Core Genes Evolve Rapidly in the Long-term Evolution Experiment with *Escherichia coli*

Rohan Maddamsetti,*[,1,2] Philip J. Hatcher,[3] Barry L. Williams,[1] and Richard E. Lenski[1,2]

[1]Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI 48824, USA.

[2]BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI, 48824, USA.

[3]Department of Computer Science, University of New Hampshire, Durham, NH 03824, USA

***Corresponding author:** E-mail: maddamse@msu.edu.

Abstract

Conserved genes evolve slowly in nature, by definition, but we find that some conserved genes are among the fastest-evolving genes in the long-term evolution experiment with *Escherichia coli* (LTEE). We identified the set of almost 2000 core genes shared among sixty clinical, environmental, and laboratory strains of *E. coli*. During the LTEE, these core genes accumulated significantly more nonsynonymous mutations than did flexible (i.e., noncore) genes after accounting for the mutational target size. Furthermore, the core genes under strongest positive selection in the LTEE are more conserved in nature than the average core gene based both on sequence diversity among *E. coli* strains and divergence between *E. coli* and *Salmonella enterica*. We conclude that the conditions of the LTEE are novel for *E. coli*, at least in relation to the long sweep of its evolution in nature. We suggest that what is most novel about the LTEE for the bacteria is the constancy of the environment, its biophysical simplicity, and the absence of microbial competitors, predators, and parasites.

Introduction

By combining experimental evolution and genomic technologies, researchers can study in great detail the genetic underpinnings of adaptation in the laboratory (Barrick and Lenski 2013). However, questions remain about how the genetic basis of adaptation might differ between experimental and natural populations (Bailey and Bataillon 2016).

To explore that issue, we examined whether the genes that evolve most rapidly in the long-term evolution experiment with *Escherichia coli* (LTEE) also evolve and diversify faster than typical genes in nature. If so, the genes involved in adaptation in the LTEE might also be involved in local adaptation to diverse environments in nature. On the other hand, it might be the case that the genes involved in adaptation during the LTEE diversify more slowly in nature than typical genes. Perhaps these genes are highly constrained in nature by purifying selection. For example, they may play important roles in balancing competing metabolic demands or fluctuating selective pressures in the complex and variable natural world, but they can be optimized to fit the simplified and stable conditions of the LTEE.

To test these alternative hypotheses, we compare the signal of positive selection across genes in the LTEE to the sequence diversity in a set of 60 clinical, environmental, and laboratory strains of *E. coli*—henceforth, the "*E. coli* collection"—and to the divergence between *E. coli* and *Salmonella enterica* genomes, respectively. We find that the genes that have evolved the fastest in the LTEE tend to be conserved core genes in the *E. coli* collection. We can exclude recurrent selective sweeps at these loci in nature as an explanation for their limited diversity because the genes and the particular amino-acid residues under positive selection in the LTEE have diverged slowly since the *Escherichia–Salmonella* split.

Results

Core Genes Are Functionally Important

To make consistent comparisons, we analyzed single-copy genes with homologs in all 60 fully sequenced genomes in the *E. coli* collection. For the purpose of our study, we define this set of panorthologous genes as the *E. coli* core genome and the set of all other genes as the flexible genome (Materials and Methods). We used published data from the Keio collection of single-gene knockouts in *E. coli* K-12 (Baba *et al.* 2006) to test whether the core genes tend to be functionally more important than the flexible genes based on

essentiality and growth yield. Core genes are indeed more essential than flexible genes (Welch's t = 6.60, d.f. = 3387.8, one-tailed $p < 10^{-10}$), and knockouts of core genes cause larger growth-yield defects than flexible genes in both rich (Welch's t = 3.79, d.f. = 3379, one-tailed $p < 0.0001$) and minimal media (Welch's t = 4.95, d.f. =3457.3, one-tailed $p < 10^{-6}$).

Core Genes Evolve Faster than Flexible Genes in the LTEE

We first examined the substitutions in single genomes sampled from each of the 12 LTEE populations after 40,000 generations. These genomes are part of a large dataset comprising 264 fully sequenced genomes from the first 50,000 generations of the LTEE (Tenaillon *et al.* 2016). Six of these populations had evolved greatly elevated point-mutation rates by 40,000 generations (Sniegowski *et al.* 1997; Wielgoss *et al.* 2013; Tenaillon *et al.* 2016). As a consequence of their much higher mutation rates, a much larger fraction of the mutations seen in hypermutable populations are expected to be neutral or even deleterious passengers (hitchhikers), as opposed to beneficial drivers, in comparison to those populations that retained the low ancestral point-mutation rate (Tenaillon *et al.* 2016). In genomes from the nonmutator populations, we observe an excess of nonsynonymous substitutions in the core genes. The core genes constitute ~48.5% of the total coding sequence in the genome of the LTEE ancestral strain, but 69% (105/152) of the nonsynonymous substitutions are in the core genes (Table 1, row 1, $p < 10^{-6}$). By contrast, the frequency of synonymous mutations does not differ significantly between the core and flexible genes (Table 1, row 2). Also, the frequencies of both nonsynonymous and synonymous substitutions in core versus flexible genes are close to the null expectations in the populations that evolved hypermutability (Table 1, rows 3 and 4).

These results indicate that core genes are evolving faster, on average, than the flexible noncore genome in the LTEE populations that retained the ancestral point-mutation rate. This faster evolution is consistent with some subset of the core genes being under positive selection to change from their ancestral state during the LTEE. To examine this issue more closely, we compared the rates of evolution of core genes observed in the LTEE with the rates of evolution of the same genes in the *E. coli* collection. As a measure of the rate of evolution of each core gene in the LTEE, we used a *G*-score, as calculated by

Tenaillon *et al.* (2016), that expresses the excess number of independent nonsynonymous mutations in nonhypermutable lineages relative to the number expected given the length of that gene's coding sequence (relative to all coding sequences) and the total number of such mutations.

We used two different measures for the rate of evolution of each core gene in nature. The first one is based on the level of nonsynonymous sequence diversity in the gene across the 60 sequenced genomes in the *E. coli* collection. There is a negative correlation between a core gene's *G*-score in the LTEE and its diversity in in the *E. coli* collection (Spearman-rank correlation $r = -0.0701$, two-tailed $p = 0.0019$; Fig. 1A). That is, core genes that evolved faster in the LTEE (i.e., higher *G*-scores) are significantly less diverse in the *E. coli* collection than those that evolved more slowly. Only 163 genes in the core genome had positive *G*-scores (i.e., one or more nonsynonymous mutations in nonhypermutable lineages) in the LTEE, and we do not find a significant correlation between the *G*-score and sequence diversity using only those genes (Spearman-rank correlation $r = -0.0470$, two-tailed $p = 0.5515$; Figure 1B). However, the 163 core genes with positive *G*-scores have significantly lower diversity in the *E. coli* collection than do the 1805 with zero *G*-scores (Mann-Whitney U = 125,660, two-tailed $p = 0.0020$; Fig. 1C). Hence, the difference between core genes with and without nonsynonymous substitutions in the nonmutator LTEE lineages largely drives the overall negative correlation.

By using segregating polymorphisms in the *E. coli* collection, our first measure of the rate of evolution of core genes in nature might be dominated by transient variation or local adaptation. By contrast, divergence between core genes found in different species has occurred over a longer timescale and should be less affected by these issues. Therefore, our second measure for the rate of evolution of core genes in nature uses the sequence divergence between *E. coli* and *Salmonella enterica*. We repeated the above analyses using the set of 2853 panorthologs—single-copy genes that map one-to-one across species (Lerat *et al.* 2003; Cooper *et al.* 2010)—for *E. coli* and *S. enterica*. We found a negative correlation across genes between their *G*-scores in the LTEE and interspecific divergence (Spearman rank-correlation $r = -0.0911$, two-tailed $p < 10^{-5}$; Figure 2A). This negative correlation remains significant even if we consider only those 210 panorthologs with positive *G*-scores in the LTEE (Spearman rank-correlation $r = -0.2567$, two-tailed $p = 0.0002$; Figure 2B). In

addition, the panorthologs with positive *G*-scores are less diverged between *E. coli* and *S. enterica* than the 2643 panorthologs with zero *G*-scores (Mann-Whitney U = 223,330, two-tailed $p < 10^{-5}$; Fig. 2C).

Taken together, these analyses contradict the hypothesis that those genes that have evolved fastest in the LTEE are ones that also evolve and diversify faster than typical genes in nature. Instead, they support the hypothesis that the genes involved in adaptation during the LTEE tend to be more conserved than typical genes in nature, presumably because they are constrained in nature by purifying selection. When the bacteria evolve under the simple and stable ecological conditions of the LTEE, these previously conserved genes undergo adaptive evolution that fits them to their new environment.

Protein Residues that Changed in the LTEE are also Conserved in Nature

It is possible that the substitutions in the LTEE occurred at highly variable sites in otherwise conserved proteins. To examine this issue, we asked whether nonsynonymous changes found in the nonmutator LTEE lineages at 40,000 generations tended to occur in fast-evolving codons. For the 66 proteins with such substitutions in the LTEE, we calculated the diversity at the mutated sites and in the rest of the protein for the 60 genomes in the *E. coli* collection. The sites that had changed in the LTEE were significantly less variable than the rest of the protein in that collection (Wilcoxon signed-rank test, $p < 10^{-5}$). In fact, only 7 of these 66 proteins had any variability at those sites in the *E. coli* collection, and they account for only 9 of the 105 amino-acid substitutions in those proteins in the 40,000-generation LTEE clones. We obtained similar results for the divergence between *E. coli* and *Salmonella.* In the 40,000-generation LTEE clones, 128 nonsynonymous substitutions occurred in 86 panorthologs, and only 5 of the LTEE substitutions were at diverged sites. These results demonstrate that particular residues under positive selection in the LTEE are ones that tend to be conserved in nature.

Discussion

It has been long known that, in nature, some genes evolve faster than others. In most cases, the more slowly evolving genes are core genes—ones possessed by most or all members of some species or higher taxon—and their relative sequence conservation reflects functional

constraints that limit the potential for the encoded proteins to change while retaining their functionality. As a consequence, the ratio of nonsynonymous to synonymous substitutions also tends to be low in these core genes. By contrast, we found that most nonsynonymous substitutions in nonmutator lineages of the LTEE occurred in core genes that are shared by all *E. coli* (Table 1). Moreover, even among the core genes, those that experienced positive selection to change in the LTEE are both less diverse over the *E. coli* species (Fig. 1) and less diverged between *E. coli* and *S. enterica* (Fig. 2) than core genes without substitutions in any of the nonmutator LTEE populations. Also, the particular sites where substitutions occurred during the LTEE are usually more conserved than the rest of the corresponding protein, excluding the possibility that substitutions occurred at a subset of fast-evolving positions in otherwise slow-evolving genes.

It is clear, then, that the specific conditions of the LTEE have favored new alleles in core genes that are usually highly conserved in nature. From one perspective, this result is surprising—the 37°C temperature of the LTEE is typical for the human and many other mammalian bodies in which *E. coli* lives; the limiting resource is glucose, which is *E. coli*'s preferred energy source, such that it will repress the expression of genes used to catabolize other resources when glucose is available; and the LTEE does not impose other stressors such as pH, antibiotics, or the like. However, the very simplicity and constancy of the LTEE conditions are presumably novel, or at least atypical, in the long sweep of *E. coli* evolution. In other words, the uniformity and simplicity of the laboratory conditions—including the absence of microbial competitors and parasites as well as host-dependent factors—stand in stark contrast to the variable and complex communities that are *E. coli*'s natural habitat (Blount 2015).

Given the importance and even essentiality of many core genes, it seems unlikely that most of the nonsynonymous mutations in the LTEE cause complete losses of function. Instead, we suspect that the mutations are beneficial because they fine-tune the regulation and expression of functions that contribute to the bacteria's competitiveness and growth in the simple and predictable environment of the LTEE. By contrast, some other genes that were repeatedly mutated in the LTEE—not by point mutations, but instead by deletions and transposable-element insertions—typically encode noncore, nonessential functions

including prophage remnants, plasmid-derived toxin-antitoxin modules, and production of extracellular structure that are probably important for host colonization (Tenaillon et al. 2016). Both types of change have been shown to be adaptive in the LTEE environment—the former by affecting a gene's function and the expression of interacting genes (Cooper et al. 2003, Philippe et al. 2007), and the latter by eliminating unused and potentially costly functions (Cooper et al. 2001).

Of course, other evolution experiments would generate different types of genomic changes, including in some cases probably a preponderance of point mutations in noncore genes. For example, if the experimental environment involves lethal agents such as phages or antibiotics, then perhaps only a few noncore genes might be the targets of selection, and the resulting mutations might be different from and even at odds with adaptation to other aspects of the environment (Scanlan et al. 2015). Similarly, adaptation to exploit novel resources—such as the ability to use the citrate that has been present throughout the LTEE, but which only one population has discovered how to use (Blount et al. 2008, Blount et al. 2012)—may produce a different genetic signature of adaptation. Yet other signatures might emerge if horizontal gene transfer from other strains or species provided another source of variation (Souza et al. 1997). Imagine a scenario in which gene flow allowed *E. coli* to obtain DNA from a diverse natural community; in that case, a transporter acquired from another bacterial species might well provide an easier pathway to use the citrate in the LTEE environment.

We can turn the question around from asking why core genes evolve so quickly in the LTEE, to asking why they usually evolve slowly in nature. Core genes encode functions that, by definition, are widely shared, and so their sequences have had substantial time to diverge across taxa (Biller *et al.* 2015) and become fine-tuned to different niches. As a consequence, there are fewer opportunities for new alleles of core genes to provide an advantage. Moreover, given the diversity of species (including transients) in most natural communities, extant species may usually fill any vacant niches that appear faster than *de novo* evolution. Nonetheless, mutations in conserved core genes might sometimes provide the best available paths for adaptation to new conditions, such as when formerly free-living or commensal bacteria become pathogens (Lieberman et al. 2011). In such cases, finding parallel or convergent changes offers a way to identify adaptive mutations when they occur

in core genes. For example, *E. coli* and *S. enterica* have been found to undergo convergent changes at the amino-acid level in core genes when strains evolve pathogenic lifestyles (Chattopadhyay *et al.* 2009; Chattopadhyay *et al.* 2012).

In summary, the genetic signatures of adaptation vary depending on circumstances including the novelty of the environment from the perspective of the evolving population, the complexity of the biological community in which the population exists, the intensity of selection, and the number and types of genes that can produce useful phenotypes. In the LTEE, nonsynonymous mutations in core genes that encode conserved and even essential functions for *E. coli* have provided a major source of the large fitness gains in the evolving populations over many thousands of generations (Wiser et al. 2013, Lenski et al. 2015).

Materials and Methods

Panortholog Identification in the *E. coli* Collection

We downloaded the nucleotide and amino-acid sequences from GenBank for 60 fully sequenced *E. coli* genome accessions (Table S1). We refer to this diverse set of clinical, environmental, and laboratory strains as the *E. coli* collection. We identified 1968 single-copy orthologous genes, or panorthologs, that are shared by all 60 strains in the *E. coli* collection using the pipeline described in Cooper *et al.* (2010). To guard against recent gene duplication or horizontal transfer events, we confirmed that none of these panorthologs had better local BLAST hits in any given genome. We refer to these panorthologs as core genes, and other genes that are present in only some of the *E. coli* collection as flexible genes.

The NCBI Refseq accession for the ancestor for the LTEE, *E. coli* B strain REL606, is NC_012967. The accession for the *S. enterica* strain used as an outgroup is NC_003197. We downloaded *E. coli* and *S. enterica* orthology information from the OMA orthology database (Altenhoff *et al.* 2015), only examining the one-to-one matches. For internal consistency, we also used the panortholog pipeline to generate one-to-one panorthologs between *E. coli* B strain REL606 and *S. enterica*. We analyzed the 2853 panortholog pairs that the pipeline and the OMA database called identically.

Analysis of the Keio Collection

We downloaded data on essentiality and growth yield in rich and minimal media for the Keio collection of single-gene knockouts in *E. coli* K-12 from the supplementary tables in the original paper describing the collection (Baba *et al.* 2006). We classified the knocked-out genes as panorthologs (i.e., core) or not (i.e., flexible), and we compared differences in essentiality and growth yield between the two sets of genes.

Nonsynonymous and Synonymous Substitutions in the LTEE at 40,000 Generations

We identified all substitutions in protein-coding genes in the genome sequences of single clones isolated from each of the 12 independently evolving populations of the LTEE at 40,000 generations. These data were reported in two previous studies (Maddamsetti *et al.* 2015; Tenaillon *et al.* 2016). Six of the 12 populations descend from REL606, and six descend from REL607 (Lenski *et al.* 1991). These ancestral strains differ by point mutations in the *araA* and *recD* genes (Tenaillon *et al.* 2016), and the two mutations were thus excluded from our analysis.

G Scores and Positive Selection on Genes in the LTEE

We use the *G*-score statistics reported in Supplementary Table 2 of Tenaillon *et al.* (2016) as a measure of positive selection at the gene level in the LTEE. The *G*-score for each gene reflects, in a likelihood framework, the number of independent nonsynonymous mutations in nonmutator lineages relative to the number expected given the length of that gene's coding sequence (relative to all coding sequences) and the total number of such mutations. In this analysis, the nonmutator lineages included the six LTEE populations that never evolved point-mutation hypermutability as well as lineages in the other populations before they became mutators. This analysis included whole-genome sequences from 264 clones isolated at 11 time points through 50,000 generations of the LTEE.

Sequence Diversity and Divergence

We adapted Nei's nucleotide diversity metric (Nei and Li 1979) for use with amino-acid sequences to reflect nonsynonymous differences. Specifically, we calculated the mean number of differences per site between all 1770 (i.e., 60 x 59 / 2) pairs of sequences in the protein alignments from the 60 genomes in the *E. coli* collection. In the site-specific analysis,

we calculated this diversity metric separately for the sites that evolved in the LTEE and those that did not, and we compared the values to see if the former also tended to vary in nature. For the sequence divergence between *E. coli* and *S. enterica*, we used the ancestral strain of the LTEE, REL606, as the representative *E. coli* genome in order to maximize the number of orthologous genes available in our analysis. The divergence for each gene was calculated as the proportion of amino-acid residues that differ between the two aligned proteins, where an amino-acid difference implies at least one nonsynonymous change in the corresponding codon since the most recent common ancestor of the two alleles.

Statistical Analyses

All data tables and analysis scripts will be deposited in the Dryad Digital Repository upon acceptance (doi:XXXX).

Acknowledgments

We thank Alita Burmeister, Michael Wiser, and Kyle Card for discussions and comments on earlier versions of our manuscript. This work was supported, in part, by a National Defense Science and Engineering Graduate Fellowship to R.M.; a grant from the National Science Foundation (DEB-1451740) to R.E.L.; and the BEACON Center for the Study of Evolution in Action (National Science Foundation Cooperative Agreement DBI-0939454).

References

Altenhoff AM, Škunca N, Glover N, Train CM, Sueki A, Piližota I, Gori K, Tomiczek B, Müller S, Redestig H, et al. 2015. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* 43:D240–249.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol.* 2:2006.0008.

Bailey SF, Bataillon T. 2016. Can the experimental evolution program help us elucidate the genetic basis of adaptation in nature? *Mol Ecol.* 25:203–218.

Barrick JE, Lenski RE. 2013. Genome dynamics during experimental evolution. *Nat Rev Genet.* 14:827–839.

Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. *Prochlorococcus:* the structure and function of collective diversity. *Nat Rev Microbiol.* 13:13–27.

Blount ZD. 2015. The unexhausted potential of *E. coli eLife* 4:e05826.

Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV. 2009. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc Natl Acad Sci U S A.* 106:12412–12417.

Chattopadhyay S, Paul S, Kisiela DI, Linardopoulou EV, Sokurenko EV. 2012. Convergent molecular evolution of genomic cores in *Salmonella enterica* and *Escherichia coli*. *J Bacteriol.* 194:5002–5011.

Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 100:1072–1077.

Cooper VS, Schneider D, Blot M, Lenski RE. 2001. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *E. coli* B. *J Bacteriol.* 183:2834–2841.

Cooper VS, Vohr SH, Wrocklage SC, Hatcher PJ. 2010. Why genes evolve faster on secondary chromosomes in bacteria. *PLoS Comput Biol.* 6:e1000732.

Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am Nat.* 138:1315–1341.

Lenski RE, Wiser MJ, Ribeck N, Blount ZD, Nahum JR, Morris JJ, Zaman L, Turner CB, Wade BD, Maddamsetti R, Burmeister AR, Baird EJ, Bundy J, Grant NA, Card KJ, Rowles M, Weatherspoon K, Papoulis SE, Sullivan R, Clark C, Mulka JS, Hajela N. 2015. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. *Proc R Soc Lond B.* 282:20152292.

Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the γ-Proteobacteria. *PLoS Biol.* 1: e19.

Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Skurnik D, Leiby N, LiPuma JJ, Goldberg JB, McAdam AJ, Priebe GP, Kishony R. 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Gen.* 43:1275–1280.

Maddamsetti R, Hatcher PJ, Cruveiller S, Médigue C, Barrick JE, Lenski RE. 2015. Synonymous genetic variation in natural isolates of *Escherichia coli* does not predict

where synonymous mutations occur in a long-term evolution experiment with *Escherichia coli*. *Mol Biol Evol.* doi:10.1093/molbev/msv161.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 76:5269–5273.

Philippe N, Crozat E, Lenski RE, Schneider D. 2007. Evolution of global regulatory networks during a long-term experiment with *Escherichia coli*. *BioEssays* 29:846–860.

Scanlan PD, Hall AR, Blackshields G, Friman VP, Davis MR Jr, Goldberg JB, Buckling A. 2015. Coevolution with bacteriophages drives genome-wide host evolution and constrains the acquisition of abiotic-beneficial mutations. *Mol Biol Evol.* 32:1425–1435.

Sniegowski PD, Gerrish PJ, Lenski RE. 1997. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387:703–705.

Souza V, Turner PE, Lenski RE. 1997. Long-term experimental evolution in *Escherichia coli*. V. Effects of recombination with immigrant genotypes on the rate of bacterial evolution. *J Evol Biol.* 10:743–769.

Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Médigue C, *et al.* 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. bioRXiv preprint http://dx.doi.org/10.1101/036806

Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D. 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A.* 110:222–227.

Wiser MJ, Ribeck N, Lenski RE. 2013. Long-term dynamics of adaptation in asexual populations. *Science* 342:1364–1367.

Table 1. Nonsynonymous Mutations Overrepresented in the Core Genome of Nonmutator LTEE Populations.

Category and Population	Core	Flexible	Odds Ratio	Significance
Nonsynonymous mutations in nonmutator populations	105	47	2.37	$p < 10^{-6}$
Synonymous mutations in nonmutator populations	10	15	0.71	$p = 0.4297$
Nonsynonymous mutations in mutator populations	2038	2247	0.96	$p = 0.2273$
Synonymous mutations in mutator populations	845	880	1.02	$p = 0.6822$

NOTE—The length of the core and flexible (i.e., noncore) portions of the coding sequences in the genome of the LTEE ancestor (*E. coli* strain REL606) are 1,944,921 and 2,066,263 bp, respectively. Data show the numbers of mutations found in the core and flexible portions in genomes sampled and sequenced at 40,000 generations from six nonmutator populations that retained the ancestral point-mutation rate and six mutator populations that evolved hypermutability. The odds ratio expresses the extent to which the category of mutation is overrepresented (>1) or underrepresented (<1) in the core genome relative to the flexible genome in the indicated populations. The *p*-value is based on a two-tailed binomial test comparing the observed numbers of mutations to the expectations based on the relative lengths of the core and flexible genomes.

FIG. 1. Relationship between positive selection in the LTEE and nonsynonymous sequence diversity of core genes in the *E. coli* collection of 60 clinical, environmental, and laboratory strains. The *G*-score provides a measure of positive selection based on the excess of nonsynonymous substitutions in the LTEE lineages that retained the ancestral point-mutation rate. The $\log_{10}$ and square-root transformations of the *G*-score and sequence diversity, respectively, improve visual dispersion of the data for individual genes, but they do not affect the nonparametric tests performed, which depend only on rank order. (A) *G*-scores and sequence diversity are negatively correlated across all 1968 core genes (Spearman-rank correlation, $p = 0.0019$). (B) The correlation becomes not significant using only the 163 genes with positive *G*-scores (Spearman-rank correlation, $p = 0.5515$). (C) The 163 core genes with positive *G*-scores in the LTEE have significantly lower nonsynonymous sequence diversity in natural isolates than the 1805 genes with zero *G*-scores (Mann-Whitney test, $p = 0.0020$). Error bars show 95% confidence intervals around the median.

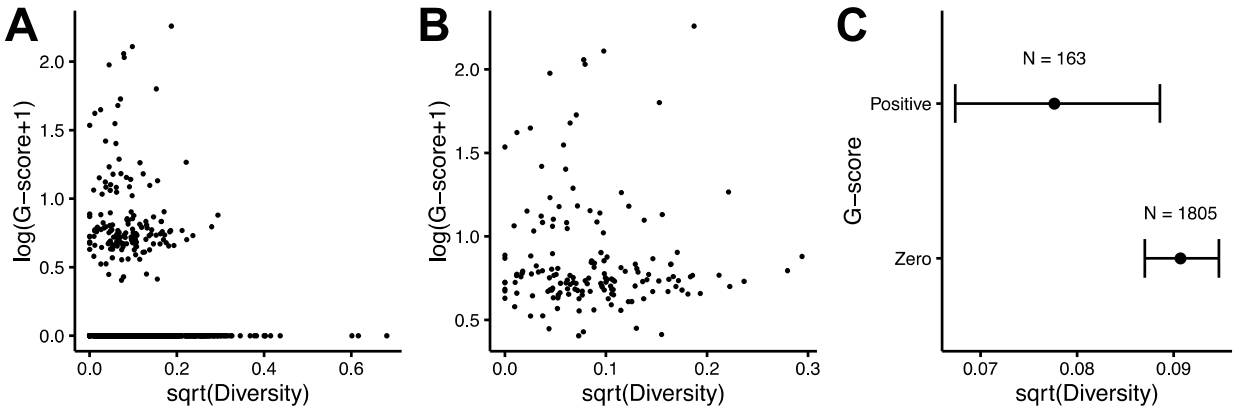
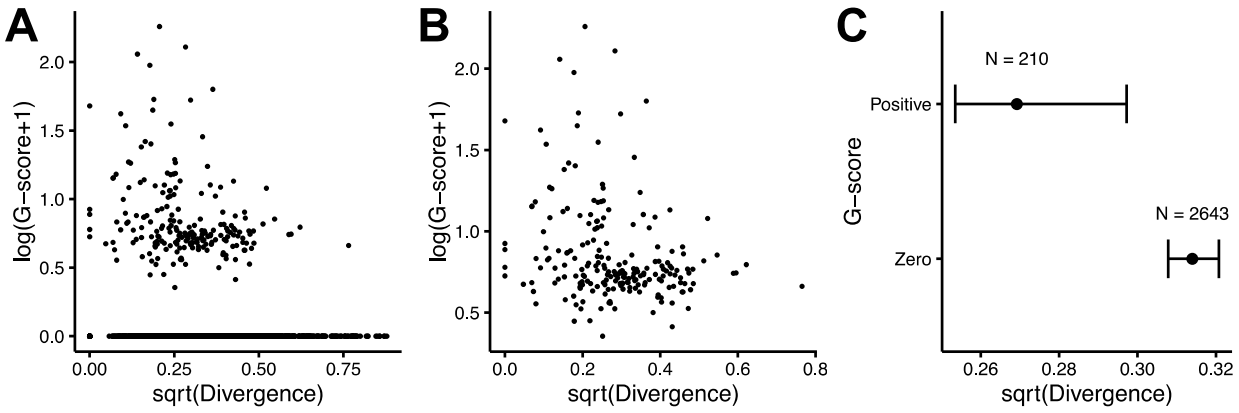


FIG. 2. Relationship between positive selection in the LTEE and nonsynonymous sequence divergence of panorthologs between *E. coli* (strain REL606) and *S. enterica*. REL606 is the common ancestor of the LTEE populations. See Fig. 1 for additional details. (A) *G*-scores and divergence are negatively correlated across all 2853 panorthologs (Spearman-rank correlation, $p < 10^{-5}$). (B) The correlation remains significant even using only the 210 panorthologs with positive *G*-scores (Spearman-rank correlation, $p = 0.0002$). (C) The 210 panorthologs with positive *G*-scores in the LTEE are significantly less diverged between *E. coli* and *S. enterica* in natural isolates than the 2643 panorthologs with zero *G*-scores (Mann-Whitney test, $p = < 10^{-5}$). Error bars show 95% confidence intervals around the median.



Supplementary File 1. NCBI Refseq ID, strain name, and lifestyle (commensal or pathogen) for the collection of 60 strains with complete genome sequences used to identify the set of panorthologs that represent the core genome of *Escherichia coli*.