset S6. The mutations are grouped by premutation codon or by premutation AA (see above). For each grouping, the observed mutation rates of all specific nonsynonymous mutations originating from codon1 can be determined for mutations found in WT and ALE strains. The WT rates are normalized and multiplied by the total number of ALE mutations in the grouping to calculate the expected distribution of mutations in the grouping ("WT exp"). Specific mutations in the grouping that are either observed (in ALE) or expected (based on the WT rate) to occur less than five times are removed from the grouping (Chi-squared test criteria, *SI Appendix*, Fig. S8*A*). The observed and expected mutation rates in groupings of two or more mutations were analyzed using a Chi-squared test (*SI Appendix*, Fig. S8*B*). Statistical metrics are provided in *SI Appendix*, Dataset S7.

This analysis is repeated for mutations in LTEE strains (*SI Appendix*, Fig. S8 *C* and *D*). This analysis is repeated for groupings of mutations originating from the same AA (*SI Appendix*, Fig. S9). Statistical metrics for these analyses are also provided in *SI Appendix*, Dataset S7.

Data, Materials, and Software Availability. All Data have been deposited in GitHub (https://github.com/EdwardCatoiu/Alleleome.git) (58). All study data are included in the article and/or *SI Appendix*. Previously published data were used for this work (25, 26).

ACKNOWLEDGMENTS. We would like to thank Adam Feist, Siddharth Chauhan, and Daniel Zielinski for helpful technical discussions and Marc Abrams for assistance with manuscript editing. This work was funded by Novo Nordisk Foundation (Grant Number NNF20CC0035580) (E.A.C. and B.O.P.) and NIH (Grant R01 GM057089) (B.O.P.).

Author affiliations: [a]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92101; [b]The Novo Nordisk Foundation (NNF) Center for Biosustainability, The Technical University of Denmark, Kongens Lyngby 2800, Denmark; and [c]Avellino Lab, Menlo Park, CA 94025

1. T. D. Read, R. C. Massey, Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: A new direction for bacteriology. *Genome Med.* **6**, 109 (2014), 10.1186/s13073-014-0109-z.
2. S. J. Forrester, N. Hall, The revolution of whole genome sequencing to study parasites. *Mol. Biochem. Parasitol.* **195**, 77–81 (2014), 10.1016/j.molbiopara.2014.07.008.
3. M. Delseny, B. Han, Y. I. Hsing, High throughput DNA sequencing: The new sequencing revolution. *Plant Sci.* **179**, 407–422 (2010), 10.1016/j.plantsci.2010.07.019.
4. N. J. Loman, M. J. Pallen, Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.* **13**, 787–794 (2015), 10.1038/nrmicro3565.
5. M. Land *et al.*, Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15**, 141–161 (2015), 10.1007/s10142-015-0433-4.
6. M. C. Maiden *et al.*, Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3140–3145 (1998), 10.1073/pnas.95.6.3140.

7. B. Spratt, Multilocus sequence typing: Molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. *Curr. Opin. Microbiol.* **2**, 312–316 (1999), 10.1016/S1369-5274(99)80054-X.
8. O. Clermont, S. Bonacrosi, E. Bingen, Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl. Environ. Microbiol.* **66**, 4555–4558 (2000), 10.1128/aem.66.10.4555-4558.2000.
9. O. Clermont, J. K. Chirstenson, E. Denamur, D. M. Gordon, The Clermont *Escherichia coli* phylo-typing method revisited: Improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* **5**, 58–65 (2013), 10.1111/1758-2229.12019.
10. S. D. Reid, C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, T. S. Whittam, Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**, 64–67 (2000), 10.1038/35017546.
11. O. Tenaillon, D. Skurnik, B. Picard, E. Denamur, The population genetics of commensal *Escherichia coli*. *Nat. Rev. Microbiol.* **8**, 207–217 (2010), 10.1038/nrmicro2298.
12. O. Clermont, D. M. Gordon, S. Brisse, S. T. Walk, E. Denamur, Characterization of the cryptic *Escherichia* lineages: Rapid identification and prevalence. *Environ. Microbiol.* **13**, 2468–2477 (2011), 10.1111/j.1462-2920.2011.02519.x.
13. S. D. Reid, R. K. Selander, T. S. Whittam, Sequence diversity of flagellin (fliC) alleles in pathogenic *Escherichia coli*. *J. Bacteriol.* **181**, 153–160 (1999).
14. D. W. Lacher, H. Steinsland, T. S. Whittam, Allelic subtyping of the intimin locus (eae) of pathogenic *Escherichia coli* by fluorescent RFLP. *FEMS Microbiol. Lett.* **261**, 80–87 (2006), 10.1128/jb.181.1.153-160.1999.
15. X. P. Koh *et al.*, Genetic and ecological diversity of *Escherichia coli* and cryptic Escherichia clades in subtropical aquatic environments. *Front. Microbiol.* **13**, 811755 (2022), 10.3389/fmicb.2022.811755.
16. H. H. Yang, R. T. Vinopal, D. Grasso, B. F. Smets, High diversity among environmental *Escherichia coli* isolates from a bovine feedlot. *Appl. Environ. Microbiol.* **70**, 1528–1536 (2004), 10.1128/AEM.70.3.1528-1536.2004.
17. R. R. Chaudhuri, I. R. Henderson, The evolution of the *Escherichia coli* phylogeny. *Infect. Genet. Evol.* **12**, 214–226 (2012), 10.1016/j.meegid.2012.01.005.
18. C. Vignaroli *et al.*, Adhesion of marine cryptic *Escherichia* isolates to human intestinal epithelial cells. *ISME J.* **9**, 508–515 (2015), 10.1038/ismej.2014.164.
19. C. Liao *et al.*, Allelic variation in outer membrane protein A and its influence on attachment of *Escherichia coli* to corn stover. *Front. Microbiol.* **8**, 708 (2017), 10.3389/fmicb.2017.00708.
20. M. Dragosits, D. Mattanovich, Adaptive laboratory evolution – principles and applications for biotechnology. *Microb. Cell Fact.* **12**, 64 (2013).
21. V. Portnoy, D. Bezdan, K. Zengler, Adaptive laboratory evolution—harnessing the power of biology for metabolic engineering. *Curr. Opin. Biotechnol.* **22**, 590–594 (2011).
22. R. A. LaCroix, B. O. Palsson, A. M. Feist, A model for designing adaptive laboratory evolution experiments. *Appl. Environ. Microbiol.* **83**, e03115-16 (2017).
23. T. Sandberg, M. Salazar, L. L. Weng, B. O. Palsson, A. M. Feist, The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. *Metab. Eng.* **56**, 1–16 (2019).
24. S. Lee, P. Kim, Current status and applications of adaptive laboratory evolution in industrial microorganisms. *J. Microbiol. Biotechnol.* **30**, 793–803 (2020).
25. P. V. Phaneuf, D. Gosting, B. O. Palsson, A. M. Feist, ALEdb 1.0: A database of mutations from adaptive laboratory evolution experimentation. *Nucleic Acids Res.* **47**, D1164–D1171 (2019).
26. Barrick Lab, LTEE-Ecoli. [Online]. Available: https://barricklab.org/shiny/LTEE-Ecoli/. [Accessed 1 April (2022)].
27. R. E. Lenski, M. R. Rose, S. C. Simpson, S. C. Tadler, Long-term experimental evolution in *Escherichia coli* adaptation and divergence during 2,000 generations. *Am. Nat.* **138**, 1315–1341 (1991).
28. J. E. Barrick *et al.*, Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243–1247 (2009).
29. R. E. Lenski, Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME J.* **11**, 2181–2194 (2017).
30. O. Tenaillon *et al.*, Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature.* **536**, 165–170 (2017).
31. S. S. Fong *et al.*, In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol. Bioeng.* **91**, 643–648 (2005).
32. A. Anand *et al.*, Pseudogene repair driven by selection pressure applied in experimental evolution. *Nat. Microbiol.* **4**, 386–389 (2019).
33. D. Choe *et al.*, Adaptive laboratory evolution of a genome-reduced *Escherichia coli*. *Nat. Commun.* **10**, 935 (2019).
34. T. E. Sandberg, C. J. Lloyd, B. O. Palsson, A. M. Feist, Laboratory evolution to alternating substrate environments yields distinct phenotypic genetic adaptive strategies. *Appl. Environ. Microbiol.* **83**, e00410-17 (2017), 10.1128/AEM.00410-17.
35. H. Mundhada *et al.*, Increased production of L-serine in Escherichia coli through adaptive laboratory evolution. *Metab. Eng.* **39**, 141–150 (2017).
36. T. Sandberg *et al.*, Evolution of *Escherichia coli* to 42 °C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. *Mol. Biol. Evol.* **31**, 2647–2662 (2014).
37. R. A. LaCroix *et al.*, Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of *Escherichia coli* K-12 MG1655 on glucose minimal medium. *Appl. Environ. Microbiol.* **81**, 17–30 (2014).
38. R. Grantham, Amino acid difference formula to explain protein evolution. *Science.* **185**, 862–864 (1974), 10.1126/science.185.4154.862.
39. C. J. Norsigian, X. Fang, B. O. Palsson, J. M. Monk, "Pangenome flux balance analysis toward panphenomes" in *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, H. Tettelin, D. Medini, Eds. (Springer, 2020), pp. 219–232.
40. T. E. Sandberg, R. Szubin, P. V. Phaneuf, B. O. Palsson, Synthetic cross-phyla gene replacement and evolutionary assimilation of major enzymes. *Nat. Ecol. Evol.* **4**, 1402–1409 (2020), 10.1038/s41559-020-1271-x.
41. B. O. Palsson, J. T. Yurkovich, Is the kinetome conserved? *Mol. Syst. Biol.* **18**, e10782 (2022), 10.15252/msb.202110782.
42. E. S. Kavvas *et al.*, Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* **9**, 4306 (2018).
43. J. Ultrilla *et al.*, Global rebalancing of cellular resources by pleiotropic point mutations illustrates a multi-scale mechanism of adaptive evolution. *Cell Syst.* **2**, 260–271 (2016), 10.1016/j.cels.2016.04.003.
44. M. Travisano, F. Vasi, R. E. Lenski, Long-term experimental evolution in *Escherichia coli*. III. Variation among replicate populations in correlated responses to novel environments. *Evolution* **49**, 189–200 (1995), 10.1111/j.1558-5646.1995.tb05970.x.
45. M. Travisano, R. E. Lenski, Long-term experimental evolution in *Escherichia coli*. IV. Targets of selection and the specificity of adaptation. *Genetics* **143**, 15–26 (1996), 10.1093/genetics/143.1.15.
46. V. S. Cooper, R. E. Lenski, The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature.* **407**, 736–739 (2000), 10.1038/35037572.
47. A. F. Bennett, R. E. Lenski, J. E. Mittler, Evolutionary adaptation to temperature. I. Fitness responses of *Escherichia coli* to changes in its thermal environment. *Evolution.* **46**, 16–30 (1992), 10.1111/j.1558-5646.1992.tb01981.x.
48. A. F. Bennett, R. E. Lenski, Evolutionary adaptation to temperature. II. Thermal niches of experimental lines of *Escherichia coli*. *Evolution* **47**, 1–12 (1993), 10.1111/j.1558-5646.1993.tb01194.x.
49. L. Noda-Garcia *et al.*, Chance and pleiotropy dominate genetic diversity in complex bacterial environments. *Nat. Microbiol.* **4**, 1221–1230 (2019), 10.1038/s41564-019-0412-y.
50. G. Kinsler, K. Geiler-Samerotte, D. A. Petrov, Fitness variation across subtle environmental perturbations reveals local modularity and global pleiotropy of adaptation. *eLife* **9**, e61271(2020), 10.7554/eLife.61271.
51. R. Wattam *et al.*, PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, D581–D591 (2014), 10.1093/nar/gkt1099.
52. N. Waters, F. Abram, F. Brennan, A. Holmes, L. Pritchard, Easy phylotyping of *Escherichia coli* via the EzClermont web app and command-line tool. *Access Microbiol.* **2**, acmi000143 (2020), 10.1099/acmi.0.000143.
53. L. S. Katz *et al.*, Mashtree: A rapid comparison of whole genome sequence files. *J. Open Source Softw.* **4**, 1762 (2019), 10.21105/joss.01762.
54. P. Letunic, Bork, Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, 293–296 (2021), 10.1093/nar/gkab301.
55. R. Overbeek *et al.*, The SEED and rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2014), 10.1093/nar/gkt1226.
56. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004), 10.1093/nar/gkh340.
57. I. M. Keseler *et al.*, EcoCyc: A comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.* **39**, D583–D590 (2011), 10.1093/nar/gkq1143.
58. E. A. Catoiu, Datasets S1-S7. The E. coli Alleleome (database). https://github.com/EdwardCatoiu/Alleleome. Deposited 2 February 2023.