Supplementary Information

Contents

List of Figures

List of Tables

1 Long-term evolution experiment

The LTEE consists of 12 replicate populations of *Escherichia coli* B. Six of the populations are founded from strains isogenic to REL606 [1, 2], and are labelled Ara−1 through Ara−6. The other six populations (Ara+1 to Ara+6) are derived from strain REL607, which differs from REL606 by a point mutation in the *araA* gene that restores the ability to grow on arabinose, and a second mutation in *recD* that has no known phenotype [3]. The 12 populations are grown with daily 100-fold dilutions in 10 ml Davis minimal medium with 0.025% glucose (DM25), and incubated in 50-ml flasks with orbital shaking at 37°C. The cells grow until they exhaust the limiting nutrients, reaching a stationary-phase cell density of $\sim 5 \times 10^7$ per ml before they are diluted into fresh media [1, 4]. This protocol results in ~ 6.67 generations per day and an effective population size of $N_e \sim 3 \times 10^7$. Aliquots from every 500 generations are mixed with glycerol as a cryoprotectant and frozen at −80°C for long-term storage. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

2 Competitive fitness trajectories

To contrast the rates of molecular and phenotypic evolution, we compared our metagenomic data with competitive fitness assays previously carried out by Wiser et al. [5] and Lenski et al. [6]. In each assay, whole-population samples from an evolved population were mixed with a reference strain with the opposite *ara* marker and propagated under the standard LTEE conditions for Δt generations; the relative frequencies of the evolved and reference subpopulations at the beginning and end of the assay were measured by plating on tetrazolium arabinose agar. Raw colony counts were downloaded from the corresponding Dryad repositories [7, 8], and relative (log) fitness was defined as

$$\Delta X \equiv \frac{1}{\Delta t} \log \left[\frac{N_e(\Delta t)}{N_r(\Delta t)} \cdot \frac{N_r(0)}{N_e(0)} \right] , \tag{1}$$

where $N_e(t)$ and $N_r(t)$ are the number of evolved and reference colonies, respectively. We chose this measure because it provides the most direct connection to the fitness parameters in population genetics, which describe how allele frequencies change timescales much longer than a single dilution cycle. For completeness, we also calculated fitness using the ratio of the competitors' realized Malthusian parameters,

$$W \equiv \frac{\log \left(\frac{N_e(\Delta t) \cdot 2^{\Delta t}}{N_e(0)} \right)}{\log \left(\frac{N_r(\Delta t) \cdot 2^{\Delta t}}{N_r(0)} \right)} , \tag{2}$$

which has traditionally been used to quantify fitness in the LTEE (see Wiser et al. [5] for more details). Though quantitatively different, the two measures are correlated, and both support the qualitative claim that fitness gains decline more rapidly than the rate of mutation accumulation.

The fitness assays in Figs 2a and S1 were carried out by Wiser et al. [5] using the ancestral reference strains and a single day of competition ($\Delta t \approx 6.7$). More precise measurements of the fitness gains after generation 40,000 were carried out by Lenski et al. [6], using a higher-fitness reference clone isolated from the Ara−5 population at generation 40,000, and a three day competition period ($\Delta t \approx 20$). We compare these late fitness gains to the corresponding mutation gains over the same period in the nonmutator populations in Fig. S2. Estimates of the average fitness gain per mutation are shown in Fig. S3.

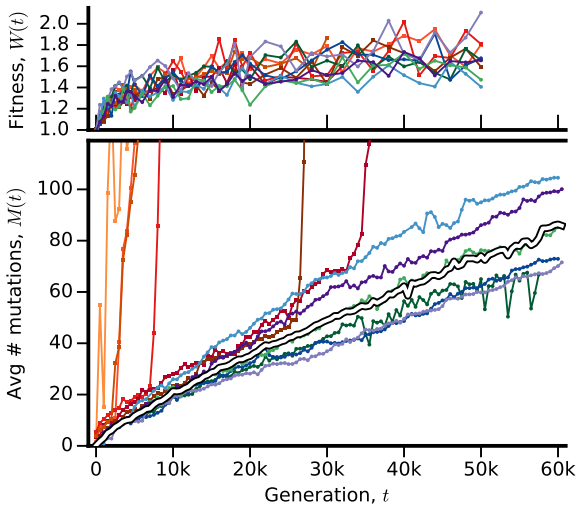

Figure S1: Fitness trajectories from Fig. 2a replotted using the traditional W measure from Eq. (2). Each population is colored as in Fig. 2. For comparison, the corresponding mutation trajectories from Fig. 2b are replotted in the bottom panel.

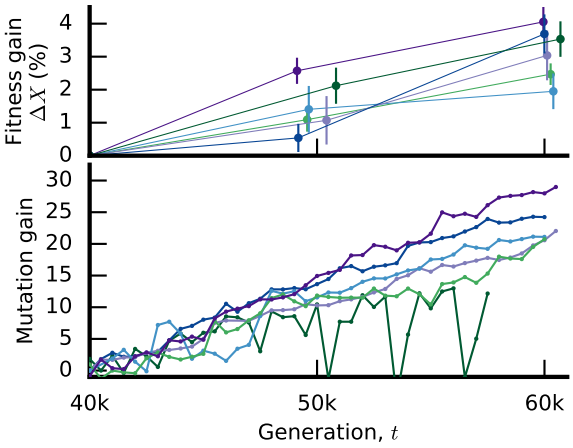

Figure S2: Fitness and mutation gains between generation 40,000 and 60,000 in the nonmutator populations. Top: fitness gains $\Delta X_p = X_p(t) - X_p(40,000)$ calculated from the high-replication assays performed by Lenski et al. [6]. The $n = 6$ independent populations are colored as in Fig. 2, and each point includes a small amount of noise on the t-axis to enhance visibility. Vertical error bars denote ±1s.e.m. intervals estimated from technical replicates with sample sizes described in Lenski et al. [6]. Bottom: corresponding mutation gains obtained by shifting the mutation trajectories in Fig. 2b by $M_{p,0} = \text{median}\{M_p(t) : 39,000 \le t \le 41,000\}$.

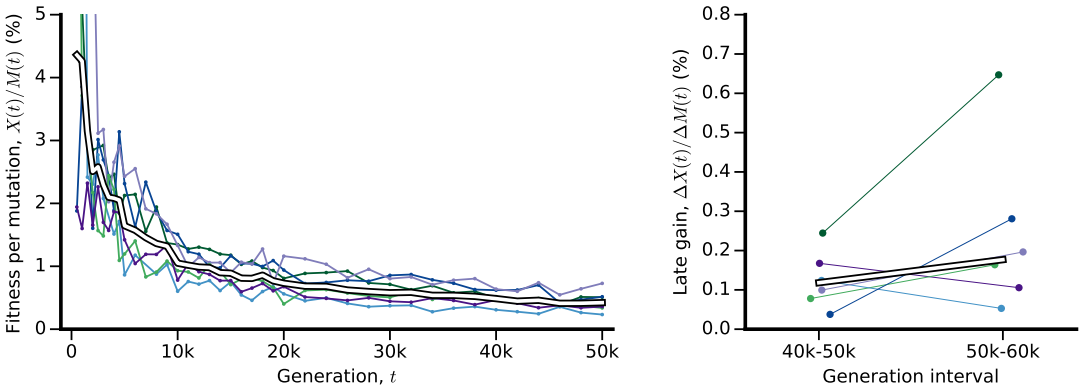

Figure S3: Estimates of the average fitness per mutation in the nonmutator populations. Left: Competitive fitness, $X(t)$, divided by the total allele frequency, $M(t)$, as a function of time. Each population is colored as in Fig. 2. Point estimates (dots) are calculated for timepoints with fitness measurements in Wiser et al. [5], and are connected by solid lines. The ratio between the ensemble averaged trajectories is shown in white. Right: Late fitness gains per mutation estimated from the fitness assays in Lenski et al. [6]. Mutation gains are estimated from the slope of $M(t)$ over the corresponding time interval. Each population includes a small amount of noise on the t-axis to enhance visibility.

3 Library construction and sequencing

3.1 Metagenomic samples

We obtained metagenomic samples from frozen gylcerol stocks for each replicate line through generation 60,000 (a full list of samples is available in Supplementary Table 1). We revived aliquots of ~ 10 μl overnight in 2 ml of DM25 in shaken 10-ml Falcon tubes, after which ~ 1 ml of culture was taken for metagenomic extraction using the GenElute Bacterial Genomic DNA Kit (Sigma). We prepared DNA sequencing libraries using the Illumina Nextera kit as described previously [9], and sequenced them using an Illumina HiSeq 2500 with 150bp paired-end reads at the Bauer Core Facility at Harvard University.

3.2 Clonal isolates

To aid in variant detection, we also supplemented our metagenomic data with data from 264 clonal isolates sequenced in an earlier study of the LTEE [3]. A full list of isolates is available in Supplementary Table 2. Raw FASTQ files were downloaded from the NCBI BioProject database (accession PRJNA294072), and were subjected to the same variant calling pipeline described below.

4 Variant calling

We called variants using a custom pipeline that extends the `breseq` software package [10]. This program identifies candidate single nucleotide variants (SNVs) and small indels by mapping sequencing reads to the reference genome of REL606; larger structural variants (SVs) are identified by looking for reads that contain a *junction* between two distinct locations in the reference genome [11]. Our pipeline uses a custom version of `breseq` v0.26 (which we term `breseq-lite`), which has been modified to reduce potential double-counting of junction-supporting reads.

4.1 Identifying candidate variants

Raw sequencing reads were first trimmed using `trimmomatic` v0.32 [12]. We then used `breseq` to align the trimmed reads to the REL606 reference genome and to generate a list of candidate junctions for each sample in each population. Candidate junctions were extracted and merged per-population using `gdtools`. We then ran `breseq` a second time for each sample, with junction candidates for each sample provided by the `--user-evidence-gd`, using the merged junction list for its respective population.

We used `samtools mpileup` and custom Python scripts to identify candidate SNVs and small indels from the BAM files produced by the second `breseq` run. By counting the reads supporting the alternate and reference alleles, we defined a trajectory of ordered pairs, (A_{pmt}, D_{pmt}), representing the alternate allele count and the total depth of coverage for mutation m in sample t from population p. As a control, we also resequenced the ancestral REL606 population using the same protocols as above, and we used this sample as an initial timepoint for all 12 populations. (The two point mutations separating the REL606- and REL607-derived populations were flagged and removed from our downstream analyses.)

For candidate indels, it is sometimes difficult for `samtools mpileup` to identify the correct alternate allele in regions of repetitive reference sequence. Thus, whenever a small (< 100bp) indel allele is detected at a site, we merged the reads of all alternate alleles in that population into a

single "compound" mutation trajectory. Indel alleles larger than 100bp were considered structural variants, and were processed with the remaining candidate junctions below.

Candidate junctions were processed using a second set of scripts. Like `samtools mpileup`, the `breseq` algorithm has similar issues identifying junction alleles in regions of repetitive reference sequence, and this problem is magnified when many timepoints are included. To address this issue, we used a custom Python script to merge similar candidate junctions from the same population into a single "compound" junction candidate, and we recorded an analogous trajectory (A_{pmt}, D_{pmt}) for each junction m in population p. To conserve space, candidate junctions were only recorded if $A_{pmt} \geq 2$ in at least two samples. We then used a second Python script to merge pairs of candidate junction trajectories into a single structural variant trajectory (e.g., an IS-mediated insertion) when possible.

To conserve space, candidate mutations were only recorded for a population if $A_{pmt} \geq 2$ in at least two samples, and if at least one of these samples had $D_{pmt} \geq 10$ and an empirical frequency $f_{pmt} \equiv A_{pmt}/D_{pmt} \geq 0.05$.

4.2 Quantifying the statistical support for each candidate variant

To distinguish true mutations from sequencing errors, we sought to assess the significance of each mutation by comparing it to an appropriate null model in which the reads supporting the alternate allele arise purely by chance. In the most idealized case, such a null model might be of the form

$$A_{pmt} \sim \text{Binomial}(D_{pmt}, p_m)\,, \tag{3}$$

where p_m is an error probability that may vary from site-to-site. Although this model is theoretically appealing, our data contain many examples of errors that are highly unlikely under this model, and would therefore be classified as real mutations. Instead, the data suggest that the correct error model is one in which the error probability can also vary between samples and between populations, leading us to consider models of the more general form:

$$A_{pmt} \sim \text{Binomial}(D_{pmt}, p_{pmt})\,. \tag{4}$$

Because this model is extremely flexible, we require additional constraints on p_{pmt} in order to distinguish these errors from true mutations.

4.2.1 Removing low-coverage timepoints

One likely source of variation in p_{pmt} arises when the error probability is correlated with the depth of coverage. There are several potential ways that this could occur. First, the coverage of the entire sample could be anomalously small due to errors in library construction or demultiplexing. The remaining reads would therefore be more likely to reflect library construction artifacts or contamination from other samples. To minimize these issues, we automatically eliminated samples in which the median depth of coverage, $\overline{D}_{pt} \equiv \text{median}(D_{pmt})_m$, was less than 5.

Mapping artifacts can also lead to error probabilities that depend on t, since they are strongly influenced by the other mutations present in the sample and typically result in additive (rather than multiplicative) errors in A_{pmt}. These errors will have the largest relative impact when D_{pmt} is low, so we also eliminated individual samples for which $D_{pmt} < 5$. In addition, mapping artifacts can be particularly problematic when the reference base or its close neighbors have been deleted from the genome, so that an anomalously large fraction of the remaining reads may actually consist of mapping errors (even when $D_{pmt} > 5$). To mitigate these issues, we attempted to "trim" the remaining timepoints of a trajectory if there was evidence for a deletion at that site.

We scanned for evidence of deletions by focusing on the *depth trajectory* of each mutation, defined as $d_{pmt} = D_{pmt}/\overline{D}_{pt}$ for all metagenomic samples for which $\overline{D}_{pt} \geq 5$. Due to systematic variation in coverage along the genome, the depth trajectory may be consistently less than unity for some m, and substantially larger for others. We did not attempt to model this variation in m. Instead, we wanted to know whether there is evidence for a sudden and consistent change in d_{pmt} as a function of t, independent of the site-specific coverage. To do so, we fit each depth trajectory to a piecewise model

$$d_{pmt} \sim \text{Normal}(\lambda_t, \sigma_t^2)\,, \tag{5}$$

$$(\lambda_t, \sigma_t^2) = \begin{cases} (\lambda_0, \sigma_<^2) & \text{if } t \leq t^*\,, \\ (\lambda_0 r, \sigma_>^2) & \text{if } t > t^*\,, \end{cases} \tag{6}$$

where λ_i, σ_i^2, and t^* are free parameters. For a given t^*, the maximum likelihood estimates of the remaining parameters are straightforward to calculate. If $n_<$ and $n_>$ denote the number of samples $\leq t^*$ or $> t^*$, then

$$\lambda_0(t^*) = \frac{1}{n_\leq} \sum_{t \leq t^*} d_{pmt}\,, \tag{7}$$

$$r(t^*) = \frac{1}{\lambda_0} \frac{1}{n_>} \sum_{t > t^*} d_{pmt}\,, \tag{8}$$

$$\sigma_<(t^*) = \sqrt{\frac{1}{n_<} \sum_{t \leq t^*} (d_{pmt} - \lambda_0)^2}\,, \tag{9}$$

$$\sigma_>(t^*) = \sqrt{\frac{1}{n_>} \sum_{t > t^*} (d_{pmt} - \lambda_0 r)^2}\,, \tag{10}$$

and t^* can be found by numerically maximizing the remaining loglikelihood,

$$\ell(t^*) = -n_> \log\left(\sigma_<(t^*)\right) - n_> \log\left(\sigma_>(t^*)\right)\,. \tag{11}$$

We carried out a likelihood ratio test between this model and one in which $t^* = \infty$, restricting the alternate hypothesis to $r < 1/2$ in order to focus on potential deletions. The test statistic is

$$\Delta\ell(\{d_{pmt}\}) = \max_{t^*,\, r \leq 1/2} \left\{ n_< \log\left(\sigma/\sigma_<\right) + n_> \log\left(\sigma/\sigma_>\right) \right\}\,, \tag{12}$$

where $\sigma^2 = \frac{1}{n_< + n_>} \sum_t d_{pmt}^2 - \left(\frac{1}{n_< + n_>} \sum_t d_{pmt} \right)^2$ is the variance of the entire depth trajectory. We estimated the null distribution of this test statistic by focusing on the distribution of $\Delta\ell(\{d_{pm\hat{\sigma}(t)}\})$, where $\hat{\sigma}(t)$ is a random permutation of the sample indices. This preserves the observed variation in d_{pmt} (which may be larger than the parametric model above), while ensuring that the consistent changes in d_{pmt} arise purely by chance. Based on this null model, we defined a P-value for each mutation trajectory,

$$P = \Pr\left[\Delta\ell(\{d_{pm\hat{\sigma}(t)}\}) \geq \Delta\ell(\{d_{pmt}\}) \right]\,, \tag{13}$$

which we estimated numerically by simulating a large number of permuted trajectories, $d_{pm\hat{\sigma}(t)}$. For the purposes of masking timepoints, we considered there to be significant evidence for a deletion if $P < 10^{-2}$. This threshold is purposely permissive in order to be conservative in the set of unmasked timepoints. When this condition was met, we removed all samples that occurred after t^* from further downstream analyses.

4.2.2 Error model for moderate-coverage timepoints

We relied on the masking scheme above to remove the most extreme cases in which the error probabilities p_{pmt} are correlated with the sample index t. For the remaining moderate-coverage timepoints, the error probabilities are still too variable for the uniform model, $p_{pmt} = p_m$, to apply. But we will now operate under the assumption that this variation is at least statistically independent of t, and can be approximated by drawing p_{pmt} from some fixed distribution.

To model p_{pmt}, we turned to an ad hoc generative model that attempts to leverage the large number of timepoints available per population. The basic idea is that for a true error, the pooled distribution of the empirical frequencies, $f_{pmt} = A_{pmt}/D_{pmt}$, for a given m and p may provide a reasonable approximation to the distribution of p_{pmt}, even though the estimates for individual t are highly overfitted. In the LTEE, we also have an additional piece of information: because the ancestral allele in REL606 is known with high certainty, we expect the average frequency, $\overline{f}_{pm} = \sum_t A_{pmt} / \sum_t D_{pmt}$, to be less than 50% (otherwise, the alternate allele would have been classified as the reference). Thus, we define a scale factor,

$$c = \max\left\{1,\, 1/2\overline{f}\right\} , \tag{14}$$

such that the renormalized frequencies cf_{pmt} are forced to satisfy this polarization condition. For a true error, this renormalization should not bias the distribution of p_{pmt}. If we again let $\hat{\sigma}(t)$ denote a random permutation of the timepoints, we can then define a model of the error probabilities,

$$\hat{p}_{pmt} = cf_{pm\hat{\sigma}_p(t)} , \tag{15}$$

which preserves the observed variation in cf_{pmt}, but erases all other temporal information. Conditioned on $\hat{p}_{pmt}$, the alternate allele count under this error model is given by

$$\hat{A}_{pmt} \sim \text{Binomial}(D_{pmt},\, \hat{p}_{pmt}) . \tag{16}$$

However, in order to sample from the null model more efficiently, we use a slightly different connection between the error probability and the read count:

$$\alpha_{pmt} \sim \text{Poisson}\left[D_{pmt}\hat{p}_{pmt}\right] , \tag{17a}$$

$$\beta_{pmt} \sim \text{Poisson}\left[D_{pmt}(1-\hat{p}_{pmt})\right] , \tag{17b}$$

$$\hat{A}_{pmt} = \text{round}\left[\frac{\alpha_{pmt}}{\alpha_{pmt}+\beta_{pmt}} \cdot D_{pmt}\right] , \tag{17c}$$

which shares the many of the basic features of the binomial model, but is faster to simulate. Together, these equations define an algorithm for simulating error mutation trajectories that are as close as possible to the observed data, but with an enforced polarization condition and scrambled temporal information.

4.2.3 Quantifying deviations from the error model

Given an error model, the next step is to quantify how the observed read counts, A_{pmt}, differ from the null distribution of $\hat{A}_{pmt}$. Some care must be taken at this point. Due to the approximate nature of the null model, we always expect some differences to arise (even for a true error), but we want to prioritize those differences that are most closely associated with a real mutation. And we want to do so in a way that the evidence for different mutation trajectories can be compared with each other. P-values provide a natural means for carrying out this comparison. In particular,

we choose a set of test statistics $T_k(\{A_{pmt}, D_{pmt}\})$, each of which is a function of the observed mutation trajectory. To combine evidence across the different test statistics, we first define a set of single-statistic P-values,

$$P_k = \Pr\left[T_k(\{\hat{A}_{pmt}, D_{pmt}\}) \geq T_k(\{A_{pmt}, D_{pmt}\})\right] , \tag{18}$$

which we calculate numerically by simulating a large number of draws from $\{\hat{A}_{pmt}, D_{pmt}\}$. We then use these individual P-values to define a composite statistic,

$$T(\{A_{pmt}, D_{pmt}\}) = \sum_k \theta(P^* - P_k) \log\left(\frac{1}{P_k}\right) , \tag{19}$$

where $\theta(\cdot)$ is the Heaviside step function and $P^* = (0.05)^{1/3}$. We chose this threshold so that the composite statistic is dominated by large deviations in at least one of the individual test statistics, rather than small or moderate deviations in all of them. Based on this definition, we then define a single composite P-value according to

$$P_{pm} = \Pr\left[T(\{\hat{A}_{pmt}, D_{pmt}\}) \geq T(\{A_{pmt}, D_{pmt}\})\right] , \tag{20}$$

which is again calculated numerically by simulating a large number of draws from $\{\hat{A}_{pmt}, D_{pmt}\}$. Increasing the number of test statistics generally leads to greater power to reject the null hypothesis, but it also makes the test more sensitive to the specific assumptions of the error model. In the present manuscript, we used three test statistics which were chosen to reflect the features of putatively real mutations in the LTEE.

Autocorrelation. If the timepoints are sampled densely enough to measure the trajectory of a mutation, we expect that the frequencies at nearby timepoints will be correlated with each other. For example, a mutation might have $f_t = 0$ for several consecutive timepoints until the mutation first rises to detectable frequencies, and then it will undergo a polymorphic phase before it either permanently survives or goes extinct. Previous work has attempted to capture these dynamics using the autocorrelation function,

$$C \equiv \sum_t (f_{t+1} - \overline{f})(f_t - \overline{f}) , \tag{21}$$

based on the assumption that C will be small (or negative) for errors and large and positive for a true mutation trajectory [13]. Here, we use a modified autocorrelation function C^* that accounts for the discreteness and uncertainty in f_t (due to finite coverage) as well as the polarization constraint ($\overline{f} \leq 50\%$) that should hold for a true error. In particular, we define

$$\overline{f} = \min\left\{\frac{\sum_t A_{pmt}}{\sum_t D_{pmt}}, \frac{1}{2}\right\} , \tag{22}$$

which is an average over $f_{pmt} = A_{pmt}/D_{pmt}$ that weights each timepoint by D_{pmt}. Similarly, in the sum in C^*, we weight each pair of timepoints by $\sqrt{D_{pm,t+1} D_{pmt}}$. To minimize rounding artifacts, we only count differences $f_t - \overline{f}$ if they correspond to at least one read count, i.e., if $|A_{pmt} - \overline{f} D_{pmt}| \geq 1$. Putting everything together, this yields a modified autocorrelation function

$$C^* = \frac{\sum_t \sqrt{D_{t+1} D_t}(f_{t+1} - \overline{f})(f_t - \overline{f})\theta(|A_{t+1} - \overline{f} D_{t+1}| - 1|)\theta(|A_t - \overline{f} D_t| - 1|)}{\overline{f}^2 \sum_t \sqrt{D_{t+1} D_t}} , \tag{23}$$

where $\theta(z)$ is the Heaviside step function, and we have normalized by $\overline{f}^2$ to obtain a coefficient of variation.

Derived allele sojourn weight. One disadvantage of the autocorrelation is that (for $\overline{f} < 1/2$), it treats positive and negative deviations from the mean symmetrically. But for true mutations, timepoints with zero or near-zero frequencies have a special interpretation — the mutation has not yet arisen or has gone extinct — while consecutive runs of positive frequency represent the sojourn path of the mutation. A larger area under the curve of one of these runs provides more evidence that the mutation is not an error. To quantify these features, we looked at runs of 2 or more timepoints, $t = t_1, \ldots, t_2$, for which f_t is larger than some threshold frequency f^* for all $t_1 \leq t \leq t_2$. We then recorded the run with the largest value of

$$I = \sum_{t=t_1}^{t_2} f_t - f^* \,. \tag{24}$$

For each trajectory, we attempted to choose f^* to be as close as possible to an unborn/extinct state, while still allowing for error rates that can be higher than $1/D_t$. To do so, let n and n_0 respectively denote the total number of timepoints and the number of timepoints for which $A_{pmt} = 0$, and let $\overline{f}$ denote the capped mean frequency in Eq. (22). We then defined the threshold f^* as

$$f^* = \begin{cases} \frac{\overline{f}}{1+e^{+(n_0-0.3n)/5}} & \text{if } n_0 > 0.3n \,, \\ \frac{\overline{f}}{1+e^{-(0.3n-n_0)/5}} & \text{if } n_0 < 0.3n \,. \end{cases} \tag{25}$$

When more than 30% of the timepoints are zero alternate alleles, this expression weights f^* closer to the median, which grows closer to zero as n_0/n increases. In the opposite case, when fewer than 30% of the timepoints are zero, this expression reverts back to the average, $\overline{f}$.

Average frequency relaxation time. Because the LTEE was founded from a clonal ancestor, a true mutation should start with zero or near-zero frequency, and only later rise to higher frequencies. This means that the average frequency for the first several timepoints should be lower than the average frequency for the entire timecourse. To quantify this tendency, we calculated the *relaxation time* T, which is defined to be the maximum number of timepoints for which the partial average allele frequencies from 5 to $T-1$ are all less than 60% of the average allele frequency, i.e.,

$$T = \max \left\{ T \,:\, \frac{\sum_{t \leq t'} A_{pmt}}{\sum_{t \leq t'} D_{pmt}} \leq 0.6 \cdot \frac{\sum_t A_{pmt}}{\sum_t D_{pmt}} \quad \forall \quad 5 \leq t' < T \right\} . \tag{26}$$

We set $T = 0$ if $n_0 > 0.3n$.

4.3 Filtering to obtain final list of mutations

After estimating a P-value for each mutation m in each population p using the algorithm above, we corrected for multiple hypothesis testing by converting these P-values into genome-wide Q-values to assess significance. Since the distribution of P-values may differ between the mutator and nonmutator populations, we calculate the Q-values separately for the two groups using the formula

$$Q_{pm} = \min_{Q > P_{pm}} \left\{ \frac{Q \sum_{p',m'} 1}{\sum_{p',m'} \theta(Q - P_{p',m'})} \right\} , \tag{27}$$

where $\theta(\cdot)$ is the Heaviside step function [14]. As expected, the Q-values increase with the total number of tests performed ($\sum_{p','m'} 1$), regardless of the evidence for a mutation at that site. To

boost power, it can be useful to refrain from performing the test for sites with insufficient variation to support a mutation, even if the P-value would seem to indicate otherwise. In other words, we are free to define a function $F(\{A_{pmt}, D_{pmt}\})$ and restrict our analysis to mutations for which $F(\{A_{pmt}, D_{pmt}\}) = 1$, provided that we modify our error model to only produce trajectories for which $F(\{A_{pmt}, D_{pmt}\}) = 1$ as well. We implemented such a restriction here, using a function $F(\{A_t, D_t\})$ that is equal to 1 if all of the following conditions are met:

1. There are at least two samples for which $A_t \geq 2$, and at least one of these has $D_t \geq 10$ and $A_t/D_t \geq 0.05$. This condition is always required for the null model (and fulfilled for the data) because it was used when compiling the original list of candidate mutations.

2. In the ancestral sample, $D_0 \geq 10$ and $A_0/D_0 \leq 0.1$.

3. There are at least three samples for which $A_t \geq 2$ and $D_t \geq 5$.

4. There is at least one sample for which $A_t \geq 3$, $D_t \geq 5$, and $A_t/D_t \geq 0.1$.

5. The difference between the maximum frequency and the capped average in Eq. (22) is at least $+10\%$.

Given this ensemble of mutation trajectories, we calculated Q-values for each candidate mutation in the mutator and nonmutator populations, and rejected all those candidates with $Q \geq 5\%$.

4.3.1 Removing mutations in repetitive regions of the genome

We excluded mutations that arose in repetitive regions of the genome, as these can be difficult to resolve using short-read 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. [3], or (3) it fell within the putative prophage element identified by Tenaillon et al. [3] (REL606 genome coordinates 880528–904682). Approximately 2×10^5 sites ($\approx 4\%$ of the reference genome) matched one of these criteria and were excluded from all downstream analyses.

4.3.2 Additional filtering using data from clonal samples

By using the read frequency A_t/D_t as an estimate of the population frequency, we are implicitly assuming that a mutation is present in $\approx 100\%$ of the reads in each mutant cell and $\approx 0\%$ of the reads in each wildtype cell. Certain types of mutations will violate this assumption, e.g., a SNV that arises after a duplication event, or residual errors with enough temporal correlations to pass through our previous filters. In either case, we wish to remove such "non-clonal" mutations from our list, since their allele frequency trajectories could interfere with our downstream analyses.

To detect non-clonal mutations, we utilized read information from the clonal samples sequenced by Tenaillon et al. [3], which we processed using the same pipeline as our metagenomic samples, but have so far neglected. In principle, we expect that true mutations should be present in either 0% or 100% of the reads in each clone, but in practice, sequencing and mapping errors (which may occur more frequently in the shorter read-lengths of the clone data) force us to employ less stringent criteria. We filtered out non-clonal mutations using the following empirically-derived thresholds, which were tuned to balance true and false positive rates for the most common types of errors in our dataset. In particular, we filtered out mutations if any of the following conditions are met:

- There were ≥ 4 clones with $0.1 \leq A/D \leq 0.7$, and of these clones, $\geq 50\%$ were sampled at timepoints when the population frequency of the mutation satisfied $0.2 \leq f_t \leq 0.7$.

- All of the clones (as well as the population frequency at the timepoints when the clones were sampled) had $A/D \leq 0.6$, but there was at least one timepoint for which the clone frequency and the population frequency was ≥ 0.1.
- There is at least one clone with $0.4 \leq A/D \leq 0.6$ that also has a depth ratio $r > 1.3$ (see Section 4.2.1), and none of the clone or population frequencies exceeded 0.9.
- There were > 10 timepoints for which the population frequency was > 0.25, the average depth ratio for these timepoints was > 1.5 times higher than the average depth ratio in the first 10 timepoints, $\geq 90\%$ of these timepoints had $f_t \leq 0.75$, and none of the clone frequencies exceeded 0.9.

4.4 Mutation annotation

Each mutation was assigned a gene and a variant type depending on its location and alternate allele. To do so, we first partitioned the *E. coli* genome into genes (including tRNA and rRNA genes) and intergenic regions according to the annotations in the REL606 reference. In the case of overlapping genes, priority was given to the gene with the lowest left coordinate. In addition, we also included 100bp upstream of each gene's start codon, which we treated as a putative promoter region. In case of overlaps, we again gave priority to the gene with the lowest left coordinate, and genic sequence was always given priority over promoter sequence.

Based on these annotations, we assigned each mutation to a gene (or classified it as "intergenic") according to its location in the genome. For deletions and other junction candidates that span multiple bases, the location was defined to be the left-most genome coordinate. This convention allows us to assign a unique gene to each mutation, but it throws away information about the other genes that were modified, even if they might have been the true "target" of selection. For example, several IS-mediated deletions with endpoints in *yieO* are classified to that gene, though previous work shows that the likely target was the *rbs* operon [15]. Similar issues apply to larger duplications and inversions. Because we can only identify the influenced bases in a metagenomic sample if the mutation rises to a sufficiently high frequency, it is difficult for us to estimate which genes are influenced by a particular structural variant without additional information (e.g. from clone sequences [3]). We therefore chose not to account for these "off-target" genes in our present analysis. The resulting errors (in addition to errors induced by incorrect promoter annotation and overlapping genes) will generally reduce our power to detect genetic parallelism in Section 6, but they avoid introducing spurious signals due to misidentified mutations.

We then assigned a variant type to each mutation depending on its gene and alternate allele. Single-nucleotide changes in coding regions were classified as nonsense, missense, and synonymous if they resulted in an early stop codon, an amino-acid change, or no amino-acid change, respectively, and the remaining point mutations were classified as noncoding. Indels (< 100 bp) and larger structural variants were annotated accordingly, regardless of whether they occurred in genes or intergenic regions.

5 Mutation trajectory inference

After the initial filtering step, we assume that the remaining read count trajectories (A_{pmt}, D_{mt}) provide noisy readouts of the true allele frequencies f_{pmt} in each population through time. On average, we expect that

$$\left\langle \frac{A_{pmt}}{D_{pmt}} \right\rangle \approx f_{pmt} \,, \tag{28}$$

so in the absence of additional information, the naive estimator $\hat{f}_{pmt} = A_{pmt}/D_{pmt}$ is our best guess for the true frequency, f_{pmt}. We use this as our default estimator throughout the remainder of the text, unless stated otherwise.

5.1 Rate of mutation accumulation

The expected allele frequency in Eq. (28) is also the expected probability that the mutation m is present in a randomly sampled individual from the population. A natural measure of mutation accumulation in each population is therefore given by

$$M_p(t) \equiv \sum_m \hat{f}_{pmt} \,, \tag{29}$$

which is plotted in Fig. 2b in the main text. When averaged across populations, the derivative of Eq. (29) also provides a natural measure of the rate of mutation accumulation. To estimate the derivative in Fig. 2c, we performed local linear regression of $\overline{M}(t) = \sum_p M_p(t) / \sum_p 1$ in sliding $5,000$ generation windows, and we estimated uncertainties by randomly resampling six nonmutator populations with replacement and recalculating this measure.

There is also variability in $M_p(t)$ among the six nonmutator populations in Fig. 2b. The overall magnitude of variation in any given time interval depends on the detailed parameters of the underlying population genetic process [16], which we do not try to estimate here. Instead, we want to ask whether there are systematic differences in the rate of mutation accumulation between populations that are correlated over multiple time intervals.

To investigate this question, we turned to an empirically-derived null model of mutation accumulation, in which the populations are assumed to be identical, while still controlling for certain features of the observed distribution of $M_p(t)$. Because different mutation trajectories are expected to be correlated for times less than the typical fixation timescale, we divided the full timecourse into six non-overlapping windows of $\Delta t = 10,000$ generations, and we estimated the average slope $\partial_t M_p$ for each population within each window using linear regression. This allows us to represent each population p by a vector of *mutation gains*, $\Delta M_{p,k} = (\Delta t \cdot \partial_t M_p)|_{t=t_k}$, for each window $k = 1, \ldots, 6$ (Extended Data Fig. 1).

We then define the null model of mutation accumulation by randomly permuting the population labels within each time interval. Each permutation creates a new bootstrapped dataset of 6 trajectories in which the mutation gains are effectively uncorrelated, while still preserving the distribution of gains in any given interval. A systematic difference in the rate of mutation accumulation between lines would then be manifested in a larger-than-expected variance at the final timepoint, e.g. arising from a population with systematically higher ΔM_k. To test this hypothesis, we calculated the final between-line variance,

$$\sigma_M^2 = \frac{1}{6}\sum_{p=1}^{6}\left(\sum_{k=1}^{6}\Delta M_{p,k}\right)^2 - \left(\frac{1}{6}\sum_{p=1}^{6}\sum_{k=1}^{6}\Delta M_{p,k}\right)^2 \,, \tag{30}$$

for both the observed and bootstrapped trajectories, and estimated a P-value according to the number of times that the bootstrapped variance exceeded the observed value (Extended Data Fig. 1). The observed value is significantly higher than expected under the null model ($P \approx 10^{-3}$), indicating that the additional mutations in each window are correlated with the identity of the population. This excess variability is still significant (though less strongly so) if we remove the Ara+1 population ($P \approx 10^{-2}$), which we expect to be an outlier since it has been shown to harbor an excess of IS-mediated mutations [3].

5.2 Simplest hidden Markov model

While Eq. (28) provides a reasonable estimate of the average allele frequency, there is often substantial uncertainty in this estimate, particularly near $f = 0$ and $f = 1$, where sequencing errors can swamp the true signal. In order to detect fixation or extinction events in a robust manner, it is therefore useful to combine information across multiple timepoints.

To do so, we require a model for the true allele frequency as a function of time, as well as an error model connecting f_{mt} with the read counts (A_{mt}, D_{mt}). (In the following sections, we drop the population index p, though it is implicit in all of the equations.) In simple settings, we can obtain a model for f_{mt} from population genetic considerations (e.g., Ref. [17]). However, because we do not know the correct population genetic model for the LTEE, we turned to an ad hoc Markov model that contains some of the minimal features of a true mutation trajectory, while attempting to remain as flexible as possible.

In the simplest version of this model, all mutations start in an ancestral state **A** where $f_{mt} = 0$, so that alternate reads in this state are assumed to arise from sequencing errors. At each timepoint, there is a finite probability that the mutation transitions to a polymorphic state **P**, where it can spend several timepoints with frequency $0 < f_{mt} < 1$. From this polymorphic state, the mutation can transition to either a fixed state **F**, where all reference reads are sequencing errors, or an extinct state **E** that is similar to the ancestral state above. In rare cases, mutations are allowed to transition from **E** back to **A**, after which it can be reborn as a recurrent mutation event.

For mutations in the polymorphic state **P**, we must also define how the allele frequencies $0 < f_{mt} < 1$ are chosen. In the absence of additional information, we assume that the new allele frequency is drawn uniformly between 0 and 1, independent of the previous allele frequency. This allows us to coarse-grain over the allele frequency f_{mt} and obtain a simple Markov chain between the **A**, **P**, **F**, and **E** "macrostates" which is illustrated in Fig. S4. We fixed the transition rates empirically based on the observed properties of the experiment:

$$\Pr[\mathbf{A} \to \mathbf{P}] \approx 10^{-2}\,, \quad \Pr[\mathbf{P} \to \mathbf{F}] \approx \Pr[\mathbf{P} \to \mathbf{E}] \approx 0.5\,, \quad \Pr[\mathbf{E} \to \mathbf{P}] \approx 10^{-6}\,, \tag{31}$$

though the results are relatively insensitive to the precise values.

To connect this Markov model with the observed data, we also require a model for generating read counts (A_{mt}, D_{mt}) based on the current state. As in Section 4, we assume that these follow a binomial mixture,

$$A_{mt} \sim \text{Binomial}(D_{mt}, p_{mt})\,, \tag{32}$$

where the p_{mt} are independent random variables whose distribution depends on the current state. For example, in the polymorphic state we assume that $p_{mt} \approx f_{mt}$, so that p_{mt} is uniformly distributed in the interval $(0, 1)$. After marginalizing over p_{mt}, we find that

$$\Pr[A_{mt}|\mathbf{P}] = \frac{1}{D_{mt} + 1}\,. \tag{33}$$

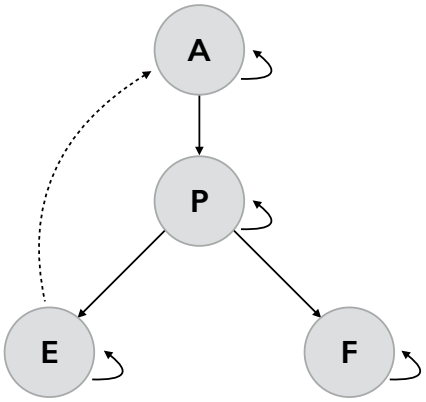

Figure S4: Schematic diagram of the possible transitions between the ancestral (**A**), polymorphic (**P**), fixed (**F**), and extinct (**E**) states in the simplest mutation trajectory HMM. The relative weights of each transition are specified in Eq. (31). The dashed line from **E** $\rightarrow$ **A** denotes a particularly low-probability transition that models a recurrent mutation event.

Similarly, in the **A** and **E** states, the true frequency is zero, so p_{mt} represents the contribution from sequencing errors. For simplicity, we assume that these are drawn uniformly between 0 and some number $2p_m^{\text{err},0}$, which is specific to each mutation m and which is capped at some low frequency $p_{\max}^{\text{err},0} \approx 0.025$. After marginalizing over p_{mt}, we obtain

$$\Pr[A_{mt}|\mathbf{A}] = \frac{1}{D_{mt}+1} \int_0^{2p_m^{\text{err},0}} \frac{\Gamma(D_{mt}+2)p^{A_{mt}}(1-p)^{D_{mt}-A_{mt}}}{\Gamma(A_{mt}+1)\Gamma(D_{mt}-A_{mt}+1)} \frac{dp}{2p_m^{\text{err},0}}\,, \tag{34}$$

and similarly for $\Pr[A_{mt}|\mathbf{E}]$. An analogous situation holds for the fixed state **F**, except with the roles of A_{mt} and $D_{mt} - A_{mt}$ reversed:

$$\Pr[A_{mt}|\mathbf{F}] = \frac{1}{D_{mt}+1} \int_0^{2p_m^{\text{err},1}} \frac{\Gamma(D_{mt}+2)p^{D_{mt}-A_{mt}}(1-p)^{A_{mt}}}{\Gamma(A_{mt}+1)\Gamma(D_{mt}-A_{mt}+1)} \frac{dp}{2p_m^{\text{err},1}}\,, \tag{35}$$

where we have allowed for the possibility that $p_{pm}^{\text{err},0} \neq p_{pm}^{\text{err},1}$. Thus, the polymorphic state allows the mutation to travel beyond $2p_m^{\text{err},0}$ and $1-2p_m^{\text{err},1}$, but at the cost of a higher state-space entropy.

Together, these equations define a simple hidden Markov model (HMM) that consists of a sequence of hidden states $L_{mt} \in \{\mathbf{A}, \mathbf{P}, \mathbf{F}, \mathbf{E}\}$ and an observed sequence of emissions, (A_{mt}, D_{mt}). We fit this model to the observed data using standard dynamic programming techniques [18]. The primary quantity of interest is the matrix of posterior state probabilities,

$$P_{mt\ell} \equiv \Pr[L_{mt} = \ell|\text{data},\, L_{m,0} = \mathbf{A},\, L_{m,t_f} \neq \mathbf{A}]\,, \tag{36}$$

which can be can expressed in the form

$$P_{mt\ell} \propto \underbrace{\Pr[L_{mt} = \ell,\, \text{data}_{\leq t}|L_{m,0} = \mathbf{A}]}_{F_{mt\ell}} \cdot \underbrace{\Pr[L_{m,t_f} \neq \mathbf{A},\, \text{data}_{>t}|L_{mt} = \ell]}_{B_{mt\ell}}\,, \tag{37}$$

where F_{mtl} and $B_{mt\ell}$ are the canonical forward and backward tables. These satisfy the recursion relations:

$$F_{mt\ell} = \Pr[A_{mt}|\ell] \sum_{\ell'} F_{m,t-1,\ell'} \Pr[\ell' \rightarrow \ell]\,, \quad F_{m,0,\ell} \propto \delta_{\ell,\mathbf{A}}\,, \tag{38a}$$

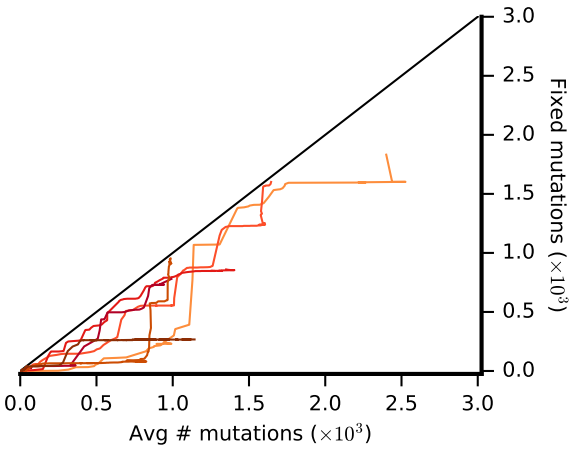

Figure S5: An analogous version of Fig. 2d replotted for the six mutator populations. Each population is colored according to the same color scheme as Fig. 2.

and

$$B_{mt\ell} = \sum_{\ell'} \Pr[\ell \to \ell'] \Pr[A_{m,t+1}|\ell'] B_{m,t+1,\ell'} \,, \quad B_{m,t_f,\ell} \propto 1 - \delta_{\ell,\mathbf{A}} \,, \tag{39a}$$

where the initial conditions ensure that all mutations must start in the **A** state and spend at least one timepoint in the **P** state.

The recursion relations for $F_{mt\ell}$, $B_{mt\ell}$, and $P_{mt\ell}$ above depend on the unknown error rates $p_m^{\mathrm{err},0}$ and $p_m^{\mathrm{err},1}$. We estimate these using an iterative scheme similar to the familiar expectation maximization (EM) algorithm [18]. We first calculate $P_{mt\ell}$ using an initial guess $p_m^{\mathrm{err},0} = p_m^{\mathrm{err},1} = 0.01$. Then, using the fact that

$$\langle A_{mt} | \, \mathbf{A} \rangle = \langle A_{mt} | \, \mathbf{E} \rangle = p_m^{\mathrm{err},0} D_{mt} \,, \tag{40}$$

we obtain a new estimate of $p_m^{\mathrm{err},0}$ via the weighted average,

$$p_m^{\mathrm{err},0} = \min \left\{ \frac{\sum_m A_{mt}(P_{m,t,\mathbf{A}} + P_{m,t,\mathbf{E}})}{\sum_m D_{mt}(P_{m,t,\mathbf{A}} + P_{m,t,\mathbf{E}})}, p_{\max}^{\mathrm{err},0} \right\} , \tag{41}$$

and similarly for $p_m^{\mathrm{err},1}$. The process is repeated for ~ 10 iterations, which is usually sufficient for convergence. The most-likely sequence of states, $\hat{L}_{mt}$, is estimated using an analogous implementation of the Viterbi algorithm [18].

5.2.1 Fixed mutation trajectories.

After estimating the most-likely sequence of states, $\hat{L}_{mt}$, we estimate the number of fixed mutations in each population through time using the formula:

$$M_{\mathrm{fixed}}(t) = \sum_m \delta_{\hat{L}_{mt},\mathbf{F}} \,. \tag{42}$$

The corresponding fixed mutation trajectories are compared with the average mutation trajectories from Eq. (29) in Figs. 2d (nonmutators) and S5 (mutators).

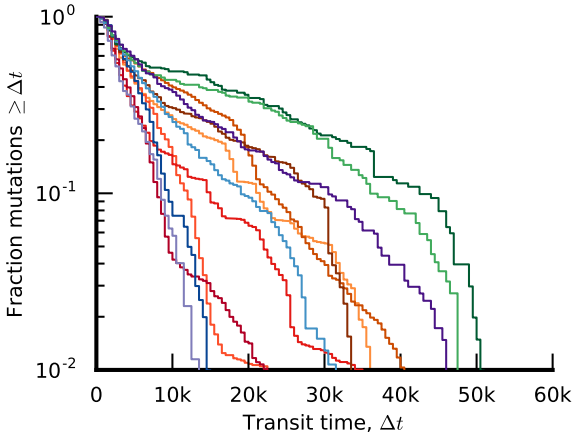

Figure S6: The distribution of transit times estimated from Eq. (45) for each of the twelve populations. Each population is colored according to the same color scheme as Fig 2.

5.2.2 Appearance and transit times

We also use the most-likely sequence of states, $\hat{L}_{mt}$, to assign an *appearance time* $T_{0,m}$ to each detected mutation. Roughly speaking, this is defined to be the first non-zero timepoint before the mutation attains its maximum frequency. To estimate $T_{0,m}$, we first define the reference time t^* to be the point at which the mutation achieved its highest frequency. If the mutation finished the experiment in the polymorphic (**P**) or extinct (**E**) states, we set t^* to be

$$t^* = \text{argmax}\left\{\hat{f}_{mt} \,:\, \hat{L}_{mt} = \mathbf{P}\right\}, \tag{43}$$

while we set t^* to the final timepoint if the mutation fixed. The appearance time $T_{0,m}$ was then defined as

$$T_{0,m} = \max_{t<t^*}\left\{t + 250 \,:\, \hat{L}_{mt} = \mathbf{A}\right\}. \tag{44}$$

We also use $\hat{L}_{mt}$ to estimate the time that the mutation spent in the polymorphic state before transitioning to fixation or extinction. We refer to this as the *transit time*, ΔT_m, which is defined as

$$\Delta T_m = \min_{t>T_{0,m}}\left\{t - T_0 - 250 \,:\, \hat{L}_{mt} \in \{\mathbf{F}, \mathbf{E}\}\right\}. \tag{45}$$

Note that this operational definition of the transit time only accounts for the of the polymorphic phase that is spent at observable frequencies (e.g. between $\sim 1\%$ and 99%). It neglects the additional time required for the mutation to transit from a single copy to $f \sim 1\%$, or from $f \sim 99\%$ to fixation, which can be much larger than ΔT_m. The distribution of transit times for each of the twelve populations is depicted in Fig. S6. The excess of very long transit times in several of the populations provides further evidence that the "missing" fixation events in Figs. 2d and S5 persist at intermediate frequencies for many generations.

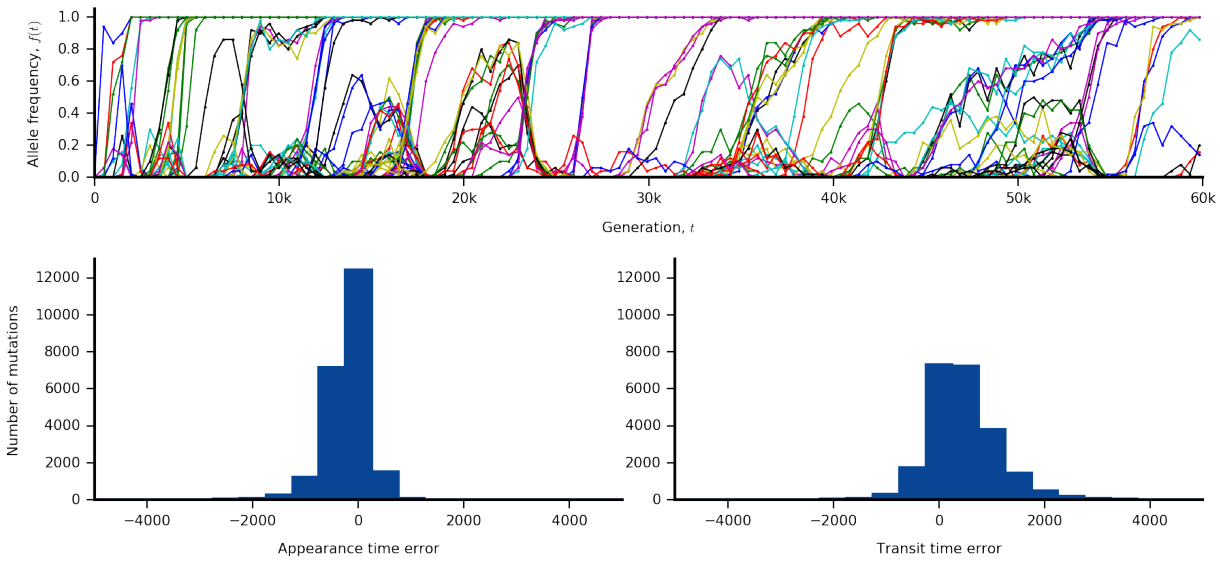

Figure S7: Validation of the single-clade HMM using forward-time simulations. Top: Allele frequency trajectories for a simulated population with $U_n = 6 \times 10^{-4}$, $U_b = 7.5 \times 10^{-5}$, $s_0 = 7 \times 10^{-3}$, and $X_c = 5 \times 10^{-2}$. Allele frequencies are binomially downsampled to a coverage of 50x. Bottom left: Distribution of the difference between the estimated appearance time from the HMM and the time at which the mutation first exceeds 5% frequency, using mutations from 108 simulated populations with the same parameters as above. Bottom right: the corresponding distribution of the difference between the estimated transit time from the HMM and the total time spent between 5% and 95% frequency.

5.2.3 Validation with forward-time simulations

To evaluate the accuracy of the HMM classification scheme, we applied it to a synthetic set of mutations obtained from forward-time simulations of the LTEE, which are described in Good and Desai [19]. In these simulations, individuals acquire neutral and beneficial mutations at rate U_n and U_b, respectively. The beneficial mutations are drawn from a fitness-dependent distribution of fitness effects,

$$\rho(s) = \frac{1}{s_0 e^{-X/2X_c}} \exp\left(\frac{s}{s_0 e^{-X/2X_c}}\right), \tag{46}$$

which has been shown to capture the leading-order effects of diminishing returns epistasis in this experiment [3, 5, 19, 20]. Each mutation creates a SNP at a unique site in the genome, whose frequency can be tracked over time. Every 500 generations, the simulation records the frequencies of all mutations that are present above 1% frequency. These true frequencies are also binomially sampled at 50x coverage to produce a synthetic metagenomic dataset similar to those analyzed in this work. The allele frequency trajectories for one such simulated population are shown in Fig. S7.

We then classified the mutations in this synthetic dataset using the HMM algorithm described above, and calculated the corresponding appearance and transit times, T_0 and ΔT. As ground truth values, we used the first and last time the mutation spent in the frequency range $(5\%, 95\%)$. The error distributions for T_0 and ΔT are shown in Fig. S7, based on mutations pooled across twelve simulated populations. The estimates are generally accurate to a precision of ± 1000 generations.

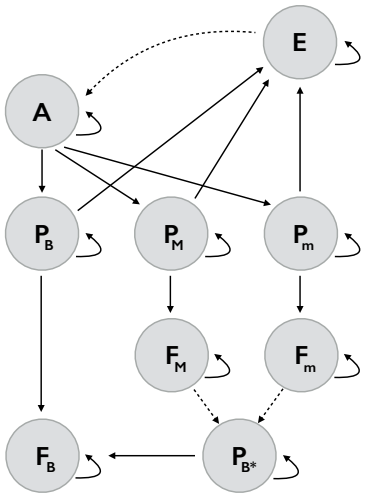

Figure S8: Schematic diagram of the possible transitions between the states in the clade-aware HMM. **B**, **M**, and **m** denote the basal, major, and minor clades, respectively, while $\mathbf{B}^*$ represents a recurrent mutation that occurs independently in both clades.

5.3 Adding a pair of subclades

In many of the LTEE populations, the average number of mutations $M(t)$ is much larger than $M_{\text{fixed}}(t)$ (Figs. 2d and S5). In several of these cases, the mutations appear to accumulate in a pair of intermediate-frequency clades that coexist for thousands of generations (Fig. S6). We therefore extended our HMM to account for the clade background of each mutation in this scenario.

In the clade-aware version of the HMM, all mutations again start in the ancestral state **A**. However, we must now differentiate between mutations that are fixed/polymorphic in the basal clade ($\mathbf{F_B}/\mathbf{P_B}$) or in the major ($\mathbf{F_M}/\mathbf{P_M}$) or minor ($\mathbf{F_m}/\mathbf{P_m}$) subclades. Polymorphic mutations in all three clades can transition to a common extinct state **E**, which can re-transition to one of the polymorphic states in rare cases through recurrent mutation. In addition, recurrent mutation can also spread mutations from major to minor clades or vice versa. We model these rare cases by introducing a final polymorphic state $\mathbf{P_{B^*}}$ that cannot transition to extinction. The network of transitions between the 9 macrostates is illustrated in Fig. S8.

We continue to model the emission probabilities using the binomial mixture in Eq. (32). The distributions of p_{mt} for the **A**, **E**, and $\mathbf{F_B}$ states remain unchanged from the simple model above. To model the emission probabilities in the other states, we must introduce two additional parameters, $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$, which denote the sizes of the major and minor clades through time. (For concreteness, we adopt the convention that the major clade is the one with the highest frequency at the final timepoint.) Mutations that are fixed within the major clade must have p_{mt} close to $f_t^{\mathbf{M}}$,

$$p_{mt}|\mathbf{F_M} \sim \text{Uniform}\left(f_t^{\mathbf{M}} - p_m^{\text{err},0},\, f_t^{\mathbf{M}} + p_m^{\text{err},0}\right) , \tag{47}$$

while polymorphic mutations in this clade can fall anywhere between 0 and $f_t^{\mathbf{M}}$:

$$p_{mt}|\mathbf{P_M} \sim \text{Uniform}\left(0,\, f_t^{\mathbf{M}} + p_m^{\text{err},0}\right) . \tag{48}$$

Analogous equations apply for the minor clade. For polymorphic mutations in the basal clade

(which is ancestral to both $\mathbf{M}$ and $\mathbf{m}$), the frequencies are constrained to remain above $f_t^{\mathbf{M}} + f_t^{\mathbf{m}}$:

$$p_{mt}|\mathbf{P_B} \sim \text{Uniform}\left(f_t^{\mathbf{M}} + f_t^{\mathbf{m}} - p_d^{\text{err},0}, 1\right) , \tag{49}$$

while recurrent mutations in both clades are free of this constraint:

$$p_{mt}|\mathbf{P}_{\mathbf{B}^*} \sim \text{Uniform}\,(0,\, 1) \, . \tag{50}$$

After marginalizing over p_{mt}, we obtain a set of emission probabilities $\Pr[A_{mt}|\ell]$ that define our new HMM.

If the clade trajectories $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$ were known a priori, it would be straightforward to extend our earlier inference scheme to assign mutations to their corresponding clades. However, we typically must estimate the clade trajectories simultaneously from the data. To do so, we again turn to an iterative scheme, loosely based on the EM algorithm. Starting with initial guesses for $f_t^{\mathbf{M}}$, $f_t^{\mathbf{m}}$, and the error probabilities $p_m^{\text{err},0}$ and $p_m^{\text{err},1}$, we infer the matrix of posterior state probabilities $P_{mt\ell}$ and the sequence of most likely states $\hat{L}_{mt}$ using the recursion relations above, and we update the error estimates using the formulae in Eq. (41). To update the clade frequencies $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$, we make use of the fact that

$$\langle A_{mt}|\, \mathbf{F_M}\rangle = f_t^{\mathbf{M}} D_{mt} \, , \tag{51}$$

and estimate $f_t^{\mathbf{M}}$ using the weighted average,

$$f_t^{\mathbf{M}} = \frac{\sum_m A_{mt} P_{m,t,\mathbf{F_M}} \delta\left(\mathbf{F_M}, \text{argmax} P_{m,t_f,\ell}\right)}{\sum_m D_{mt} P_{m,t,\mathbf{F_M}} \delta\left(\mathbf{F_M}, \text{argmax} P_{m,t_f,\ell}\right)} \, , \tag{52}$$

where the δ-function restricts the sum to mutations that finished in the $\mathbf{F_M}$ state. An analogous equation applies for the minor clade. Based on these updated parameter values, we re-estimate the posterior state probabilities $P_{mt\ell}$, and this process is iterated ~ 10 times until convergence.

By convention, we only retained clades that persisted at intermediate frequencies for a sufficiently long period of time. In particular, we required that there was at least one timepoint for which $0.2 \leq f_t^{\mathbf{M}} \leq 0.8$, $0.2 \leq f_t^{\mathbf{m}} \leq 0.8$, and $f_t^{\mathbf{M}} + f_t^{\mathbf{m}} > 0.8$ (i.e., at least one timepoint where the major and minor clades were both large enough to be clearly observed and together accounted for the bulk of the population). The duration of intermediate-frequency coexistence was defined to be the difference between the latest such timepoint and the first timepoint for which $f_t^{\mathbf{M}} + f_t^{\mathbf{m}} > 0.8$. Note that this is only a lower bound on the true duration of coexistence: there are populations like Ara-2 where the minority clade is known to persist at frequencies much lower than 20% for thousands of generations, without rising above this threshold again. Our algorithm is capable of inferring these rare clade frequencies in many cases (see e.g., Fig. 3B). However, in order to be conservative when declaring that a given population shows signs of frequency-dependence, we only included clades where the duration of intermediate-frequency coexistence was at least $10,000$ generations; this was the case for 9 of the 12 populations in Fig. 3B. By adopting this convention, we may miss examples of coexistence that did not spend sufficient time at intermediate frequencies, in addition to those that never reached sufficient frequency to be detected at our present sequencing depth.

We note that while the clade trajectory estimators in Eq. (52) appear to give reasonable results for our data, they are still suboptimal because they ignore information from polymorphic mutations in the $\mathbf{P_B}$, $\mathbf{P_M}$ or $\mathbf{P_m}$ states. Thus, when estimating $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$, the algorithm does not properly penalize the clade frequencies if there are polymorphic mutations in those clades with frequencies

that exceed $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$ (or conversely, mutations in the $\mathbf{P_B}$ state that fall below $1 - f_t^{\mathbf{M}} - f_t^{\mathbf{m}}$). This makes it easier for the algorithm to get trapped in a suboptimal region of parameter space if the initial estimates for $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$ are not chosen carefully. To minimize such stability issues, we excluded indels and structural variants from this iterative procedure, estimating their clade membership only after the estimates for $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$ were obtained. We also developed an additional heuristic algorithm to obtain the initial estimates of $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$ that are used to initialize the HMM algorithm.

Heuristic algorithm for initial clade frequency estimates. To obtain an initial estimates of $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$, we first calculate a measure of the heterozygosity in the population,

$$H_t \equiv \sum_m \left(\hat{f}_{mt} - 0.5\right)\left(0.8 - \hat{f}_{mt}\right)\theta\left(\hat{f}_{mt} - 0.5\right)\theta\left(0.8 - \hat{f}_{mt}\right) , \tag{53}$$

where $\theta(z)$ is the Heaviside step function, and look for the time t^* where H_t attains its maximum. This point is likely to have the largest number of mutations segregating at frequencies between 0.5 and 0.8, many (but not all) of which are likely to be fixed in the major clade, if it exists. We use this subset of putative major mutations,

$$I_{\mathbf{M}} \equiv \left\{m : 0.5 < \hat{f}_{m,t^*} < 0.8\right\} , \tag{54}$$

to estimate the frequency of the major clade for $t \geq t^*$:

$$f_t^{\mathbf{M}} \equiv \underset{m \in I_{\mathbf{M}}}{\text{median}} \left\{\hat{f}_{mt}\right\} , \quad (t \geq t^*) . \tag{55}$$

We naively set $f_t^{\mathbf{m}} = 1 - f_t^{\mathbf{M}}$ for the minor clade. If the major clade at t^* is no longer in the majority at the final timepoint t_f (i.e., if $f_{t_f}^{\mathbf{m}} > f_{t_f}^{\mathbf{M}}$), we permute the labels $\mathbf{M} \leftrightarrow \mathbf{m}$.

We then use these partial trajectories $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$ for $t \geq t^*$ to obtain a better set of putative major and minor mutations by forcing the mutations to match $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$ at multiple timepoints, rather than just at t^*. In particular, for each mutation with $0.2 \leq \hat{f}_{mt} \leq 0.8$, we calculate the set of distances

$$D_m^{\mathbf{M}} = \sqrt{\sum_{t \geq t^*} (\hat{f}_{mt} - f_t^{\mathbf{M}})^2} , \tag{56a}$$

$$D_m^{\mathbf{m}} = \sqrt{\sum_{t \geq t^*} (\hat{f}_{mt} - f_t^{\mathbf{m}})^2} , \tag{56b}$$

$$D_m^{\mathbf{E}} = \sqrt{\sum_{t \geq t^*} (\hat{f}_{mt} - 0)^2} , \tag{56c}$$

$$D_m^{\mathbf{F}} = \sqrt{\sum_{t \geq t^*} (\hat{f}_{mt} - 1)^2} , \tag{56d}$$

and we redefine our sets of putative major and minor mutations according to

$$I_{\mathbf{M}} \equiv \left\{m : D_m^{\mathbf{M}} = \max\left\{D_m^{\mathbf{M}}, D_m^{\mathbf{m}}, D_m^{\mathbf{E}}, D_m^{\mathbf{F}}\right\}\right\} , \tag{57a}$$

$$I_{\mathbf{m}} \equiv \left\{m : D_m^{\mathbf{M}} = \max\left\{D_m^{\mathbf{m}}, D_m^{\mathbf{m}}, D_m^{\mathbf{E}}, D_m^{\mathbf{F}}\right\}\right\} . \tag{57b}$$

For times before t^*, we can no longer use the simple formula in Eq. (55) to estimate $f_t^{\mathbf{M}}$, since the mutations in $I_{\mathbf{M}}$ and $I_{\mathbf{m}}$ arise and fix within their respective clades at different times, and have frequency $\hat{f}_{mt} = 0$ before that point. To estimate $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$, we would like to focus on the subset of $I_{\mathbf{M}}$ and $I_{\mathbf{m}}$ that are fixed within the clade at a given timepoint t. We estimate these subsets according to

$$I^*_{\mathbf{M},t} \equiv \left\{ m : m \in I_{\mathbf{M}},\, \hat{f}_{mt} > \max_{m \in I_{\mathbf{M}}} \frac{\hat{f}_{mt}}{2} \right\}, \tag{58a}$$

$$I^*_{\mathbf{m},t} \equiv \left\{ m : m \in I_{\mathbf{m}},\, \hat{f}_{mt} > \max_{m \in I_{\mathbf{m}}} \frac{\hat{f}_{mt}}{2} \right\}, \tag{58b}$$

and use these subsets to reestimate $f_t^{\mathbf{M}}$ and $f_t^{\mathbf{m}}$ using the formulae

$$f_t^{\mathbf{M}} \equiv = \frac{1}{|I^*_{\mathbf{M},t}|} \sum_{m \in I^*_{\mathbf{M},t}} \hat{f}_{mt}\,, \tag{59a}$$

$$f_t^{\mathbf{m}} \equiv = \frac{1}{|I^*_{\mathbf{m},t}|} \sum_{m \in I^*_{\mathbf{m},t}} \hat{f}_{mt}\,. \tag{59b}$$

If the sum of the inferred clade frequencies never exceeds 0.2, we set $f_t^{\mathbf{M}} = f_t^{\mathbf{m}} = 0$. These final estimates are then used to seed the HMM algorithm above.

5.3.1 Appearance, fixation, and transit times

After estimating the most-likely sequence of states, $\hat{L}_{mt}$, we again define an appearance time $T_{0,m}$ and a transit time ΔT_m for each mutation using analogous versions of Eqs. (44) and (45):

$$T_{0,m} = \max_{t<t^*} \left\{ t + 250 : \hat{L}_{mt} = \mathbf{A} \right\}, \tag{60}$$

$$\Delta T_m = \min_{t>T_{0,m}} \left\{ t - T_0 - 250 : \hat{L}_{mt} \in \{\mathbf{F_B}, \mathbf{F_M}, \mathbf{F_m}, \mathbf{E}\} \right\}. \tag{61}$$

For mutations that fix in the basal and major clades, we can obtain an analogous fixation trajectory for each population using an extension of Eq. (42):

$$M^M_{\text{fixed}}(t) = \sum_m \delta_{\hat{L}_{mt},\mathbf{F_B}} + \delta_{\hat{L}_{mt},\mathbf{P_{B^*}}} + \delta_{\hat{L}_{mt},\mathbf{F_M}}\,. \tag{62}$$

These measures are depicted for the six nonmutator populations in Fig. 4a,b. In principle, we can obtain an analogous fixation trajectory for the minor clade as well:

$$M^m_{\text{fixed}}(t) = \sum_m \delta_{\hat{L}_{mt},\mathbf{F_B}} + \delta_{\hat{L}_{mt},\mathbf{P_{B^*}}} + \delta_{\hat{L}_{mt},\mathbf{F_m}}\,. \tag{63}$$

In practice, however, it is more difficult to detect mutations in the minor clade, especially when its frequency becomes very small. In Fig. S9, we compare the major and minor fixation trajectories for the two nonmutator populations (Ara+5 and Ara−6) where the clades persisted at intermediate frequencies through the end of the experiment. In both cases, the minor clade fixes ~ 10 fewer mutations over the course of the experiment, though future work will be required to assess the significance of this observation in light of the ascertainment bias described above.

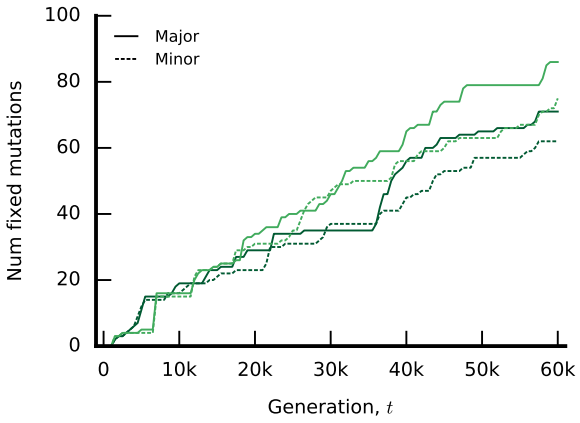

Figure S9: The number of fixed mutations in the major and minor clades in the Ara+5 and Ara−6 populations estimated from Eqs. (62) and (63). The two populations are colored according to the same color scheme as Fig. 2.

5.3.2 Quantifying clonal interference within clades

We investigated the extent of clonal interference in the LTEE by estimating the survival probability of a mutation, p_{survive}, as a function of its current allele frequency f. In the absence of clonal interference, a lineage can fluctuate to extinction if it remains below the drift barrier, $f_{\text{drift}} \sim 1/N_e s$, where s is the net fitness benefit of the lineage in question. In the LTEE, this drift barrier is far less than the $\sim 10\%$ detection threshold, so this classical hitchhiking model would predict that

$$p_{\text{survive}}(f) \approx 1 \tag{64}$$

for all frequencies greater than 10% (and even for much lower frequencies as well).

On the other hand, when clonal interference is pervasive, a mutation will survive only if it is lucky enough to be linked to a future common ancestor of the population. This can only occur if the lineage contains individuals in the high-fitness "nose" of the population fitness distribution, which share a roughly equal probability of producing a future common ancestor [21]. For mutations in our observable frequency range, the population-wide frequency f is a good approximation for the frequency of that mutation within the nose [22]. This yields an alternative prediction,

$$p_{\text{survive}}(f) \approx f\,, \tag{65}$$

in which mutations that reach majority frequency still have a substantial probability of going extinct. In practice, a real population will likely lie somewhere between these two extreme limits, with lower values of $p_{\text{survive}}(f)$ indicating a higher degree of clonal interference.

To estimate $p_{\text{survive}}(f)$ from our metagenomic data, we first use $\hat{L}_{mt}$ to split each mutation trajectory into consecutive runs of polymorphic timepoints, each of which is terminated by a single fixation or extinction event (or the end of the timecourse). Each run then represents an independent survival event that we wish to sum over. We let r index the independent runs for each mutation m, and consider the sub-trajectories $\hat{f}_{m,r,t}$ and $\hat{L}_{m,r,t}$ belonging to each run.

In the absence of frequency-dependent selection, it is straightforward to estimate $p_{\text{survive}}(f)$ from these sub-trajectories. For a given frequency range f, we calculate the fraction of polymorphic

timepoints sufficiently close to f that finish their run above some critical frequency f^*. One can compute this average using a Gaussian kernel,

$$p_{\text{survive}}(f) \approx \frac{\sum_{m,r,t} \exp\left[-\left(\frac{f_{m,r,t}-f}{\Delta f}\right)^2\right] \theta(f_{m,r,t_f} - f^*)}{\sum_{m,r,t} \exp\left[-\left(\frac{f_{m,r,t}-f}{\Delta f}\right)^2\right]}, \tag{66}$$

for a given kernel width Δf.

In practice, however, long-lived coexistence of the kind observed in Fig. 3 can cause problems for the estimator in Eq. (66). Most of the intermediate-frequency mutations in populations like Ara−6 are fixed within their parent clade, and will therefore survive as long as the clade persists to the final timepoint. These mutations will tend to bias Eq. (66) towards $p_{\text{survive}}(f) \approx 1$, even if there is substantial clonal interference.

Strictly speaking, the predictions in Eqs. (64) and (65) are no longer valid in the presence of frequency-dependent selection. However, if we assume that the fixation process is roughly independent of the competition between the clades, then the predictions should approximately hold as long as we replace the population-wide frequencies, $\hat{f}_{mt}$, with the corresponding *within*-clade frequencies, $\tilde{f}_{mt}$. We estimate these using the output of our clade inference algorithm above.

If the run of polymorphic states ends in state $\mathbf{F_M}$ (or if the experiment ends while it is in state $\mathbf{P_M}$), we set

$$\tilde{f}_{mt} = \min\left\{\frac{f_{mt}}{f_t^{\mathbf{M}}}, 1\right\}, \tag{67}$$

and similarly for the minor and basal clades (where $f_t^{\mathbf{B}} \equiv 1$ by definition). Runs that terminate in extinction are more problematic, since it is harder to assign them to the correct clade background while they are at low frequency. In order to be conservative with respect to the amount of clonal interference, we renormalize f_{mt} by the size of the largest possible background the mutation could have arisen on, so that $\tilde{f}_{mt}$ is as small as possible:

$$\tilde{f}_{mt} = \min\left\{\frac{f_{mt}}{f_t^{\mathbf{M}}}, \frac{f_{mt}}{f_t^{\mathbf{m}}}, \frac{f_{mt}}{1 - f_t^{\mathbf{M}} - f_t^{\mathbf{m}}}, 1\right\}. \tag{68}$$

Based on these definitions, we estimated the survival probability for each of the twelve LTEE populations using Eq. (66) with $\Delta f = 0.05$ and $f^* = 0.5$ (Fig. 4c,d). We excluded indels and structural variants from this calculation in the mutator populations, so that the excess of difficult-to-resolve homopolymer indels did not downwardly bias our results.

5.3.3 Validation with clonal samples

In constrast to the single-clade HMM in Section 5.2, there is no established model that reproduces the long-lived coexistence observed in many of the LTEE populations. To validate the clade-aware HMM, we therefore turn to an empirical test using data from clonal isolates sequenced by Tenaillon et al. [3].

For each isolate, we calculated the number of mutations supported by $\geq 50\%$ of the reads that were also classified as fixed within the $\mathbf{F_B}$, $\mathbf{F_M}$, or $\mathbf{F_m}$ states at that timepoint. (Mutations that were still segregating within a clade were not included due to the difficulty in assigning clades to low-frequency mutations.) The resulting mutation profiles of the clones are shown in Fig. S10. As expected, each clone is primarily composed of either $\mathbf{F_B} + \mathbf{F_M}$ mutations, or $\mathbf{F_B} + \mathbf{F_m}$ mutations,

with only a few misclassified mutations scattered throughout. However, given the sparse coverage of the clades, a more densely sampled panel of clones would be required to judge the inferences from the HMM.

5.4 Beyond pairwise coexistence

So far, we have only considered cases where a single pair of clades persists for substantial periods of time. In many of the LTEE populations, this appears to be a much better model for the lineage dynamics than the single-clade model in Section 5.2. However, our clade-aware HMM does not detect all instances of frequency-dependent selection in the LTEE: a notable example includes the $\mathrm{Cit}^+/\mathrm{Cit}^-$ cross-feeding interaction in the Ara$-$3 population [23]. Nor can we rule out more complicated scenarios such as multi-way coexistence, or additional pairs of sub-clades that arise after one of the initial clades fixes. Given our limited sequencing coverage, it can be difficult to resolve examples of multi-way coexistence, since more lineages must be squeezed into a smaller frequency range. Occasionally, however, the different lineages may be subject to sufficiently large shifts in frequency that they may be resolved by their temporal behavior, even if they overlap in frequency for many timepoints. One striking example is provided by the Ara+1 population, where three major lineages persist for $\sim 20,000$ generations before one eventually comes to dominate the population. While they coexist, these three lineages undergo a series of dramatic reversals illustrated in Fig. S11, which allow us to resolve the clades by eye. A more rigorous analysis of this example, as well as extensions of our HMM algorithm to enable more general haplotype estimates, remain interesting avenues for future work.

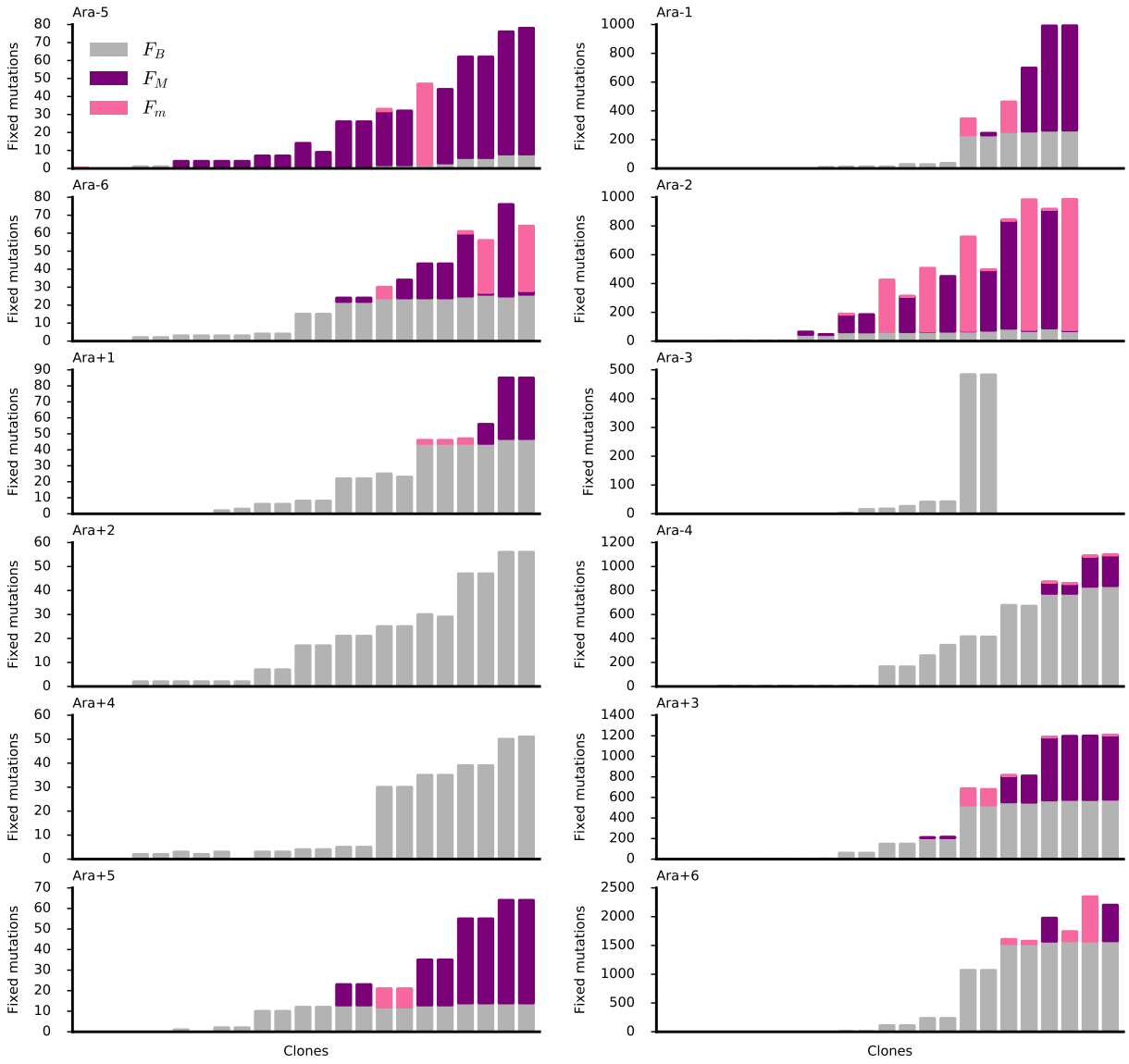

Figure S10: The number of mutations assigned to the $\mathbf{F_B}$, $\mathbf{F_M}$, and $\mathbf{F_m}$ states that are present in $\sim$ 250 clones isolated from the LTEE. Each bar denotes a separate clonal isolate sequenced by Tenaillon et al. [3], and the clones are ordered according to the generation at which they were sampled (starting with two clones from generation 500 and ending with two from generation 50,000). Nonmutator populations are shown at left; populations that evolved hypermutability are on the right.

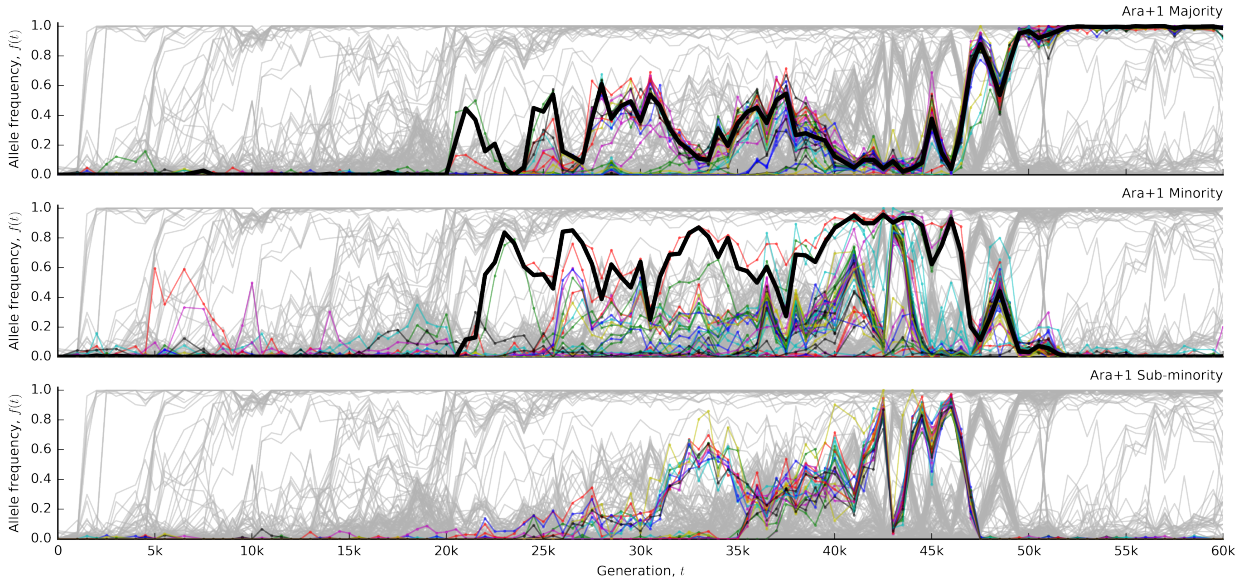

Figure S11: An example of lineage dynamics that are not captured by the pairwise clade model in Section 5.3. Three distinct lineages persist in the Ara+1 population for $\sim 20,000$ generations. The latter two cluster within the minor clade identified by the HMM.

6 Parallelism and contingency analysis

In this section, we investigate the targets of selection in the LTEE by incorporating information about the identities of the detected mutations, focusing on their distribution through time and across replicate populations. This serves as a companion to Figs. 5 and 6 in the main text.

6.1 Parallelism at the variant type level

At the most coarse-grained functional level, we classified mutations based on the variant types assigned in Section 4.4. The cumulative number of detected mutations of each type is plotted in Fig. 5a,b as a function of their appearance time. In the absence of natural selection, the total number of mutations of each type should be proportional to the target size and the rate at which they occur. These quantities are difficult to estimate for indels and structural variants, but they can be estimated for the synonymous and nonsynonymous (including both missense and nonsense) sites by counting the fraction of base-pair mutations that result in each variant type. After excluding repetitive regions, the REL606 reference has $L_{\rm s} \approx 8.9 \times 10^5$ synonymous sites and $L_{\rm ns} \approx 3 \times 10^6$ nonsynonymous sites. Thus, in the nonmutator populations, we see an enrichment of nonsynonymous over synonymous mutations (Extended Data Fig. 2), suggesting that a significant fraction of the observed point mutations in these classes are driven to detectable frequencies by positive selection.

In the mutator populations, this ratio is much lower (Extended Data Fig. 2), suggesting a role for purifying selection. However, the spread among the six mutator populations is quite large, with two populations (Ara−1 and Ara+6) having dN/dS values much larger than one. We note that Ara−1 and Ara+6 are the only two populations with the *mutT* mutator phenotype, which suggests that their anomalously high dN/dS values may reflect a biased mutational spectrum rather than a much larger fraction of beneficial mutations. To check this hypothesis, we recalculated $L_{\rm s}$ and $L_{\rm ns}$ based on the observed single-nucleotide substitution frequencies in each population, thereby obtaining substitution-specific estimates of dN/dS (Extended Data Fig. 2). As expected, this correction makes the six mutator populations much more tightly distributed, with $dN/dS \lesssim 1$ in all cases.

We next sought to investigate temporal patterns in the accumulation of different variant types, after controlling for the observed number of mutations in each class. We focused on the six nonmutator populations, since the temporal patterns in the mutator lines are already known to depend very strongly on the fixation times of mutator and antimutator alleles [24]. To examine the temporal patterns in the nonmutator lines, we compared the cumulative distribution of appearance times in each variant class with the pooled distribution of appearance times across all classes (Fig. 5c). We quantified the differences between these distributions using the scaled Kolmogorov-Smirnov distance,

$$D_i = \sqrt{\frac{n_i n_{\rm tot}}{n_i + n_{\rm tot}}} \cdot \max_t \|F_i(t) - F_{\rm tot}(t)\| \tag{69}$$

where n_i and $n_{\rm tot}$ are the numbers of mutations in class i and the entire pool, and $F_i(t)$ and $F_{\rm tot}(t)$ are the corresponding empirical CDFs. We assessed significance of this statistic relative to a null model in which the observed appearance times and variant types are randomly permuted. We observe a small but significant enrichment of missense mutations early in the experiment ($P < 0.01$); the other variant types are indistinguishable from the pooled distribution.

6.2 Parallelism at the nucleotide level

At the most fine-grained functional level, we also searched for signatures of parallelism by looking for independent mutations that occurred at the same site in the genome. Because it is difficult to resolve independent appearances of an allele in the same population, we focused on sites that were mutated in multiple replicate populations. We also excluded indels and structural variants, due to the difficulty in assigning a consistent nucleotide site to these complex mutation events. Our analysis therefore represents a lower bound on the amount of nucleotide-level parallelism in the LTEE.

Nucleotide multiplicity. For each site, we defined the *multiplicity*, m_i, as the number of populations with a point mutation detected at that site. We calculated the multiplicity separately for both the mutator and nonmutator populations, so that the multiplicity could range from 1 to 6. Each mutation was then assigned a multiplicity score according to the site in which it occurred, and the distribution of these multiplicity scores is shown in Extended Data Fig. 3.

To put these observations in context, we can compare them to a null model in which mutations are uniformly distributed across the sites in the genome. In this model, the expected fraction of mutations with multiplicity $\geq m$ in a sample of size n_{tot} is given by

$$S(m) \approx \sum_{n \geq m} \frac{n}{n_{\text{tot}}} \cdot L_{\text{tot}} \cdot \frac{\left(\frac{n_{\text{tot}}}{L_{\text{tot}}}\right)^n}{n!} e^{-n_{\text{tot}}/L_{\text{tot}}}, \tag{70}$$

where $L_{\text{tot}} \approx 4.4 \times 10^6$ is the total number of annotatable sites in the reference genome. These predictions are illustrated in Extended Data Fig. 3 as well. Although the data show an excess of nucleotide multiplicity compared with this simple null model, multi-hit sites still constitute a relatively small fraction of all observed point mutations ($\sim 5\%$ and 10% in the nonmutator and mutator populations, respectively). Maddamsetti et al. [25] have recently presented evidence that many of these multi-hit sites in the nonmutator populations are beneficial mutations that modify protein function, rather than knocking them out. This finding may partially explain why these sites are enriched relative to other locations in the same gene.

6.3 Parallelism at the gene level

Given the limited extent of nucleotide-level parallelism, we focused on patterns of genetic parallelism at the gene level, clustering mutations based on the gene assigned in Section 4.4.

6.3.1 Gene multiplicity, assessing individual and global significance

If selection pressures and mutation rates did not vary between genes, the number of mutations in each gene should be proportional to the target size. While it is difficult to estimate the local target size for beneficial, deleterious, and neutral mutations in any particular gene, we assume that gene length is a primary driver of the target size. Similar to our nucleotide-level analysis above, we then define a *multiplicity* for each gene according to

$$m_i = n_i \cdot \frac{\overline{L}}{L_i}, \tag{71}$$

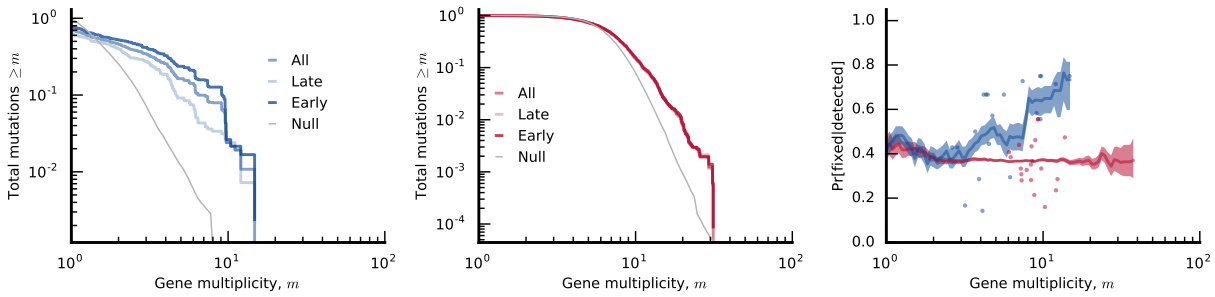

Figure S12: An analogous version of Fig. 5 excluding structural variants. Left and center: The fraction of all non-synonymous mutations in the nonmutator (left) and mutator (center) populations that are found in genes with multiplicity ($m_i = n_i\overline{L}/L_i$) greater than or equal to m. The grey line is the null distribution from Eq. (72). Right: Average conditional fixation probability of a mutation as a function of its gene multiplicity (in sliding windows of 0.2 log10 units) in nonmutator (blue) and mutator (red) populations. Shaded confidence intervals denote the 14th and 84th percentiles of the beta posterior distribution of each window. Fixation probabilities of the 20 most-frequently mutated genes are shown as dots.

where n_i is the number of mutations in gene i across all replicate populations (including indels and structural variants,[1] but excluding synonymous mutations), L_i is the total number of nonsynonymous and noncoding sites in gene i, and $\overline{L}$ is the average value of L_i across all genes in the genome. This definition ensures that under the null hypothesis, all genes have the same expected multiplicity $\overline{m} = n_{\text{tot}}/n_{\text{genes}}$. As above, we calculated the multiplicity separately for the mutator and nonmutator populations. In this case however, we have the power to resolve independent mutations in a gene within the same population, so n_i can be much larger than 6.

To quantify the amount of gene-level parallelism in the LTEE, we assigned each mutation a multiplicity score according to the multiplicity of the gene in which it arose. The distribution of these scores is shown in Fig. 5d,e in the main text, while Fig. S12 shows analogous distributions after excluding structural variants. In both cases, the null distribution is now given by a generalization of Eq. (70),

$$S(m) \approx \sum_{i=1}^{N_{\text{genes}}} \sum_{n=0}^{\infty} \frac{n}{n_{\text{tot}}} \cdot \theta\left(n_i \cdot \frac{\overline{L}}{L_i} - m\right) \cdot \frac{\left(\frac{n_{\text{tot}} L_i}{\overline{L} N_{\text{genes}}}\right)^n}{n!} e^{-\frac{n_{\text{tot}} L_i}{\overline{L} N_{\text{genes}}}}, \tag{72}$$

which accounts for variation in gene length. As described in the main text, we observe an excess of high-multiplicity mutations in both the mutator and nonmutator populations. In the nonmutator populations, approximately half of all mutations occurred in genes with $m_i \geq 2$, though only half as many would be expected under the null model.

This suggests that we should replace the null model with an alternative where mutations are assigned to genes with probability

$$p_i \propto L_i r_i \,, \tag{73}$$

[1] We chose to include structural variants in the multiplicity (despite the fact that their target size may be more strongly influenced by factors other than gene length) because they account for a significant fraction ($\approx 40\%$) of all genic mutations in the nonmutator populations. For completeness, we also repeat our analysis excluding structural variants below.

for some set of enrichment factors r_i that are not all equal to 1. This enrichment can be factored into a local change in mutation rate $\mu_i/\overline{\mu}$ and a function that depends on the effective selection coefficient of the gene. For our purposes here, we will focus on the set of compound parameters $\{p_i\}$, which we will refer to as the *realized gene mutation spectrum* (or the *realized mutation spectrum* for short).

In the alternative model, the maximum likelihood estimators for the enrichment factors are simply the ratios of observed and expected multiplicities, $r_i = m_i/\overline{m}$, and the net increase in log-likelihood compared to the null model ($r_i = 1$) is given by

$$\Delta\ell = \sum_i n_i \log\left(\frac{m_i}{\overline{m}}\right) . \tag{74}$$

This likelihood ratio coincides with the total "G-score" used by Tenaillon et al. [3]. Consistent with their results from clonal isolates, our metagenomic data shows a statistically significant G-score in both the mutator and nonmutator populations ($P < 10^{-4}$), indicating that we must reject the simple null model in favor of the alternative.

However, if we wish to go beyond rejection and infer the underlying p_i we must remember that the maximum likelihood estimate $r_i = m_i/\overline{m}$ for an arbitrary gene may still substantially overfit the data, particularly in the nonmutator populations. For example, since the expected multiplicity in the nonmutators is $\overline{m} \approx 0.3$, a single neutral hitchhiker in an otherwise unmutated gene would lead to an apparent enrichment factor $r_i > 1$, even if there might be many such events genome-wide. A more appropriate alternative model would therefore be one in which only a subset I of the genes have $r_i \neq 1$, while the remaining genes have $r_i = 1$.

To estimate the subset I, we searched for genes that are significantly different from the null hypothesis at an individual level. For any particular gene, the P-value for a likelihood ratio test is given by

$$P_i = \sum_{n \geq n_i} \frac{\left(\frac{n_{\rm tot} L_i}{\overline{L} N_{\rm genes}}\right)^n}{n!} e^{-\frac{n_{\rm tot} L_i}{\overline{L} N_{\rm genes}}} , \tag{75}$$

and we wish to estimate I using the subset of genes with sufficiently low P-values. To be conservative, we also restricted our attention to genes with $n_i \geq 3$, in order to limit low P-values that are driven primarily by gene length. Under the null hypothesis, the expected number of genes with $n_i \geq 3$ and $P_i \leq p$ is given by

$$\overline{N}(P) \approx \sum_{i=1}^{N_{\rm genes}} \sum_{n=3}^{\infty} \theta\left(P - P_i(n, L_i)\right) \cdot \frac{\left(\frac{n_{\rm tot} L_i}{\overline{L} N_{\rm genes}}\right)^n}{n!} e^{-\frac{n_{\rm tot} L_i}{\overline{L} N_{\rm genes}}} , \tag{76}$$

and we can compare this to the observed number of genes, $N(P)$, with the same properties (Fig. S13). In particular, for a given FDR α, we define a critical P-value, P^*, such that

$$\frac{\overline{N}(P^*)}{N(P^*)} \leq \alpha . \tag{77}$$

For this value of $P^*(\alpha)$, we then define the set of significantly enriched genes as

$$I = \{i \,:\, P_i \leq P^*(\alpha)\} . \tag{78}$$

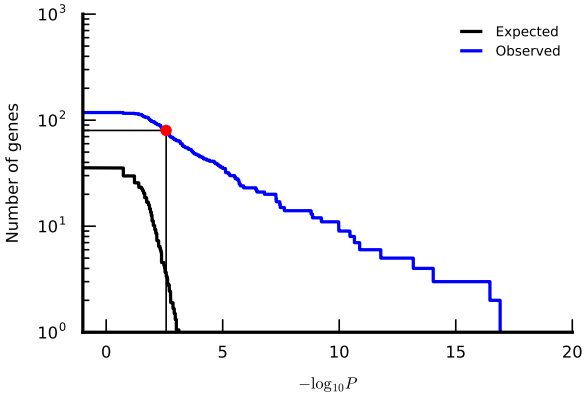

Figure S13: The observed (N) and expected number of genes ($\overline{N}$) with $n_i \geq 3$ and $P_i \leq P$, as a function of P. The symbol denotes the genome-wide significance threshold P^* defined in Eq. (77) for $\alpha = 0.05$.

The corresponding enrichment factors are given by

$$r_i = \begin{cases} \frac{m_i}{\overline{m}} \left(\frac{1 - \frac{\sum_{i \in I} L_i}{L N_{\text{genes}}}}{1 - \frac{\sum_{i \in I} n_i}{n_{\text{tot}}}} \right) & \text{if } i \in I, \\ 1 & \text{else.} \end{cases} \tag{79}$$

The significantly enriched genes with $\alpha = 0.05$ are listed in Supplementary Table 3. These account for $\approx 35\%$ of the total mutations but only $\approx 2\%$ of the total target size. After removing these individually significant examples, the resulting G score is still significantly higher than expected by chance ($P < 10^{-4}$), which implies that other genes were also targeted more often than expected under the null model, even though they are not in that list.

6.3.2 Changing signatures of parallelism over time

After confirming an overall signature of parallelism in the LTEE, we next sought to quantify how these patterns changed over evolutionary time as the experiment progressed. To do so, we compared the observed appearance times of the genic mutations against a null model in which the appearance times are randomly assigned to genes, while still preserving the overall amount of parallelism in each gene and the non-uniform distribution of appearance times (Fig. S14).

There are several ways to perform this comparison. First, at a global level, we can compare the multiplicity distributions for mutations that arose in the first half of the experiment versus the second half (Fig. 5d,e), or more generally, before or after some *partition time* t^*. In both cases, we continue to use the same multiplicity scores calculated for the entire set of mutations, so that they can be compared to the pooled multiplicity distribution as well. To quantify how the overall levels of parallelism differ between these two distributions, we defined a scaled G-score change,

$$\Delta g_< = \frac{\sum_{i=1}^{N_{\text{genes}}} \left(n_i^< - \frac{n_i n_{\text{tot}}^<}{n_{\text{tot}}} \right) \log\left(\frac{m_i}{\overline{m}}\right)}{\sum_{i=1}^{N_{\text{genes}}} n_i \log\left(\frac{m_i}{\overline{m}}\right)}, \quad \Delta g_> = \frac{\sum_{i=1}^{N_{\text{genes}}} \left(n_i^> - \frac{n_i n_{\text{tot}}^>}{n_{\text{tot}}} \right) \log\left(\frac{m_i}{\overline{m}}\right)}{\sum_{i=1}^{N_{\text{genes}}} n_i \log\left(\frac{m_i}{\overline{m}}\right)}, \tag{80}$$

where $n_i^<$ and $n_i^>$ are the numbers of mutations in gene i that appeared before and after t^*. We plot these changes in Fig. S15 as a function of t^*. We find that, regardless of the choice of

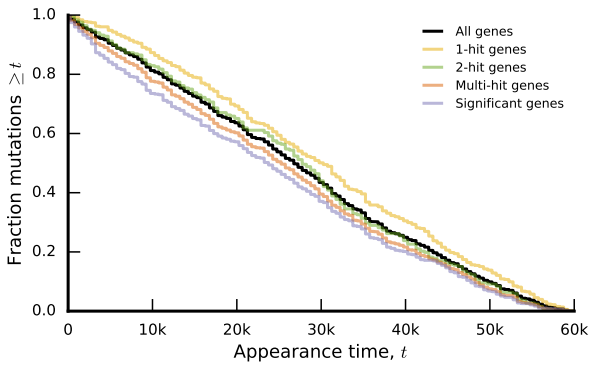

Figure S14: The pooled distribution of appearance times for mutations in different sets of genes.

t^*, the early epochs have an excess of parallelism compared to the total timecourse ($\Delta g_< > 0$) while the later epochs have a parallelism deficit ($\Delta g_> < 0$). The largest difference between $\Delta g_<$ and $\Delta g_>$ occurs near $t^* \approx 30,000$, and the corresponding multiplicity distributions are shown in Fig. S15. To assess the significance of $\Delta g_< - \Delta g_>$ at this timepoint, we calculated the distribution of $\max_{t^*} \{\Delta g_<(t^*) - \Delta g_>(t^*)\}$ across all bootstrap samples, which shows that the observed value is highly significant (one-sided $P < 10^{-4}$). In terms of the effect size, however, this difference accounts for only $\sim 10\%$ of the total G-score observed in the nonmutator populations.

Though the overall levels of parallelism decline only modestly through time, this global signal could mask significant temporal non-uniformity in individual genes. To investigate this hypothesis further, we first compared the realized mutation spectrum, $\{p_i\}$, for the set of enriched genes in Supplementary Table 3, estimated before or after some threshold time t^* (Extended Data Fig. 7c). Some of the differences in $\{p_i\}$ are expected to occur by chance given the finite number of mutations. We assessed the significance of the pre- and post-t^* mutation spectra through the ratio of their multinomial likelihoods,

$$\Delta\ell = \sum_{i \in I} \left[n_i^< \log\left(\frac{n_i^< n_{\text{tot}}}{n_{\text{tot}}^< n_i}\right) + n_i^> \log\left(\frac{n_i^> n_{\text{tot}}}{n_{\text{tot}}^> n_i}\right) \right], \tag{81}$$

and we calculated a corresponding null distribution by randomly permuting the appearance times among all the mutations in Supplementary Table 3. As shown in Extended Data Fig. 7, there is a wide range of t^* where the difference between the pre- and post-t^* mutation spectra exceeds the predictions of the null model. The maximum value of $\Delta\ell$ occurs at $t^* \approx 12,000$ and is highly significant ($P < 10^{-4}$).

Though the shape of the fitness trajectories may suggest a model of two evolutionary epochs [19], this division is somewhat arbitrary [5]. We can gain a more complete picture of the temporal patterns of individual genes by comparing the distribution of appearance times within each gene (Fig. 6a) against the pooled distribution (Fig. S14), similar to our analysis of the different variant types in Section 6.1. As above, we quantified the differences between the distributions using the scaled Kolmogorov-Smirnov distance in Eq. (69), and we calculated P-values numerically by randomly permuting the appearance times among the genes in Supplementary Table 3. We also calculated a corresponding Q-value using an analogous version of Eq. (27) to correct for multiple hypothesis testing, rejecting the null hypothesis if $Q_i < 0.05$. This yielded a handful of candidate genes in which mutations arose non-randomly during the experiment. The mutation trajectories for the early- and late-arising examples in the text (*hslU* and *atoS*, respectively) are shown in Extended

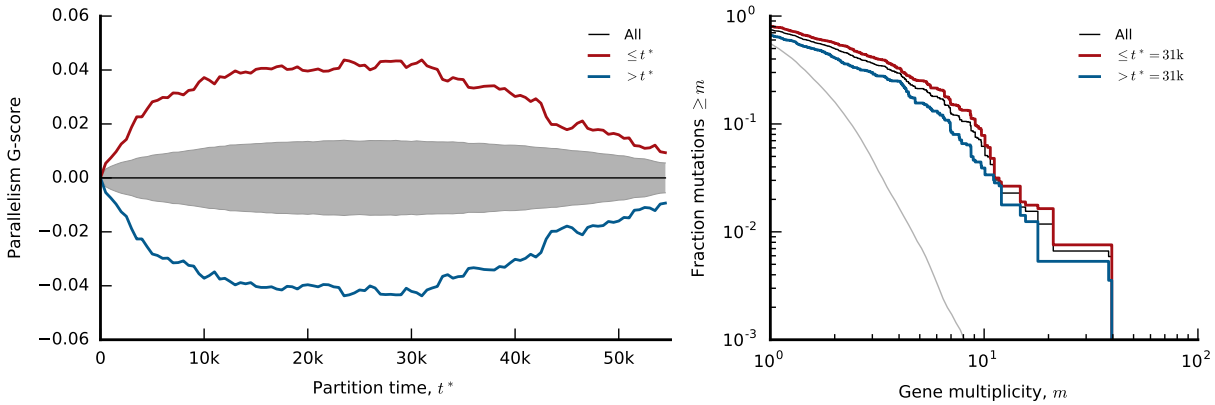

Figure S15: Overall levels of parallelism in the nonmutator populations as a function of time. Left panel: scaled G-score changes in Eq. (80) as a function of the partition time, t^*. Lines denote observed values, while the shaded regions represent 95% confidence intervals obtained by randomly permuting appearance times across genes. Right panel: fraction of non-synonymous mutations that are found in genes with multiplicity greater than or equal to m and which occurred either before (red) or after (blue) the value of t^* (see legend) which maximizes the difference between $\Delta g_<$ and $\Delta g_>$ in the left panel. For comparison, the distribution of all mutations is shown in black, while the null distribution from Eq. (72) is shown in grey.

Data Figs. 4 and 5. After removing the individually significant genes from the pool, we assessed the global signal of temporal nonuniformity by summing the KS distances for each remaining gene, and comparing this observed value to the null distribution obtained by permutation. As described in the text, the summed KS distance for the remaining genes was significantly larger than expected by chance ($P < 10^{-3}$), which implies that some of the other genes in Supplementary Table 3 also arose non-randomly in time, even though they were not individually significant.

Finally, to determine whether this signal of gene-specific appearance times extended beyond the set of enriched genes in Supplementary Table 3, we calculated the average difference in appearance times of mutations in 2-hit genes, and compared this to the difference in appearance times for a random pair of mutations in this set by permutation ($P < 10^{-3}$, Extended Data Fig. 6).

Each of these tests produced significant statistical evidence that the repertoire of mutated genes has shifted during the LTEE.

6.3.3 Evidence for historical contingency

There are several potential explanations for a changing spectrum of adaptive mutations. The simplest is a "coupon-collecting" model of evolution, in which natural selection favors a mutation in one of several functional units (or modules), each of which could comprise one or many genes. Benefits associated with mutations in different modules are assumed to combine additively, but once a given module has been mutated, subsequent mutations in that module are assumed to be neutral or deleterious. Beneficial loss-of-function mutations are the canonical example of this behavior. Under the coupon-collecting model, replicate populations will first tend to fix mutations in the modules with the largest selective advantage / target size combination, and the spectrum of adaptive mutations will then begin to change as these strongly beneficial mutations are exhausted. This model provides a simple and biologically plausible explanation for the preferentially early

genes (like *hslU*) in Fig. 6a, and it is consistent with the results of previous studies [26–28]. In principle, it also predicts that mutations in preferentially early genes should arise (and fix) in many of the populations (as observed for *hslU*), but this prediction is not required to hold in the likely event that modules comprise multiple genes whose identities are not known beforehand.

Though somewhat counterintuitive, this coupon-collecting model could also account for preferentially late genes like *atoS* when there is clonal interference. In the absence of clonal interference, the substitution rate of a particular mutation is proportional to its own fitness advantage s, and it is independent of the fitness effects at other sites. However, in the presence of clonal interference, population genetic theory predicts that the substitution rate for moderately beneficial mutations scales like $\sim e^{T_c s}$, where the coalescent timescale T_c depends on the distribution of fitness effects across the genome [29]. In a model where these fitness effects are distributed exponentially with a cutoff at some $s_{\max}$, one can show that $T_c \propto 1/s_{\max}$. Thus, if the cutoff decreases over time (e.g., due to the depletion of strongly beneficial mutations), the substitution rate of moderate-effect mutations can suddenly increase.[2] In addition to this dynamical explanation, the preferentially late-evolving genes could reflect global changes in selection pressures as the fitness of the population increases, as well as subtle shifts in the evolution environment over time.

Alternatively, the late-evolving genes could represent new evolutionary paths that are opened up by previous substitutions in one or more populations. We refer to this as the "historical contingency" model. Although such synergistic interactions play a central role in evolutionary theory, relatively few examples have been observed in experimental evolution. Among these, one example is the evolution of citrate-utilization that evolved in the Ara−3 population [23]. This strongly beneficial phenotype, which has yet to evolve in any of the other 11 LTEE populations even after 65,000 generations, is only evolutionarily accessible in the presence of one or more specific potentiating mutations [23, 30–32]. Another study in which *E. coli* adapted to high temperature found a statistical association between mutations in *iclR* and *cls*, and between mutations in *rpoB* and a handful of other genes [26].

With only six nonmutator populations and vastly more potential gene combinations, we lack the power to scan for such interactions directly. But if these interactions are sufficiently common, we expect their patterns of historical contingency to be reflected in the distribution of mutations across these six populations. In contrast to the explanations described above, which affect all populations equally, we expect mutations in contingent genes to be clustered in a smaller number of replicate populations that already fixed the unknown potentiating mutation. We also expect such genes to be mutated later in the experiment, since it will take some time for the initial potentiating mutation to arise and fix. We note, however, that if the new pathways are not mutually exclusive, there may only be a limited time window in which the signature of contingency is strongest. Given sufficient time, all six replicates may acquire the relevant potentiating mutation, and the new evolutionary path would then resemble one of the global explanations above.

In addition, if we could observe only the successful mutations, it would be difficult to differentiate between a strongly beneficial mutation that appears in a subset of the replicates due to contingency, or a more weakly selected mutation that occurred less often in the finite length of the experiment. However, in a large population, a new strongly beneficial pathway will often be mutated in multiple genetic backgrounds in the same population before one of them manages to fix; this effect is also enhanced by the presence of long-lived clades. These unfixed variants provide an additional signature of historical contingency, which occurs when the mutations in a multi-hit gene are clustered in a smaller number of populations than expected by chance. By contrast, genes

[2] This behavior depends on the assumption that the strong-effect mutations are depleted, and is less likely to arise in global diminishing returns models where the fitness effects of *all* mutations are reduced by a common factor [5, 19, 20, 28].

in the coupon-collecting model will tend to be over-dispersed across the populations, given the total number of times they have been mutated. Scanning through the genes in Supplementary Table 3, we can find anecdotal examples of late-hit genes that suggest historical contingency, e.g., the *argR* gene discussed in the main text (see Extended Data Fig. 8). However, the small sample sizes prevent any of these examples from attaining statistical significance on their own.

Missed opportunitites. To quantify the global signature of contingency in our data, we calculated the between-population dispersion for all genes that were mutated in the nonmutator populations. For each gene i, we recorded the number of populations k_i that had a mutation in that gene. This number ranges from 1 to 6, but cannot exceed the total number of mutations n_i. The distribution of observed (k_i, n_i) pairs is shown in Fig. 6b in the main text.

In the absence of epistasis, we expect the mutations to be distributed across the replicate populations in a multinomial fashion, with weights proportional to the total number of mutations in each population. In other words, if $n_{i,p}$ denotes the number of mutations in gene i in population p, then under the null model, we expect that

$$n_{i,p} \sim \text{Multinomial}(n_i, p_p)\,, \tag{82}$$

where $p_p = \sum_i n_{i,p} / \sum_{i,p} n_{i,p}$ is the relative fraction of mutations that fall in population p. By drawing random samples from this model, we obtain a null distribution, $P(k|n)$, for the various (k_i, n_i) pairs in Fig. 6b, which we compare to the observed distribution

$$\hat{P}(k|n) = \frac{\sum_i \delta_{n_i,n}\delta_{k_i,k}}{\sum_i \delta_{n_i,n}}\,. \tag{83}$$

While the difference between $\hat{P}(k|n)$ and $P(k|n)$ provides a direct readout of the between-population dispersion, the uncertainty in any individual element of $\hat{P}(k|n)$ is substantial. For this reason, we also used an aggregate measure of dispersion that allows us to combine the various entries in Fig. 6b.

To define this measure, we note that the quantity $n_i - k_i$ represents the number of "redundant" mutations in gene i, i.e., the number of times a mutation appears in a population that already produced a mutation in that same gene. Under the null model, these redundant mutations could have equally well occurred in one of the populations where the gene was not mutated. With this in mind, we define a gene-specific probability,

$$p_{0,i} = \sum_p p_p \delta_{n_{i,p},0}\,, \tag{84}$$

which represents the total probability that a random mutation would appear in a population that did not already have a mutation in gene i. The *number of missed opportunities*, m_i, is then defined as the number of redundant mutations that would be expected to appear in one of the unmutated populations:

$$m_i = p_{0,i}\,(n_i - k_i)\,. \tag{85}$$

According to this definition, a 2-hit gene that occurred in the same population has $\approx 5/6$ missed opportunities, a 3-hit gene in the same population has $\approx 10/6$ missed opportunities, and a 10-hit gene spread across all six populations has no missed opportunities (despite having a few "redundant" mutations).

Of course, we cannot attribute all such missed opportunities to historical contingency: many will arise simply by chance under the null model in Eq. (82). By drawing many samples from this

model, we obtain a null distribution $P_i(m|n_i)$ for the number of missed opportunities in each gene. We can then calculate the *net* missed opportunities by comparing the observed and expected totals across all genes:

$$\Delta m = \sum_i \left[m_i - \int m P_i(m|n_i)\, dm \right] . \tag{86}$$

As described in the main text, the distribution in Fig. 6b suggests an excess of both under- and over-dispersed mutations, such that while there is a net excess of missed opportunities ($\Delta m \approx 8$), this value is not statistically significant ($P \approx 0.1$). This is not surprising, since we know that any signal of historical contingency must compete with the coupon-collecting genes that are known to exist in the LTEE [33].

To disentangle these effects, we sought to exploit their opposing temporal trends. As described above, coupon-collecting genes are expected to be mutated early, while historically-contingent genes are expected to be mutated later. Thus, if we divide the genes according to whether their median appearance time is before or after some critical time t^*, we expect to see a stronger signature of coupon-collecting in the pre-t^* genes, and a stronger signature of historical contingency in the post-t^* genes. This hypothesis is confirmed in Extended Data Fig. 9, where we plot the net missed opportunities for the pre- and post-t^* genes as a function of t^*.

Since Δm depends on the sample size, we expect that the number of missed opportunities will eventually decline for larger t^* when the sample size becomes small. To balance these competing demands, we focused on a single value of t^* where the differences in the pre- and post-t^* values was as large as possible:

$$t^* = \text{argmax} \left\{ \Delta m^{>} - \Delta m^{<} \right\} . \tag{87}$$

The net missed opportunities and the corresponding dispersion distributions for this value of t^* are shown in Fig. 6c,d. To assess the statistical significance of these values, we calculated the distribution of the maximal difference, $\max_{t^*} \{\Delta m^{>} - \Delta m^{<}\}$, across each of our bootstrapped datasets, and compared it to the observed value above. The corresponding P-value is $P \approx 3 \times 10^{-3}$.

6.4 Parallelism at higher levels of organization

In principle, genetic parallelism is likely to be present at higher levels of organization, which might increase our power to resolve temporal and population-specific changes [26]. At the same time, it is increasingly challenging to define a proper collection of modules given the complex set of interactions that take place in the cell. Here, we focus only on the next-highest level of organization, repeating our parallelism and contingency analyses at the operon level.

We obtained a list of operons for the REL606 reference genome from the Database of prOkaryotic OpeRons (DOOR) [34] and clustered the genes accordingly.[3] To ensure uniqueness, genes that were annotated in multiple operons were assigned to the operon with the largest number of genes. Genes without an operon assignment were ignored. Each operon was labelled by its list of constituent genes.

Based on these annotations, several of the most frequently mutated genes (e.g., *malT* and *pykF*) were assigned to operons that contained only a single gene, so that the clustering had little effect on these entries. Other multi-hit genes, like *atoS* and *atoC*, were merged into much larger multi-hit operons. To examine how often such merging occurred, we calculated the fraction of mutations in

[3] The list of operons was originally downloaded from http://csbl.bmb.uga.edu/DOOR/downloadNCoperon.php?NC_id=NC_012967, and is included in the Github repository described in Section 7.

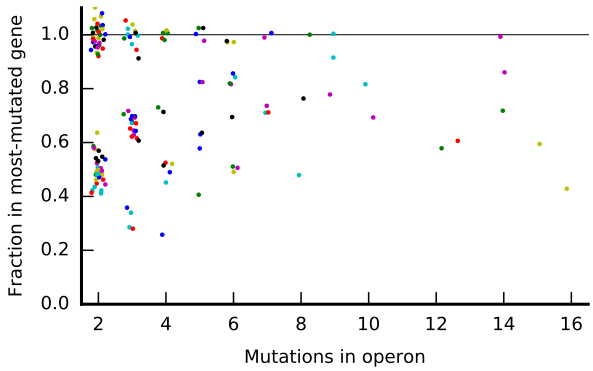

Figure S16: The fraction of mutations in the most-frequently mutated gene in an operon as a function of the total number of mutations in that operon, based on data from the six nonmutator populations. Each point represents an operon with at least two genes. The points are colored for contrast, and a small amount of noise has been added to enhance readability.

the most-frequently mutated gene for all multi-hit operons in the nonmutator lines (Fig. S16). The clustering step produced ~ 10 potentially interesting operons that were mutated 6 or more times in total while each of their constituent genes had 5 or fewer mutations.

We next repeated our analyses in Fig. 6, Fig. S15, and Extended Data Figs. 7 and 9 to examine the overall levels of parallelism among the operons and how these patterns change over time (Figs. S17, S18, S19). The results were largely similar to the gene-level analysis above. This suggests that, compared to genes, operons are not necessarily a better predictor of the targets of selection in the LTEE, at least not without further biological refinement. Exploration of parallelism at higher levels of organization remains an interesting avenue for future work.

7 Data and code availability

Raw sequencing reads have been deposited in the NCBI BioProject database under accession number PRJNA380528. All associated metadata, as well as the source code for the sequencing pipeline, downstream analyses, and figure generation, are available at GitHub (`https://github.com/benjaminhgood/LTEE-metagenomic`). The repository also contains the final list of mutations obtained after the variant calling steps in Section 4, so that the downstream analyses can be reproduced without the computationally intensive steps in the sequencing pipeline.

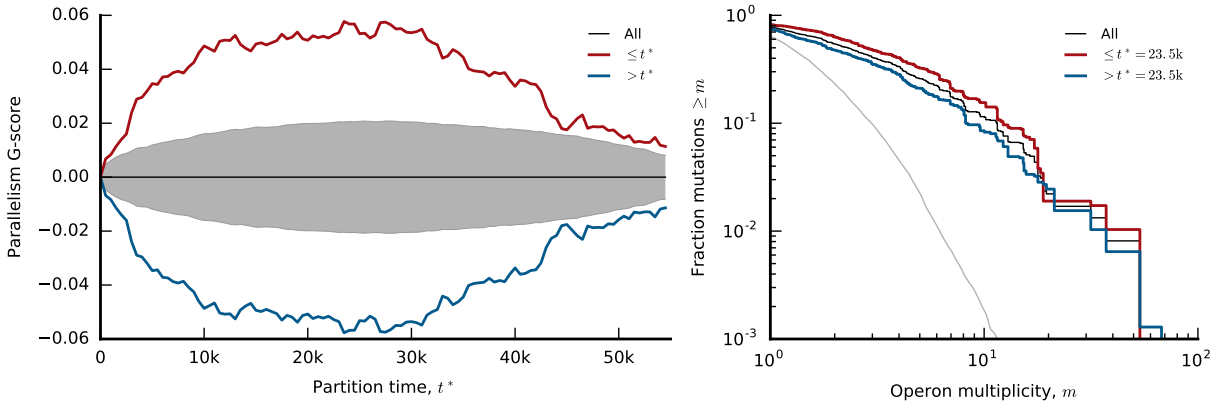

Figure S17: An analogous version of Fig. S15 calculated at the operon level. Left panel: scaled G-score changes in Eq. (80) as a function of the partition time, t^*. Lines denote observed values, while the shaded regions represent 95% confidence intervals obtained by randomly permuting appearance times across operons. Right panel: fraction of non-synonymous mutations that are found in operons with multiplicity greater than or equal to m and which occurred either before (red) or after (blue) the value of t^* (see legend) that maximizes the difference between $\Delta g_<$ and $\Delta g_>$ in the left panel. For comparison, the distribution of all mutations is shown in black, while the null distribution from Eq. (72) is shown in grey.

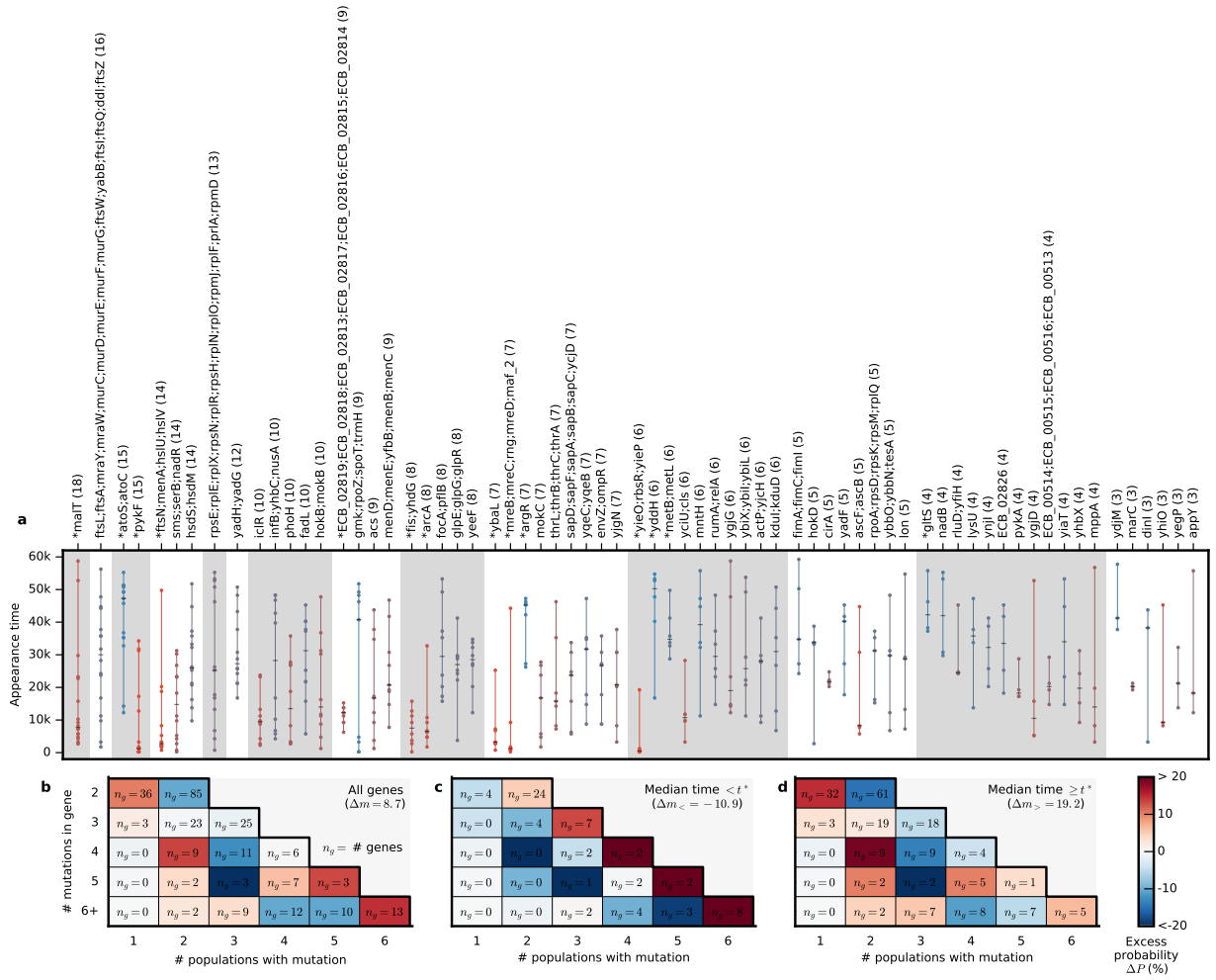

Figure S18: An analogous version of Fig. 6 calculated at the operon level. (a) Operons with three or more independent mutations in the nonmutator populations and whose multiplicity is significant at an FDR of 5%. Circles denote the estimated appearance time of each mutation, and they are connected by a vertical line for visualization. Each operon is colored according to its median appearance time, which is indicated by a dash. Operons with significantly non-random (i.e., non-uniform) appearance times are indicated by an asterisk. b, c, d, The distribution of possible dispersion configurations for (b) all mutations and (c, d) those in operons with median appearance time before or after t^*, as defined by Eq. (87).

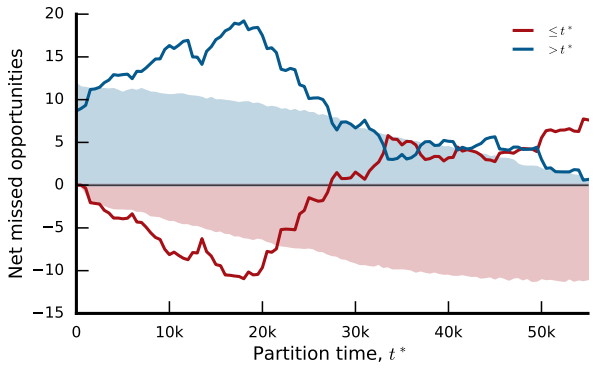

Figure S19: Net missed opportunities at the operon level in the nonmutator populations as a function of the partition time t^*. Lines denote the net missed opportunities for pre- and post-t^* operons calculated from Eq. (86). Shaded regions denote one-sided 95% confidence intervals obtained by sampling from the null model in Eq. (82).

Supplementary tables

Supplementary Table 1: **List of metagenomic samples used in this study.** 1512 mixed-population samples are summarized according to their REL freezer identifier, population, and timepoint of origin, and associated sequencing batch metadata. Twenty-eight samples were removed from further analysis due to insufficient coverage or demultiplexing errors. These excluded samples are indicated in the "Flagged?" column.

Supplementary Table 2: **List of clonal isolates used in this study.** For completeness, the 264 clonal isolates sequenced by Tenaillon et al. [3] are listed in the same format as Supplementary Table 1. Ten samples were removed from further analysis because they consumed too much memory at the variant calling step. These excluded samples are indicated in the "Flagged?" column.

Supplementary Table 3: **List of genes showing significant parallelism in the nonmutator populations.** Genes are summarized by their name, estimated target size, the observed and expected number of mutations across the six nonmutator populations, the corresponding multiplicity score, and the P-value describing the probability of observing an equal or larger of mutations under the null model.

Supplementary Table 4: **List of operons showing significant parallelism in the nonmutator populations.** Operons are summarized by their constituent genes, estimated target size, the observed and expected number of mutations across the six nonmutator populations, the corresponding multiplicity score, and the P-value describing the probability of observing an equal or larger of mutations under the null model.

References

[1] R E Lenski, M R Rose, S C Simpson, and S C Tadler. Long-term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *American Naturalist*, 138:1315–1341, 1991.

[2] H Jeong, V Barbe, C H Lee, D Vallenet, D S Yu, et al. Genome sequences of *Escherichia coli* B strains rel606 and bl21(de3). *J Mol Biol*, 394:644–652, 2009.

[3] O Tenaillon, J E Barrick, N Ribeck, D E Deatherage, J L Blanchard, et al. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536:165–170, 2016.

[4] F Vasi, M Travisano, and R E Lenski. Long-term experimental evolution in *Escherichia coli*. ii. changes in life-history traits during adaptation to a seasonal environment. *American Naturalist*, 144:432–456, 1994.

[5] M J Wiser, N Ribeck, and R E Lenski. Long-term dynamics of adaptation in asexual populations. *Science*, 342:1364–1367, 2013.

[6] R E Lenski, M J Wiser, N Ribeck, Z D Blount, J R Nahum, et al. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. *Proceedings of the Royal Society, London B*, 282:20152292, 2015.

[7] M J Wiser, N Ribeck, and R E Lenski. Data from: Long-term dynamics of adaptation in asexual populations. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.0hc2m, 2013.

[8] R E Lenski, M J Wiser, N Ribeck, Z D Blount, J R Nahum, et al. Data from: Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.gd3dq, 2015.

[9] M Baym, S Kryazhimskiy, T D Lieberman, H Chung, M M Desai, and R Kishony. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE*, 10:e0128036, 2015.

[10] D E Deatherage and J E Barrick. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol*, 1151:165–188, 2014.

[11] D E Deatherage, C C Traverse, L N Wolf, and J E Barrick. Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq. *Front Genet*, 5: 468, 2015.

[12] A M Bolger, M Lohse, and B Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30:2114–2120, 2014.

[13] G I Lang, D P Rice, M J Hickman, E Sodergren, G M Weinstock, D Botstein, and M M Desai. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500:571–574, 2013.

[14] J D Storey and R Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, 100:9440–9445, 2003.

[15] V S Cooper, D Schneider, M Blot, and R E Lenski. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J Bacteriol*, 183: 2834–2841, 2001.

[16] M McCandlish, J Otwinowski, and J B Plotkin. Detecting epistasis from an ensemble of adapting populations. *Evolution*, 69:2359–2380, 2015.

[17] S F Levy, J R Blundell, S Venkataram, D A Petrov, D S Fisher, and G Sherlock. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, 519:181–186, 2015.

[18] R Durbin, S R Eddy, A Krogh, and G Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK, 1998.

[19] B H Good and M M Desai. The impact of macroscopic epistasis on long-term evolutionary dynamics. *Genetics*, 199:177–190, 2015.

[20] S Kryazhimskiy, G Tkačik, and J B Plotkin. The dynamics of adaptation on correlated fitness landscapes. *Proc Natl Acad Sci USA*, 106:18638–18643, 2009.

[21] K Kosheleva and M M Desai. The dynamics of genetic draft in rapidly adapting populations. *Genetics*, 195:1007–1025, 2013.

[22] M M Desai, A M Walczak, and D S Fisher. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 193:565–585, 2013.

[23] Z D Blount, C Z Borland, and R E Lenski. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc Natl Acad Sci USA*, 105: 7899–7906, 2008.

[24] S Wielgoss, J E Barrick, O Tenaillon, M J Wiser, J Dittmar, S Cruveiller, B Chane-Woon-Ming, C Médigue, R E Lenski, and D Schneider. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci USA*, 110: 222–227, 2013.

[25] R Maddamsetti, P J Hatcher, A G Green, B L Williams, D S Marks, and R E Lenski. Core genes evolve rapidly in the long-term evolution experiment with *Escherichia coli*. *Genome Biology and Evolution*, in press, 2017.

[26] O Tenaillon, A Rodríguez-Verdugo, R L Gaut, P McDonald, A F Bennett, A D Long, and B S Gaut. The molecular diversity of adaptive convergence. *Science*, 335:457–461, 2012.

[27] D J Kvitek and G Sherlock. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genetics*, 9:e1003972, 2013.

[28] S K Kryazhimskiy, D P Rice, E R Jerison, and M M Desai. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344:1519–1522, 2014.

[29] B H Good, I M Rouzine, D J Balick, O Hallatschek, and M M Desai. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc Natl Acad Sci USA*, 109:4950–4955, 2012.

[30] Z D Blount, J E Barrick, C J Davidson, and R E Lenski. Genomic analysis of a key innovation in an experimental escherichia coli population. *Nature*, 489:513–518, 2012.

[31] E M Quandt, D E deatherage, A D Ellington, G Georgiou, and J E Barrick. Recursive genomewide recombination and sequencing reveals a key refinement step in the evolution of a metabolic innovation in *Escherichia coli*. *Proc Natl Acad Sci USA*, 111:2217–2222, 2014.

[32] E M Quandt, J Gollihar, Z D Blount, A D Ellington, G Georgiou, and J E Barrick. Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the lenski evolution experiment. *Elife*, 4:e09696, 2015.

[33] R Woods, D Schneider, C L Winkworth, M A Riley, and R E Lenski. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc Natl Acad Sci USA*, 103:9107–9112, 2006.

[34] F Mao, P Dam, J Chou, V Olman, and Y Xu. DOOR: a database of prokaryotic operons. *Nucl Acids Res*, 37:D459–D463, 2009.