doi:10.1038/nature24287

The dynamics of molecular evolution over 60,000 generations

Benjamin H. Good[1,2,3,4,5]*, Michael J. McDonald[1,2,6]*, Jeffrey E. Barrick[7,8], Richard E. Lenski[8,9] & Michael M. Desai[1,2,3]

The outcomes of evolution are determined by a stochastic dynamical process that governs how mutations arise and spread through a population. However, it is difficult to observe these dynamics directly over long periods and across entire genomes. Here we analyse the dynamics of molecular evolution in twelve experimental populations of *Escherichia coli*, using whole-genome metagenomic sequencing at five hundred-generation intervals through sixty thousand generations. Although the rate of fitness gain declines over time, molecular evolution is characterized by signatures of rapid adaptation throughout the duration of the experiment, with multiple beneficial variants simultaneously competing for dominance in each population. Interactions between ecological and evolutionary processes play an important role, as long-term quasi-stable coexistence arises spontaneously in most populations, and evolution continues within each clade. We also present evidence that the targets of natural selection change over time, as epistasis and historical contingency alter the strength of selection on different genes. Together, these results show that long-term adaptation to a constant environment can be a more complex and dynamic process than is often assumed.

Evolutionary adaptation is driven by the accumulation of mutations, but the temporal dynamics of this process are difficult to observe directly. Recently, time-resolved sequencing of microbial evolution experiments[1–6], viral and bacterial infections[7–9], and cancers[10] has begun to illuminate this process. These studies reveal complex dynamics, characterized by rapid adaptation, competition between beneficial mutations, diminishing-returns epistasis, and extensive genetic parallelism. These forces can alter patterns of polymorphism[11] and influence which mutations ultimately fix[12–15]. However, it is unclear whether these dynamics are general or, instead, reflect the short timescales and novel environmental conditions of previous studies.

To address this question, we turned to an experiment with the longest frozen 'fossil record': the *E. coli* long-term evolution experiment (LTEE)[16]. The twelve LTEE populations have been serially propagated in the same medium for more than 60,000 generations, with samples preserved every 500 generations (Supplementary Information 1). Previous work has shown that the competitive fitness of each population continues to increase through 60,000 generations, despite a decline in the rate of improvement[17,18]. The genome sequences of evolved clones have shown that these fitness gains are accompanied by steady accumulation of mutations[3,4]. Parallel genetic changes across replicate populations suggest that there is a common pool of adaptive mutations that has yet to be exhausted in any single population[4].

These previous findings show that the LTEE populations have not yet reached a fitness peak, even after tens of thousands of generations in the same environment. However, the existing data provide only limited information about the population genetic processes that drive these changes. Does the supply of adaptive mutations eventually diminish enough that evolution proceeds via discrete selective sweeps? Are the populations still approaching the same fitness peak as they accumulate mutations from a common pool? Or do more complicated dynamics arise that require more complex models? To answer these questions, we require more finely resolved information about the genetic diversity within each population through time. This will allow us to analyse when and in what order the successful mutations occur, the dynamics by which they spread through a population, and what other competing mutations have arisen and been eliminated.

Reconstructing the molecular fossil record

To measure the dynamics of molecular evolution, we sequenced mixed-population samples taken at 500-generation intervals across 60,000 generations of evolution in each of the twelve LTEE populations (Ara+1 to Ara+6 and Ara−1 to Ara−6) (Supplementary Information 3). This yielded a total of 1,431 samples with a median coverage of about 50× (Supplementary Table 1). To distinguish mutations from sequencing errors, we developed a pipeline that leverages the temporal correlations expected in a true mutation trajectory (Supplementary Information 4). This approach allows us to identify a subset of the mutations that reached of order 10% frequency in at least two sampled time points, and to track the frequency of the derived alleles through the rest of the timecourse. Our pipeline identifies both point mutations and indels, including many events mediated by insertion sequence elements (Supplementary Information 4).

Figure 1 shows the allele frequency trajectories of all mutations identified in each population. Although previous work has shown that fitness gains across the replicate populations are largely similar to one another[17,18] (Fig. 2a), Fig. 1 reveals a wide range of dynamics at the genetic level.

We analysed the rate at which mutations accumulate through time by calculating the total derived allele frequency $M_p(t) = \Sigma f_{p,m}(t)$ for all mutations m in population p at time t (Fig. 2b, Supplementary Information 5.1). This quantity approximates the expected number of mutations in a randomly sampled individual, neglecting mutations that never rise above our detection threshold. Consistent with earlier

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [2]FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA. [4]Department of Physics, University of California Berkeley, Berkeley, California 94720, USA. [5]Department of Bioengineering, University of California Berkeley, Berkeley, California 94720, USA. [6]Centre for Geometric Biology, School of Biological Sciences, Monash University, Clayton, Victoria 3800, Australia. [7]Department of Molecular Biosciences, The University of Texas, Austin, Texas 78712, USA. [8]BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, Michigan 48824, USA. [9]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA.
*These authors contributed equally to this work.

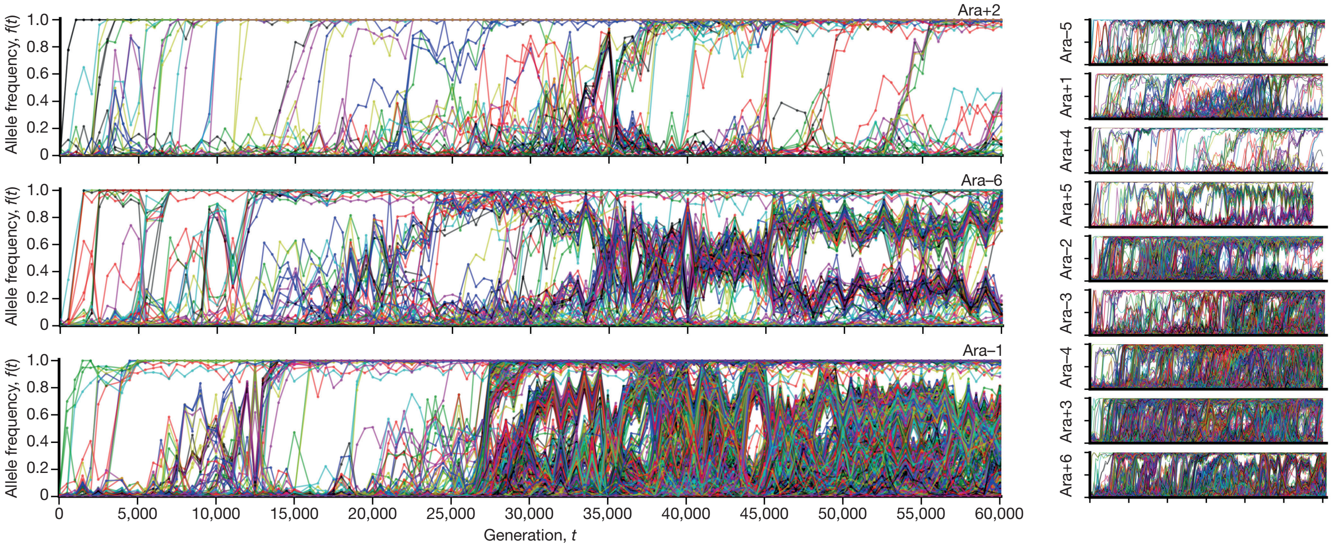


Figure 1 | The dynamics of molecular evolution. Allele frequency trajectories of all *de novo* mutations detected in the twelve LTEE populations.

work[3,4], Fig. 2 shows that the pace of molecular evolution has remained rapid throughout the experiment, even as the rate of fitness improvement has declined[17,18].

The high temporal resolution of the data reveals striking differences in the rate of molecular evolution over time and across replicate populations. Six populations evolved a mutator phenotype[4,19], producing a sudden jump in total derived allele frequency (Fig. 2b). In some of these mutator populations, the rate of molecular evolution later declined (Fig. 2 inset), consistent with evidence from sequenced clones[4]. In Ara−1, previous work has shown that this deceleration is driven by ‘antimutator’ alleles that arise after the fixation of the initial mutator[20]. Our results suggest that a similar process also occurs in other populations.

In contrast to the mutator lines, the six ‘nonmutator’ populations accumulate mutations at a steadier pace. Their average rate of molecular evolution declines modestly over time, decreasing from about 20 mutations in the first 10,000 generations to about 10 mutations in the last 10,000 (Fig. 2c). There are also systematic differences between populations that persist over 10,000-generation intervals, suggesting that the populations acquired mutations at slightly different rates (Extended Data Fig. 1).

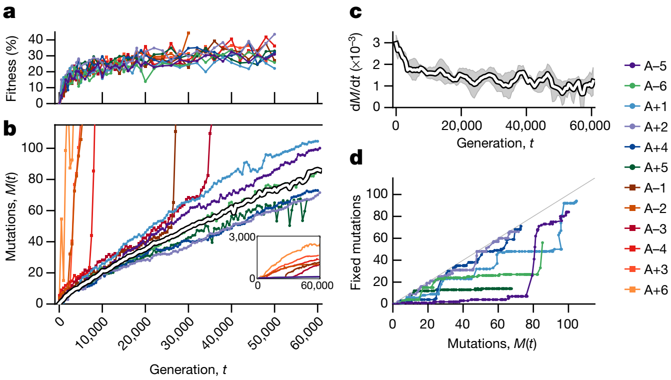


Figure 2 | Rates of molecular evolution. a, Competitive fitness through time (Supplementary Information 2). **b**, Number of mutations in each population as a function of time, measured by total derived allele frequency, $M_p(t)$. The average of the nonmutator populations is shown in white. **c**, Average rate of change of $M_p(t)$ for nonmutators in 5,000-generation sliding windows. Shaded region depicts a 95% confidence interval obtained by bootstrapping replicate populations 10,000 times. **d**, Number of fixed mutations versus $M_p(t)$ in nonmutators.

The rates at which mutations accumulate in nonmutator lineages are comparable to previous estimates of bacterial mutation rates[21]. However, they are incompatible with the timescale of neutral evolution. With an effective population size of $N_e \sim 10^7$, new mutations would require $\Delta t \sim 0.1 N_e \sim 10^6$ generations to reach the 10% detection threshold by genetic drift alone[22]. Thus, the mutations in Fig. 1 must have reached observable frequencies through the direct or indirect action of natural selection.

Emergence of quasi-stable coexistence

Once a mutation reaches detectable frequencies, the shape of its allele frequency trajectory contains information about selective forces. The trajectories in Fig. 1 are inconsistent with a ‘periodic selection’ model in which individual driver mutations fix in a sequence of discrete selective sweeps. This model predicts that a driver mutation with fitness benefit *s* (along with any nearly-neutral hitchhikers) should quickly and deterministically fix after reaching observable frequency, which greatly exceeds the drift barrier $1/N_e s$. By contrast, many mutations in Fig. 1 persist at intermediate frequencies for long periods, often undergoing reversals in frequency that sometimes result in extinction.

Part of this complexity is driven by clonal interference. When beneficial mutations are common, mutations that would otherwise drive selective sweeps can be outcompeted by other lineages carrying superior beneficial mutations[23]. Further beneficial mutations can draw out this battle, resulting in allele-frequency trajectories with multiple inflection points[12,24,25]. But models of clonal interference predict that one lineage must eventually win, and so on long timescales the number of fixed mutations should grow at the same rate as the total allele frequency $M_p(t)$.

To test this expectation, we developed a hidden Markov model (HMM; Supplementary Information 5.2) to estimate the ‘fixation time’ of each mutation from its allele frequency trajectory, allowing us to estimate the number of fixed mutations through time (Fig. 2d). The number of fixed mutations closely tracks $M_p(t)$ in some populations (for example, Ara+2 and Ara+4), but there is a marked deficit of fixations in others (for example, Ara−6). Instead of fixing, the ‘missing’ mutations segregate into at least two intermediate-frequency clades that coexist for long periods (Fig. 1).

To investigate these clades, we extended our mutation-trajectory HMM to assign mutations to basal, major or minor clades, and to infer the frequencies of these clades through time (Fig. 3a, Supplementary Information 5.3). This approach leverages correlations in the trajectories of many independent mutations, while accounting for noise in each sample. The results confirm that long-lived clades are common

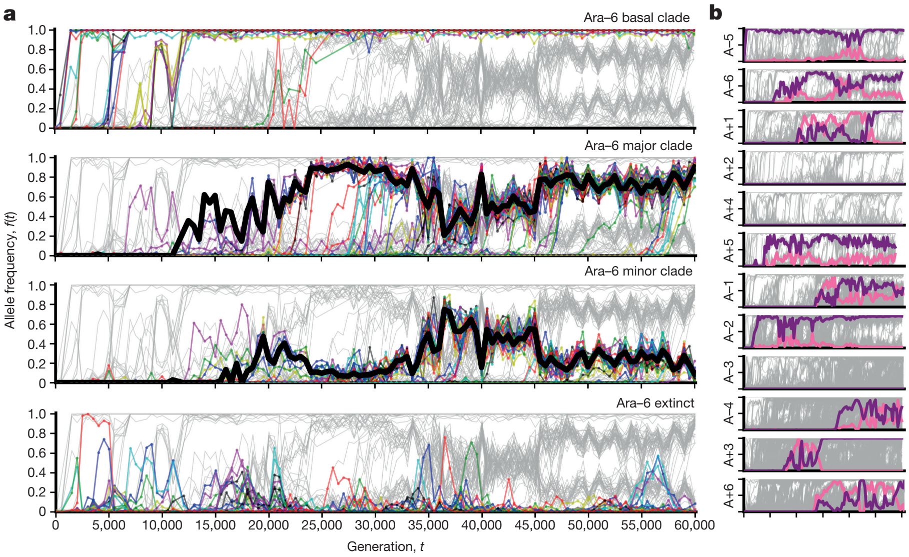


Figure 3 | Long-term coexistence of competing clades. a, Output of the clade-aware HMM for population Ara−6. Major and minor clades (solid black lines) are defined by the clade frequencies at the final time point, and the basal clade contains mutations shared by major and minor clades. Coloured lines indicate mutations within the corresponding clade in each panel; all other mutations are shown in grey. **b**, Estimated clade frequencies for all twelve populations (major clade in purple, minor clade in pink). Individual mutations are shown in grey.

in the LTEE. Figure 3b shows that nine of the twelve populations have clades that coexist for more than 10,000 generations, often persisting through to generation 60,000. By partitioning the mutations into clades (Fig. 3a), we also see that fixations continue to accumulate within each clade, even when population-wide fixation events have ceased.

This striking separation of timescales between inter-clade and intra-clade fixations cannot be explained by clonal interference[26]. Instead, long-term coexistence is likely to be maintained by negative frequency-dependent selection, as has been demonstrated in Ara−2[27,28]. It is not known whether these additional examples of coexistence revealed by our data involve the same glucose/acetate cross-feeding interaction as was seen in Ara−2, or whether these populations have exploited other opportunities for ecological diversification.

Regardless of the mechanism of coexistence, the metagenomic data show that the balance between the two clades does not remain constant over long timescales. Instead, their relative abundance can shift by at least an order of magnitude during their coexistence. The timing and magnitudes of these shifts vary from population to population; they could reflect ongoing selection on the mechanism of coexistence or a general coupling between the ecologically divergent phenotypes and ordinary fitness gains[28–30]. Further work is needed to distinguish between these scenarios.

Dynamics and fates of new mutations

Most models of molecular evolution do not account for frequency-dependent selection, which complicates efforts to understand the evolutionary dynamics using population-wide data. To overcome this problem, we focused on the dynamics within each clade.

First, using the clade-aware HMM, we estimated the appearance and fixation times of all mutations that fixed in basal or majority clades in the nonmutator populations (Supplementary Information 5.3.1). These are upper and lower bounds, respectively, as they exclude time outside the observable frequency range. From these measurements, we calculated the number of fixed mutations in the basal or majority clade through time (Fig. 4a). These data show that within-clade fixations continue at a steady pace, consistent with the $M_p(t)$ trajectories in Fig. 2b. Although the average rate of fixations declines only modestly during the experiment, there is noticeable temporal variability as mutations often fix in cohorts of multiple linked mutations. These cohorts have been observed previously[1,29] and are expected in models of clonal interference[31,32]. However, they could also reflect transiently stable frequency-dependent interactions, as previously observed in Ara−1[29].

The difference between the appearance and fixation times of each successful mutation (the transit time) is a proxy for the strength of selection acting on a lineage. Despite the declining rate of fitness gain (Fig. 2a), we observe a broad distribution of transit times throughout the experiment (Fig. 4b). Even after 50,000 generations, some mutations appear to fix nearly as rapidly as those that occurred in the first 5,000 generations of evolution. This observation suggests that fitness differences between cohorts of mutations can remain high, with selection coefficients $s > 2\log|1-\Delta f|/\Delta t \sim 1\%$, even after many beneficial mutations have fixed.

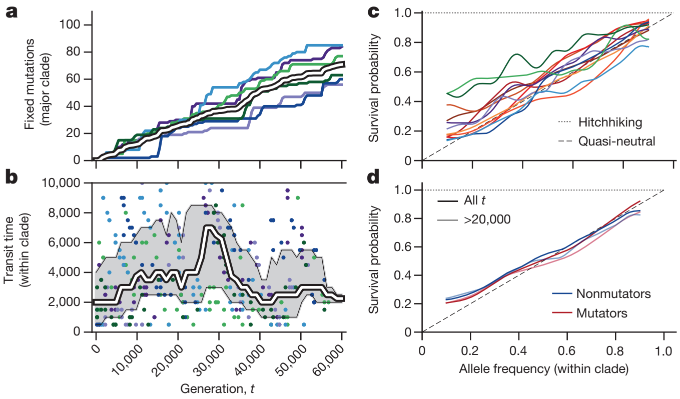


Figure 4 | Evolutionary dynamics within clades. a, Number of mutations fixed within the basal or major clade through time in the nonmutator populations. Colours are the same as in Fig. 2, and the ensemble average is in white. **b**, The transit time of each mutation in **a** as a function of its appearance time. White line shows the median across the six populations in non-overlapping five-percentile windows, and the interquartile range of each window is in grey. **c**, Fixation probability as a function of current mutation frequency within its parent clade, along with expectations under quasi-neutral and hitchhiking models. Fixation probabilities are estimated using sliding frequency windows (Supplementary Information 5.3.2). **d**, Pooled version of **c** for mutator and nonmutator populations. Lighter lines include only time points from generation 20,000 onwards.

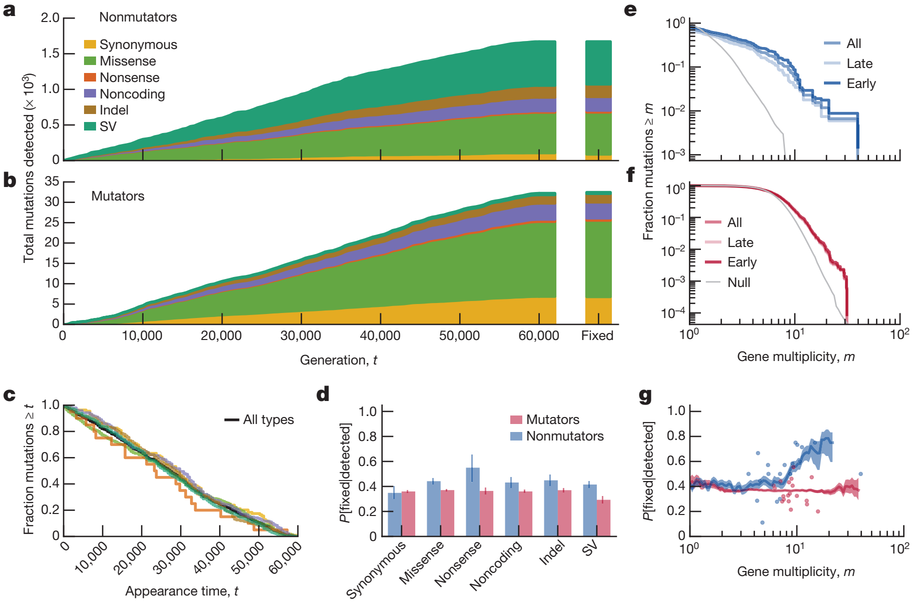


Figure 5 | Parallelism. a, **b**, Cumulative distribution of detected mutations of each type in nonmutator (**a**) and mutator (**b**) populations over time. The SV class denotes structural variants, including insertion sequence-mediated mutations. Bars at right depict the distribution of mutations that fixed within their respective clades. **c**, Distribution of appearance times for each variant type in nonmutators. **d**, Fraction of detected mutations of each type that fixed in nonmutator and mutator populations (blue and red, respectively). Error bars denote the 16th and 84th percentiles of the beta posterior distribution; sample sizes are indicated by the final timepoint in **a**. **e**, **f**, Fraction of all mutations (excluding synonymous mutations) in nonmutator (**e**) and mutator (**f**) populations in genes with multiplicity $m_i \geq m$. The grey line is the null distribution, obtained by randomly distributing the mutations across genes. **g**, Average conditional fixation probability of a mutation as a function of its gene multiplicity (in sliding windows of 0.2 $\log_{10}$ units) in nonmutator (blue) and mutator (red) populations. Shaded confidence intervals denote the 16th and 84th percentiles of the beta posterior distribution of each window. Fixation probabilities of the 20 most frequently mutated genes are shown as dots.

In addition to mutations that fix, many others reach substantial frequencies before going extinct, consistent with clonal interference. To quantify this effect, we estimated the fixation probability of a mutation as a function of its (within-clade) frequency (Fig. 4c, d). As explained above, a mutation can reach observable frequencies only if it is linked to a beneficial driver mutation or is a driver itself. Thus, without clonal interference, all observed mutations should fix in their clade with nearly 100% probability. By contrast, the fixation probabilities in Fig. 4c, d are substantially lower, even when restricted to mutations that arose in later generations. Instead, the observed fixation probabilities are more consistent with the quasi-neutral limit, $p_{\text{fix}}(f) \approx f$, which arises when clonal interference is strong[13,25] (Supplementary Information 5.3.2). This quasi-neutrality implies that adaptation in the LTEE is not mutation-limited; instead, clonal interference and hitchhiking remain important even after tens of thousands of generations in the same environment.

Parallelism at the genetic level

Allele frequency trajectories provide evidence for pervasive adaptation in the LTEE, but the dynamics alone provide limited information about which mutations are beneficial drivers and which are neutral or deleterious passengers. However, we can leverage the identities of mutations to learn about the targets of selection, and to investigate whether these targets change through time or differ across populations.

Figure 5a, b shows the cumulative distribution of all detected variant types through time. In the mutator populations, this distribution reflects the mutational biases and appearance times of mutator phenotypes. By contrast, we see few temporal changes in the types of mutations in nonmutators, apart from a slight early enrichment of missense mutations (Fig. 5c). Consistent with previous studies[3,4], we observe an excess of nonsynonymous relative to synonymous mutations in nonmutators ($dN/dS > 1$; Extended Data Fig. 2), indicating that many observed mutations are adaptive (even those driven extinct by clonal interference). By contrast, $dN/dS \leq 1$ in mutators, reflecting a higher proportion of passenger mutations.

Because we observe the fates of mutations through time, we can examine how the distribution of variant types differs between the entire pool of detected mutations and the subset that fixed in their respective clades (a generalization of the McDonald–Kreitman test[33]). This approach allows us to estimate a fixation probability for each class of mutations, conditioned on reaching detectable frequency (Fig. 5d). In nonmutator lines, synonymous mutations have a smaller conditional fixation probability than other variant types (Fig. 5d), as expected if the latter are more likely to be beneficial. Nevertheless, the ratio of conditional fixation probabilities is smaller than dN/dS, suggesting that mutations are strongly influenced by genetic draft (that is, linkage and associated hitchhiking) once they reach observable frequencies. Consistent with this interpretation, conditional fixation probabilities in mutator lines meet (or slightly exceed) the synonymous expectation, even though $dN/dS \leq 1$.

Parallel genetic changes can reveal targets of selection on more finely resolved scales. Although we find some parallelism at the nucleotide level (Extended Data Fig. 3), more information is obtained by grouping mutations into genes and their respective promoter regions. We quantified parallelism in a gene by its effective multiplicity, m_i, defined as the observed number of non-synonymous changes n_i (including indels and structural variants), normalized by gene length. Consistent with previous studies[4,34], we find significantly more multi-hit mutations than expected by chance (Supplementary Information 6.3.1), though the excess is more pronounced in nonmutators (Fig. 5e, f).

This excess parallelism could be driven by natural selection or local increases in mutation rate (for example, due to a nearby insertion sequence element). However, we find that multiplicity is positively correlated with conditional fixation probability in nonmutators ($P \approx 0.001$; logistic regression) and essentially uncorrelated in mutators ($P \approx 0.4$), suggesting that much of the excess parallelism in nonmutators is driven

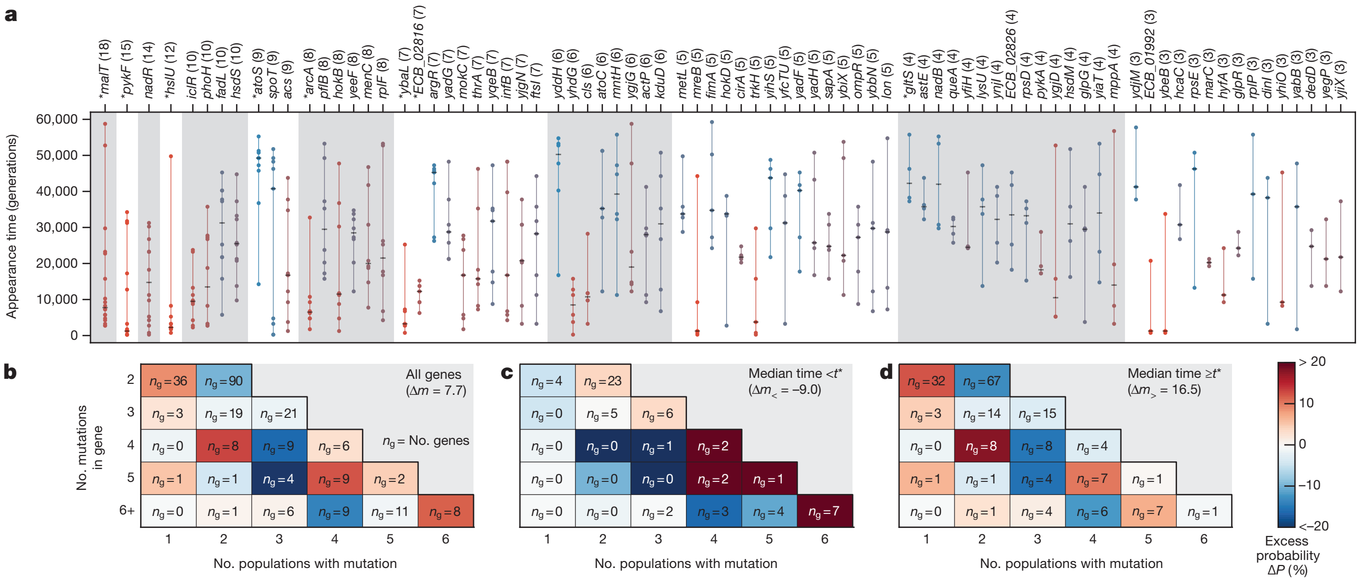


Figure 6 | Epistasis and contingency. a, Genes mutated three or more times in nonmutators with multiplicities significant at 5% FDR. Circles indicate the appearance time of each mutation, connected by a vertical line for visualization. Each gene is coloured according to its median appearance time (hatch-mark). Genes with significantly non-random appearance times are marked by asterisks. **b**–**d**, The distribution of dispersion configurations of a gene (that is, the total number of mutations versus the number of different populations in which they appeared) for all genes (**b**) and those with median mutation appearance times before or after $t^* = 17{,}500$ generations, which was chosen to maximize the number of 'missed opportunities' (**c**, **d**; Supplementary Information 6.3.3).

by selection (Fig. 5g). However, there is substantial variation around this trend, and even for the most recurrently mutated genes the fixation probability rarely rises above 80%. Thus, although selection plays a large role in driving mutations to detectable frequencies, stochastic forces and interactions among competing lineages also are important in determining the fates of mutations.

Signatures of epistasis and historical contingency

We next quantified how signatures of parallelism vary over time and across populations. We first focused on genes mutated three or more times in nonmutators with multiplicities that were significant at 5% false discovery rate (FDR) (Fig. 6a, Supplementary Information 6.3.1). These genes include many previously identified targets of parallel evolution[3,4,34]. By permuting the appearance times of mutations across these genes (Supplementary Information 6.3.2), we find that mutations in many individual genes are distributed non-randomly (KS test, $q < 0.05$). Some genes (for example, *hslU*; Extended Data Fig. 4) are mutated early in the experiment but almost never late, whereas others (for example, *atoS*; Extended Data Fig. 5) show the opposite tendency. Moreover, there is a global enrichment of non-random appearance times, even after individually significant cases are removed (summed KS test, $P < 0.001$). This temporal bias is not restricted to high-multiplicity genes: mutations in two-hit genes also tend to happen closer together in time than mutations in different genes ($P < 0.001$; Extended Data Fig. 6). As a result, the observed repertoire of adaptive mutations changes over time (Extended Data Fig. 7, Supplementary Information 6.3.2).

Genes that accumulate mutations early are expected under a 'coupon collecting' model, in which genes with the most strongly beneficial mutations (or with higher mutation rates) are depleted once each population has acquired that mutation. Preferentially late genes might also be consistent with this model in the presence of clonal interference: weakly beneficial mutations that are usually outcompeted early can become successful once their stronger counterparts have fixed (Supplementary Information 6.3.3).

Preferentially late mutations could also reflect global changes in selection pressures with increasing fitness, or new evolutionary paths opened up by earlier substitutions. An example of the latter scenario is the evolution of citrate utilization in Ara−3, in which key mutations became beneficial only after earlier mutations[35–37]. We lack the statistical power to scan for such interactions directly, but this signal of contingency might still be reflected in the distribution of mutations across nonmutator populations (Supplementary Information 6.3.3). Specifically, we expect mutations in a contingent gene to be clustered in a subset of the populations (that is, those that fixed an unknown potentiating mutation). By contrast, genes in the coupon-collecting model should be over-dispersed, as additional mutations in the same lineage are no longer beneficial[33].

We find a few under-dispersed genes that are candidates for historical contingency (for example, *argR* has seven mutations clustered in three populations; Extended Data Fig. 8). However, these examples cannot reach genome-wide significance in our limited sample, so we instead focused on the global distribution of dispersion configurations (Fig. 6b). We find a trend towards under-dispersion in genes that were mutated four times or fewer, and signatures of both under- and over-dispersion in genes mutated five or more times. This pattern suggests a combination of historical contingency and coupon collecting, with the latter expected to decline over time as targets are depleted, and the former expected to increase as potentiating mutations arise. Consistent with this hypothesis, over-dispersion declines when we focus on genes with later appearance times, and under-dispersion becomes more pronounced (Fig. 6c, d). When summed across genes, this under-dispersion amounts to at least 16 'missed opportunities' (populations that would be expected to have produced a mutation in a target gene but did not), more than expected by chance ($P \approx 0.003$; Extended Data Fig. 9). Similar results are obtained after clustering genes into operons (Supplementary Information 6.4).

Together, these results support the hypothesis that new routes for adaptation are sometimes opened up by earlier mutations. Although purely statistical, this evidence implies that some adaptive mutations should be less beneficial (or even deleterious) when transplanted to genetic backgrounds without the corresponding potentiating mutations. This prediction might be tested directly in future work.

Discussion

The evolutionary dynamics that characterize long-term adaptation to a constant environment remain poorly documented empirically. Here, we observed this process directly by sequencing metagenomic samples from 60,000 generations of an ongoing experiment with *E. coli*. Our time-resolved molecular fossil record reveals a complex adaptive process, with clonal interference, genetic draft, and eco-evolutionary feedback playing important roles. Our data also suggest that the targets of selection shift over time, as emergent ecological interactions and changing genetic backgrounds create new genetic opportunities for adaptation that were not initially available. Such effects help to explain why the rate of molecular evolution remains so high through 60,000 generations.

Together, our results demonstrate that long-term adaptation to a fixed environment can be characterized by a rich and dynamic set of population genetic processes, in stark contrast to the evolutionary desert expected near a fitness optimum. Rather than relying only on standard models of neutral mutation accumulation and mutation–selection balance in well-adapted populations, these more complex dynamical processes should also be considered and included more broadly when interpreting natural genetic variation.

Data 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).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 April; accepted 21 September 2017.
Published online 18 October 2017.

1. Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500,** 571–574 (2013).
2. Kvitek, D. J. & Sherlock, G. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet.* **9,** e1003972 (2013).
3. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461,** 1243–1247 (2009).
4. Tenaillon, O. *et al.* Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536,** 165–170 (2016).
5. Miller, C. R., Joyce, P. & Wichman, H. A. Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics* **187,** 185–202 (2011).
6. McDonald, M. J., Rice, D. P. & Desai, M. M. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* **531,** 233–236 (2016).
7. Zanini, F. *et al.* Population genomics of intrapatient HIV-1 evolution. *eLife* **4,** e11282 (2015).
8. Luksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507,** 57–61 (2014).
9. Lieberman, T. D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* **43,** 1275–1280 (2011).
10. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
11. Neher, R. A. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annu. Rev. Ecol. Evol. Syst.* **44,** 195–215 (2013).
12. Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O. & Desai, M. M. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl Acad. Sci. USA* **109,** 4950–4955 (2012).
13. Schiffels, S., Szöllosi, G. J., Mustonen, V. & Lässig, M. Emergent neutrality in adaptive asexual evolution. *Genetics* **189,** 1361–1375 (2011).
14. Good, B. H. & Desai, M. M. Deleterious passengers in adapting populations. *Genetics* **198,** 1183–1208 (2014).
15. Kryazhimskiy, S., Tkacik, G. & Plotkin, J. B. The dynamics of adaptation on correlated fitness landscapes. *Proc. Natl Acad. Sci. USA* **106,** 18638–18643 (2009).
16. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-term experimental evolution in *Escherichia coli.* I. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138,** 1315–1341 (1991).
17. Wiser, M. J., Ribeck, N. & Lenski, R. E. Long-term dynamics of adaptation in asexual populations. *Science* **342,** 1364–1367 (2013).
18. Lenski, R. E. *et al.* Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. *Proc. R. Soc. B* **282,** 20152292 (2015).
19. Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* **387,** 703–705 (1997).
20. Wielgoss, S. *et al.* Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl Acad. Sci. USA* **110,** 222–227 (2013).
21. Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl Acad. Sci. USA* **88,** 7160–7164 (1991).
22. Ewens, W. *Mathematical Population Genetics* (Springer, 2004).
23. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102-103,** 127–144 (1998).
24. Desai, M. M. & Fisher, D. S. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176,** 1759–1798 (2007).
25. Kosheleva, K. & Desai, M. M. The dynamics of genetic draft in rapidly adapting populations. *Genetics* **195,** 1007–1025 (2013).
26. Desai, M. M., Walczak, A. M. & Fisher, D. S. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics* **193,** 565–585 (2013).
27. Rozen, D. E. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli.* VIII. Dynamics of a balanced polymorphism. *Am. Nat.* **155,** 24–35 (2000).
28. Plucain, J. *et al.* Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science* **343,** 1366–1369 (2014).
29. Maddamsetti, R., Lenski, R. E. & Barrick, J. E. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics* **200,** 619–631 (2015).
30. Frenkel, E. M. *et al.* Crowded growth leads to the spontaneous evolution of semistable coexistence in laboratory yeast populations. *Proc. Natl Acad. Sci. USA* **112,** 11306–11311 (2015).
31. Park, S. C. & Krug, J. Clonal interference in large populations. *Proc. Natl Acad. Sci. USA* **104,** 18135–18140 (2007).
32. Fisher, D. S. Asexual evolution waves: fluctuations and universality. *J. Stat. Mech.* **2013,** P01011 (2013).
33. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
34. Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A. & Lenski, R. E. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **103,** 9107–9112 (2006).
35. Blount, Z. D., Borland, C. Z. & Lenski, R. E. 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).
36. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* **489,** 513–518 (2012).
37. Quandt, E. M. *et al.* Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the Lenski evolution experiment. *eLife* **4,** e09696 (2015).

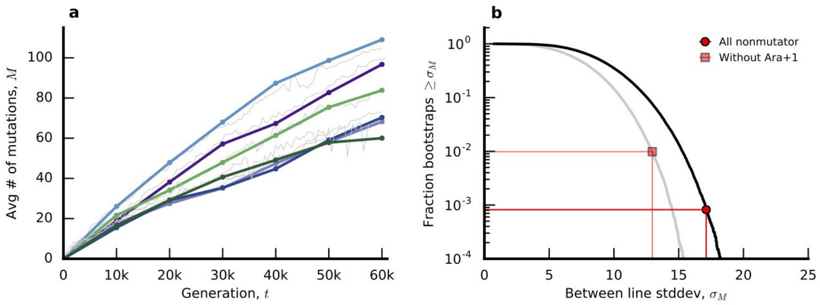
Supplementary Information is available in the online version of the paper.

Acknowledgements We thank N. Hajela, A. N. Nguyen Ba, and E. Jerison for assistance. B.H.G. acknowledges support from the US National Science Foundation (DEB-1501580) and the Miller Institute for Basic Research in Science at the University of California Berkeley. R.E.L. acknowledges support from the US National Science Foundation (DEB-1451740) and BEACON Center for the Study of Evolution in Action (DBI-0939454). M.M.D. acknowledges support from the Simons Foundation (grant 376196), the US National Science Foundation (PHY-1313638), and the US National Institutes of Health (GM104239). Computational work was performed on the Odyssey cluster supported by the Research Computing Group at Harvard University.

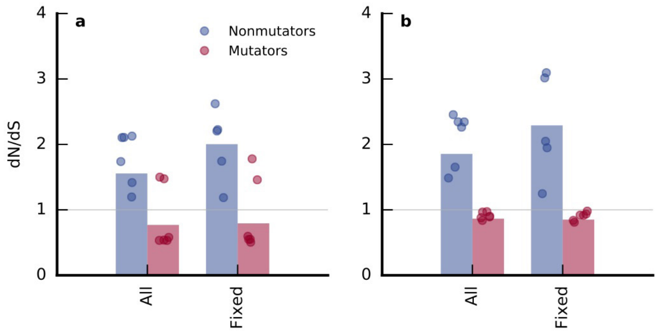
Author Contributions B.H.G., M.J.M., R.E.L. and M.M.D. designed the project; B.H.G. and M.J.M. conducted the experiments and generated the sequence data; B.H.G. and J.E.B. designed and conducted the bioinformatics analyses; B.H.G. developed theory and statistical methods; B.H.G., M.J.M., J.E.B., R.E.L. and M.M.D. analysed the data and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.M.D. (mmdesai@fas.harvard.edu).

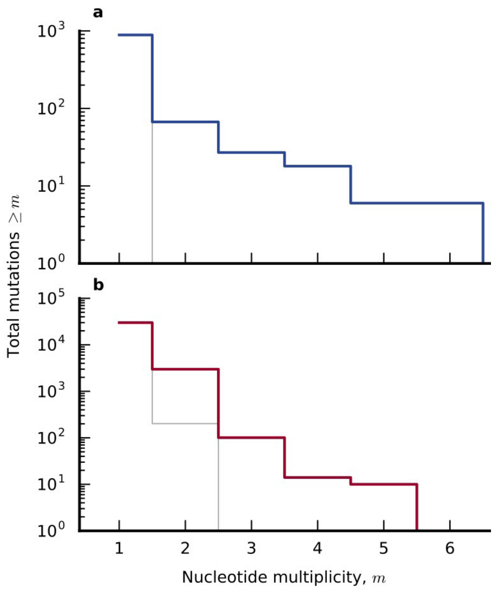
Reviewer Information *Nature* thanks R. Kishony, J. Plotkin, G. Sherlock and the other anonymous reviewer(s) for their contribution to the peer review of this work.



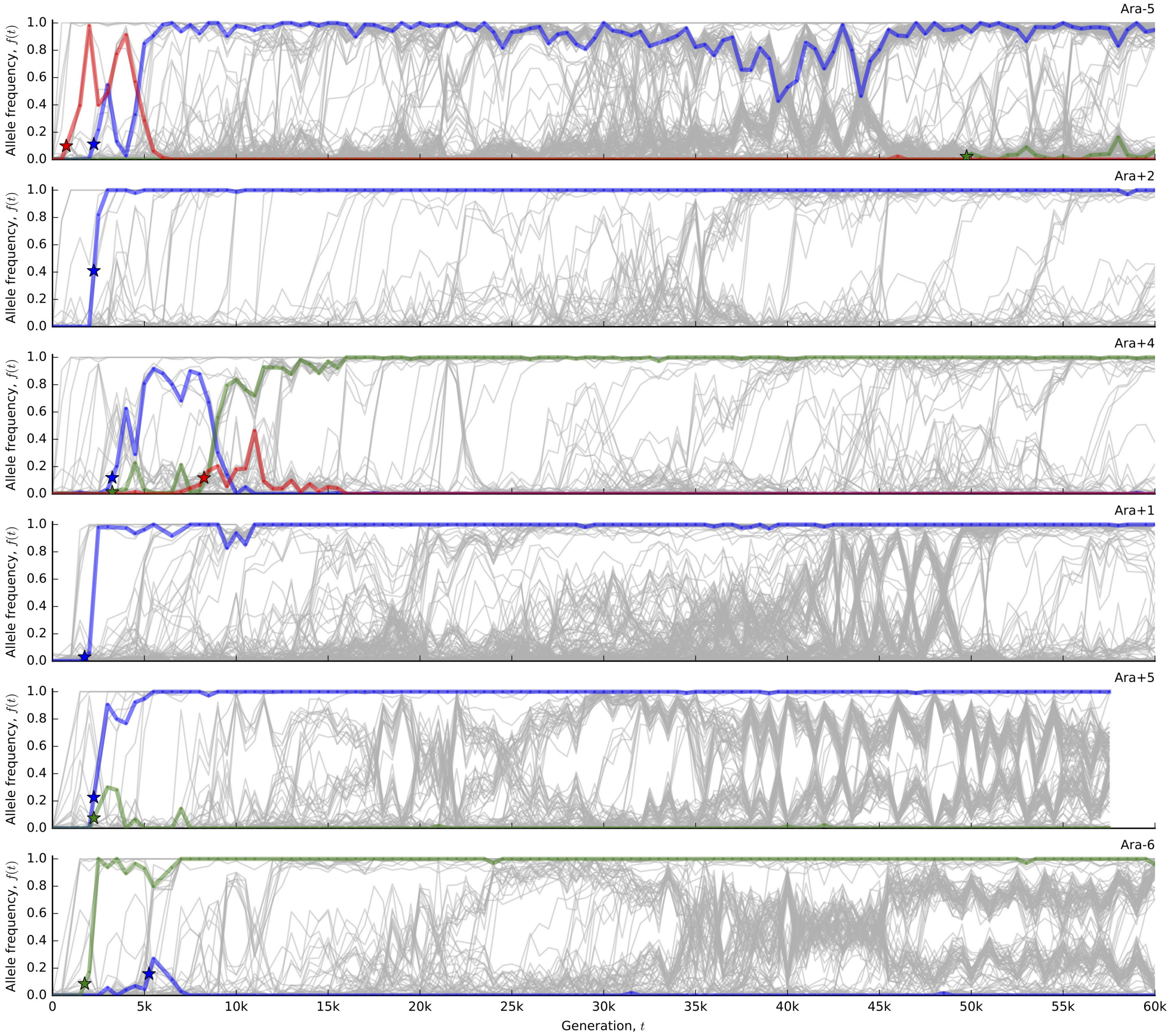
Extended Data Figure 1 | Between-line variability in the rate of mutation accumulation. a, Coarse-grained mutation gains $\Delta M_{p,k}$ (Supplementary Information 5.1) for the six nonmutator populations, plotted using the same colour scheme as in Fig. 2. For comparison, the original mutation trajectories $M_p(t)$ are shown in light grey. **b**, Between-line variability in $\sum_k \Delta M_{p,k}$, with and without the Ara+1 population. Observed values are indicated as symbols; solid lines show the corresponding null distribution obtained by randomly permuting $\Delta M_{p,k}$ across the six populations.



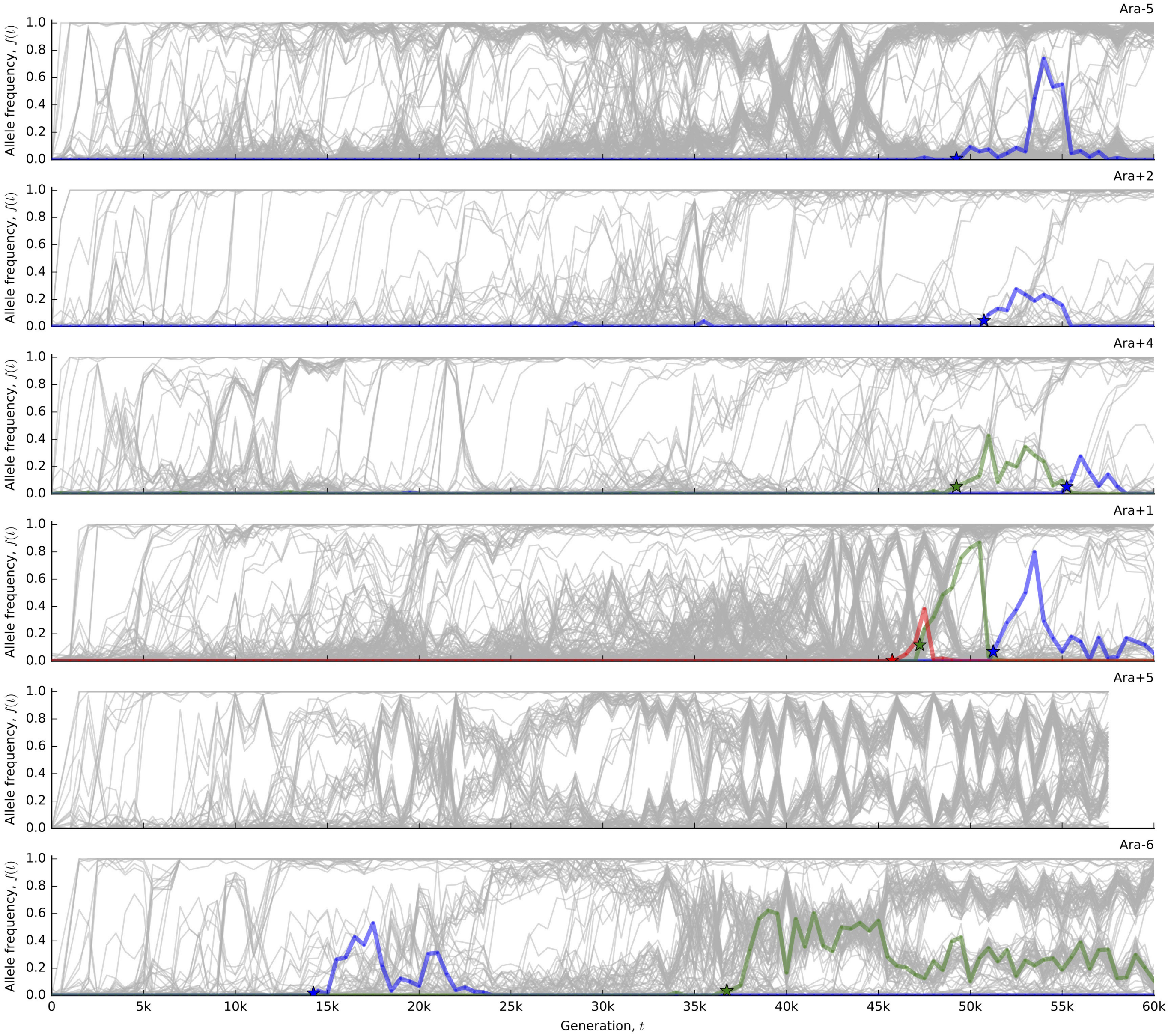
Extended Data Figure 2 | Nonsynonymous versus synonymous mutations. The ratio of nonsynonymous to synonymous mutations (d*N*/d*S*) in the entire pool of detected mutations, as well as the subset that fixed within their respective clades. Symbols denote individual populations; bars denote pooled estimates across either the nonmutator or mutator populations. In **a**, this ratio is normalized by the relative number of synonymous and nonsynonymous sites. Panel **b** corrects for the observed spectrum of single-nucleotide mutations in each population.



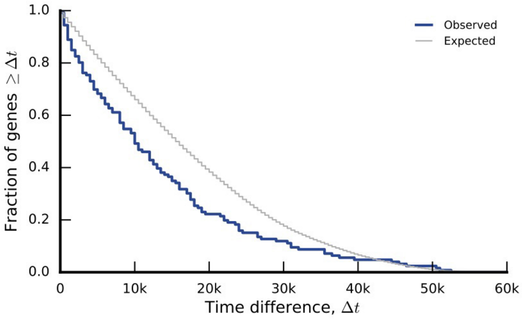
Extended Data Figure 3 | Parallelism at the nucleotide level. a, **b**, The distribution of nucleotide multiplicity (Supplementary Information 6.2) for the nonmutator (**a**) and mutator (**b**) populations. Observed data are shown in coloured lines, and the null expectations are shown in grey for comparison.



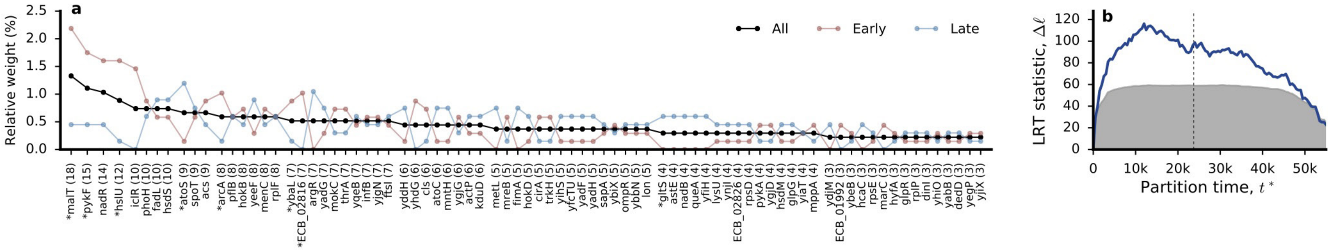
Extended Data Figure 4 | Mutations in *hslU*. Mutations that arose in the *hslU* gene in the six nonmutator populations. The inferred appearance times are indicated by the star symbols.



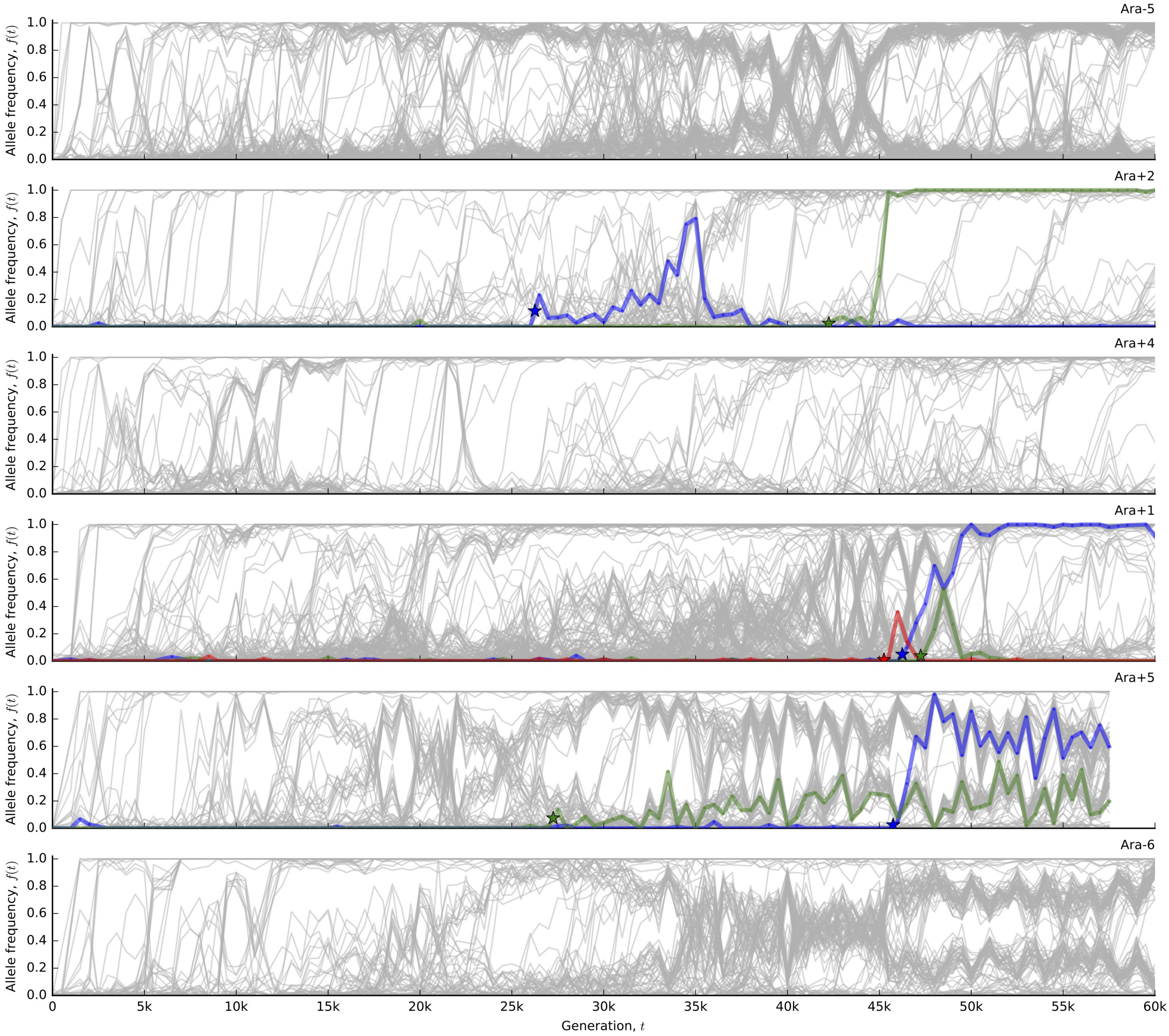
Extended Data Figure 5 | Mutations in *atoS*. Mutations that arose in the *atoS* gene in the six nonmutator populations. The inferred appearance times are indicated by stars.



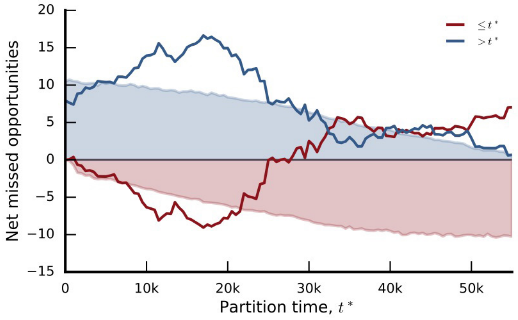
Extended Data Figure 6 | Temporal similarity among two-hit genes. The distribution of the difference between the earliest and latest appearance times in genes with exactly two detected mutations in the nonmutator lines. The null distribution is obtained by randomly permuting appearance times among the two-hit genes for 10,000 bootstrap iterations.



Extended Data Figure 7 | Realized mutation spectrum in different time windows. a, Fraction of mutations contributed by each gene in Fig. 6a, including time windows before and after the median appearance time of all mutations in those genes. **b**, Differences between the early and late distributions in **a** as a function of the partition time *t**. Dashed line denotes the median appearance time used to divide in **a**. Solid line shows the value of the likelihood ratio test (LRT) between these two distributions for different choices of *t** (Supplementary Information 6.3.2). Shaded region represents a one-sided 95% confidence interval obtained by randomly permuting appearance times across the subset of genes in **a** for 10,000 bootstrap iterations.



Extended Data Figure 8 | Mutations in *argR*. Mutations that arose in the *argR* gene in the six nonmutator populations. The inferred appearance times are indicated by stars.



Extended Data Figure 9 | Missed opportunities. Net missed opportunities in the nonmutator populations as a function of the partition time *t**. Lines denote the net missed opportunities for genes with median appearance times before and after *t**, as defined by the formula in Supplementary Information 6.3.3. Shaded regions denote one-sided 95% confidence intervals obtained by bootstrap resampling from the corresponding null model 10,000 times (see Supplementary Information 6.3.3).

nature research

Corresponding author(s): Michael Desai

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Our study analyzes samples from a laboratory evolution experiment begun in 1988, which consists of 12 biological replicates. We analyze the complete set of frozen samples from all 12 replicates.

2. Data exclusions

Describe any data exclusions.

Twenty-eight mixed-population samples were removed from further analysis due to insufficient coverage or demultiplexing errors. Ten clonal samples were removed from further analysis because they consumed too much memory at the variant calling step.

3. Replication

Describe whether the experimental findings were reliably reproduced.

No additional replication was performed.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

All 12 biological replicates have been subjected to identical experimental conditions.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not relevant to our study because all 12 biological replicates were subjected to identical experimental conditions.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

□ ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ A statement indicating how many times each experiment was replicated

□ ☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)

□ ☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

□ ☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted

□ ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

□ ☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

We used the freely available third-party software trimmomatic v0.32 as well as custom scripts available on Github (https://github. com/benjaminhgood/LTEE-metagenomic).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All unique materials are readily available from the authors.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study did not involve human research participants.