A simple test to infer mode of selection in metagenomics time series of evolving asexual populations

Rohan Maddamsetti[1, †*] and Nkrumah A. Grant[2,3,†*]

[1]Department of Biological Sciences, Old Dominion University, Norfolk VA
[2]BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI
[3]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI
[†]Joint first author
[*]Correspondence to: rohan.maddamsetti@gmail.com, grantnkr@msu.edu

ABSTRACT

We introduce a simple test to infer mode of selection (STIMS) in metagenomic time series of evolving asexual populations. STIMS compares the tempo of molecular evolution for a gene set of interest against a null distribution that is bootstrapped on random gene sets. We test STIMS on metagenomic data spanning 62,750 generations of Lenski's long-term evolution experiment with *E. coli* (LTEE). Our method successfully recovers signals of purifying selection and positive selection on gold standard sets of genes. We then use STIMS to study the evolution of genetic modules in the LTEE. We find strong evidence of ongoing positive selection on key regulators of the *E. coli* gene regulatory network. Key regulatory genes show evidence of positive selection over the entire time series, even in some hypermutator populations. By contrast, we found no signal of selection on the genetic modules that show the strongest transcriptional responses to changes in growth conditions. In addition, the cis-regulatory regions of key regulators are evolving faster than the cis-regulatory regions of their downstream regulatory targets. These results indicate that one mechanistic cause for ongoing fitness gains in the LTEE is ongoing fine-tuning of the gene regulatory network.

Keywords: positive selection, purifying selection, experimental evolution, metagenomics, gene regulatory network

INTRODUCTION

Our understanding of the tempo and mode of genetic and phenotypic change in evolving populations remains incomplete (Simpson 1944). In part, our incomplete understanding reflects the fragmentary evolutionary record that has been pieced together from fossils, ancient DNA, and the biodiversity of extant species. New progress has been driven by technologies such as high-throughput genome sequencing and phenotyping, which allow scientists to observe and measure rates of evolutionary change in real-time. In particular, the ability to sample genetic and phenotypic diversity that lie off the (eventual) line of descent in evolving populations gives unprecedented power to synthesize our understanding of the proximate causes of evolutionary change (the molecular, structural, and systems biology of viruses, cells, and organisms) with the ultimate causes of evolutionary change (forces such as natural selection, mutation, recombination, and gene flow) in the broader context of ecological communities and ecosystems.

In this work, we present STIMS: a Simple Test to Infer Mode of Selection in metagenomic time series of evolving asexual populations of microbes. Our idea is an old one: we can use measurements of the tempo of evolutionary change to make inferences about the mode of evolutionary change (e.g. positive selection or purifying selection). We developed STIMS in the context of an ongoing long-term evolution experiment with *Escherichia coli* (LTEE) (Lenski, et al. 1991; Lenski 2017).

The LTEE has become an important test bed for many fundamental questions in evolutionary biology, afforded by its "frozen fossil record" on which new technologies can be applied to study bacterial populations frozen decades ago. Previous studies have used the LTEE as a model system to study the tempo and mode of both genomic (Barrick and Lenski 2009; Barrick, et al. 2009; Maddamsetti, et al. 2015; Tenaillon, et al. 2016; Good, et al. 2017) and phenotypic evolution (Lenski, et al. 1991; Vasi, et al. 1994; Elena, et al. 1996; Lenski, et al. 1998; Wiser, et al. 2013) (Grant et al. 2020a on aerobic/anaerobic fitness; Grant et al. 2020b on cell size).

Many of those previous studies focused on the evolutionary dynamics underlying adaptive evolution. It has been noted that the evolution of hypermutator strains with elevated mutation rates in many LTEE populations has obscured the genomic signatures of adaptation (Barrick and Lenski 2009; Tenaillon, et al. 2016; Couce, et al. 2017; Good, et al. 2017;

Maddamsetti, et al. 2017). This thinking is based on the finding in large asexual populations that large numbers of neutral and nearly-neutral variants hitchhike to high frequency by virtue of being linked with highly beneficial driver mutations. In this regime, called "emergent neutrality" (Schiffels, et al. 2011), it is difficult to tell the difference between genes evolving neutrally from those under weak to moderate selection. Conventional wisdom holds that it is challenging, if not impossible, to detect any signal of natural selection operating in the hypermutator populations of the LTEE. One goal of this paper is to overturn this consensus. In particular, we wanted to rigorously understand earlier observations that suggested the action of purifying selection on hypermutator populations of the LTEE (Grant et al. 2020a).

In that work, which examines how the fitness of LTEE strains in an anoxic environment has changed over 50,000 generations of relaxed selection on anaerobic growth, we compared the number of mutations occurring in aerobic-specific genes to the number of mutations in anaerobic-specific genes across the LTEE populations. In the non-mutator populations, aerobic-specific genes are enriched in mutations compared to anaerobic-specific genes, which is reasonable as the LTEE has taken place in a strictly oxic environment for more than three decades (Grant et al. 2020a). Surprisingly, we also found that aerobic-specific genes were depleted in mutations compared to anaerobic-specific genes in 4 out of the 6 hypermutator populations. Furthermore, aerobic-specific genes were depleted in mutations compared to a bootstrapped null distribution in three hypermutator populations: Ara−2, Ara+3, and Ara+6. This finding suggests that aerobic-specific genes may have transitioned from positive selection to purifying selection in those hypermutator LTEE populations.

In this paper, we fully describe and generalize the method that we developed for studying the evolution of aerobic-specific and anaerobic-specific genes in the LTEE (Grant et al. 2020a). We asked whether we could use STIMS to examine the *tempo* of molecular evolution in different functional modules of genes, and generate testable hypotheses about the *mode* (i.e. purifying or positive selection) of evolution on those gene modules. In order to explore these ideas, we examine the distribution of mutations over time per module, and how that distribution compares to a null distribution generated by sampling random gene sets.

Since we are interested in inferring the action of natural selection, we need to take the effects of local mutation biases into account (Bailey, et al. 2018). One can control for local mutation biases by constructing null distributions that model chromosomal variation in mutation

rates and biases. This can be quite complicated, given that mutational biases (Couce, et al. 2017) and regional mutation rates over the chromosome have evolved idiosyncratically across the replicate LTEE populations (Maddamsetti and Grant 2020). In this paper, we sample gene sets uniformly over the *E. coli* genome. We also implemented a sampling procedure that take the wave pattern of mutation rate variation over the *E. coli* chromosome into account (Foster, et al. 2013; Niccum, et al. 2019) (Maddamsetti and Grant 2020); in any case, the use of this more complicated procedure produced the same results as sampling genes uniformly over the chromosome (Materials and Methods).

RESULTS

Conceptual framework for how evolutionary mode affects evolutionary tempo in asexual populations

The key premise of STIMS is that evolutionary rates (i.e. tempo of evolution) differ across different gene sets, and that those differences reflect variation in the underlying selection pressures (i.e. mode of evolution) in those gene sets. In this section, we describe the patterns that we expect different modes of evolution to leave in metagenomics time-series of evolving asexual populations (Figure 1).

Each panel of Figure 1 must be interpreted in relation to the background rate of evolution in a given population. Evolutionary rates depend on population genetic parameters, such as mutation rates and population size. By resampling evolutionary rates in random sets of genes over the genome, a null distribution of evolutionary rates can be generated. The grey area in each figure represents this null distribution of evolutionary rates, integrated over time. Gene modules with trajectories above the grey region have significantly higher evolutionary rates compared to the background distribution, while gene modules with trajectories below the grey region have significantly lower evolutionary rates compared to the background distribution. Figure 1 specifically assumes that the background distribution is for a large asexual population that has evolved a hypermutator phenotype, as that is the case most relevant for this work. We note that the majority of mutations occurring in non-mutator populations are expected to be driven by positive selection, and so the background distribution for non-mutator populations should reflect weak to moderate positive selection.

Model 1: obscured selection. Gene modules that show this pattern are evolving in a manner that is consistent with the null distribution of evolutionary rates (Figure 1A). For this reason, one cannot infer the mode of evolution affecting this module.

Model 2: purifying selection. Purifying selection at a locus will restrict the effective mutational target size of that locus. Therefore, we expect gene modules under strong purifying selection to accumulate fewer mutations over time than expected under the null distribution (Figure 1B).

Model 3: positive selection. Loci under strong positive selection in the LTEE often show multiple beneficial mutations at the same time in the population. For example, multiple alleles of *pykF* occur on distinct competing lineages in Ara−1, in a process called clonal interference (Barrick and Lenski 2009; Maddamsetti, et al. 2015) For this reason, loci under strong positive selection should show an excess number of observed mutations in comparison to the null distribution. Once mutations at a locus under selection go to fixation however, the distribution of fitness effects at that locus may significantly change. We depict three separate scenarios of positive selection, depending on how the distribution of fitness effects changes after the fixation of mutations in that gene or gene module.

Model 3A: Transition from positive selection to effective neutrality. Imagine a gene in which knockouts are highly beneficial. In that case, we would expect a high rate of parallel knockout mutations at that locus in the population, higher than the background expected under mutation accumulation. However, after one of those knockout mutations has fixed, then the vast majority of further mutations would be completely neutral, and that locus would be under relaxed selection. In this scenario, we expect an early burst of observed mutations if knockouts are highly beneficial, followed by a transition, after fixation of one of those knockouts, to the molecular clock set by mutation accumulation (Figure 1C). We call this scenario "coupon collecting for knockouts", as no further beneficial mutations remain at the locus once its knockout has fixed in the population. This particular inference may not be clear cut, however. A dramatic increase in mutation rates could also put a gene that is still undergoing positive selection into the "effective neutrality" regime discussed earlier, causing the pattern seen in Figure 1C.

Model 3B: positive selection for a finite number of fine-tuning mutations. In this scenario, imagine a gene in which a small number of mutations are highly beneficial, and no

further beneficial mutations remain at that locus once those have fixed in the population (Figure 1D). We call this scenario "coupon collecting for fine-tuning mutations", as once beneficial mutations have fine-tuned the function of a protein, any further mutations are deleterious. In this case, we expect an early burst in tempo at that locus, then very slow tempo due to a transition from positive to purifying selection.

Model 3C: ongoing positive selection for continual fine-tuning mutations. Imagine a gene under positive selection, where new beneficial mutations continuously become available (Figure 1E). We call this scenario continual fine-tuning. There are a number mechanisms that could underpin this scenario. First, it may be that further mutations optimize gene function, or affect additional functions of the gene. Second, further beneficial mutations could compensate for deleterious mutations at the locus or elsewhere in the genome. Third, further beneficial mutations could repair deleterious consequences of antagonistic pleiotropic mutations in the genome (that is, mutations that have a net fitness benefit despite some fitness cost). Under continual fine-tuning, we expect to see a consistently faster tempo of mutation accumulation over time in comparison to the null distribution. Under this model, positively selected genes or modules continue to acquire mutations, even after fixations within the locus or module.

Finally, it is possible that some genes become the target of positive selection contingent upon a particular set of mutations elsewhere in the genome. We term this scenario "historically contingent synergistic epistasis". Under this scenario, we expect idiosyncratic changes of tempo in some populations but not others, reflecting windows of opportunities that may open or close unpredictably, based on the consequences of mutations occurring elsewhere in the genome. We do not draw a picture of this scenario as it is more complicated and less relevant for our findings. Discrete change points in the trajectory of observed mutations in a gene module would reflect the action of epistasis and historical contingency.

In practice, all such patterns need to be interpreted in the context of appropriate positive control experiments to calibrate inferences. In the following sections, we describe the results of three computational control experiments, using three gold standard sets of genes under relaxed, purifying, and positive selection in the LTEE.

Control 1: Resolving relaxed selection in the LTEE

As a first positive control, we examined 63 genes that encode cryptic peptides, show little gene expression, or have experimental evidence for neutral fitness effects in REL606 (A. Couce, personal communication). Figure 2 shows the results of applying STIMS on these genes, which we suppose are evolving under relaxed selection in the LTEE. In 5 out of 6 non-mutator populations (top two rows), the trajectory of this gene set falls within the null distribution. The remaining non-mutator population, Ara+5, has no mutations in this gene set, and so falls significantly below the null distribution (two-tailed non-parametric bootstrap: approximate $p < 0.005$). In 5 out of 6 hypermutator populations (bottom two rows), the trajectory of this gene set falls within the null distribution. The remaining hypermutator population, Ara+6 lies at the margin of the null distribution (two-tailed non-parametric bootstrap: approximate $p = 0.06$).

These findings are consistent with previous work. In the non-mutator populations, the global tempo of observed mutations is best explained by the occurrence and fixation of beneficial mutations. In contrast, the global tempo of observed mutations in the hypermutator populations is best explained by large numbers of nearly neutral mutations hitchhiking to high frequency with a much smaller number of beneficial mutations (Tenaillon, et al. 2016).

Control 2: Detecting purifying selection in the LTEE

As a second control, we examined genes that were experimentally shown to be essential or near essential in REL606 under LTEE conditions (Couce, et al. 2017). Many of the genes under strongest positive selection in the LTEE are also essential genes (Maddamsetti, et al. 2017), thus we excluded 24 essential genes with evidence of parallel evolution in the LTEE (Materials and Methods). We hypothesized that the remaining 491 essential genes would show evidence of purifying selection in the hypermutator populations, such that those loci would be depleted in mutations in comparison to the null distribution. Indeed, these 491 loci show significant depletion of observed mutations in all 6 hypermutator populations (one-tailed non-parametric bootstrap: $p < 0.0001$ for Ara−1; $p < 0.01$ for Ara−2; $p < 0.01$ for Ara−3; $p < 0.005$ for Ara−4; $p < 0.02$ for Ara+3; $p < 0.0001$ for Ara+6). This result shows that STIMS can indeed detect the action of purifying selection in populations with a hypermutator phenotype. Of the six non-mutator populations, Ara+1 shows a marginally significant depletion of mutations at these loci (one-tailed non-parametric bootstrap: $p = 0.048$).

Despite the global signal of purifying selection on these 491 essential genes, it is not clear which ones in particular are under purifying selection. The best candidates are the ones that have no mutations whatsoever in the LTEE, while the second-best candidates are those that were only affected by synonymous mutations. We found 65 genes with no observed mutations in the metagenomics data, 18 of which are essential. 105 have only synonymous mutations, 30 of which are essential. We examined 490 essential genes, out of 3948 genes in the genome that passed our filters (Materials and Methods). Based on these numbers, there is a significant association between essentiality and having no observed mutations (Fisher's exact test: $p = 0.0007$) and between essentiality and only having synonymous mutations (Fisher's exact test: $p < 10^{-5}$). The first set of candidates for purifying selection in the LTEE are reported in Supplementary Table S1, and the second set of candidates are reported in Supplementary Table S2. As one would expect, many of these genes encode proteins that catalyze key biological reactions, such as ATP synthase and ribosomal proteins.

Control 3: Verifying positive selection in the LTEE

As a final control, we examined genes that have previously been shown to be evolving under positive selection in the LTEE. We examined the genes with the strongest signal of parallel evolution in sampled over the first 50,000 generations of the LTEE (Tenaillon, et al. 2016). We first consider the top 50 G-scoring genes in the non-mutator populations, then we consider the top 50 G-scoring genes in the hypermutator populations.

STIMS successfully recovered a very strong signal of positive selection in all 12 populations, as can be seen in Figure 4 (one-tailed non-parametric bootstrap: $p < 0.03$ in every population). We can additionally ask how these genes evolve between 50,000 and 62,750 generations. Are these genes under continued positive selection, or do they switch over to purifying selection? Or, do they simply evolve at a rate indistinguishable from the background rate?

The results are mixed (Figure 5). After 50,000 generations, three non-mutator populations, Ara+1, Ara+2, and Ara+4 continue to show an excess of mutations, indicative of positive selection even after 50,000 generations. Two hypermutator populations, Ara−3 and Ara−4, show a significant depletion of mutations. The pattern, if any, in the remaining populations, is indistinguishable from background.

We also examined the genes with the strongest signal of parallel evolution in the hypermutator genomes in the same fashion (Supplementary Figures S1 and S2). In the first 50,000 generations, the non-mutator populations accumulate mutations in the top 50 G-scoring genes in hypermutator genomes at a rate similar to background. As expected, 5 out 6 hypermutator populations show a strong excess of mutations above background (Supplementary Figure S1; one-tailed non-parametric bootstrap: $p < 0.005$ in Ara−1, Ara−2, Ara−4, Ara+3, Ara+6). In the remaining population, Ara−3, mutations in this gene set accumulate at the background rate. It is unclear why this population is an outlier, but this finding may be related to the evolution of a citrate metabolic innovation that modified the ecology of this population (Turner, et al. 2015). After 50,000 generations, there are few to no mutations in this gene set in the non-mutator populations, while mutations accumulate at the background rate in 5 out of 6 hypermutator populations (Supplementary Figure S2).

Application to module decompositions of the E. coli genome

Our control experiments demonstrate that STIMS is powerful enough to recover signals of purifying and positive selection. In this section, we use STIMS to examine how different modules of genes in the *E. coli* genome are evolving in the LTEE. Our goal is to develop testable hypotheses for which genetic modules and biochemical pathways underlie the ongoing fitness gains observed in the LTEE (Lenski, et al. 2015). We examined three different module decompositions of the *E. coli* genome. The first partitioned the gene regulatory network into independent components based on gene expression (Sastry, et al. 2019); the second partitioned the *E. coli* proteome into sectors based on quantitative mass spectrometry (Hui, et al. 2015); while the third reported sets of genes ("eigengenes") whose expression best predicted *E. coli* growth rates (Wytock and Motter 2019).

<u>Regulators of I-modulons are under positive selection.</u> Sastry et al. (2019) used independent component analysis on the *E. coli* transcriptome, sampled across diverse conditions and strains, to infer 92 independent modules (called I-modulons) in the *E. coli* gene regulatory network (GRN). 61 of the 92 I-modulons correspond to known transcription factors and regulators in the GRN. We asked two questions: first, is there a difference between how the regulators and regulated genes in the I-modulons evolve in the LTEE? Second, are any I-modulons enriched with or depleted of mutations in the LTEE?

We find that transcriptional regulators of the I-modulons are under strong positive selection in the LTEE, while the genes in the I-modulons tend to accumulate mutations at the expected genomic background rate over time (Figure 6). In all 12 populations, regulators of I-modulons are under stronger selection than the genes in the I-modulons themselves. I-modulon regulators are under very strong selection in all 6 non-mutator populations (one-tailed non-parametric bootstrap: $p < 0.001$ in all non-mutator populations), and show evidence of ongoing positive selection in some hypermutator populations as well (one-tailed non-parametric bootstrap: $p = 0.03$ in Ara−3, $p \leq 0.0001$ in Ara+3 and Ara+6). By contrast, the evolution of genes within I-modulons fits the null distribution well. These results indicate that the upper levels of the regulatory hierarchy of the *E. coli* GRN are under strong selection in the LTEE, while the effector genes at the lower levels of the gene regulatory hierarchy are under much weaker selection.

Figure 6 also shows that the tempo of I-modulon regulator evolution is quite variable across populations. Consider the number of mutations in I-modulon regulators in the non-mutator populations. After 40,000 generations, Ara+1 has 5, Ara+2 has 4, Ara+4 has 2, and Ara+5 has 1, while Ara−5 and Ara−6 have none at all. Variation in the tempo of evolution of I-modulon regulators is also seen in the hypermutator populations, in particular the strong signal of ongoing positive selection found in Ara+3 and Ara+6 but not in the others.

While the overall trend is that I-modulons evolve at background rates, some I-modulons show evidence of idiosyncratic selection in the hypermutator populations. While these findings are not significant due to the large number of statistical tests, we hypothesize that some of these cases represent historical contingency and epistasis in the LTEE. These hypotheses could be tested in future studies. The results for all I-modulons are provided in Supplementary File 1, while particular I-modulons of interest are listed in Supplementary Table 3.

<u>One proteome sector shows evidence of purifying selection.</u> We also examined proteome sectors that were identified by quantitative mass spectroscopy during carbon-, nitrogen-, and ribosome-limiting growth conditions (Hui, et al. 2015). Figure 7 shows our findings. No proteome sectors showed evidence of positive selection. One proteome sector was significantly depleted in mutations in Ara−1, Ara+3, and Ara+6, suggesting purifying selection. This one, called the U-sector, contains genes that were not upregulated by any of the growth-limitation treatments in (Hui, et al. 2015). Rather, the expression of the proteins in the U-sector show a

generic positive correlation with growth rate across all of the conditions tested by (Hui, et al. 2015).

No evidence of selection on *E. coli* eigengenes. Finally, we examined the sets of genes (called eigengenes), that were most predictive of *E. coli's* growth rate (and presumably fitness under LTEE conditions), using different *E. coli* strains and growth environments (Wytock and Motter 2019). No eigengenes showed any evidence of selection in the LTEE (Figure 8).

Summary of module decomposition results. Altogether, we find that the protein-coding genes with the strongest transcriptional response to changing conditions show little to no evidence of selection in the LTEE. By contrast, top-level transcriptional regulators showed strong signatures of positive selection in all non-mutator populations, as well as in some hypermutator populations. Based on these findings, we hypothesize that the genes that show the strongest transcriptional response to changing conditions tend to be at the lower levels of the gene regulatory network, while natural selection in the LTEE is largely operating on regulators of those genes.

Cis-regulatory regions of key regulators show significantly more parallel evolution than the cis-regulatory regions of downstream targets

If upper levels of the gene regulatory network are under stronger selection in the LTEE, then cis-regulatory regions of I-modulon regulators should also show more evidence of positive selection than the genes within I-modulons. We tested this prediction by examining non-coding mutations associated with I-modulon regulators and their downstream targets. These non-coding mutations occurred within 100 bp upstream of their annotated gene (Good, et al. 2017). There are 48 non-coding mutations associated with 70 I-modulon regulators, and 542 non-coding mutations associated with 1394 genes regulated within I-modulons. These data are consistent with our prediction (Binomial test with 48 successes out of 590 trials and expected probability of success = 70/1394: $p = 0.001$).

DISCUSSION

In this paper, we show that one can infer the *mode* of genetic change in genetic modules of interest by analyzing their *tempo* of genetic change, as recorded in metagenomic time series of

evolving asexual populations. We test STIMS using data from Lenski's long-term evolution experiment with *E. coli* (LTEE), but we expect that our basic idea could be more rigorously justified by deeper theoretical work, and further extended to study metagenomics time series of evolving populations and communities in both the laboratory and in the wild. In particular, we anticipate that STIMS could be applied to clinical time-series, in order to discover gene modules that are under selection in pathogens as they adapt to their host (Chung, et al. 2017; Cooper, et al. 2020).

Our findings add to the body of work on the relationship between the tempo and the mode of evolution in the LTEE. Barrick et al. (2019) reported that the tempo of genomic change in Ara–1 was roughly linear (which is a prediction of the neutral theory of molecular evolution), despite strong evidence of adaptive evolution in the LTEE. Barrick *et al*. also suggested that some mutations, such as those in global regulators, those affecting DNA topology, or the stringent response mediated by (p)ppGpp concentrations in the cell, may be highly pleiotropic and may thus cause maladaptive side effects, which could introduce new opportunities for compensatory adaptations. By considering the evolution of genomes along the phylogeny of the LTEE, Tenaillon *et al.* (2016) found that the tempo of genome evolution in the LTEE was well fit by a model accounting for both the linear accumulation of neutral mutations and a power-law accumulation of beneficial mutations. Good *et al.* (2017) produced the LTEE metagenomics data that we analyze here, reported extensive evidence for eco-evolutionary dynamics, and found that evidence for epistasis and historical contingency in the evolution of several genes under positive selection in the LTEE.

In contrast to these previous studies, which examined global patterns in the tempo and mode of evolution in the LTEE, our analysis focuses on functional modules encoded by the *E. coli* genome. We ask whether the tempo of evolution in particular molecular subsystems can give us insight into how those modules evolve (for example, relaxed, purifying, or positive selection) in real time. Our control experiments confirm that STIMS works on genes that are hypothesized to be under relaxed (Figure 2) and purifying selection (Figure 3), and genes known to be under positive selection (Figure 4).

Our work also contrasts with previous work that has shown that the signal of selection is obscured in hypermutator genomes due to genomic draft (Couce, et al. 2017). We find that the accelerated pace of genomic evolution in the hypermutators, combined with that the fine detail of

data on molecular evolution provided by metagenomic time series, may open new opportunities for understanding the genomic basis of adaptive evolution. How can this be so? First, the vast number of nearly neutral hitchhikers that are observed in hypermutator populations provides insight into how mutations in modifier alleles affect genome-wide and local mutation rates and biases (Maddamsetti and Grant 2020). Second, regions of the genome that are highly depleted in mutations in the hypermutator populations are strong candidates for purifying selection. Third and most importantly— metagenomic sequencing gives deep sampling of genetic variation off the line of descent.

We found evidence of purifying selection in the hypermutator populations of the LTEE. Many of the genes that are strongest candidates for purifying selection are deeply conserved over evolutionary time, such as those encoding ribosomal subunits. The action of purifying selection on the hypermutator populations is also consistent with the observation that anti-mutator alleles have fixed in several populations of the LTEE (Wielgoss, et al. 2013) (Maddamsetti and Grant 2020). Together, these observations suggest that hypermutator lineages are accumulating some deleterious passenger mutations, even as their fitness continues to increase. We caution that careful work is needed to understand the evolution of antimutators in the LTEE, as it is possible that the evolution of anti-mutator alleles is at least in part being driven by genomic draft, as the fitness consequences of hypermutator and antimutator alleles depends on the magnitude of the mutation load affecting hypermutator lineages (Good and Desai 2016; Couce and Tenaillon 2019; Ramiro, et al. 2020). In any case, our analysis shows that the depletion of mutations in particular genes is not due to the evolution of anti-mutator alleles: by bootstrapping a null distribution of background rates, STIMS controls for the genome- and population-wide effects of anti-mutator alleles.

When we applied STIMS to different module decompositions of the *E. coli* genome, we found compelling evidence of strong positive selection on key global regulators of the *E. coli* gene regulatory network, especially in comparison to the genes that they regulate. Furthermore, we found an excess of mutations in the cis-regulatory regions of those regulators in comparison to the genes that they regulate. One explanation could be that mutations that affect the cis-regulation and structure of global regulators at higher levels of the GRN cause a cascade of effects on downstream targets, and so are more effective targets for fine-tuning the *E. coli* GRN.

Strikingly, we found evidence of continued strong positive selection on key regulatory genes in the two populations with the most mutations in the LTEE: Ara+3 and Ara+6 (Figure 6).

Further work will be needed to determine whether there is a significant mutation load in the LTEE, and if there is a connection between purifying selection and ongoing evolution of key regulators in Ara+3 and Ara+6. Ongoing positive selection on the *E. coli* GRN could also be due to compensatory evolution that is not specifically for hitchhiking deleterious mutations. For instance, it is possible that early beneficial mutations become deleterious due to further mutations (Zee, et al. 2014; Quandt, et al. 2015). Furthermore, natural selection may greedily favor mutationally accessible but suboptimal trajectories (Rodrigues and Shakhnovich 2019; Venkataram, et al. 2019) that then open new, idiosyncratic paths for further refinement.

Our work has some critical limitations. By testing multiple hypotheses on the same dataset, we find a number of patterns with an unknown rate of false positives. Therefore, future work is needed to test the hypotheses that we have generated, either by using additional time-course data from the LTEE, by analyzing related evolution experiments, or by direct experimental means. Another limitation is that STIMS is not able to distinguish loci evolving under relaxed selection from those that are evolving at the background rate in the LTEE. That said, one may still be able to make some inferences by comparing the rank-order of module trajectories across replicate populations. Finally, a critical caveat is that deletions that fix in the LTEE can lead to spurious inferences of purifying selection, if not handled properly. By that same token, gene duplications, amplifications, or other forms of copy number variation could lead to spurious inferences of positive selection, or elevated mutation rates.

Overall, our work provides greater insight into the mechanistic connection between genotypic and phenotypic evolution in the LTEE. Previous work has found that the majority of mutations are found in a very small fraction of the genome, indicating that most fitness gains in the LTEE can be attributed to evolution at a small number of loci (Tenaillon, et al. 2016; Good, et al. 2017; Maddamsetti, et al. 2017). In particular, it has been shown that phylogenetically-defined core genes are under strong selection in the LTEE, including global regulators of cellular physiology, transcription, and translation (Maddamsetti, et al. 2017). Consistent with those reports, we find little evidence of specific genetic modules under positive selection across LTEE populations: none of the modules that we examined coherently accounted for the loci under selection in the LTEE. By contrast, we found evidence of strong positive selection on global

regulators of the *E. coli* gene regulatory network, and as well as evidence indicating purifying selection on the ribosome and other macromolecules encoded by essential genes in the hypermutator populations of the LTEE.

Based on these observations, the genetic architecture underlying fitness improvements made in the LTEE likely involves a small number of loci that control global aspects of cellular physiology (ppGpp, cAMP, redox potential, and DNA topology and supercoiling). These factors may control many downstream pathways, which are being modulated in response to selection in the LTEE. Hence, we hypothesize that the regulation of these pathways is being rewired during the LTEE, without significantly changing their downstream effectors. This would explain why the modules that show the strongest changes in expression, and best predict growth rate and fitness in response to environmental conditions (Hui, et al. 2015; Sastry, et al. 2019; Wytock and Motter 2019) show little evidence of adaptive evolution in the LTEE. Overall, the genotype-phenotype map of *E. coli*, at least in the context of the LTEE, resembles the hierarchical "supervisor-worker" gene architecture proposed by (Chen, et al. 2018) to explain the genetic architecture of quantitative traits in *Saccharomyces cerevisiae*.

Of the models of evolutionary mode that we introduced, coupon-collecting (Good, et al. 2017; Wei and Zhang 2019) best explains the evolution of the genes under positive selection in the LTEE that were identified by Tenaillon et al. (2016). In other words, the genes with the strongest signal of selection in the first 50,000 generations, show an attenuated signal of positive selection after 50,000 generations (Figure 5 and Supplementary Figure 2). Modeling has shown that the coupon-collecting process is sufficient to cause diminishing returns epistasis in fitness, assuming a modular genotype-phenotype map in which the (environment-dependent) fitness benefit of each genetic module has an upper bound (Wei and Zhang 2019). This model is consistent with the findings of Tenaillon *et al.* (2016), in which the rate of molecular evolution due to beneficial mutations reflects diminishing returns epistasis (Chou, et al. 2011; Khan, et al. 2011; Kryazhimskiy, et al. 2014). Even though the model by Wei and Zhang is sufficient to predict diminishing returns epistasis, it also assumes the existence of a fitness optimum in which all coupons (i.e., beneficial mutations) have been collected. It is interesting to note that I-modulon regulators show evidence of coupon-collecting in two non-mutator populations, Ara−5 and Ara−6 (i.e., no mutations in I-modulon regulators in the last 20,000 generations), while simultaneously showing evidence on ongoing positive selection in the hypermutator populations

with the most mutations, Ara+3 and Ara+6 (Figure 6). This suggests that the GRN may initially evolve close to some local fitness maximum (subject to pleiotropic constraints), but then evolve further to compensate for the effects of mutations elsewhere in the genome. Ongoing compensatory adaptation could explain the decelerating yet ongoing rate of fitness improvement in the LTEE, and would serve as a proximate, mechanistic explanation to complement ultimate explanations for open-ended evolution, such as the intractable computational complexity of finding the optimum of a frustrated energy landscape (Kaznatcheev 2019). Of course, much more work is needed to explore and test these hypotheses.

As the costs of protein production exerts a strong contrast on cellular energetics, the proper allocation of proteome resources in order to maximize growth, rather the evolution of particular genetic modules, may be the cellular phenotype under strongest selection in the LTEE (Scott, et al. 2010; Klumpp and Hwa 2014; Basan, et al. 2015; Erickson, et al. 2017). Future work could test this hypothesis by trying to mimic evolutionary changes in the LTEE by directly perturbing the balance of global physiological and metabolic state variables (chromosome conformation, ppGpp, cAMP levels and redox potential) to maximize growth in the ancestral REL606 strain. Coevolution between the function of those proteome resources, pleiotropic constraints on optimal allocation, as well as feedback with the environment could all play a part in causing the open-ended increases in fitness seen in the LTEE (Wiser, et al. 2013; Lenski, et al. 2015).

MATERIALS AND METHODS

Description of method

In brief, we count the cumulative number of mutations occurring over time in a module of interest, and compare that number to a null distribution that is constructed by subsampling random sets of genes over the genome, in which the cardinality of these random gene sets is fixed to the cardinality of the module of interest. We normalize the number of observed mutations in a gene set by the total length of that gene set in base-pairs.

We hypothesized that modules with significantly more mutations would be under positive selection, while modules with significantly fewer mutations would be under purifying selection. In the Results, we show that these assumptions are reasonable, by examining how STIMS

behaves on sets of genes that are expected to evolve under relaxed, purifying, and positive selection in the LTEE.

By bootstrapping a null distribution based on random gene sets, STIMS implicitly controls for genome-wide variation in the mutation rate (since this will affect all genes). This is a key advantage over models that assume particular parameters (such as mutation rates and population sizes) that define the evolutionary dynamics, because those parameters (and thus the evolutionary dynamics) are themselves subject to unexpected change during experimental evolution (Sniegowski, et al. 1997; Blount, et al. 2008; Blount, et al. 2018).

This approach, however, does not directly control for local variations in mutation rate within the genome (Maddamsetti and Grant 2020). An approach that one could take would be to weight the sampling probabilities for choosing random genes sets based on local mutation rates. Such an approach would require the availability of good estimates of local mutation rates in each LTEE population. For that reason, we took a different approach. In order to control for local variations in mutation rate, we subsampled random modules from the same chromosomal neighborhood as the module of interest, under the assumption that nearby genes have similar mutation rates. This approach gave the same results as sampling genes uniformly across the genome, despite a higher computational cost. For this reason, the results reported here use the simplest method to construct the null distribution, in which genes are sampled without considering their position on the chromosome.

In the visualizations shown in the figures, the top 2.5% and bottom 2.5% of points in the null distribution are omitted, such that each panel can be directly interpreted as a randomization test with a false-positive (type I error) probability $\alpha = 0.05$. The variance in the bootstrapped null distributions depends on the cardinality of the gene module of interest: larger gene sets have smaller sampling variance, and smaller gene sets have larger sampling variance. When multiple gene modules with different cardinalities are shown in the same panel, the null distribution depicted is always the minimum cardinality, which maximizes the sampling variance, and so gives the most conservative statistical comparison. The statistics reported in the text however, always keep the cardinality fixed between the random gene sets of the bootstrapped null distribution and the gene module of interest. As in the figures, statistics are calculated separately for each population. Approximate p-values for two-sided tests against the bootstrapped null distribution were calculated as $min[2p, 2(1 - p)]$, where min is a function that returns the

minimum of its arguments, and p is the upper-tail probability of the event of sampling a normalized cumulative number of mutations under the null distribution that is greater than the normalized cumulative number of mutations of the module of interest. While p-values for one-sided tests are exact estimates of the tail of the null distribution, p-values for two-sided tests are approximations, as the factor of two is based on the assumption that the null distribution is symmetric.

Filtering of genes in REL606 genome

As we analyzed the pre-processed data published by (Good, et al. 2017), genes that were excluded from that analysis were also excluded from this analysis. See Section 4.3.1 "Removing mutations in repetitive regions of the genome" in the supplementary information of (Good, et al. 2017) for more details. In particular, pseudogenes and genes in repetitive regions of the genome were excluded, as these can be difficult to resolve using short-read sequencing data. A site was marked as repetitive if (1) it was annotated as a repeat region in the REL606 reference, (2) it was present in the set of masked regions compiled by (Tenaillon, et al. 2016), or (3) it fell within the putative prophage at REL606 genome coordinates 880528–904682. This last region was explicitly added to the masked region file published by (Tenaillon, et al. 2016), and saved it as: *REL606.L20.G15.P0.M35.RM-edited.mask.gd*. This masked region file corresponds to the Genbank reference genome annotation file, *REL606.7.gbk*.

These filters were applied by a python script called *printEcoliIDs.py*. In addition, 4 pairs of genes with duplicate gene names (8 genes in total) were excluded from the analysis. These were: *alr*, *bioD*, *ddl*, and *maf*.

Essential genes excluded from purifying selection control experiment

Previous work has shown that many of the genes under strongest positive selection in the LTEE are also essential genes (Maddamsetti, et al. 2017). For this reason, we excluded essential genes that showed parallel evolution in (Tenaillon, et al. 2016). In that paper, a G-score was calculated to measure parallel evolution. G-scores were calculated separately for the non-mutator and hypermutator populations. We ranked all genes in REL606 that passed our filters (*Materials and Methods: Filtering of genes in REL606 genome*) by their non-mutator and hypermutator G-scores. Essential genes that were in the top 50 of either list were excluded from the purifying

selection control experiment. 21 out of the top 50 G-scoring non-mutator genes were essential in REL606, based on the transposon mutagenesis and sequencing experiments carried out by (Couce, et al. 2017). These genes were: *arcA*, *arcB*, *crp*, *fabF*, *ftsI*, *hflB*, *infB*, *infC*, *mrdA*, *mreB*, *mreC*, *mreD*, *nusA*, *rne*, *rplF*, *rpoB*, *rpsD*, *sapF*, *spoT*, *topA*, and *yabB*. 3 out of the top 50 G-scoring hypermutator genes were essential in REL606. These genes were: *sapA*, *tilS*, and *yhbH*.

ACKNOWLEDGEMENTS

We thank Alejandro Couce for sharing the list of gold-standard neutral genes in *E. coli* B strain REL606.

REFERENCES

1. Bailey SF, Guo Q, Bataillon T. 2018. Identifying drivers of parallel evolution: A regression model approach. Genome biology and evolution 10:2801-2812.
2. Barrick JE, Lenski RE editors. Cold Spring Harbor symposia on quantitative biology. 2009.
3. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature 461:1243-1247.
4. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T. 2015. Overflow metabolism in Escherichia coli results from efficient proteome allocation. Nature 528:99-104.
5. Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proceedings of the National Academy of Sciences 105:7899-7906.
6. Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: Replaying life's tape. Science 362:eaam5979.
7. Chen H, Wu C-I, He X. 2018. The genotype–phenotype relationships in the light of natural selection. Molecular biology and evolution 35:525-542.
8. Chou H-H, Chiu H-C, Delaney NF, Segrè D, Marx CJ. 2011. Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332:1190-1192.
9. Chung H, Lieberman TD, Vargas SO, Flett KB, McAdam AJ, Priebe GP, Kishony R. 2017. Global and local selection acting on the pathogen Stenotrophomonas maltophilia in the human lung. Nature communications 8:14078.
10. Cooper VS, Honsa E, Rowe H, Deitrick C, Iverson AR, Whittall JJ, Neville SL, McDevitt CA, Kietzman C, Rosch JW. 2020. Experimental Evolution in vivo to Identify Selective Pressures During Pneumococcal Colonization. Msystems 5.
11. Couce A, Caudwell LV, Feinauer C, Hindré T, Feugeas J-P, Weigt M, Lenski RE, Schneider D, Tenaillon O. 2017. Mutator genomes decay, despite sustained fitness gains,

in a long-term experiment with bacteria. Proceedings of the National Academy of Sciences 114:E9026-E9035.
12. Couce A, Tenaillon O. 2019. Mutation bias and GC content shape antimutator invasions. Nature communications 10:1-9.
13. Elena SF, Cooper VS, Lenski RE. 1996. Punctuated evolution caused by selection of rare beneficial mutations. Science 272:1802-1804.
14. Erickson DW, Schink SJ, Patsalo V, Williamson JR, Gerland U, Hwa T. 2017. A global resource allocation strategy governs growth transition kinetics of Escherichia coli. Nature 551:119-123.
15. Foster PL, Hanson AJ, Lee H, Popodi EM, Tang H. 2013. On the mutational topology of the bacterial genome. G3: Genes, Genomes, Genetics 3:399-407.
16. Good BH, Desai MM. 2016. Evolution of mutation rates in rapidly adapting asexual populations. Genetics 204:1249-1266.
17. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017. The dynamics of molecular evolution over 60,000 generations. Nature 551:45-50.
18. Hui S, Silverman JM, Chen SS, Erickson DW, Basan M, Wang J, Hwa T, Williamson JR. 2015. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. Molecular systems biology 11.
19. Kaznatcheev A. 2019. Computational complexity as an ultimate constraint on evolution. Genetics 212:245-265.
20. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. Science 332:1193-1196.
21. Klumpp S, Hwa T. 2014. Bacterial growth: global effects on gene expression, growth feedback and proteome partition. Current opinion in biotechnology 28:96-102.
22. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. 2014. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science 344:1519-1522.
23. Lenski RE. 2017. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. The ISME journal 11:2181-2194.
24. Lenski RE, Mongold JA, Sniegowski PD, Travisano M, Vasi F, Gerrish PJ, Schmidt TM. 1998. Evolution of competitive fitness in experimental populations of E. coli: what makes one genotype a better competitor than another? Antonie van Leeuwenhoek 73:35-47.
25. Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. The American Naturalist 138:1315-1341.
26. Lenski RE, Wiser MJ, Ribeck N, Blount ZD, Nahum JR, Morris JJ, Zaman L, Turner CB, Wade BD, Maddamsetti R. 2015. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with Escherichia coli. Proceedings of the Royal Society B: Biological Sciences 282:20152292.
27. Maddamsetti R, Hatcher PJ, Green AG, Williams BL, Marks DS, Lenski RE. 2017. Core genes evolve rapidly in the long-term evolution experiment with Escherichia coli. Genome biology and evolution 9:1072-1083.
28. Maddamsetti R, Lenski RE, Barrick JE. 2015. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with Escherichia coli. Genetics 200:619-631.

29. Niccum BA, Lee H, MohammedIsmail W, Tang H, Foster PL. 2019. The Symmetrical Wave Pattern of Base-Pair Substitution Rates across the Escherichia coli Chromosome Has Multiple Causes. mBio 10.
30. Quandt EM, Gollihar J, Blount ZD, Ellington AD, Georgiou G, Barrick JE. 2015. Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the Lenski evolution experiment. Elife 4:e09696.
31. Ramiro RS, Durão P, Bank C, Gordo I. 2020. Low mutational load and high mutation rate variation in gut commensal bacteria. PLoS Biology 18:e3000617.
32. Rodrigues JV, Shakhnovich EI. 2019. Adaptation to mutational inactivation of an essential gene converges to an accessible suboptimal fitness peak. Elife 8.
33. Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, Choudhary KS, Yang L, King ZA, Palsson BO. 2019. The Escherichia coli transcriptome mostly consists of independently regulated modules. Nature communications 10:1-14.
34. Schiffels S, Szöllősi GJ, Mustonen V, Lässig M. 2011. Emergent neutrality in adaptive asexual evolution. Genetics 189:1361-1375.
35. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. 2010. Interdependence of cell growth and gene expression: origins and consequences. Science 330:1099-1102.
36. Simpson GG. 1944. Tempo and mode in evolution: Columbia University Press.
37. Sniegowski PD, Gerrish PJ, Lenski RE. 1997. Evolution of high mutation rates in experimental populations of E. coli. Nature 387:703-705.
38. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Médigue C. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536:165-170.
39. Turner CB, Blount ZD, Mitchell DH, Lenski RE. 2015. Evolution and coexistence in response to a key innovation in a long-term evolution experiment with Escherichia coli. bioRxiv:020958.
40. Vasi F, Travisano M, Lenski RE. 1994. Long-term experimental evolution in Escherichia coli. II. Changes in life-history traits during adaptation to a seasonal environment. The American Naturalist 144:432-456.
41. Venkataram S, Monasky R, Sikaroodi SH, Kryazhimskiy S, Kaçar B. 2019. The Onset of Evolutionary Stalling and the Limit on the Power of Natural Selection to Improve a Cellular Module. bioRxiv:850644.
42. Wei X, Zhang J. 2019. Patterns and mechanisms of diminishing returns from beneficial mutations. Molecular biology and evolution 36:1008-1021.
43. 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. Proceedings of the National Academy of Sciences 110:222-227.
44. Wiser MJ, Ribeck N, Lenski RE. 2013. Long-term dynamics of adaptation in asexual populations. Science 342:1364-1367.
45. Wytock TP, Motter AE. 2019. Predicting growth rate from gene expression. Proceedings of the National Academy of Sciences 116:367-372.
46. Zee PC, Mendes-Soares H, Yu YTN, Kraemer SA, Keller H, Ossowski S, Schneeberger K, Velicer GJ. 2014. A shift from magnitude to sign epistasis during adaptive evolution of a bacterial social trait. Evolution 68:2701-2708.

FIGURES

Figure 1. Evolutionary mode affects the evolutionary rate of a gene set of interest. The scenarios shown here assume a population that has evolved a hypermutator phenotype. M(t) is the number of cumulative mutations in the gene set, normalized by total length, as a function of time. The grey area represents the distribution of background evolutionary rates in this hypothetical hypermutator population, constructed with randomly assembled gene sets. The black line represents the trajectory M(t) for the gene set of interest. A) A gene set evolving at the background evolutionary rate will fall within the grey area. B) A gene set evolving under purifying selection evolves more slowly than the background rate. C) A gene set evolving under positive selection to knock out those genes will evolve rapidly, and then accumulate mutations once those genes have been knocked out. D) A gene set under fine-tuning positive selection will evolve rapidly, then will eventually evolve more slowly than background to preserve the fine-tuned functions of those genes. E) A gene set under continual fine-tuning selection will always evolve more rapidly than background, as new opportunities for beneficial mutations arise over time.

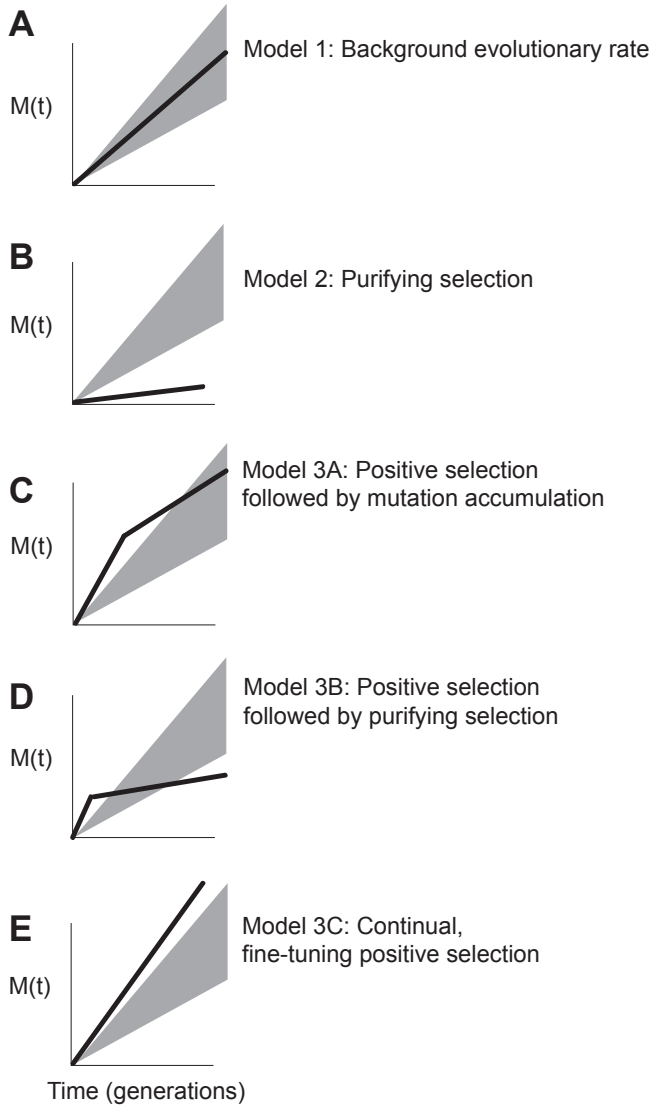


Figure 2. Gold-standard genes under relaxed selection. Each panel shows the cumulative number of mutations on gold-standard neutral genes (black) in the 12 LTEE populations. For comparison, random sets of 57 genes were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray, in order to show a two-tailed statistical comparison of the cumulative number of mutations in the gold-standard neutral gene set to the null distribution at $\alpha = 0.05$. The top six panels are populations with the ancestral mutation rate (nonmutator populations) and the bottom six panels are those that evolved elevated mutation rates (hypermutator populations).

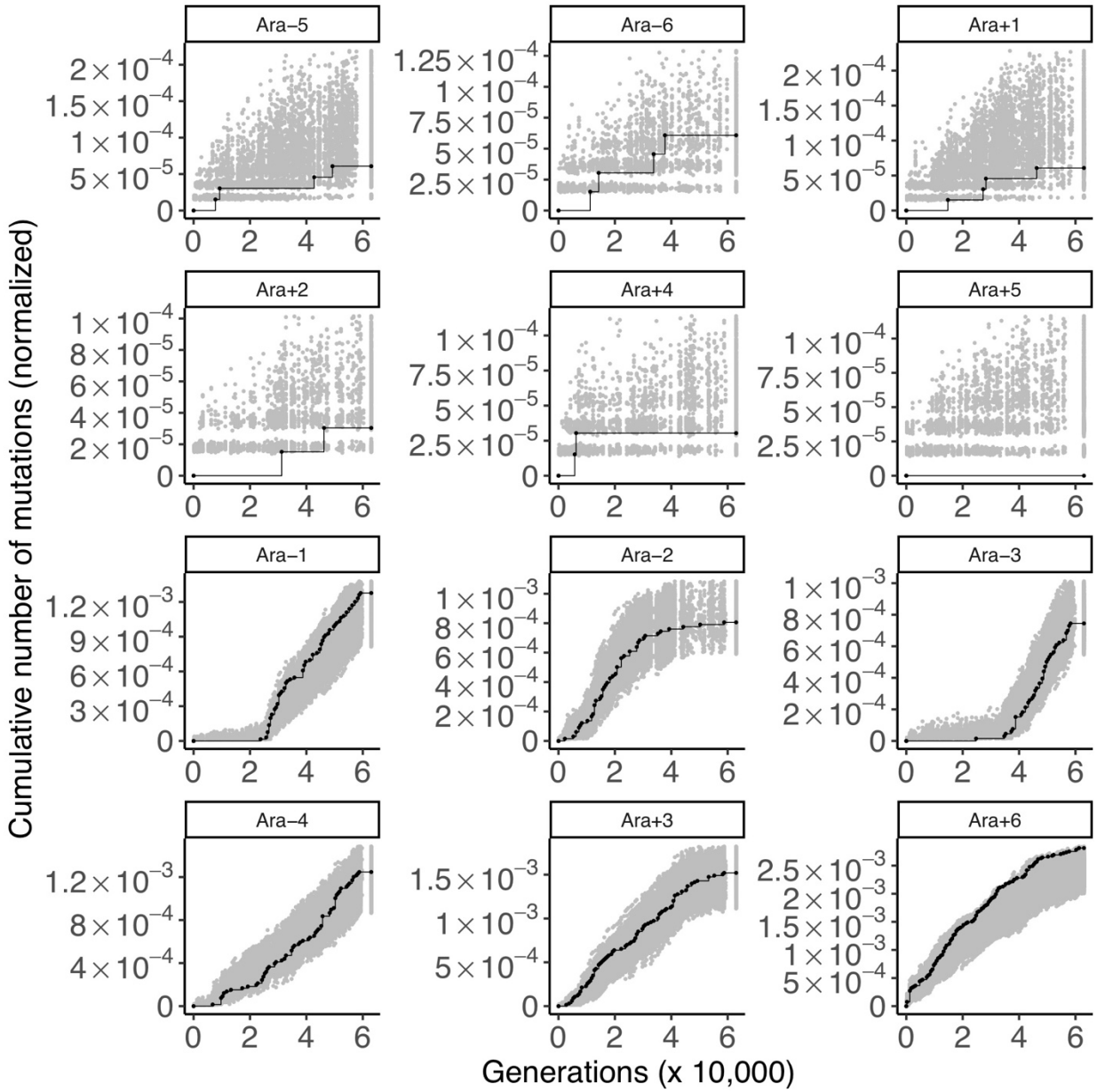


Figure 3. Gold-standard genes under purifying selection. Each panel shows the cumulative number of mutations on the gold standard set of genes under purifying selection (black) in the 12 LTEE populations. For comparison, random sets of 490 genes were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray. The top six panels are the nonmutator populations and the bottom six panels are the hypermutator populations.

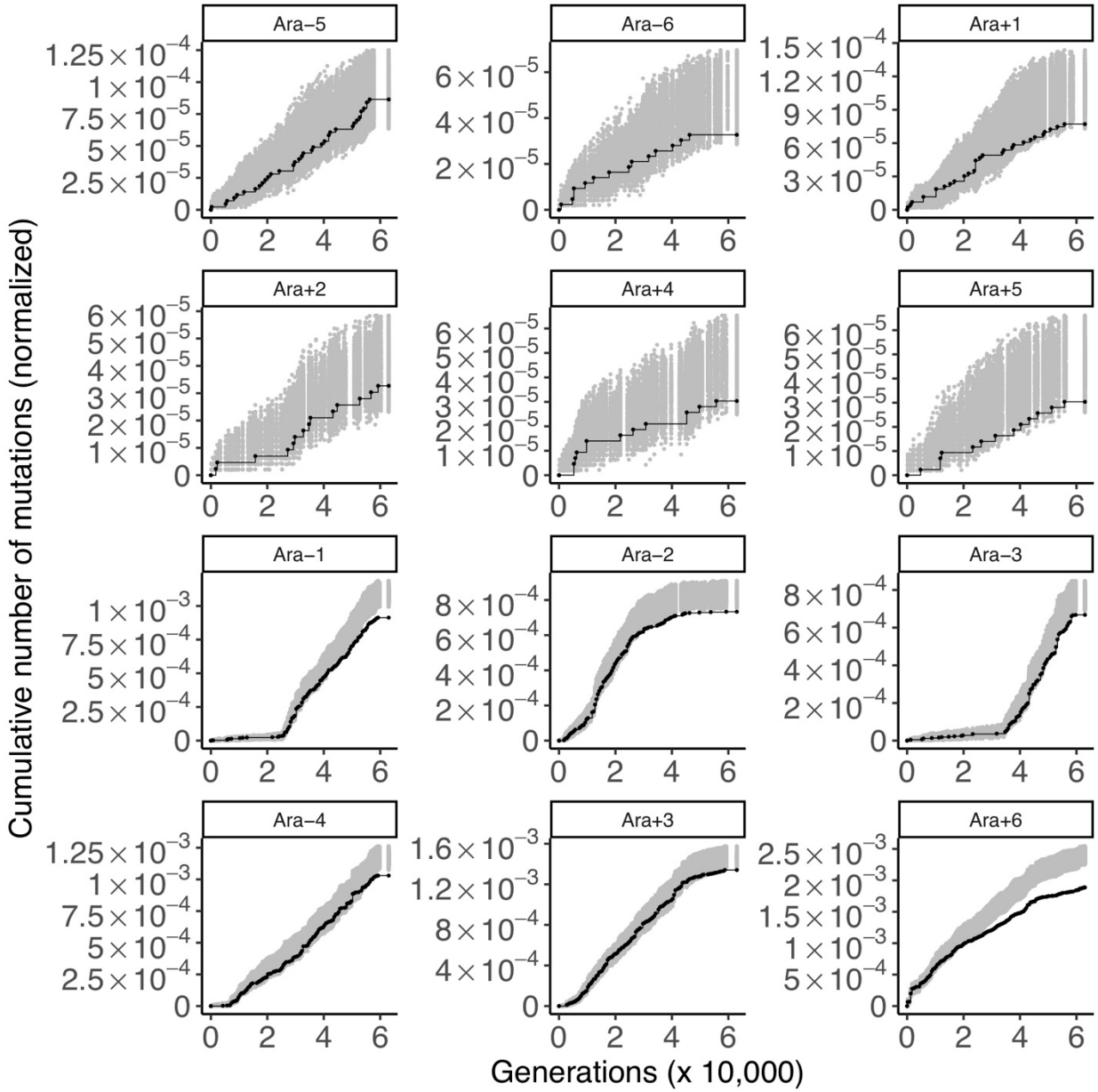


Figure 4. Gold-standard genes under positive selection before 50,000 generations. Each panel shows the cumulative number of mutations on the gold standard set of genes under positive selection (black) in the 12 LTEE populations, up to 50,000 generations. These genes are the 50 genes with the strongest signal of parallel evolution in clones isolated from the six non-mutator populations, as scored by Tenaillon et al. (2016). For comparison, random sets of 50 genes were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray. The top six panels are the nonmutator populations and the bottom six panels are the hypermutator populations.

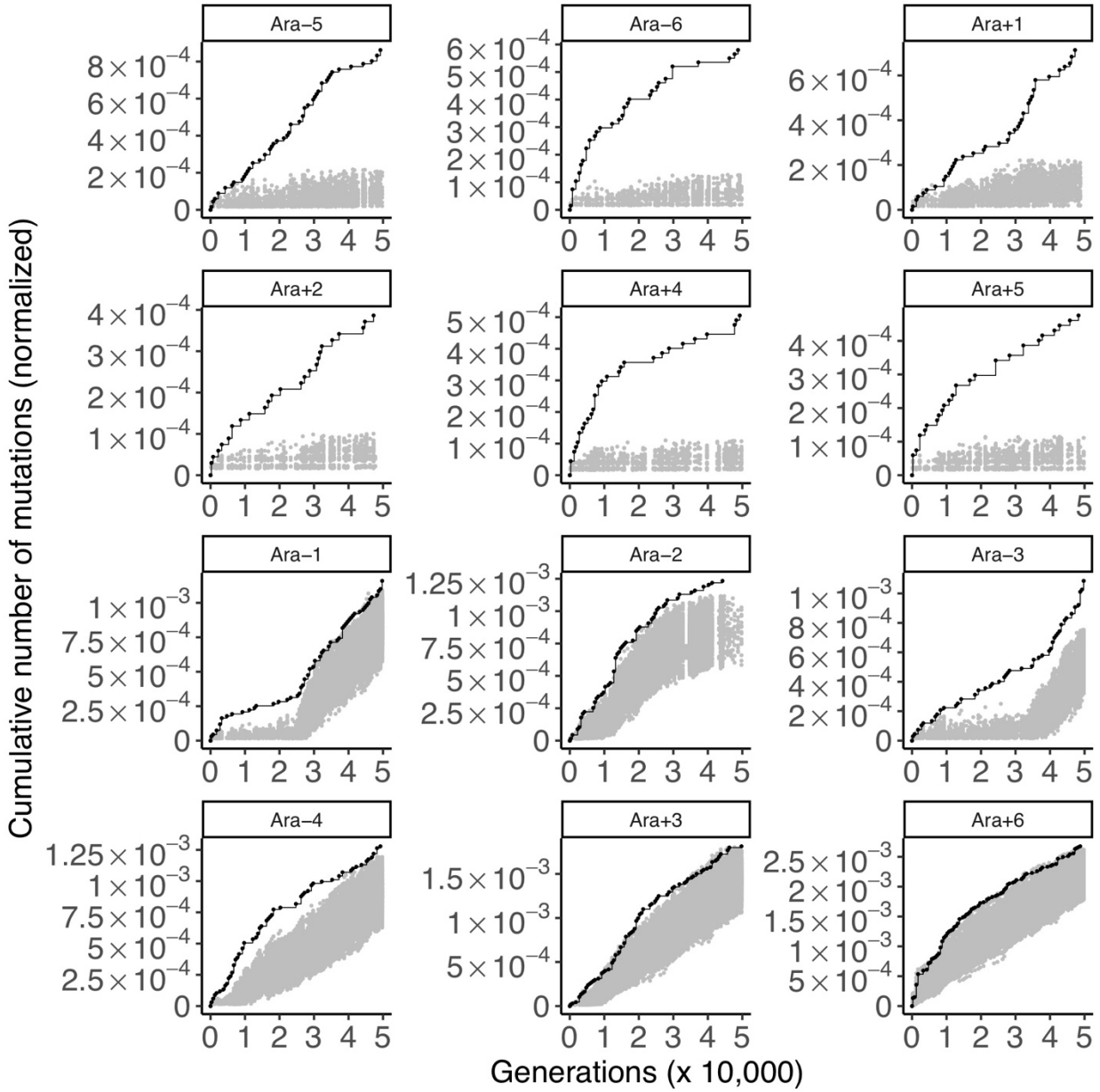


Figure 5. Gold-standard genes under positive selection after 50,000 generations. Each panel shows the cumulative number of mutations on the gold standard set of genes under positive selection (black) in the 12 LTEE populations, between 50,000 to 62,750 generations. These genes are the 50 genes with the strongest signal of parallel evolution in clones isolated from the six non-mutator populations, as scored by Tenaillon et al. (2016). For comparison, random sets of 50 genes were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray. The top six panels are the nonmutator populations and the bottom six panels are the hypermutator populations.

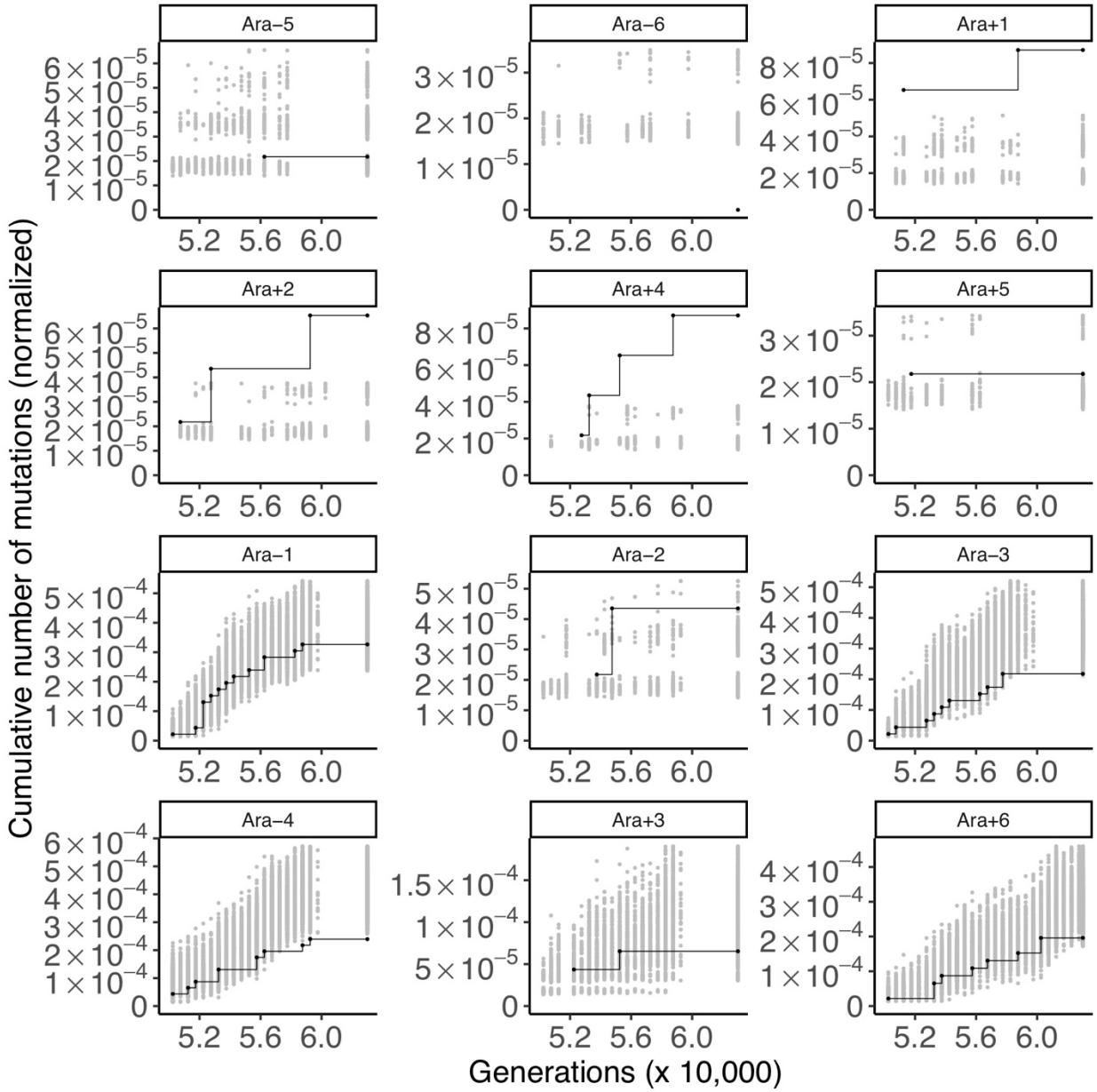


Figure 6. I-modulon regulators evolve under strong positive selection, while genes regulated within I-modulon evolve at the background rate. Each panel shows the cumulative number of mutations in 70 I-modulon regulators (black) and in 1,394 genes regulated by I-modulon regulators (red) in the 12 LTEE populations. For comparison to the I-modulon regulators, random sets of 70 genes were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray. A similar procedure was used with random sets of 1,394 genes to make a comparable distribution in pink to compare to the genes regulated by I-modulon regulators. The top six panels are the nonmutator populations and the bottom six panels are the hypermutator populations.

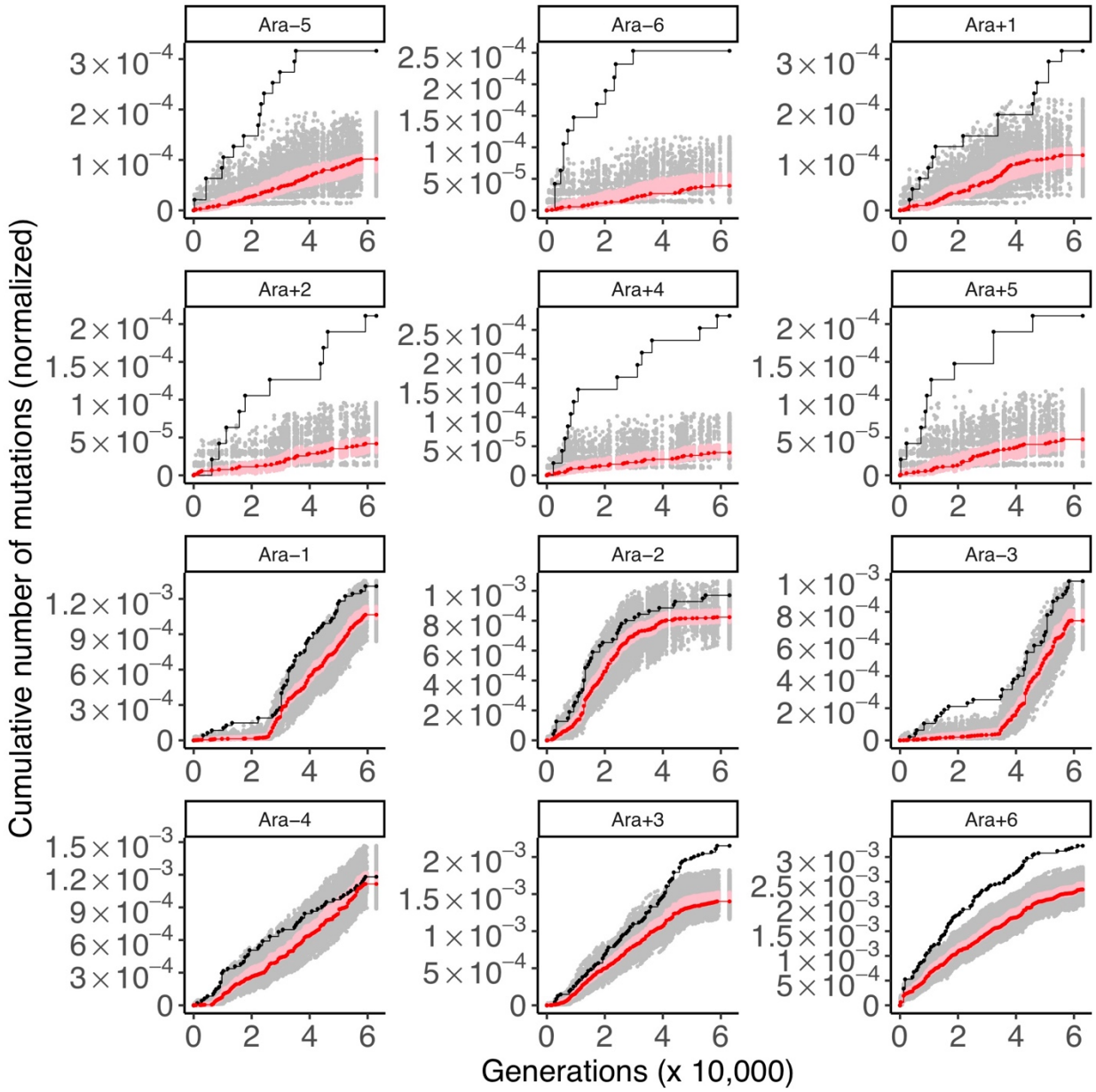


Figure 7. Purifying selection on the proteome U-sector. Each panel shows the cumulative number of mutations in the A-sector (black), S-sector (red), O-sector (blue), U-sector (green), R-sector (yellow), and C-sector (orange) in the 12 LTEE populations. For comparison, random sets of 92 genes (i.e. the smallest proteome sector cardinality) were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray. The top six panels are the nonmutator populations and the bottom six panels are the hypermutator populations.

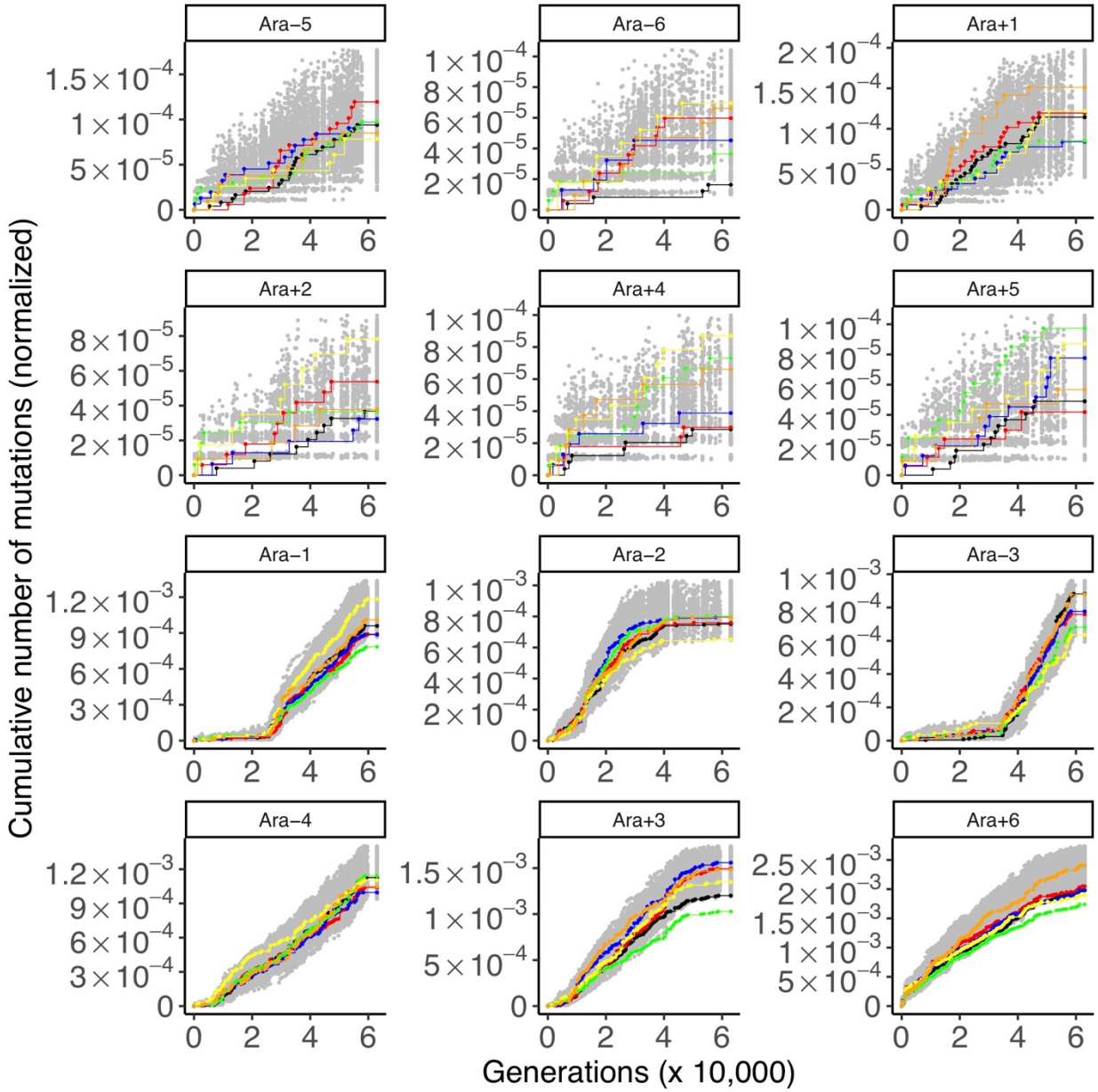
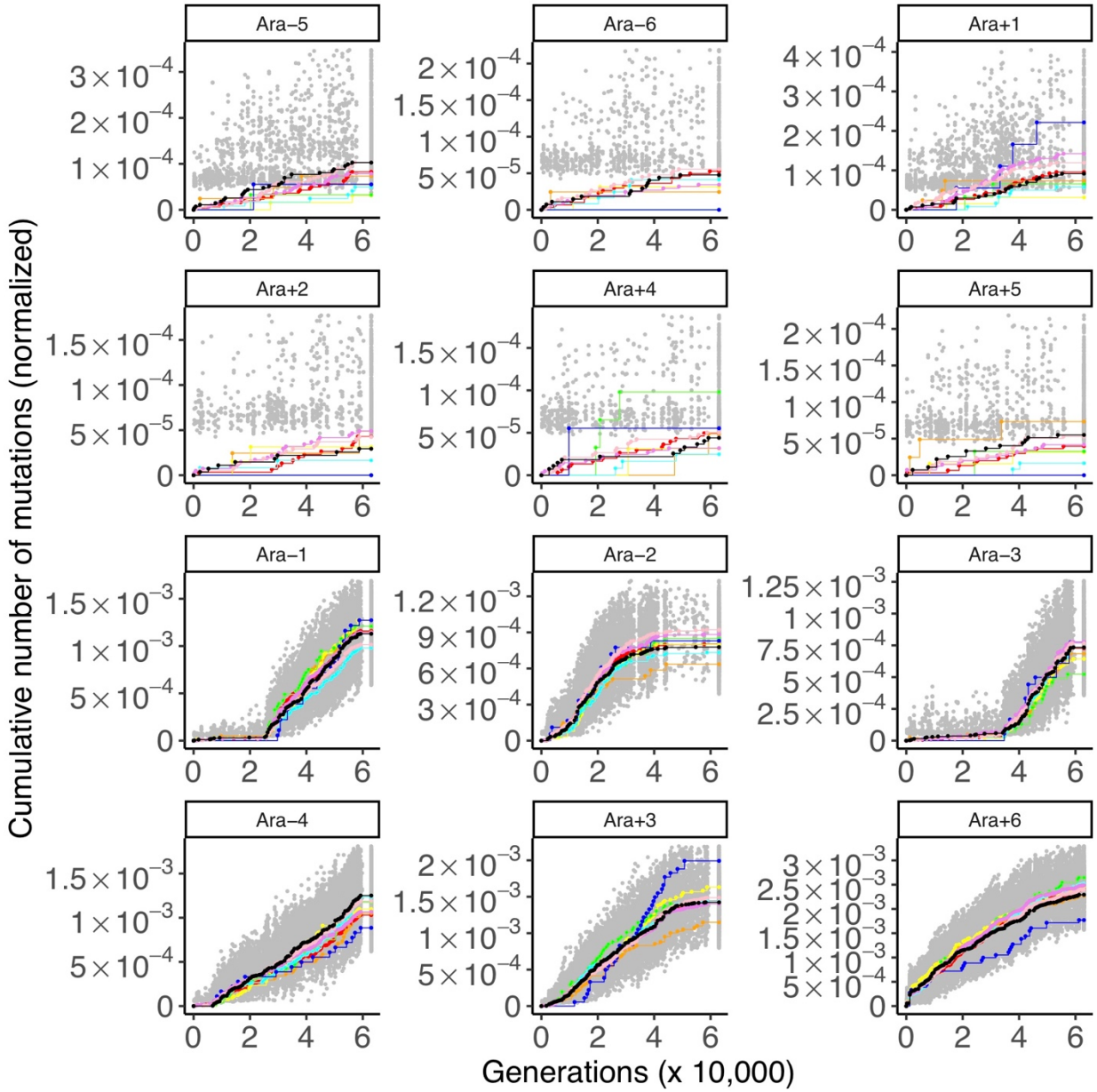


Figure 8. No evidence of selection on *E. coli* eigengenes. Each panel shows the cumulative number of mutations in eigengene 1 (red), eigengene 2 (orange), eigengene 3 (yellow), eigengene 4 (green), eigengene 5 (cyan), eigengene 6 (blue), eigengene 7 (violet), eigengene 8 (pink), and eigengene 9 (black) in the 12 LTEE populations. For comparison, random sets of 15 genes (i.e. the smallest eigengene cardinality) were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray. The top six panels are the nonmutator populations and the bottom six panels are the hypermutator populations.



SUPPLEMENTARY INFORMATION:

Supplementary File 1. Results of running method on each I-modulon in Sastry et al. (2019).

Supplementary Table S1. Essential genes in REL606 with no observed mutations in the LTEE metagenomics data.

Gene	Locus tag	Gene length	Product
rlpB	ECB_00610	582	minor lipoprotein
rplJ	ECB_03861	498	50S ribosomal protein L10
ECB_01539	ECB_01539	408	protein similar to DicA regulator of DicB encoded by prophage CP-933O
sdhD	ECB_00682	348	succinate dehydrogenase cytochrome b556 small membrane subunit
rplV	ECB_03166	333	50S ribosomal protein L22
rpsJ	ECB_03172	312	30S ribosomal protein S10
rpsN	ECB_03158	306	30S ribosomal protein S14
yrbB	ECB_03056	294	hypothetical protein
groES	ECB_04012	294	co-chaperonin GroES
rpsS	ECB_03167	279	30S ribosomal protein S19
rpsT	ECB_00027	264	30S ribosomal protein S20
rpsQ	ECB_03162	255	30S ribosomal protein S17
atpE	ECB_03621	240	F0F1 ATP synthase subunit C
rpmC	ECB_03163	192	50S ribosomal protein L29
ydaE	ECB_01329	171	conserved protein
rpmJ	ECB_03150	117	50S ribosomal protein L36
ybgT	ECB_00694	114	hypothetical protein
rhoL	ECB_03660	102	rho operon leader peptide

Supplementary Table S2. Essential genes in REL606 with only synonymous mutations in the LTEE metagenomics data.

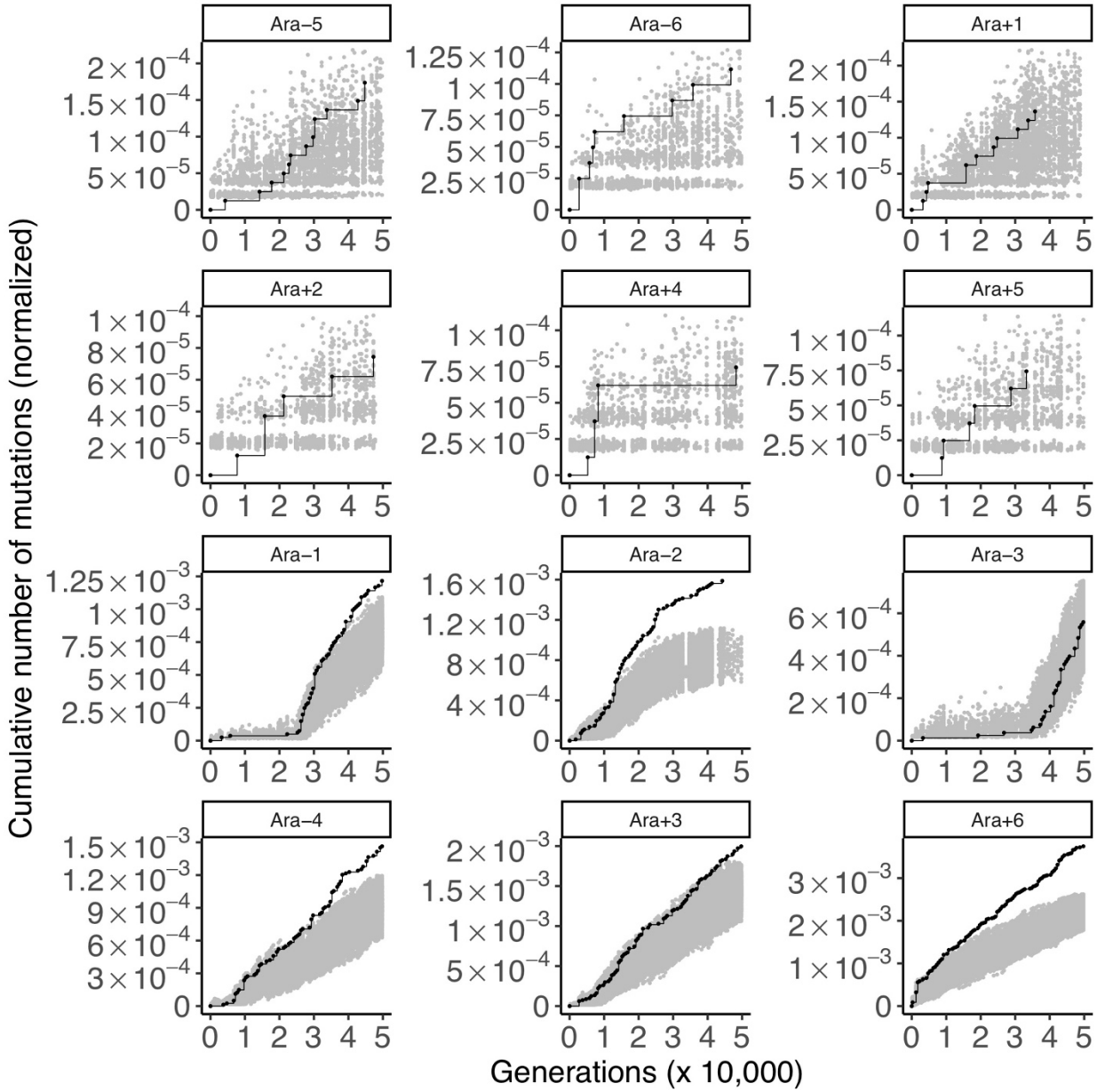
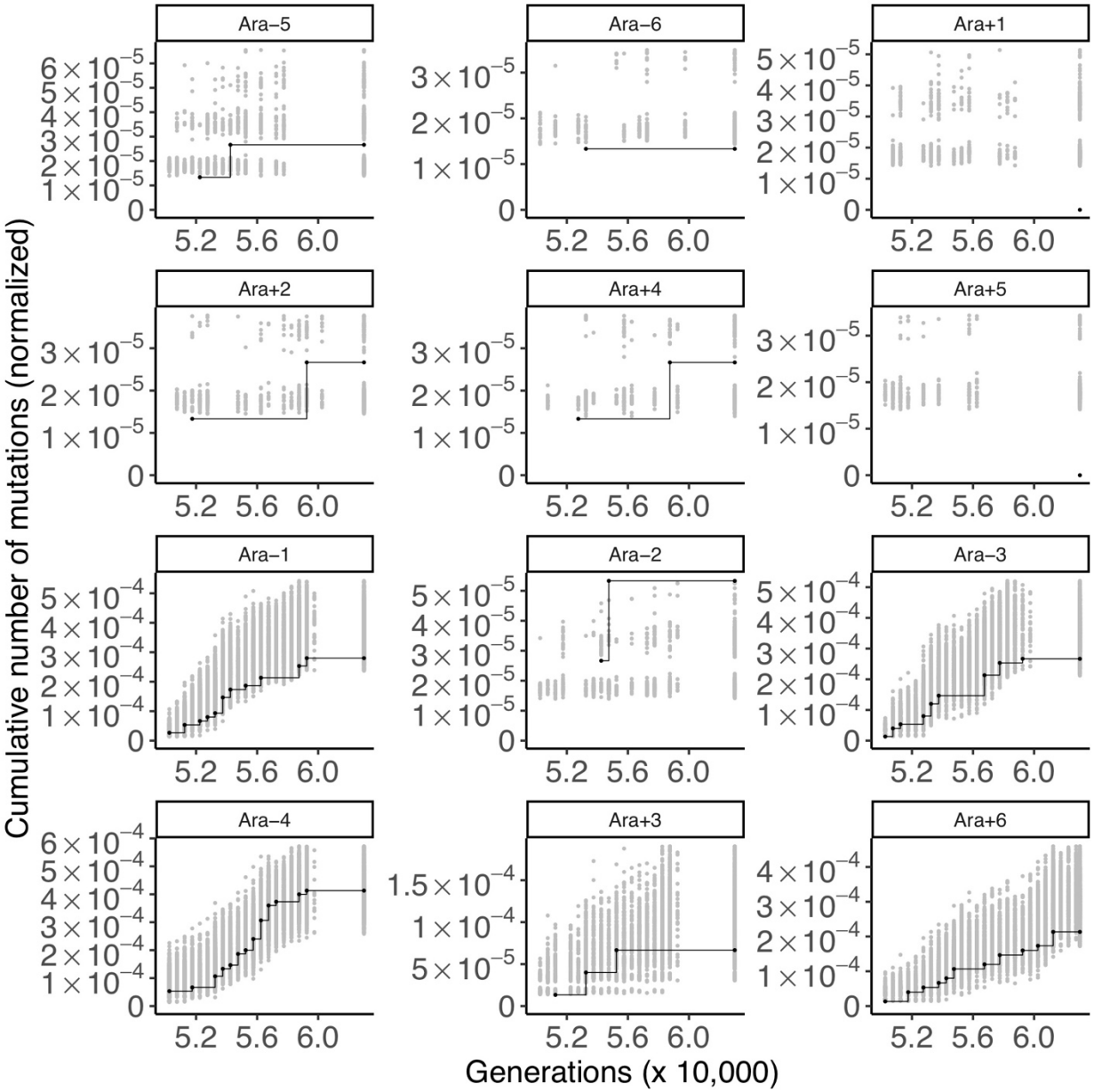
Gene	Locus tag	Gene length	Product
atpD	ECB_03616	1383	F0F1 ATP synthase subunit beta
icdA	ECB_01134	1251	isocitrate dehydrogenase
thyA	ECB_02675	795	thymidylate synthase
plsC	ECB_02890	738	1-acyl-sn-glycerol-3-phosphate acyltransferase
pdxH	ECB_01608	657	pyridoxamine 5'-phosphate oxidase
rplD	ECB_03170	606	50S ribosomal protein L4
lpcA	ECB_00217	579	phosphoheptose isomerase
rimM	ECB_02497	549	16S rRNA-processing protein
rplE	ECB_03159	540	50S ribosomal protein L5
sixA	ECB_02264	486	phosphohistidine phosphatase
folK	ECB_00141	480	2-amino-4-hydroxy-6-hydroxymethyldihyropteridine pyrophosphokinase
accB	ECB_03113	471	acetyl-CoA carboxylase
holD	ECB_04247	414	DNA polymerase III subunit psi
rpsI	ECB_03090	393	30S ribosomal protein S9
iscU	ECB_02421	387	scaffold protein
rplQ	ECB_03145	384	50S ribosomal protein L17
panD	ECB_00130	381	aspartate 1-decarboxylase precursor
rpsL	ECB_03193	375	30S ribosomal protein S12
rplN	ECB_03161	372	50S ribosomal protein L14
rnpA	ECB_03587	360	ribonuclease P
rplT	ECB_01685	357	50S ribosomal protein L20
yadR	ECB_00155	345	hypothetical protein
fdx	ECB_02417	336	[2Fe-2S] ferredoxin
secG	ECB_03040	333	protein-export membrane protein
ybaB	ECB_00422	330	hypothetical protein
acpP	ECB_01090	237	acyl carrier protein
rpmI	ECB_01686	198	50S ribosomal protein L35
rpmH	ECB_03586	141	50S ribosomal protein L34
pheL	ECB_02487	48	pheA gene leader peptide
trpL	ECB_01239	45	trp operon leader peptide

Supplementary Table 3. Candidate I-modulons for historical contingency and epistasis in the LTEE.

Page*	I-modulon	Populations of Interest	Hypothesis
3	ArcA-2	Ara−2, Ara−3	purifying selection
11	CpxR	Ara−3, Ara−4, Ara+6	positive and purifying selection
13	Crp-1	Ara−5, Ara+1	positive selection
18	CysB	Ara−1, Ara+6	purifying selection
19	Deletion-1	Ara−3	positive selection
22	Duplication-1	Ara−2, Ara−3, Ara−4, Ara+3	strong purifying selection
23	e14-deletion	Ara−3	strong positive selection
24	efeU-repair	Ara+3, Ara+6	strong positive selection, purifying selection
30	fimbriae	Ara−2	strong positive selection
31	FlhDC	Ara−2, Ara+6	positive selection
37	fur-KO	Ara−3	strong positive selection
40	galWX-KO	Ara+6	positive selection
46	insertion	Ara−1, Ara+6	strong positive selection
49	lipopolysaccharide	Ara−1	strong positive selection
50	Lrp	Ara+6	purifying selection
63	PurR-1	Ara−1	purifying selection
77	translation	Ara+1, Ara+2, Ara+5, Ara−3, Ara+6	positive and purifying selection
79	uncharacterized-1	Ara+2, Ara+5, Ara+3	positive selection
81	uncharacterized-3	Ara−1	positive selection
82	uncharacterized-4	Ara−6, Ara+1, Ara+2, Ara+4, Ara+5, Ara−2, Ara−3, Ara−4, Ara+3, Ara+6	strong positive and purifying selection
90	YieP	Ara−2, Ara−4	positive selection

*Page numbers refers to Supplementary File 1.

Supplementary Figure 1. Genes evolving in parallel in hypermutator populations before 50,000 generations. Each panel shows the cumulative number of mutations in the 50 genes with the strongest signal of parallel evolution in clones isolated from the six hypermutator populations, as scored by Tenaillon et al. (2016). These curves are drawn in black, and this time series covers the first 50,000 generations. For comparison, random sets of 50 genes were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray. The top six panels are the nonmutator populations and the bottom six panels are the hypermutator populations.



Supplementary Figure 2. Genes evolving in parallel in hypermutator populations after 50,000 generations. Each panel shows the cumulative number of mutations in the 50 genes with the strongest signal of parallel evolution in clones isolated from the six hypermutator populations, as scored by Tenaillon et al (2016). These curves are drawn in black, and this time series covers generations 50,000–62,750. For comparison, random sets of 50 genes were sampled 1000 times, and the cumulative number of mutations in those random gene sets, normalized by gene length, were calculated. The middle 95% of this null distribution is shown in gray. The top six panels are the nonmutator populations and the bottom six panels are the hypermutator populations.