

The evolution of genetic drift over 50,000 generations

Joao A. Ascensao^{1,2}, QinQin Yu^{3,4}, and Oskar Hallatschek^{3,5,6,†}

¹Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA

²Present affiliation: Department of Organismal and Evolutionary Biology, Harvard University, Cambridge, MA, USA

³Department of Physics, University of California Berkeley, Berkeley, CA, USA

⁴Present affiliation: Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

⁵Department of Integrative Biology, University of California Berkeley, Berkeley, CA, USA

⁶Peter Debye Institute for Soft Matter Physics, Leipzig University, 04103 Leipzig, Germany

[†]Corresponding author: ohallats@berkeley.edu

ABSTRACT

Random variation in reproductive success—genetic drift—profoundly shapes genetic diversity and evolutionary trajectories. The strength of drift depends on the variance in descendant number, σ_d^2 , which governs key evolutionary outcomes: for instance, the establishment probability of a beneficial mutation scales inversely with σ_d^2 . However, whether σ_d^2 itself evolves over long timescales has remained unclear, because allele-frequency fluctuations depend on drift only through the effective population size, $N_e = N/\sigma_d^2$, which blends census population size with descendant-number variance. Here, we disentangle these components by using model-based Bayesian inference combined with joint tracking of (i) frequency fluctuations of neutrally barcoded lineages and (ii) census population sizes across growth cycles in the *E. coli* Long-Term Evolution Experiment. Analyzing 33 clones spanning the ancestor through 50,000 generations in two replicate populations (Ara-2 and Ara+2), we find that the strength of genetic drift evolved markedly—and divergently—between the two replicate populations. Both census size and σ_d^2 changed substantially through time, with most variation in N_e driven by shifts in σ_d^2 rather than census size. After approximately 2,000 generations, the σ_d^2 of the two populations diverged sharply: Ara+2 generally remained close to a bottleneck-only null expectation, whereas Ara-2 exhibited 1.5–5× stronger drift, consistent with an evolved increase in stochasticity during growth. Because establishment probability scales as s/σ_d^2 , a beneficial mutation of given effect is roughly twice as likely to establish in Ara+2 as in Ara-2. Our results demonstrate that the key parameter governing genetic drift can itself evolve, with direct consequences for adaptation.

1 Introduction

Genetic drift is a stochastic evolutionary force that arises from random variation in reproductive success^{1–3}. When drift is strong, it can erode neutral diversity⁴ and accelerate the accumulation of deleterious mutations through Muller’s ratchet^{5,6}. Drift also matters in settings where adaptation is common, because it controls the rate at which a newly arising beneficial mutation escape early stochastic loss and establish in the population^{2,7,8}.

For many evolutionary questions, drift is governed primarily by the variance σ_d^2 of the descendant-number distribution. This is evident from standard branching-process approximations for the fate of a weakly beneficial mutation: if a single mutant has mean descendant number $1 + s$ with $s \ll 1$, its establishment probability satisfies^{2,7,9} $p_{\text{est}} \approx 2s/\sigma_d^2$. Consequently, doubling σ_d^2 halves the chance that a mutation of given effect s establishes, implying that such mutations must arise about twice as frequently to maintain the same establishment rate. By contrast, for processes in which drift accumulates over many generations, higher moments of the descendant-number distribution typically contribute little. Quantifying drift therefore largely amounts to measuring the descendant-number variance.

Descendant-number stochasticity has several observable consequences, most directly fluctuations in allele frequencies. For a neutral allele at frequency f , the variance of the frequency change per generation is approximately $f(1-f)/N_e$, where the effective population size³ is $N_e = N/\sigma_d^2$. Frequency fluctuations therefore probe drift only through N_e , a compound parameter that entangles census size N with descendant-number variance σ_d^2 . The same fluctuation magnitude can thus arise from a small population with modest reproductive variance or from a large population with highly variable reproduction. Inferring σ_d^2 from frequency trajectories therefore requires independent information about N .

The concept of effective population size has a complex history in population genetics, with different con-

ceptualizations arising from different modeling assumptions and frameworks^{3,10–12}. In a strictly neutral setting over long timescales, N_e determines typical coalescence times (in units of generations) and hence scales with the pairwise diversity^{13,14}. However, in populations experiencing strong selection, patterns of genetic diversity reflect a combination of drift and “genetic draft”—the stochastic hitchhiking of alleles linked to selected variants^{15–17}. Under such conditions, coalescence times and pairwise diversity can have vastly different relationships to the census population size and forward-in-time descendant number variance^{17–19}. Indeed, it has long been noted that effective population size can have very little relation to the census population size^{20–23}. This has led to calls for caution in using diversity-based N_e estimates to infer population sizes or predict evolutionary dynamics, especially in adapting populations^{11,16}.

Here, rather than inferring genetic drift indirectly from patterns of sequence diversity, we directly measure the short-term descendant number variance by tracking frequency fluctuations of neutral genetic markers over short timescales. By observing how the frequencies of thousands of neutrally barcoded lineages fluctuate from one growth cycle to the next, and measuring census population sizes, we can estimate both σ_d^2 and N independently, without relying on assumptions about selection or coalescent structure.

We apply this approach to two replicate populations from the *Escherichia coli* Long-Term Evolution Experiment (LTEE), tracking how the strength of genetic drift has changed over 50,000 generations of evolution. The LTEE was initiated by dividing a single *E. coli* strain into twelve replicate cultures, which were then propagated in a constant environment through daily 1:100 dilutions²⁴. In this setting, the natural unit of time is the growth cycle, and we define σ_d^2 as the variance in descendant number from one dilution to the next for cells sampled immediately after a bottleneck. Random sampling at the bottleneck contributes a baseline variance of one to σ_d^2 ; any excess variance reflects additional stochasticity during growth and reduces the effective population size below the bottleneck size, N_b . Although the establishment probability of a new mutant depends on when during the cycle it arises because population size changes over time⁹, the rate of establishment is governed by the classical expression $2sN_e\mu$, where μ is the mutation rate and s is the per-cycle excess in relative growth (see Supplementary Note S2). We therefore quantify drift on a per-cycle basis.

We focus on two LTEE populations—Ara-2 and Ara+2—that represent distinct eco-evolutionary outcomes: Ara-2 diversified into two coexisting ecotypes around 6,500 generations and evolved hypermutability^{25–28}, while Ara+2 retained the ancestral mutation rate and did not form a stable ecological structure.

We isolated 33 clones from the two replicate populations of the *E. coli* Long-Term Evolution Experiment, spanning the ancestral strain through 50,000 generations of evolution. For each clone, we generated thousands to tens of thousands of neutrally barcoded variants and tracked their frequency fluctuations over several days of growth. Using a Bayesian inference framework applied to barcoded lineage trajectories, we find that both the census population size and descendant number variance change substantially through evolutionary time. Notably, while the census population sizes of the two lineages follow similar trajectories, their descendant number variances diverge markedly after approximately 2,000 generations. This divergence in σ_d^2 has direct consequences for adaptation: because establishment probability scales as s/σ_d^2 , a beneficial mutation of given effect is generally roughly twice as likely to establish in Ara+2 as in Ara-2. Our results demonstrate that the descendant number variance is a highly evolvable trait, with important implications for predicting evolutionary dynamics and interpreting patterns of genetic diversity.

Results

Measuring genetic drift parameters

To map out the evolution of genetic drift in the Ara-2 and Ara+2 lines, we first isolated sixteen clones from each line, from eight timepoints spread between two thousand and fifty thousand generations of evolution (Figure 1A). For Ara+2, we used the previously isolated²⁹ A and B clones from each time point; the labels were arbitrarily assigned and do not reflect phylogenetic structure. For Ara-2, we used the A and B clones from 2k generations. Around 6.5k generations, the Ara-2 population split into two coexisting lineages that have coevolved for the duration of the experiment—S and L. Thus, for all timepoints 6.5k generations and after, we used one S clone and one L clone. We also included an LTEE ancestor (REL606) in our study. We sequenced the genomes of the four clones that had not previously been sequenced. We see that both lines accumulate mutations over time relative to the REL606 ancestor; the Ara-2 lineage has a much faster mutation accumulation rate due to its hypermutator phenotype (Figure 1). As expected, we see that the number of mutations separating the S and L clones increases over evolutionary time, while the genetic distance between Ara+2 A and B clones does not systematically change.

We sought to measure the strength of genetic drift—captured by the values of parameters σ_d^2 , N_b , and N_e —for

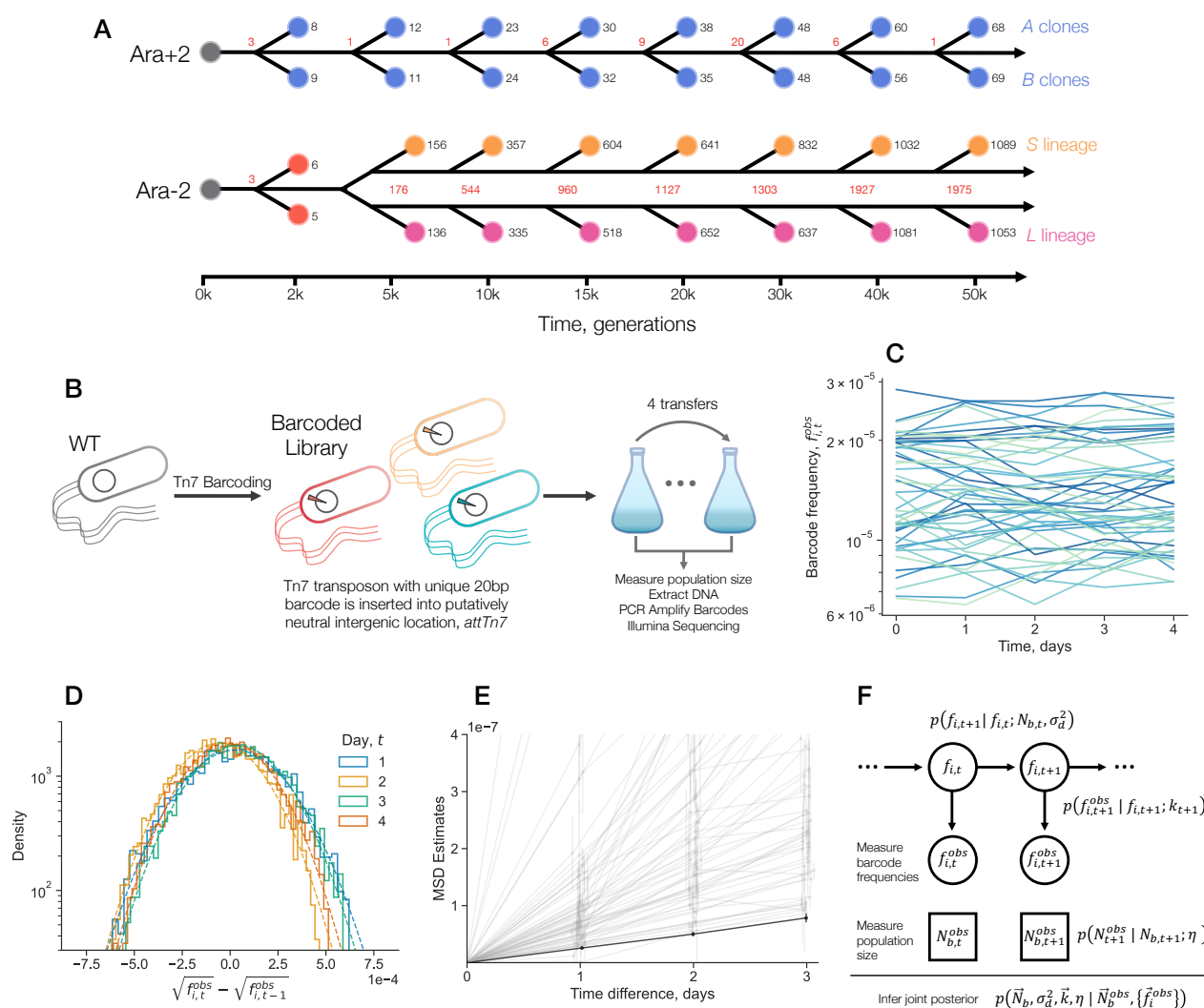


Figure 1. Measuring the strength of genetic drift over evolutionary time. (A) Schematic of the phylogenies and clone selection for the two LTEE lines examined in this study, Ara+2 and Ara-2. We sampled two clones from each timepoint in each line. The Ara-2 line diversified into the S and L lineages early in the experiment, while the Ara+2 line did not diversify into multiple long-lasting coexisting lineages. Black numbers represent the total number of mutations relative to the ancestor for each clone; red numbers represent number of different mutations between clones at the same time point. (B) We created many neutral variants of each LTEE clone by inserting random barcodes into a single intergenic genomic location with a Tn7 transposon system. We passaged the libraries in the LTEE evolutionary condition, taking samples every day. (C) Example barcode frequency trajectories, from the LTEE ancestor (REL606) library (replicate 1). (D) Transition densities of barcode frequencies for the REL606 library, replicate 1. Dotted lines represent Gaussian fits to the empirical densities. (E) Estimates of the mean squared displacements (MSD) of the square-root transformed barcode frequencies. The black line represents the MSD for the REL606 library, replicate 1; the grey lines are the MSDs for all other libraries. Error bars represent 95% confidence intervals. (F) Statistical model and inference strategy to infer genetic drift parameters. Unobserved, “true” barcode frequencies $f_{i,t}$ evolve according to a specified transition density. We additionally model the technical noise arising from measurement of the barcode frequencies and total population size. Given the data, we estimate a joint posterior of all parameters through an MCMC approach.

each of the 33 LTEE clones. Our approach to measure the three parameters is based on observing the frequency trajectories of many neutral variants of a clone, as the frequency fluctuations are related to the strength of genetic drift. For neutral lineages that are composed of sufficiently many individuals such that the central limit theorem applies, the lineage frequency at time $t + 1$, f_{t+1} , can be written as the following Markov chain,

$$f_{t+1}|f_t \sim \mathcal{N}\left(f_t, \frac{f_t(1-f_t)}{N_{e,t}}\right). \quad (1)$$

We measure lineage frequencies at the end of the 24-hour growth cycle, effectively coarse-graining over the within-cycle dynamics. Accordingly, each time step corresponds to one dilution-to-dilution interval, and σ_d^2 denotes the bottleneck-to-bottleneck variance in descendant number for cells sampled immediately after dilution. We only consider small lineage frequencies, $f_t \ll 1$, so that we can ignore the factor of $(1 - f_t)$ in the variance. We allow the effective population size to vary in time, as we have previously observed large abundance fluctuations in LTEE clones³⁰, so it is necessary to account for potentially changing population sizes. We assume that the descendant number variance stays approximately constant over the timecourse, so we can write the (potentially time-dependent) effective population size as,

$$N_{e,t} = \frac{N_{b,t}}{\sigma_d^2}. \quad (2)$$

Thus, we must measure the bottleneck population size in addition to the frequency trajectories to isolate the contribution of σ_d^2 to $N_{e,t}$.

To experimentally generate many neutral variants of each LTEE clone, we used a previously developed method³¹ where random 20bp barcodes are inserted into a putatively neutral genomic location, *attTn7*, by a Tn7 transposon system (Figure 1B). We aimed for approximately 1,000-10,000 barcodes per strain, which we found was a sufficient number to accurately resolve the magnitude of frequency fluctuations, while allowing us to sequence many experiments on a single Illumina lane. To measure the frequency of each neutrally barcoded variant, we amplified the barcode region via PCR (polymerase chain reaction), and used Illumina amplicon sequencing to count the number of times each barcode appears (Figure 1C). We propagated the Tn7 barcoded libraries for each clone in the LTEE evolutionary condition (without any other strains present in culture), sampling the population every day to (i) extract genomic DNA for later barcode sequencing and (ii) measure the bottleneck population size by counting colony forming units (CFUs). For each strain, we performed two biological replicate experiments, on different sets of days.

Following barcode sequencing, we quantified barcode reads and performed error correction on the sequences. Subsequently, we filtered out barcodes in each time course that appeared to have non-neutral fitness effects (relative to the other barcodes), putatively caused by mutations that arose on the background of the barcoded lineage. Generally, only a small percentage of barcodes displayed non-neutral trajectories. We simulated the data generating process and computational pipeline to ensure that the filtering process did not inadvertently cause biases or other pathologies to appear (Figures S7, S8).

As we are measuring barcode frequencies by counting sequencing reads, there must be measurement error in the frequencies that is at least as strong as Poisson noise; any additional noise in the sequencing process will cause overdispersion. But again, if the read counts are sufficiently large, then the central limit theorem applies, and we can model the measurement error as Gaussian,

$$f_t^{obs}|f_t \sim \mathcal{N}\left(f_t, \frac{k_t}{R_t}f_t\right), \quad (3)$$

where k_t is the measurement noise overdispersion, and R_t is the total number of sequencing reads. We use a simple square-root transformation on both the true and observed frequencies as a variance stabilizing transformation³²; i.e. the square-root transformed frequencies still have Gaussian densities, but the variance is no longer dependent on the mean. The observed transition densities result from a convolution between genetic drift and measurement noise. By employing technical replicates, we show that the measurement noise of square-root transformed frequencies is very well fit by a Gaussian (Figure S6). Together, this would imply that observed transition densities between the square-root transformed observed frequencies at neighboring times should follow a Gaussian distribution, providing a check on our model. Indeed, we see that the observed transition densities are very well fit by Gaussian densities, well into the tails of the distribution (Figure 1D).

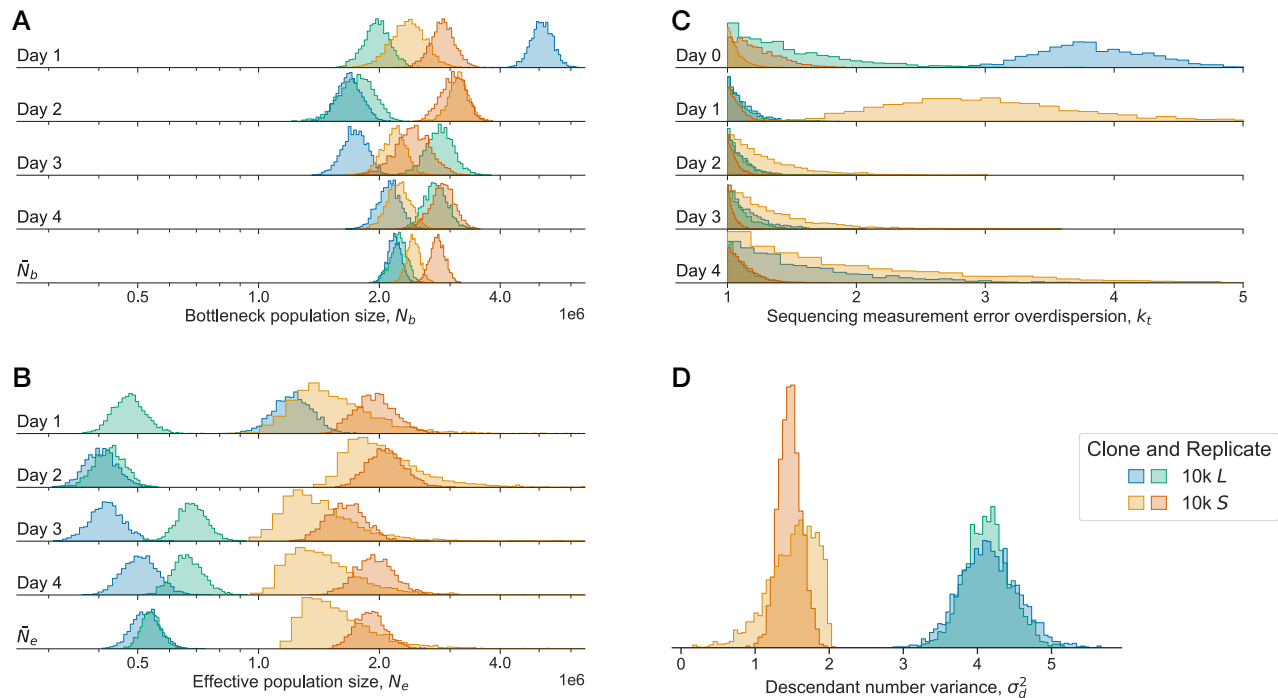


Figure 2. Examples of posterior densities. Data from the *S* and *L* clones sampled from 10k generations; different colors represent different libraries and biological replicates. Posterior estimates of the (A) bottleneck and (B) effective population sizes over the experimental time points. We also show the harmonic mean of the population sizes over the four experimental time points, $\bar{N}_b$ and $\bar{N}_e$. Posterior estimates for the (C) sequencing measurement error overdispersion and (D) descendant number variance.

Our model also predicts that variance in a barcode frequency trajectory should increase over time (Figure S1). Specifically, we expect that the mean squared displacement (MSD) between square-root transformed barcode frequencies should increase linearly over time, with a slope proportional to N_e^{-1} ,

$$\text{MSD}(\Delta t) = \langle (\sqrt{f_t} - \sqrt{f_{t-\Delta t}})^2 \rangle_{\Delta t} \quad (4)$$

$$= \frac{\Delta t}{4N_e}. \quad (5)$$

This simple relationship assumes that N_e is constant over time, and does not include the effects of measurement noise. However, we were able to use a simple procedure to account for measurement noise (see Methods), and obtained MSD estimates for the barcoded libraries (Figure 1E). We see that the MSD across strains and experiments generally increases approximately linearly over time (Figure 1E); slight deviations from linearity may be due to sampling noise or changes in N_e over time. This observation of accumulating variance serves as another validation of our model of genetic drift.

In order to more rigorously estimate the genetic drift parameters for each time series, we developed a Bayesian inference approach. Following Yu et al. (2024)³², we combine the likelihoods of equations 1 and 3, and integrate over the hidden “true” barcode frequencies to define a Kalman Filter (Hidden Markov Model) of the observed barcode frequencies (Figure 1F). We also define a likelihood to quantify the measurement error of the observed bottleneck population size, $N_{b,t}^{\text{obs}}$. We combine the likelihoods with non-informative/weakly informative priors, and use a Hamiltonian Monte Carlo (HMC) approach to sample from a joint posterior of all the parameters (see Methods).

The strength of genetic drift varies across evolutionary time

We obtained joint posterior estimates for all of the genetic drift and measurement noise parameters, across all experiments and biological replicates. We show representative examples of marginal posteriors in Figure 2. Here and throughout, $\bar{N}_b$ and $\bar{N}_e$ represent the harmonic mean over experimental timepoints of the bottleneck and

effective population size, respectively; the harmonic mean is the evolutionarily relevant parameter³³. We see that the bottleneck population size is generally not constant over time, or across biological replicates (Figure 2A), consistent with prior experiments that have observed large abundance fluctuations in LTEE populations³⁰. We see that the 10k *S* and *L* clones show consistently different effective population sizes and descendant number variances (Figures 2B,D and S2). Additionally, the sequencing measurement error overdispersion, k_t is oftentimes close to one—indicating noise dominated by Poissonian sampling—but the magnitude is sometime uncertain, and sometimes almost certainly greater than one (Figure 2C). This underlines the importance of allowing k_t to vary across samples, and not constraining the measurement noise to follow a Poisson distribution.

We then systematically compared the genetic drift parameters—the bottleneck and effective population size, and the descendant number variance—across biological replicates and strains (Figure 3). As the bottleneck and effective population sizes are allowed to vary over time, we focus on the harmonic mean across time points for both parameters. The correlation between biological replicates is generally quite good and greater than zero for all three parameters, ranging from about 0.65 to 0.9 (Figure 3A-C).

We combined the data of the two biological replicates for each strain to obtain final posterior estimates for all parameters. We see that the LTEE ancestor has a mean bottleneck population size around $4 \cdot 10^6$, which is the highest among all strains (Figure 3D-E). This is consistent with the fact that nearly all MSD lines are higher than the ancestor (Figure 1E). We observe that the mean bottleneck population size quickly starts to decrease as evolution proceeds, consistent with prior measurements³⁴. The rate of decrease is initially faster in Ara+2 compared to Ara-2—at 2,000 generations, the Ara-2 clones had mean bottleneck population sizes around $3.5 \cdot 10^6$, while the Ara+2 clones were around $2.5 \cdot 10^6$. After 2k generations, the bottleneck population size fluctuates around $1 - 2 \cdot 10^6$ for clones in both lines. The effective population size also varies widely over the course of evolution, but notably it has markedly distinct dynamics compared to the bottleneck population size (Figure 3F-G). This difference is caused by the complex changes in descendant number variance over evolutionary time.

The changes in the descendant number variance over time shows stark differences between the Ara+2 and Ara-2 lines (Figure 3H-I). For both lines, we see that the descendant number variance declines slightly between zero and two thousand generations. The LTEE ancestor has a descendant number variance around 1.3, indicating that while the primary source of variance is indeed the once-per-cycle bottlenecks, there is some additional unexplained variance, putatively caused by stochasticity in the growth process. All four clones at 2k generations have a descendant number variance indistinguishable from 1, potentially pointing towards parallelism in this initial evolution of genetic drift. After two thousand generations, the Ara+2 and Ara-2 lines diverge sharply. The majority of clones in Ara+2 show descendant number variances that hover around one, the exceptions being the clones at the 20k and 50k timepoints. All four clones in those timepoints display descendant number variances significantly above one. The 50k *B* clone shows especially outlying behavior, with a σ_d^2 about 12 fold larger than expected from bottlenecking alone; the excess strength of fluctuations can be visually seen in the barcode trajectories (Figure S3). In Ara-2, the descendant number variance sharply increases immediately following diversification into the *S* and *L* lineages. At 6.5k generations (immediately after diversification), both *S* and *L* show descendant number variances around 2.5, around twice as large as the ancestor. It may be that the initially similar descendant number variances of *S* and *L* is due to the fact that they have not yet had the chance to significantly genetically diverge. After 6.5k generations, the descendant number variances start to become more uncoupled from each other, varying significantly over evolutionary time, while almost always staying larger than one.

Quantifying systematic differences in the evolution of genetic drift between lineages

From the marginal genetic drift posteriors, we saw systematic differences in the evolution of genetic drift between Ara+2 and Ara-2. We sought to quantify the magnitude of such differences. In Bayesian inference, any inference question can be framed in terms of manipulations of posteriors, so we obtain posteriors for several summary statistics for each LTEE line (Figure 4).

As a simple first question, we wished to compare the typical bottleneck and effective population size across Ara-2 and Ara+2. To that end, we computed the posterior of the median $\bar{N}_b$ and $\bar{N}_e$ across all clones within a line (Figure 4A); we also separated clones from the *S* and *L* lineages. We see that typically, $\bar{N}_b$ is higher in Ara-2 clones compared to Ara+2 clones, by a factor of about 25%. However, once we separate the Ara-2 by ecotype lineage, we see that *L* clones actually typically have the lowest $\bar{N}_b$, while *S* clones have the highest. The pattern changes when we shift our attention to the effective population size. In contrast to the case with $\bar{N}_b$, the Ara+2 clones typically have larger $\bar{N}_e$ than all Ara-2 clones, as well as when compared to the *S* and *L* clones (when the lineages are separated). This discrepancy between the typical bottleneck and effective population sizes between the LTEE lines is due to the fact that Ara-2 clones typically have a larger descendant number variance compared to Ara+2

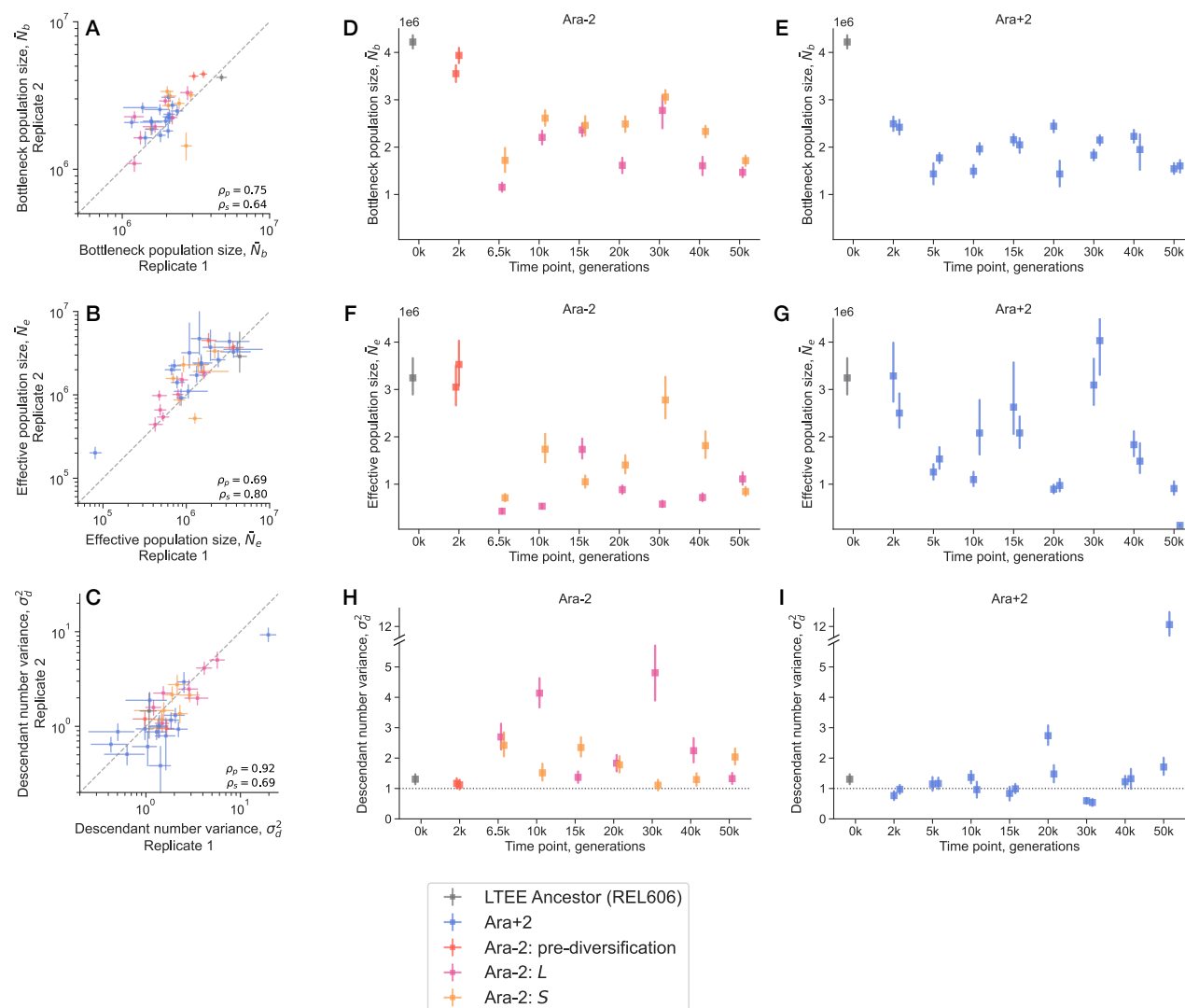


Figure 3. Estimates of genetic drift parameters over evolutionary time. (A-C) Correlations of inferred parameters between biological replicates; ρ_p and ρ_s represent Pearson's and Spearman's correlation coefficient, respectively. (D-I) Estimates of the inferred parameters, for each sampled clone, over evolutionary time. All point estimates represent posterior medians, and error bars represent 95% credible intervals.

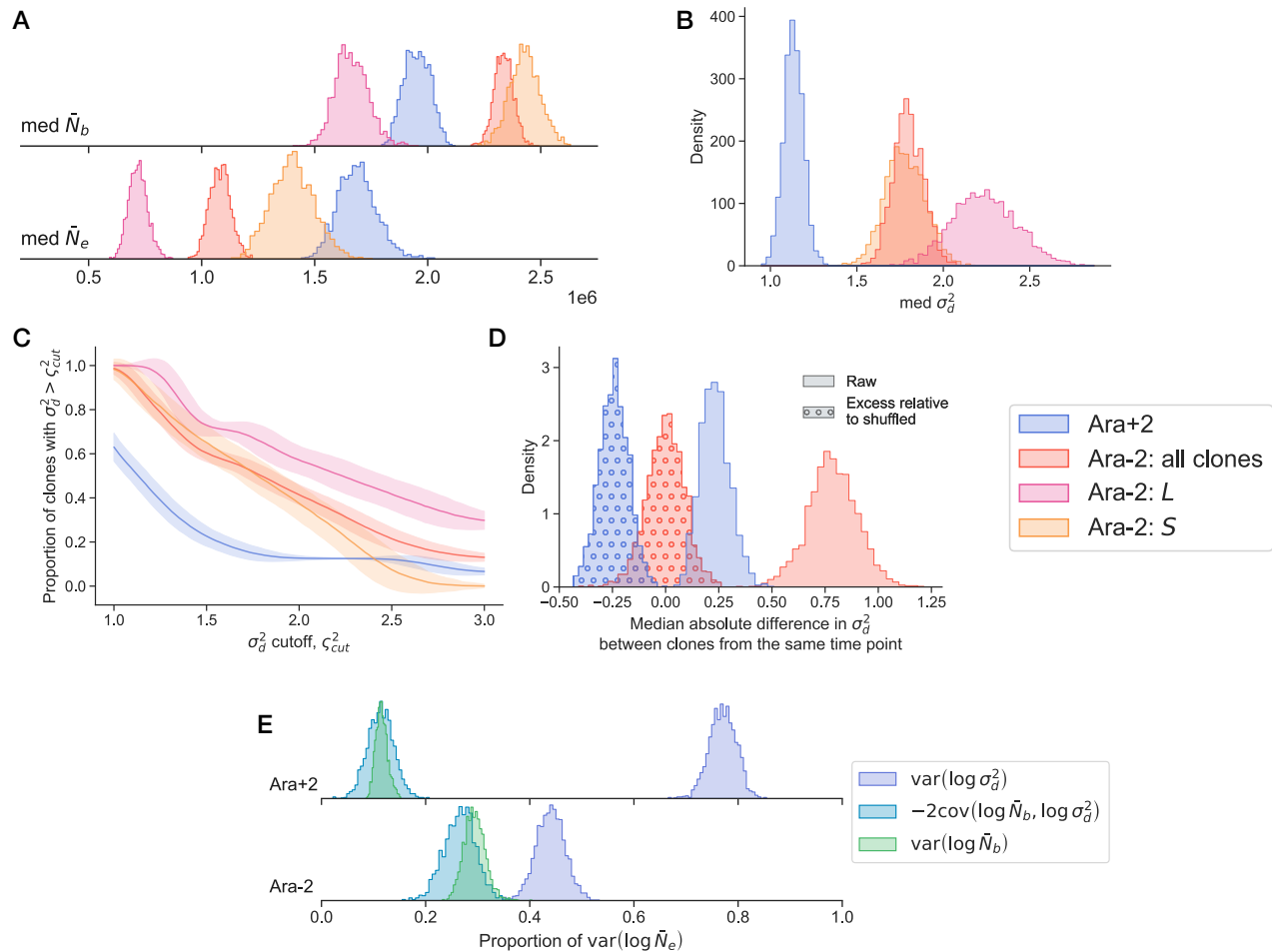


Figure 4. Patterns of the strength of genetic drift across lineages. (A-B) Posterior density estimates of the median (A) bottleneck and effective population sizes and (B) the descendant number variance, over all clones within a group. (C) We computed the proportion of clones within a group that have an inferred descendant number variance larger than a cutoff value. Error bars represent the posterior standard deviation. (D) For each pair of clones within a lineage (Ara+2 or Ara-2), we compute the absolute difference between their descendant number variances, and then compute the median across all clone pairs within the line (solid posterior densities). Then, we shuffled the identity of the clones, repeated the process, and subtracted the shuffled differences from the raw differences (circle hatched posterior densities). (E) Posterior densities of the contributions to the variance of log effective population sizes, across clones in a lineage.

clones (Figure 4B). We see that the typical σ_d^2 for Ara+2 clones is around 1, while it is above 1 for Ara-2 clones. In a slightly different way of quantifying differences in descendant number variance, we computed the proportion of clones with σ_d^2 larger than some cutoff value (Figure 4C). We see that nearly all Ara+2 clones have a descendant number variance lower than 1.5, while the Ara-2 clones span a larger range of values. Together, these observations underscore how differences in effective population size between populations can be decoupled from their relative census population sizes by factors of order 1.

We next sought to understand if clones from the same evolutionary time point displayed similar strengths of genetic drift. For each pair of clones (A and B) at the same evolutionary time point t , we computed the absolute difference in their inferred descendant number variance, and then took the median over all clone pairs (Figure 4D),

$$d_0 = \text{med}_t |\sigma_{d,A(t)}^2 - \sigma_{d,B(t)}^2|. \quad (6)$$

We see that the median difference in σ_d^2 is generally larger in the Ara-2 line, compared with the Ara+2 line. This could reflect that fact that Ara+2 clones at the same evolutionary time point are more genetically related to each other than Ara-2 clones at the same time point (Figure 1A). We were curious to see if clones at the same evolutionary time point were more or less similar to each other than the random expectation. To that end, we shuffled the identities of the A and B clones, applied equation 6, repeated this process one thousand times, and then subtracted the original difference from the averaged shuffled difference. We see that the excess difference in σ_d^2 relative to the shuffled data is centered around zero for Ara-2, but it is below zero for Ara+2. This indicates that Ara+2 clones at the same evolutionary time point are generally more similar to each other than would be expected if there was no time point stratification. On the other hand, we do not see the same similarity between Ara-2 clones at the same time point.

We observed that the effective population size varied strongly over evolutionary time, for both the Ara+2 and Ara-2 lines (Figure 3F-G). However, as $\bar{N}_e$ is a compound parameter of $\bar{N}_b$ and σ_d^2 , it is not immediately clear which constituent parameter causes most of the variance in $\bar{N}_e$ across clones. To assess the contribution of the two parameters to $\bar{N}_e$, the variance of $\log \bar{N}_e$ across all clones in a line can be decomposed into three components,

$$\text{var}(\log \bar{N}_e) = \text{var}(\log \bar{N}_b) + \text{var}(\log \sigma_d^2) - 2\text{cov}(\log \bar{N}_b, \log \sigma_d^2). \quad (7)$$

We thus compute the three components across all clones within a line, and normalize each component to $\text{var}(\log \bar{N}_e)$ (Figure 4E). We focused on $\log \bar{N}_e$ as it has a simple additive relationship to the bottleneck population size and descendant number variance. We see that variance in $\log \sigma_d^2$ has the largest contribution to variability in $\log \bar{N}_e$ across clones, for both lines. A portion of this pattern is a result of the outlying σ_d^2 of the Ara+2 50k B clone, and the relatively large $\bar{N}_b$ of REL606; however, even when we excluded either or both of those clones from our analysis, we saw the same qualitative pattern (Figure S4). This analysis supports the idea that most of the variability in $\bar{N}_e$ across evolutionary time arises from changes to the descendant number variance, not from changes to the census population size.

Discussion

In this work, we measured how genetic drift changed over 50,000 generations in two *E. coli* populations from the Long-Term Evolution Experiment (LTEE). By pairing neutral genetic barcodes with a Bayesian inference framework, we could disentangle the two components of the effective population size, N_e : the census population size (N_b) and the descendant-number variance (σ_d^2). Both N_b and N_e fluctuated substantially through time, with most of the variation in N_e driven by shifts in σ_d^2 . After about 2,000 generations, the populations began to diverge: their census sizes followed similar paths, but their descendant number variances—and thus the strength of genetic drift—separated markedly. These differences in descendant number variances have important evolutionary consequences—because the establishment probability of a beneficial mutant is $p_{\text{est}} \approx s/\sigma_d^2$, a mutant of a given fitness effect is roughly twice as likely to establish in Ara+2 as in Ara-2.

From a theoretical perspective, our work highlights the need to consider the potential for the descendant number variance to evolve over time. Most classical population genetic models assume that while the population size can freely vary, the descendant number variance remains constant^{2,37–39}. However, we have demonstrated that in real, evolving microbial populations, the descendant number variance is a highly evolvable trait. Models of evolutionary processes, such as adaptation, may need to account for this additional source of variation to accurately capture the dynamics of natural populations.

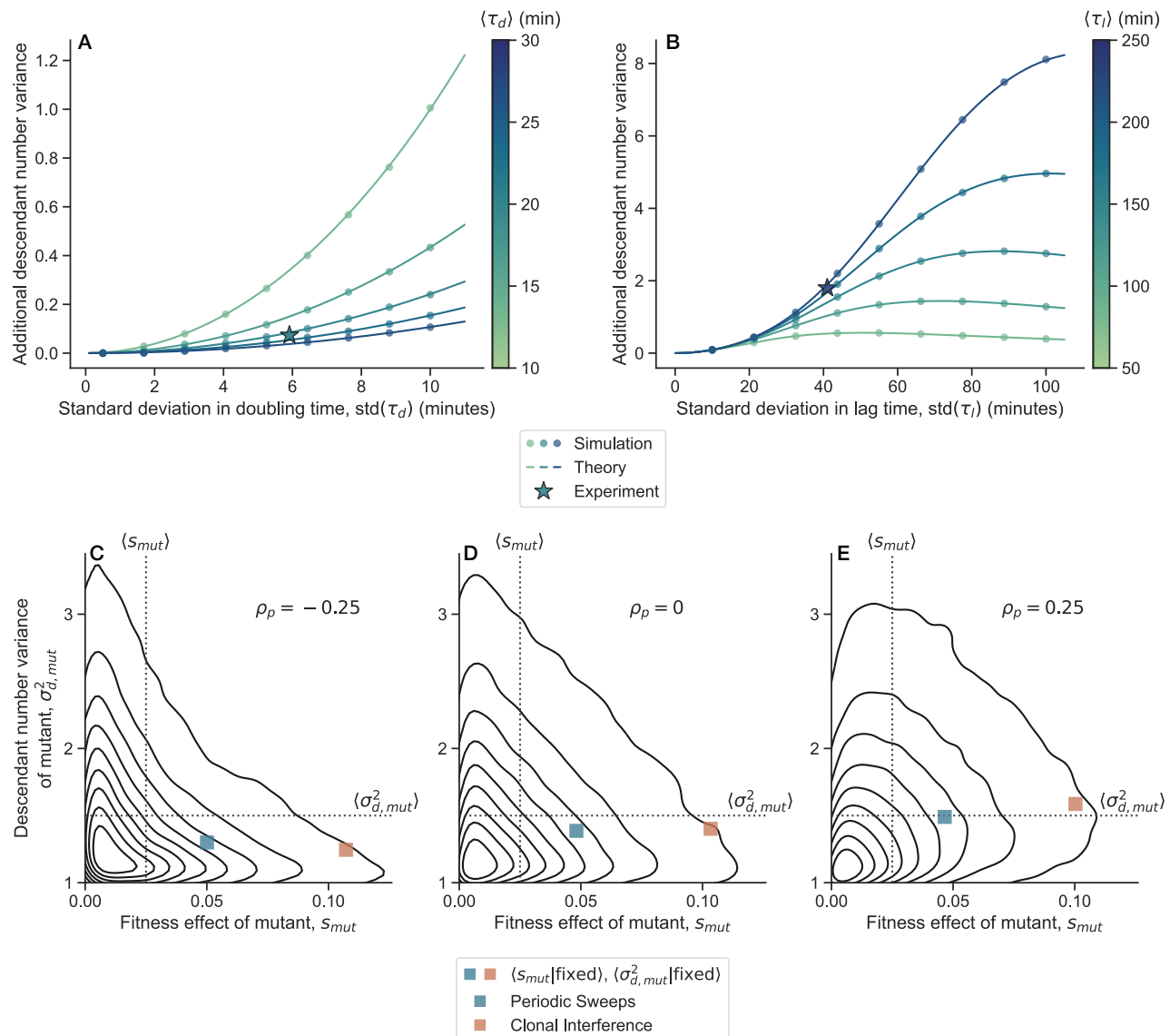


Figure 5. Possible physiological and evolutionary causes of changes to descendant number variance. (A-B) Additional contributions to the descendant number variance beyond that from bottlenecks can arise from noise in (A) the exponential doubling time, τ_d , or (B) the time in lag phase, τ_l . We model both τ_d and τ_l as gamma random variables, simulating the stochastic growth process (circle symbols) and obtaining analytical predictions (SI section S3). We use the theoretical results to predict the additional descendant number variance that would arise from experimentally measured distributions of doubling times³⁵ and lag times³⁶ in other strains of *E. coli* (star symbols). Note that the axes for each panel are on different scales. (C-E) We hypothesize that for any given genetic background, there is a joint distribution of mutant fitness effects and mutant descendant number variances, $p(s_{mut}, \sigma_{d,mut}^2)$ (black contour lines). Contour levels correspond to iso-proportions of the density. Dotted lines represent marginal means of the distribution. When an adaptive mutant fixes in the population, it can change both the fitness and descendant number variance. The probability of mutant fixation depends on both the mutant's fitness effect and descendant number variance. Evolutionary dynamics simulations demonstrate that on average, adaptation will generally bias fixed mutants towards larger fitness effects, and smaller descendant number variance (square symbols). However, the magnitude and direction of such biases is affected by both the evolutionary regime (periodic sweeps versus clonal interference) and the correlation structure of $p(s_{mut}, \sigma_{d,mut}^2)$. Here, ρ_p represents Pearson's correlation coefficient of $p(s_{mut}, \sigma_{d,mut}^2)$.

For empirical population geneticists, our results underscore the challenges in interpreting patterns of genomic variation. Many demographic inference methods aim to reconstruct the history of the effective population size from genomic data^{14,40}. However, as we have shown, changes in the effective population size can arise from changes in either the census population size or the descendant number variance (or both). Thus, inferred changes in effective population size over time cannot be assumed to reflect changes in the census population size alone. This may partially explain the common observation that the effective population size can have little relationship to the census population size^{16,20–22}.

Our work also has implications for the predictability of evolution. We observed that two populations evolved under the same conditions and from the same ancestor can have quite different evolutionary histories of effective population size changes. This was primarily due to divergence in the descendant number variance between the populations, rather than differences in census population sizes. As the strength of genetic drift can impact important evolutionary outcomes, such as the rate of adaptation^{21,41}, this suggests that even populations evolving under similar conditions may have distinct evolutionary fates. Predicting evolution may thus require understanding not only the population size, but also the distribution of descendant numbers.

The physiological cause of the changes in descendant number variance in the *E. coli* LTEE remains unclear. We hypothesize that noise in growth traits could give rise to additional effective descendant number variance. Previous studies have documented stochastic variability in both the exponential growth rate and lag time among sister microbes grown in the same environment^{35,36,42–44}. To investigate this, we used simulations and analytic theory to study asexual growth processes with noise in the doubling time and lag time, and connected such noise to effective additional descendant number variance (Figure 5A–B, SI section S3). Our findings suggest that both stochastic lag and doubling times can contribute to increases in descendant number variance. Using our theoretical results, we predicted the contributions to descendant number variance based on previously measured distributions of lag³⁶ and doubling times³⁵ in *E. coli* populations. The experimentally measured noise in lag time is predicted to cause a much larger increase in descendant number variance compared to the experimentally measured noise in doubling time.

What evolutionary forces caused the descendant number variance to change over evolutionary time in the *E. coli* LTEE? Theory predicts that differences in σ_d^2 between two otherwise neutral variants would cause quasi-selective effects that favor the variant with lower descendant number variance⁴⁵. However, this effect is likely weak, as the effective selective effect is proportional to $1/N$. In adapting populations, theory predicts that evolvability depends on the compound parameter U_b/σ_d^2 , so variants that decrease σ_d^2 can be favored for the same reason as mutation-rate modifiers—they generate beneficial mutations that establish more frequently and thus hitchhike with adaptation⁴⁶. Additionally, we hypothesize that changes in σ_d^2 can occur as a side effect of adaptation. We propose extending the distribution of fitness effects (DFE) of spontaneous mutations to a joint distribution of mutant fitness effects and descendant number variances, $p(s_{mut}, \sigma_{d,mut}^2)$ (Figure 5C–E). Changes in descendant number variance between genotypes could arise due to differences in growth trait distributions; even changes solely to the distribution mean can cause shifts in $\sigma_{d,mut}^2$ (Figure 5A–B). When an adaptive mutant fixes in the population, it can simultaneously alter both the mean fitness and descendant number variance. The probability of an individual mutant fixing depends on its fitness effect and descendant number variance relative to the rest of the population. We used evolutionary dynamics simulations to study the fixation of mutants drawn from $p(s_{mut}, \sigma_{d,mut}^2)$, which we modeled as a bivariate exponential distribution. On average, the fixation of adaptive mutants will be biased towards those conferring larger fitness gains and decreases in descendant number variance. However, the strength of this bias depends on the evolutionary regime (periodic selective sweeps versus clonal interference) and the correlation between fitness effects and descendant number variances among newly arising mutations. Selection will generally favor mutants with lower $\sigma_{d,mut}^2$, as the mutant establishment probability is approximately $p_{est} \approx s_{mut} / \sigma_{d,mut}^2$. However, if s_{mut} is positively correlated with $\sigma_{d,mut}^2$, on average, mutants with lower s_{mut} and higher $\sigma_{d,mut}^2$ will fix in the population (Figure 5E). The correlation structure of the joint distribution of mutant fitness effects and mutant descendant number variances, $p(s_{mut}, \sigma_{d,mut}^2)$, thus plays an important role in shaping the evolution of descendant number variance over time, along with the overall rate of adaptation.

In summary, tracking the strength of genetic drift in experimental microbial populations over 50,000 generations revealed that the descendant number variance is highly evolvable, with important consequences for evolutionary dynamics. To fully understand and predict evolution, both theoretical and empirical studies should consider this (potentially mutable) additional source of variation alongside classical parameters like the population size. We hope this work spurs further investigation into both the causes and consequences of evolutionary changes to the strength of genetic drift.

Methods

Construction of Tn7 barcoded LTEE-derived clones

To make barcoded libraries for each LTEE-derived clone, we first determined the optimal agar plating media for each one that we wanted to barcode. We tested 1% agar plates made with either DM2000 (Davis Minimal Media supplemented with 2000mg/L glucose) or LB (Lysogeny Broth). Davis Minimal Media base (DM) consists of 5.36g/L potassium phosphate (dibasic), 2g/L potassium phosphate (monobasic), 1g/L ammonium sulfate, 0.5g/L sodium citrate, 0.01% Magnesium sulfate, 0.0002% Thiamine HCl. We chose the plating media that supported the most robust growth (Table S1), while also avoiding using LB for *S* and *L* clones, which we previously found to cause issues⁴⁷. We generally used previously isolated and sequenced LTEE clones from Tenaillon et al. (2016)²⁹, except for the *S* and *L* clones from 6.5k, 10k, and 50k generations. We used the 6.5k clones isolated from a different set of experiments²⁷. For the 10k and 50k clones, we streaked whole population aliquots (REL4537 and REL11319, respectively) directly from glycerol stocks onto DM2000 plates. We then isolated several single colonies, resuspended them in DM2000 media, and used Sanger sequencing to sequence the *arcA* and *aspS* loci (Table S2), which each have a single nucleotide polymorphism that unambiguously distinguishes the *S* and *L* lineages. After confirming the lineage identity of the picked clones, we moved forward with a single *S* and *L* clone for each of the two populations.

We used a previously developed method³¹ to insert random DNA barcodes at the *attTn7* locus in the *E. coli* genome using a Tn7 transposon system. We used an older version of the Tn7 barcoding system that had a duplicated GFP-kanamycin resistance cassette. To make the barcoded Tn7 libraries for each LTEE-derived clone, we first streaked the LTEE strain for single colonies on the appropriate agar media, directly from glycerol stock. We also streaked the Tn7 “helper” donor strain (AMD1663) on LB/agar supplemented with 100µg/mL Carbenicillin and 0.3mM Diaminopimelic acid (DAP). The AMD1663 strain is an *E. coli* WM3064 donor with a pTNS2 plasmid (Addgene 64968), which encodes the Tn7 transposase genes⁴⁸. We then picked single colonies for the LTEE clone and AMD1663, and grew them overnight in DM2000 and LB + 100µg/mL Carbenicillin + 0.3mM DAP, respectively. Additionally, we thawed an entire 1mL glycerol stock aliquot of AMD1386, a WM3064 donor containing the barcoded Kan-Tn7 transposon plasmid library, and grew overnight in LB + 100µg/mL Carbenicillin + 0.3mM DAP + 50µg/mL Kanamycin. The next day, we washed the two donor cultures (to eliminate the antibiotics) by centrifugation for 3 minutes at 2500 xg, aspirating the supernatant, and resuspending in DM0 (Davis Minimal Media without a carbon source) three times. We then combined and plated approximately 1 OD·mL of the LTEE culture, AMD1663, and AMD1386 on either LB + 0.3mM DAP plates or DM2000 + 50mM casamino acids (ThermoFisher; cat. no. 223050) + 0.3mM DAP plates, as appropriate. We grew the agar plates overnight to allow the two donor strains to conjugate with the LTEE clones, and initiate transposition of the Tn7 barcoded transposon. The following day, we scraped up the lawn on the agar plates into approximately 1.5mL of DM0, resuspended, and washed the culture by centrifugation for 3 minutes at 2500xg, aspirating the supernatant, and resuspending in DM0 three times (to eliminate residual DAP). We added sterile glycerol to a final concentration of 20%, and then serially diluted the mixture and plated on DM2000/LB + 50µg/mL Kanamycin agar plates, growing the plates overnight at 37°C. We saved the mixtures as glycerol stocks in the -80°C freezer. The following day, we determined the approximate total number of successful transconjugants by counting colonies. We then thawed the glycerol stock from the day prior, diluted the stock in DM0 so that we would expect to get approximately 10,000 colonies (and thus uniquely barcoded transconjugants), and plated the diluted mixture on DM2000/LB + 50µg/mL Kanamycin agar plates, growing the plates overnight at 37°C. The following day, we scraped up the colonies on the plate, resuspended in DM0 + 20% glycerol, and distributed the resuspended final library into four 1mL glycerol stock aliquots, saving them in the -80°C freezer.

Genomic sequencing

The majority of genomes from LTEE clones used in this work have previously been sequenced—REL606⁴⁹, the majority of both Ara-2 and Ara+2 clones²⁹, and the 6.5k *S* and *L* clones²⁷. That left four clones that have not yet been sequenced—both the *S* and *L* clones from 10k and 50k generations. To sequence these clones, we grew them directly from glycerol stock in DM2000 media overnight. Then, we pelleted 2mL of the culture, and extracted genomic DNA with the DNeasy Blood and Tissue Kit (Qiagen 69504), according to the manufacturer’s protocol. Samples were prepared for whole genome sequencing using the Illumina DNA Prep tagmentation kit (cat. no 20018704) and IDT For Illumina Unique Dual Indexes (cat. no. 20027213-16). Sequencing occurred on the Illumina NextSeq2000 system using a 300 cycle flow cell kit, generating 2x150bp paired-end reads. To enhance base calling accuracy, 1-2% PhiX control was added to the sequencing run. Read demultiplexing, trimming, and analytical operations were conducted using DRAGEN v3.10.12, the integrated analysis software on the NextSeq2000. Reads were mapped to the REL606 genome⁴⁹ and mutations called with *breseq* (v0.38.3)⁵⁰, all with default settings.

BarSeq experiments

We performed two independent biological replicates of the BarSeq experiments for most strain libraries, done on different sets of days. For REL606, we performed three biological replicates. To start a BarSeq experiment, we first unfroze 1mL glycerol stock of the relevant transposon libraries and transferred the entirety to 10mL DM2000 media (supplemented with 50μL/mL Kanamycin) in a 50mL glass erlenmeyer flasks, which were grown for 16-24hrs at 37°C, shaken at 120rpm. Biological timecourse replicates using the same strain libraries were always performed on different sets of days (Table S1). All cultures for all experiments were grown with the same shaker, in the same 37°C warm room. The next day, we washed the cultures by pelleting via centrifugation for 3 minutes at 2500 xg, aspirating the supernatant, and resuspending in DM0 three times. After thoroughly vortexing the cultures, we transferred them 1:1000 to 20mL DM25 (Davis Minimal Media supplemented with 25mg/L glucose) in a 50mL glass erlenmeyer flask. Subsequently, every 24 hours, we transferred 200μL of culture into 20mL of fresh DM25 in a 50mL glass erlenmeyer flask. Again, the cultures were grown at 37°C, shaken at 120rpm. We did this for three cycles to ensure that the cultures have sufficiently physiologically adapted to the culture conditions. Then, we repeated the procedure for five more days (Days 0-4), where we collected samples for later sequencing and colony counting. Pellets were harvested every day by centrifugation at approximately 20,000xg for 10min, using all of the culture remaining after transferring. Subsequently, the pellets were stored at -20°C until the experiment was finished. To measure population size for each culture at each time point via colony forming units (CFUs), we generally diluted cultures twice at 1:100 in 1mL of DM0, then we plated 100μL of the final diluted culture on an agar plate. This process was repeated so that each culture had two independent “technical” plating replicates. We used either LB or DM2000 agar plates, again depending on which media allowed the strain in question to form the most robust colonies (Table S1).

After the completion of the experiment, pellets were retrieved from the -20°C freezer, and genomic DNA was extracted using the Qiagen DNeasy blood and tissue extraction kit (cat no. 69504), with elution in double-distilled water yielding approximately 50ng/μL of DNA. DNA Tn7 barcodes were amplified from genomic DNA samples using PCR with Q5 Hot Start Polymerase (NEB, cat. no. M0493S); the 50μL reactions included 5μL PCR primers, 5μL genomic DNA, 10μL 5x buffer, 10μL GC enhancer, 1μL dNTPs, 0.5μL Q5 polymerase, and 18.5μL water. Custom dual-indexed primers were utilized, incorporating binding sites adjacent to the barcode region, along with required Illumina read/index binding sites^{51,52}. The forward primer sequence was AATGAT ACGGCG ACCACC GAGATC TACACT CTTTCC CTACAC GACGCT CTTCCG ATCT N_nXXXXXX GTCGAC CTGCAG CGTACG, where X represents the custom forward 6bp index, and N_n indicates 1-4 random nucleotides, varying by primer pair; the reverse primer sequence was CAAGCA GAAGAC GGCATA CGAGAT XXXXXX GTGACT GGAGTT CAGACG TGTGCT CTTCCG ATCTGA TGTCCA CGAGGT CTCT, where X denotes the standard Illumina 6bp IT index. A distinct primer pair was used for each genomic DNA sample from different experiments, replicates, or time points to enable demultiplexing after sequencing. The PCR protocol involved 4 minutes at 95°C, followed by 25 cycles of [30 seconds at 95°C, 30 seconds at 55°C, 30 seconds at 72°C], and a final extension of 5 minutes at 72°C. The correct PCR products were confirmed through agar gel electrophoresis. All PCR reactions were subsequently pooled and purified using the Zymo DNA Clean and Concentrator kit (cat. no. D4013), eluted in double distilled water. The final pooled sample was sequenced on an Illumina HiSeq 4000 (50SR) at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley.

Barcode sequencing data preprocessing

To process the raw Illumina sequence reads for each sample, we first employed a previously described Perl script (MultiCodes.pl)^{51,52} to process the demultiplexed sequencing reads. This script extracts the barcode sequences by trimming the areas associated with the sequencing primers and the regions flanking the barcode. It also excludes reads that fail to align with the secondary sequencing index or that do not meet the requisite quality scores. Subsequently, we tally the counts of unique barcodes to compile a table linking each barcode sequence with its corresponding counts.

However, errors during PCR and sequencing can lead to mutations in some barcode reads, which can be incorrectly interpreted as distinct and real barcodes. Therefore, it is necessary to correct these sequencing errors by merging mutated barcodes with their original versions and consolidating their read counts. We do this by creating a graph where each node is a barcode, drawing an edge if the Levenshtein (edit) distance between the two barcodes is ≤ 3; to populate the nodes of the graph, we use the union of all barcodes from each time point in an experiment. We excluded barcodes with highly skewed GC contents—those lower than 20% or higher than 80%. The previously mentioned Perl script detects barcode pairs that differ by one mutation; we initialize the graph by drawing edges between these barcode pairs. For all remaining pairs of barcodes, to avoid excessively long run times, we first

computed their Hamming distance. We then only computed the Levenshtein distance between a barcode pair if the Hamming distance ≤ 10 . We then populate the edges of the barcode similarity graph using `networkx`⁵³. Finally, we pull out all connected components of the graph, and treat each connected component as a “superbarcode,” summing together the read counts for each time point from each constituent barcode. For all subsequent analyses, we proceed to use the “superbarcodes” read trajectories, and will generally refer to them simply as “barcodes.”

Outlier detection and barcode filtering

We observed some barcodes appear to have non-neutral fitness effects, likely due to spontaneously occurring selected mutations on the background of the barcoded lineage. We thus sought to detect and filter out barcodes with putatively non-neutral fitness effects, using a simple likelihood-based model. Using an approach similar to that presented in Ba et al. (2019)⁵⁴, we model sequencing read trajectories for barcode i , $r_{i,t}^{obs}$, as a negative binomial random variable parameterized such that the variance is proportional to the mean,

$$r_{i,t}^{obs} | r_{i,t-1}^{obs} \sim \text{NB} \left(r_{i,t-1}^{obs} \frac{R_t}{R_{t-1}} e^{(s_i - \bar{x}_t) \Delta t}, \kappa_t \right), \quad (8)$$

$$\langle r_{i,t}^{obs} | r_{i,t-1}^{obs} \rangle = r_{i,t-1}^{obs} \frac{R_t}{R_{t-1}} e^{(s_i - \bar{x}_t) \Delta t}, \quad (9)$$

$$\text{var} \left(r_{i,t}^{obs} | r_{i,t-1}^{obs} \right) = \kappa_t \langle r_{i,t}^{obs} | r_{i,t-1}^{obs} \rangle. \quad (10)$$

We chose to use a negative binomial density to more accurately capture the discreteness of sequencing reads, which is important for more accurately modeling barcodes with low read counts. Here, R_t is the total number of sequencing reads (across all barcodes) at timepoint t , κ_t is total overdispersion arising from a combination of sequencing measurement error and genetic drift, s_i is the fitness effect of barcode i , $\bar{x}_t$ is the average fitness of the population at time t , and $\Delta t = 6.64$ is the number of generations per cycle.

Before inferring the fitness effect for each barcode, we estimate the overdispersion parameters κ_t and the population-averaged fitness $\bar{x}_t$ from all barcodes in the experiment. We first roughly filter the barcodes so that their frequency is (i) large enough so that the central limit theorem applies to both read counts and the number of cells per barcode at the bottleneck, and (ii) is small enough that they are less likely to have a fitness effect (larger lineages appear more likely to be non-neutral). Thus, for inference of κ_t and $\bar{x}_t$ we chose to focus on barcodes with observed frequencies ($f_{i,t}^{obs} = r_{i,t}^{obs} / R_t$) at time $t - 1$ above $\max[30/R_{t-1}, 30/N_{b,t-1}^{obs}]$ and below $7 \cdot 10^{-4}$. Here, $N_{b,t}^{obs}$ is the bottleneck size estimated from the CFU data (averaged between the two technical replicates), $N_{b,t}^{obs} = \langle \text{cfus}_{j,t} \rangle_j / D_t$. The inferred overdispersion and mean fitness parameters are not sensitive to changes in the filtering cutoffs, and we use robust statistics to infer the two parameters. Using the remaining barcodes after filtering, we estimated the mean fitness for each timepoint as,

$$\hat{x}_t = \frac{1}{\Delta t} \text{med}_i \left(\log f_{i,t-1}^{obs} - \log f_{i,t}^{obs} \right). \quad (11)$$

We again used a robust estimation method for κ_t ⁵⁵. First, we used a square root variance-stabilizing transformation on the observed frequencies to eliminate the mean-variance dependence, and took the difference between the two neighboring time points, $\Delta \phi_{i,t}^{obs} = \sqrt{f_{i,t}^{obs}} - \sqrt{f_{i,t-1}^{obs}}$. We use the median absolute deviation to robustly estimate the variance and then scale appropriately to recover κ_t ,

$$\hat{\kappa}_t = 4R_t \cdot 2.1981 \cdot \left(\text{med}_i |\Delta \phi_{i,t}^{obs} - \text{med}_i \Delta \phi_{i,t}^{obs}| \right)^2. \quad (12)$$

We used standard bootstrapping to estimate the sampling distribution of both parameters. The relative parameter standard errors were generally $\lesssim 1\%$, so we proceeded with the parameter point estimates. We note that we only use these two inferred parameters for the purpose of outlier detection, and not for any other downstream application.

We can now proceed to infer fitness effects independently for each barcode via maximum likelihood. In order to capture both constant fitness effects and time-varying fitness effects (arising due to e.g. fluctuating selection, or multiple different selected lineages within a barcode), we obtain maximum likelihood estimates of s_i when all

timepoints are combined,

$$\mathcal{L}_i^{all}(s_i | \vec{r}_i) = \prod_{t=1}^T p(r_{i,t}^{obs} | r_{i,t-1}^{obs}, s_i, \hat{\kappa}_t, \hat{x}_t), \quad (13)$$

$$\hat{s}_i^{all} = \underset{s_i}{\operatorname{argmax}} \log \mathcal{L}_i^{all}(s_i | \vec{r}_i), \quad (14)$$

along with the estimate for each time point pair,

$$\mathcal{L}_i^t(s_i | \vec{r}_i) = p(r_{i,t}^{obs} | r_{i,t-1}^{obs}, s_i, \hat{\kappa}_t, \hat{x}_t), \quad (15)$$

$$\hat{s}_i^t = \underset{s_i}{\operatorname{argmax}} \log \mathcal{L}_i^t(s_i | \vec{r}_i), \quad (16)$$

in both cases using the density from equation 8. We then compute a p-value for the fitness effect of each barcode under the null hypothesis $s_i = 0$, for all estimates, $\hat{s}_i^t$ and $\hat{s}_i^{all}$. For ease of computation and generality we compute the p-value as the posterior probability that the likelihood ratio between null and alternative hypotheses is greater than 1, i.e. the probability that the data more strongly support the null hypothesis over the alternative⁵⁵,

$$p_i = P_{s_i | \vec{r}_i} \left(\frac{\mathcal{L}(0 | \vec{r}_i)}{\mathcal{L}(s_i | \vec{r}_i)} > 1 \right), \quad (17)$$

$$P(s_i | \vec{r}_i) \propto \mathcal{L}(s_i | \vec{r}_i). \quad (18)$$

This convenient definition has been shown to be equivalent to the frequentist definition of the p-value using a likelihood ratio test statistic (if the distribution is invariant under transformation)^{56,57}, and does not require asymptotic approximations. We computed separate p-values using $\mathcal{L}_i^{all}(s_i | \vec{r}_i)$ and all $\mathcal{L}_i^t(s_i | \vec{r}_i)$ for the likelihoods. In practice, this p-value can be calculated by first, finely discretizing the likelihood curve along s_i and normalizing it to get the posterior,

$$P_{i,j}(s_{i,j} | \vec{r}_i) = \frac{\mathcal{L}(s_{i,j} | \vec{r}_i)}{\sum_j \mathcal{L}(s_{i,j} | \vec{r}_i)}. \quad (19)$$

Then, calculating the log-likelihood ratio along all discretized s values,

$$\text{LLR}_{i,j} = \log \mathcal{L}(0 | \vec{r}_i) - \log \mathcal{L}(s_{i,j} | \vec{r}_i). \quad (20)$$

And finally, summing to get the posterior probability that the data supports the null hypothesis more than the alternative, where $I[\cdot]$ is the indicator function,

$$p_i = \sum_j I[\text{LLR}_{i,j} > 0] P_{i,j}(s_{i,j} | \vec{r}_i). \quad (21)$$

We then want to exclude barcodes that have likely non-neutral fitness effects from further analysis. We first performed an initial (liberal) barcode filtering where we filter all barcodes where the observed frequency at day 0 is above the 50th percentile, and below the 10th percentile; we did this because we had observed that high-frequency barcodes are more likely to have non-neutral fitness effects, and low-frequency barcodes are more likely to be sequencing artifacts. We then choose to filter out barcode i if it had any p-values (p_i^{all} and p_i^t for all t) less than 0.05. This p-value cutoff appears to perform sufficiently well in simulations of the outlier detection and genetic drift inference pipeline (Figures S7-S9).

Following filtering, we sum the remaining low-frequency barcodes together so that each (combined) barcode is always at sufficiently high frequency such that the central limit theorem applies to the trajectory statistics. We do this by first computing a time-dependent cutoff frequency, $\text{cutoff}_t = \max[50/R_t, 50/N_{b,t}^{obs}]$. For each time point, we sum the lowest barcode below the cutoff with the trajectory of the second-lowest barcode. We iteratively continue this process until no barcodes remain under any of the cutoffs.

Estimates of mean squared displacements

We expect that the mean squared displacement (MSD) of the square-root transformed barcode frequency trajectories (Figure 1E) should increase linearly over time increments with a slope determined by N_e (assuming that N_e does not change over time),

$$\text{MSD}(\Delta t) = \langle (\sqrt{f_{i,t}} - \sqrt{f_{i,t-\Delta t}})^2 \rangle \quad (22)$$

$$= \frac{\Delta t}{4N_e}. \quad (23)$$

We estimated the MSD for all pairs of time points that were either 1, 2, or 3 time increments (Δt) apart, using barcodes that passed filtering (from the prior section). We require a method to distinguish between measurement error (which is uncorrelated in time) and genetic drift (which accumulates over time). We applied a method previously employed^{30,55} to differentiate between these two types of noise. In brief, the variance between two time points represents the sum of the decoupling noise and the measurement noise (ζ_t) at those time points,

$$\kappa_{t,t-\Delta t} = \frac{\Delta t}{4N_e} + \zeta_t + \zeta_{t-\Delta t}. \quad (24)$$

For each barcode, we calculate the difference of square-root transformed frequencies between considered pairs of time points. We then obtain a robust estimate of $\hat{\kappa}_{t,t-\Delta t}$ by using the median absolute deviation (equation 12), across all barcodes. We estimate the standard error, $\text{std}, \hat{\kappa}_{t,t-\Delta t}$, through standard bootstrapping techniques. Using the relationship outlined in equation 24 between $\kappa_{t,t-\Delta t}$ and the noise parameters, we can determine the noise parameters based on our measured values of $\hat{\kappa}_{t,t-\Delta t}$. This is achieved by numerically minimizing the weighted squared differences, expressed as $\sum [(\kappa_{t,t-\Delta t} - \hat{\kappa}_{t,t-\Delta t}) / \text{std}, \hat{\kappa}_{t,t-\Delta t}]^2$. Our final estimate of the MSD is obtained by adjusting each $\hat{\kappa}_{t,t-\Delta t}$ for the measurement noise parameters, and then averaging the adjusted values across all instances with the same time increment Δt ,

$$\widehat{\text{MSD}}(\Delta t) = \langle \hat{\kappa}_{t,t-\Delta t} - \hat{\zeta}_t - \hat{\zeta}_{t-\Delta t} \rangle_{\Delta t}. \quad (25)$$

We obtain MSD confidence intervals via bootstrapping.

Bayesian inference of genetic drift parameters

Following Yu et al. (2024)³², we model the frequency trajectories of (post-filtering) lineages within a population using a Cannings-type model. If the number of individuals within lineage i is sufficiently large such that the central limit theorem applies, and each individual within the lineage behaves identically and independently, we can model the lineage frequency dynamics, $f_{i,t}$, as the following Markov chain,

$$f_{i,t+1} | f_{i,t} \sim \mathcal{N} \left(e^{(s_i - \bar{x}_t)\Delta t} f_{i,t}, \frac{f_{i,t}(1 - f_{i,t})}{N_{e,t}} \right). \quad (26)$$

Here, s_i is the relative fitness effect of barcode i , $\bar{x}_t$ is the average fitness of the population at time t , $\Delta t \approx 6.64$ is the number of generations per cycle, and $N_{e,t}$ is the effective population size at time t . The effective population size can be decomposed into two components—the bottleneck population size, N_b , and the descendant number variance, σ_d^2 . We assume throughout that the bottleneck population size can vary over time, but the descendant number variance is constant, i.e. $N_{e,t} = N_b / \sigma_d^2$. We further simplify the model by restricting ourselves to small lineage frequencies, $f_{i,t} \ll 1$, such that we can ignore the $1 - f_{i,t}$ nonlinearity, and only considering neutral lineages, $s_i = 0$,

$$f_{i,t+1} | f_{i,t} \sim \mathcal{N} \left(e^{-\bar{x}_t \Delta t} f_{i,t}, \frac{f_{i,t}}{N_{e,t}} \right). \quad (27)$$

In our experiments, we never directly observe the frequency of a lineage; we indirectly measure the frequency by sequencing the barcode regions and counting the barcodes. Thus, the measurement process will introduce additional noise that will be at least as strong as Poisson sampling noise. If (i) the observed number of sequencing reads, $r_{i,t}^{\text{obs}}$, is sufficiently large such that the central limit theorem applies, (ii) the measurement process is unbiased,

and (iii) measurement noise affects every barcoded cell/sequencing read independently and identically, we can model measurement as the following,

$$r_{i,t}^{obs} | f_{i,t} \sim \mathcal{N}(f_{i,t} R_t, k_t f_{i,t} R_t) . \quad (28)$$

Here, R_t is the total number of sequencing reads (across all lineages) at timepoint t , and k_t is sequencing measurement error overdispersion. This mean-variance relationship was confirmed by analyzing the noise in BarSeq technical replicates from a previous experiment⁵⁵ with the exact same sample processing procedure that we use here (Figure S6). The measurement error is solely driven by poisson sampling noise when $k_t = 1$; any other sources of noise (e.g. associated with gDNA extraction or the amplicon PCRs) will increase k_t above one. We can rewrite equation 28 in terms of the observed frequency, $f_{i,t}^{obs} = r_{i,t}^{obs} / R_t$,

$$f_{i,t}^{obs} | f_{i,t} \sim \mathcal{N}\left(f_{i,t}, \frac{k_t}{R_t} f_{i,t}\right) . \quad (29)$$

Together, equations 27 and 29 define a hidden Markov model, which can be used to infer parameters from collected data. To ease further analysis, we apply a square-root variance stabilizing transformation on both equations, where we define $\phi_{i,t} = \sqrt{f_{i,t}}$,

$$\phi_{i,t+1} | \phi_{i,t} \sim \mathcal{N}(a_t \phi_{i,t}, \gamma_t) , \quad (30)$$

$$\phi_{i,t}^{obs} | \phi_{i,t} \sim \mathcal{N}(\phi_{i,t}, \zeta_t) . \quad (31)$$

We expect this variance stabilizing transformation to be valid as long as $\phi_{i,t}$ is sufficiently large. For notational simplicity, we defined the following compound parameters,

$$a_t = e^{-\bar{x}_t \Delta t / 2} , \quad (32)$$

$$\gamma_t = \frac{\sigma_d^2}{4a_t^2 N_{b,t}} , \quad (33)$$

$$\zeta_t = \frac{k_t}{4R_t} . \quad (34)$$

With the mean-variance dependency extinguished, we can treat the system as a standard Kalman filter, integrating out the unobserved $\phi_{i,t}$ states, and obtaining a recursively-defined likelihood,

$$p(\vec{\phi}_i^{obs}) = \prod_{t=1}^T \psi(\phi_{i,t}^{obs}; \mu_{i,t}, \zeta_t + v_{i,t}) , \quad (35)$$

$$\mu_{i,1} = a_t \phi_{i,0}^{obs} , \quad (36)$$

$$v_{i,1} = a_t^2 \zeta_0 + \gamma_0 , \quad (37)$$

$$\mu_{i,t+1} = a_t \frac{\phi_{i,t}^{obs} v_{i,t} + \mu_{i,t} \zeta_t}{v_{i,t} + \zeta_t} , \quad (38)$$

$$v_{i,t+1} = \frac{a_t^2}{v_{i,t}^{-1} + \zeta_t^{-1}} + \gamma_t . \quad (39)$$

We assumed that we obtained measurements of $\phi_{i,t}^{obs}$ at every time point from time 0 to T . We defined $\psi(x; \mu, v)$ as a Gaussian density with mean μ and variance v .

In order to separate the contribution of $N_{b,t}$ and σ_d^2 to $N_{e,t}$, we would need measurements of how the bottleneck population size $N_{b,t}$ changes over time. To that end, every day in the timecourses from days 0 to 3 we plated the end-of-cycle cultures on agar plates at a defined dilution rate, with two technical replicates for each culture. Once colonies grew, we counted them and used the colony forming units (CFUs) as a measure of the bottleneck population size. We then modeled the CFUs as a negative binomial random variable, parameterized such that the variance is proportional to the mean,

$$cfus_{j,t} \sim \text{NB}(N_{b,t} D_t, \eta_{cfus}) , \quad (40)$$

Here $j \in \{1, 2\}$ represents the technical replicate. Here D_t is the total dilution rate and η_{cfus} is the measurement error overdispersion. Again, when $\eta_{cfus} = 1$, the noise reduces to poisson sampling error. We did not obtain a technical replicate for certain time points/cultures (due to e.g. plate contamination), in which case we only used the CFU measurement from the single replicate. We experimentally verified the proposed mean-variance relationship for CFUs, and obtained a preliminary estimate of η_{cfus} ; we grew three independent cultures of REL606 at different cell densities, then we diluted and plated 16 independent technical replicates for each culture. We used the same dilution and plating protocol as the main experiments. We then computed the CFU mean and variance across technical replicates for each culture. We observed a linear mean-variance relationship for the CFUs, as expected, with a slope greater than one (Figure S5A). We sought to incorporate this experimental data into our model. Thus, we used the Hamiltonian Monte Carlo (HMC) algorithm implemented in STAN⁵⁸ to jointly estimate the posterior of each parameter in the following model,

$$cfus_{l,m} \sim \text{NB}(\mu_m, \eta_{cfus}). \quad (41)$$

Here, l is the label for each CFU measurement, $m \in \{1, 2, 3\}$ is the label for each independent culture at a different density, and μ_m is the underlying “true” culture density for each group. We used a uniform prior for each μ_m and a Pareto prior for the overdispersion parameter $p(\eta_{cfus}) \propto \eta_{cfus}^{-2}$. We found that the marginal posterior of η_{cfus} was well fit by a gamma distribution. We thus fit the posterior estimate to a gamma distribution via maximum likelihood, and obtained estimates of the gamma distribution parameters (Figure S5B),

$$(\eta_{cfus} - 1) \sim \Gamma(\hat{\alpha}, \hat{\beta}). \quad (42)$$

We can now use equations 35-40 as the likelihood for our model describing lineage frequency trajectories in our timecourse experiments. We combine the likelihoods from all barcodes in the experiment, and raise each likelihood to the power of ξ (“power posterior”), $p(\{\vec{\phi}_i^{obs}\}) = \prod_i^M p(\vec{\phi}_i^{obs})^\xi$. We use the power posterior on the suggestion of Miller and Dunson (2019)⁵⁹ to help enable robustness of the inference results against small deviations from the model. Here, $\xi = [1 + M/\alpha]^{-1}$, where we chose $\alpha = 500$, as simulations indicate that it is the largest (i.e. most conservative) value that still has a (modest) regularizing effect. To perform inference on the model, we again use Pareto priors for the relevant overdispersion parameters of the model, $p(k_t) \propto k_t^{-2}$ and $p(\sigma_d^2) \propto (\sigma_d^2)^{-2}$. We use a weakly regularizing prior for $\bar{x}_t \sim \mathcal{N}(0, 0.1^2)$, as relative fitness effects generally do not exceed 0.1/generation in LTEE strains and environment^{55,60}. We use equation 42 as the prior for η_{cfus} (with the maximum likelihood-fit parameters), allowing us to incorporate the information from the CFU technical replicates experiment, while also allowing the posterior of η_{cfus} to change based on the data for each timecourse. And we use uniform priors for all $N_{b,t}$. We then used the priors and likelihoods along with the HMC algorithm implemented in STAN to jointly estimate the posterior of each parameter. We used the same inference procedure to obtain final parameter estimates, where the data from the biological replicates is pooled together; all parameters except σ_d^2 and η_{cfus} were allowed to vary independently for data from each replicate. All HMC chains were evaluated for convergence by ensuring that the $\hat{R}$ statistic was between 1 and 1.01, and through visual inspection.

We performed all of our BarSeq experiments in a culture volume two times larger than the LTEE cultures (20mL versus 10mL), but retaining the same dilution rate and all other culture conditions. This was primarily to be able to obtain a sufficiently large sample for genomic DNA extraction. We thus divided our estimates of effective population size and bottleneck population size by a factor of two, to enable more direct comparison with the standard LTEE culture environment.

Simulations of the data generating process and inference pipeline

We ran simulations of the data generating process to validate our inference pipeline and probe the effects of the robustifying measures. We directly modeled how the number of cells n_t in a lineage changes over time, where the markov transition density is a negative binomial with the variance proportional to the mean,

$$n_t | n_{t-1} \sim \text{NB}\left(n_{t-1}e^{(s-\bar{x}_t)\Delta t}, \sigma_d^2\right). \quad (43)$$

We always initialize the number of cells with $n_0 = 500$, and use $\bar{x}_t = 0$ and $\Delta t = 6.64$ generations. We simulate the number of sequencing reads as a negative binomial random variable, again where the variance is proportional to the mean,

$$r_t^{obs} | n_t \sim \text{NB}\left(R_t \frac{n_t}{N_{b,t}}, k_t\right). \quad (44)$$

The observed frequencies are then $f_t^{obs} = r_t^{obs} / R_t$. In our simulations, we always use $N_{b,t} = 5 \cdot 10^6$, $R_t = 5 \cdot 10^6$, $\bar{x}_t = 0$, and $k_t = 2$, which are typical parameters for our experimental data. We varied the descendant number variance, $\sigma_d^2 \in \{1, 3, 5\}$. For each simulated dataset, we independently simulated 1000 barcodes. Each barcode has a 20% chance of having a non-neutral (non-zero) fitness effect; this percentage of selected barcodes is likely high relative to our experimental data, but we were interested in exploring how robust our pipeline is, even in the presence of a large amount of “contaminating” data. If the barcode is randomly selected to have a non-zero fitness effect, there is a 50% chance it will have a negative (deleterious) fitness effect, and a 50% chance it will have positive (beneficial) fitness effect. Beneficial and deleterious fitness effects, respectively, are drawn from the following two exponential distributions,

$$s_{bene} \sim \exp(\bar{s}_{bene}) , \quad (45)$$

$$-s_{del} \sim \exp(\bar{s}_{del}) . \quad (46)$$

We chose exponential distributions because previously measured distributions of fitness effects (DFEs) for LTEE strains can be well approximated by exponential distributions⁵⁵. Ascensao et al. (2023) found that the deleterious DFEs of REL606, 6.5k S and 6.5k L in monoculture (the same condition studied in this work) were all approximately this same distribution, with $\bar{s}_{del} \approx 0.015$. By contrast, the beneficial DFE in REL606 was higher, with $\bar{s}_{bene} \approx 0.025$, and the beneficial DFE of 6.5k S and L lower, with $\bar{s}_{bene} \approx 0.008$. Thus, for all simulations, to accurately reflect the spectrum of mutations that the LTEE populations likely have available, we always used $\bar{s}_{del} = 0.015$, and used either $\bar{s}_{bene} = 0.025$ or $\bar{s}_{bene} = 0.008$.

For each simulation run, we simulated CFUs as a negative binomial with the variance proportional to the mean, with two “technical replicates” per time point,

$$cfus_{j,t} \sim \text{NB}(N_{b,t} D_t, \eta_{cfus}) , \quad (47)$$

$$\eta_{cfus} = 2 . \quad (48)$$

To summarize, we simulated 100 independent “experiments” for each of six distinct parameter sets, where we varied $\sigma_d^2 \in \{1, 3, 5\}$ and $\bar{s}_{bene} \in \{0.008, 0.025\}$. For each experiment, we obtained simulated CFU data. Within each experiment, we simulated 1000 independent barcode sequencing read trajectories, 20% of which (on average) have non-neutral fitness effects.

Once we obtained all simulated datasets of observed frequency trajectories and CFUs, we pushed the data through the fitness inference/outlier detection pipeline and filtered the barcodes appropriately, as in the real experiments. The outlier detection method appears to be able to detect non-neutral lineages sufficiently well; we chose a p-value cutoff of 0.05 that balances a low false positive rate with a high enough true positive rate (Figure S9). Then we used the previously introduced Bayesian inference method to jointly infer the genetic drift parameters, i.e. $N_{b,t}$ and σ_d^2 , along with all the nuisance parameters, i.e. k_t , η_{cfus} , and $\bar{x}_t$. Additionally, we ran the Bayesian inference in the case where we did not use either or both of the robustifying methods, i.e. outlier detection or the power posterior (Figures S7-S8).

Simulations of noisy growth processes

We simulate noisy asexual splitting growth processes to compare to our analytical results (Figure 5A-B; SI section S3). To simulate variable lag times, we sampled about 10^7 independent instances of t_{lag} from either a gamma (Figure 5B) or exponential (Figure 5B) distribution, for a given set of parameters. After the lag time, each individual starts growing exponentially (deterministically), such that $N = e^{r(T-t_{lag})}$. We use $T = 12$ hrs and $r = 0.03 \text{ min}^{-1}$, which is a typical growth rate for *E. coli*^{35,42}. Thereafter, we compute the descendant number variance by appropriately averaging over the simulated dataset,

$$\sigma_d^2 = \frac{\text{var } N}{\langle N \rangle^2} . \quad (49)$$

To simulate noise in doubling times, we simulated 10^5 independent lineages, each starting from a single individual. We draw the doubling time from either a gamma (Figure 5A) or exponential (Figure 5A) distribution. At doubling time, the individual splits into two new individuals, and new random doubling times are drawn for each of the new individuals. This process is continued until time T , which we take as 6.64 generations. The simulated distribution of the total number of individuals per lineage is then used to compute the descendant number variance, by again applying equation 49.

Evolutionary dynamics simulations

We simulated the dynamics of the first mutational step, when mutations can be drawn from a joint distribution of fitness effects and descendant number variances, $p(s_{mut}, \sigma_{d,mut}^2)$ (Figure 5C-E). We took the joint distribution to be Gumbel's bivariate exponential distribution⁶¹, where the density of the distribution is,

$$p(x, y) = e^{-x-y} [1 + \alpha (2e^{-x} - 1) (2e^{-y} - 1)] . \quad (50)$$

Under this joint distribution, both x and y have exponential marginal distributions with mean of one. Here, α is a parameter that controls the strength of correlation between x and y , which is directly related to Pearson's correlation coefficient, $\rho_p = \alpha/4$. We vary $\alpha \in \{-1, 0, 1\}$. After sampling from equation 50, we scale x and y to their desired means and locations— $s_{mut} = 0.025x$ and $\sigma_{d,mut}^2 = 0.5y + 1$. We chose a mean fitness effect of 0.025, as previous work has shown that the distribution of fitness effects of REL606 can be well approximated as an exponential distribution with mean around 0.025⁵⁵.

To simulate the evolutionary dynamics in the strong selection, weak mutation regime (periodic sweeps), we can consider the fate of each mutation independently. To that end, for 10^6 independent replicates, we initialize a mutant at a population size of $n_0^{mut} = 1$ and the wildtype at $n_0^{wt} = 10^3 - 1$. For each mutant, we sample a pair of s_{mut} and $\sigma_{d,mut}^2$ from the aforementioned joint distribution. For the wildtype, we always use $s_{wt} = 0$ (by definition) and $\sigma_{d,wt}^2 = 3$. We then simulate the population using a Wright-Fisher like procedure, with an expected total population size of $N_{tot} = 10^3$. Every generation, we compute the expected population size for both the mutant and wildtype, and renormalize to obtain an expected strain frequency,

$$\langle n_t^i \rangle = n_{t-1}^i e^{s_i} \quad (51)$$

$$\langle f_t^i \rangle = \frac{\langle n_t^i \rangle}{\sum_i \langle n_t^i \rangle} . \quad (52)$$

Subsequently, for each strain, we simulate the effect of genetic drift by sampling the population size of the next generation for each strain i from a negative binomial density (similar to equation 43),

$$n_t^i | n_{t-1}^i \sim \text{NB} \left(N_{tot} \langle f_t^i \rangle, \sigma_{d,i}^2 \right) . \quad (53)$$

We iteratively continue this procedure until either the mutant or wildtype goes extinct. We then take all of the simulation runs where the mutant successfully fixed in the population, and average across the mutant parameter values to obtain $\langle s_{mut} | \text{fixed} \rangle$ and $\langle \sigma_{d,mut}^2 | \text{fixed} \rangle$.

We used a slightly different simulation procedure to simulate evolutionary dynamics in the weak selection, strong mutation regime (clonal interference). We initialize the population with solely the wildtype at $n_0^{wt} = 10^3$. Then, for each generation, each individual has a probability of $U_b = 10^{-3}$ of obtaining a beneficial mutation, drawn from the bivariate exponential distribution. We assume that fitness effects are additive, and the mutant descendant number variance replaces that of its ancestor. We then used equations 51-53 to simulate the dynamics of each strain across generations. We continue each simulation until one (non-wildtype) lineage has fixed in the population. We then take the root mutant parameter values of the lineage that eventually fixed, and average over all independent simulations to obtain $\langle s_{mut} | \text{fixed} \rangle$ and $\langle \sigma_{d,mut}^2 | \text{fixed} \rangle$.

Code and data availability

Raw genomic sequencing reads are available at BioProject PRJNA1129587. All code and data presented in this manuscript are available at: <https://github.com/joaoascensao/genetic-drift-evolution>

Acknowledgements

We thank Robin Herbert and Adam Deutschbauer for gifting us the barcoded Tn7 donor library, and providing advice on the system and protocols. We thank Adam Arkin, Kelly Wetmore, Aditya Prasad, Takashi Okada, and all members of the Hallatschek lab (past and present) for helpful comments and advice on the project. We thank Richard Lenski for sending us the LTEE-derived strains and populations, along with experimental advice and feedback. Research reported in this publication was supported by a National Science Foundation CAREER Award

(1555330). This work was supported by the National Institute of General Medical Sciences of the NIH under award R01GM115851 and by a Humboldt Professorship of the Alexander von Humboldt Foundation. JAA acknowledges support from an NSF graduate research fellowship, a Berkeley fellowship (from UC Berkeley), and Lloyd and Brodie scholarships (from UC Berkeley Dept of Bioengineering). QY acknowledges support from an NSF graduate research fellowship. Genomic sequencing was performed by SeqCoast Genomics LLC. This work used the Illumina HiSeq 4000 at the QB3 Sequencing Laboratory at UC Berkeley (RRID:SCR_022170) for barcode sequencing, supported by NIH S10 OD018174 Instrumentation Grant.

References

1. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159, DOI: [10.1093/genetics/16.2.97](https://doi.org/10.1093/genetics/16.2.97) (1931).
2. Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719, DOI: [10.1093/genetics/47.6.713](https://doi.org/10.1093/genetics/47.6.713) (1962).
3. Ewens, W. J. *Mathematical Population Genetics 1: Theoretical Introduction* (Springer Science & Business Media, 2004).
4. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer Associates, 1997). Google-Books-ID: 4ypuQgAACAAJ.
5. Muller, H. J. Some Genetic Aspects of Sex. *The Am. Nat.* **66**, 118–138, DOI: [10.1086/280418](https://doi.org/10.1086/280418) (1932). Publisher: The University of Chicago Press.
6. Muller, H. J. The relation of recombination to mutational advance. *Mutat. Res. Mol. Mech. Mutagen.* **1**, 2–9, DOI: [10.1016/0027-5107\(64\)90047-8](https://doi.org/10.1016/0027-5107(64)90047-8) (1964).
7. Haldane, J. B. S. A mathematical theory of natural and artificial selection, part V: Selection and mutation. *Math. Proc. Camb. Philos. Soc.* **23**, 838–844, DOI: [10.1017/S0305004100015644](https://doi.org/10.1017/S0305004100015644) (1927).
8. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930).
9. Wahl, L. M. & Gerrish, P. J. The Probability That Beneficial Mutations Are Lost in Populations with Periodic Bottlenecks. *Evolution* **55**, 2606–2610, DOI: [10.1111/j.0014-3820.2001.tb00772.x](https://doi.org/10.1111/j.0014-3820.2001.tb00772.x) (2001). [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0014-3820.2001.tb00772.x](https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0014-3820.2001.tb00772.x).
10. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
11. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205, DOI: [10.1038/nrg2526](https://doi.org/10.1038/nrg2526) (2009).
12. Wang, J., Santiago, E. & Caballero, A. Prediction and estimation of effective population size. *Heredity* **117**, 193–206, DOI: [10.1038/hdy.2016.43](https://doi.org/10.1038/hdy.2016.43) (2016).
13. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Probab.* **19**, 27–43, DOI: [10.2307/3213548](https://doi.org/10.2307/3213548) (1982).
14. Wakely, J. *Coalescent Theory: An Introduction* (Macmillan Learning, 2016). Google-Books-ID: x30RAGAACAAJ.
15. Gillespie, J. H. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics* **155**, 909–919, DOI: [10.1093/genetics/155.2.909](https://doi.org/10.1093/genetics/155.2.909) (2000).
16. Gillespie, J. H. Is the population size of a species relevant to its evolution? *Evol. Int. J. Org. Evol.* **55**, 2161–2169, DOI: [10.1111/j.0014-3820.2001.tb00732.x](https://doi.org/10.1111/j.0014-3820.2001.tb00732.x) (2001).
17. Neher, R. A. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annu. Rev. Ecol. Evol. Syst.* **44**, 195–215, DOI: [10.1146/annurev-ecolsys-110512-135920](https://doi.org/10.1146/annurev-ecolsys-110512-135920) (2013).
18. Desai, M. M., Walczak, A. M. & Fisher, D. S. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics* **193**, 565–585, DOI: [10.1534/genetics.112.147157](https://doi.org/10.1534/genetics.112.147157) (2013).
19. Hallatschek, O. Selection-Like Biases Emerge in Population Models with Recurrent Jackpot Events. *Genetics* **210**, 1053–1073, DOI: [10.1534/genetics.118.301516](https://doi.org/10.1534/genetics.118.301516) (2018).
20. Lewontin, R. C. *The Genetic Basis of Evolutionary Change* (Columbia University Press, 1974). Google-Books-ID: rLMTAQAAIAAJ.
21. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
22. Gillespie, J. H. *The Causes of Molecular Evolution*. Oxford Series in Ecology and Evolution (Oxford University Press, Oxford, New York, 1994).

23. Yu, Q. *et al.* Mutability of demographic noise in microbial range expansions. *The ISME J.* **15**, 2643–2654, DOI: [10.1038/s41396-021-00951-9](https://doi.org/10.1038/s41396-021-00951-9) (2021). Number: 9 Publisher: Nature Publishing Group.
24. Lenski, R. E. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME J.* **11**, 2181–2194, DOI: [10.1038/ismej.2017.69](https://doi.org/10.1038/ismej.2017.69) (2017).
25. Rozen, D. E. & Lenski, R. E. Long-term experimental evolution in escherichia coli. viii. dynamics of a balanced polymorphism. *The Am. Nat.* **155**, 24–35 (2000).
26. Rozen, D. E., Schneider, D. & Lenski, R. E. Long-Term Experimental Evolution in Escherichia coli. XIII. Phylogenetic History of a Balanced Polymorphism. *J. Mol. Evol.* **61**, 171–180, DOI: [10.1007/s00239-004-0322-2](https://doi.org/10.1007/s00239-004-0322-2) (2005).
27. Plucain, J. *et al.* Epistasis and allele specificity in the emergence of a stable polymorphism in Escherichia coli. *Science* **343**, 1366–1369, DOI: [10.1126/science.1248688](https://doi.org/10.1126/science.1248688) (2014).
28. Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of E. coli. *Nature* **387**, 703–705, DOI: [10.1038/42701](https://doi.org/10.1038/42701) (1997).
29. Tenaillon, O. *et al.* Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536**, 165–170, DOI: [10.1038/nature18959](https://doi.org/10.1038/nature18959) (2016).
30. Ascensao, J. A., Lok, K. & Hallatschek, O. Asynchronous abundance fluctuations can drive giant genotype frequency fluctuations. *Nat. Ecol. & Evol.* **9**, 166–179, DOI: [10.1038/s41559-024-02578-3](https://doi.org/10.1038/s41559-024-02578-3) (2025). Publisher: Nature Publishing Group.
31. Herbert, R. A. & Deutschbauer, A. M. BarTn7: Optimizing Bacterial Lineage Tracking at Sub-Species Resolution for Population Dynamics in Ecological and Evolutionary Studies, DOI: [10.1101/2025.11.24.690285](https://doi.org/10.1101/2025.11.24.690285) (2025). ISSN: 2692-8205 Pages: 2025.11.24.690285 Section: New Results.
32. Yu, Q. *et al.* Lineage frequency time series reveal elevated levels of genetic drift in SARS-CoV-2 transmission in England. *PLOS Pathog.* **20**, e1012090, DOI: [10.1371/journal.ppat.1012090](https://doi.org/10.1371/journal.ppat.1012090) (2024). Publisher: Public Library of Science.
33. Karlin, S. Rates of Approach to Homozygosity for Finite Stochastic Models with Variable Population Size. *The Am. Nat.* **102**, 443–455 (1968). Publisher: [University of Chicago Press, American Society of Naturalists].
34. Vasi, F., Travisano, M. & Lenski, R. E. Long-Term Experimental Evolution in Escherichia coli. II. Changes in Life-History Traits During Adaptation to a Seasonal Environment. *The Am. Nat.* **144**, 432–456, DOI: [10.1086/285685](https://doi.org/10.1086/285685) (1994). Publisher: The University of Chicago Press.
35. Wang, P. *et al.* Robust Growth of *Escherichia coli*. *Curr. Biol.* **20**, 1099–1103, DOI: [10.1016/j.cub.2010.04.045](https://doi.org/10.1016/j.cub.2010.04.045) (2010).
36. Niven, G. W., Fuks, T., Morton, J. S., Rua, S. A. C. G. & Mackey, B. M. A novel method for measuring lag times in division of individual bacterial cells using image analysis. *J. Microbiol. Methods* **65**, 311–317, DOI: [10.1016/j.mimet.2005.08.006](https://doi.org/10.1016/j.mimet.2005.08.006) (2006).
37. Fisher, R. A. XXI.—On the Dominance Ratio. *Proc. Royal Soc. Edinb.* **42**, 321–341, DOI: [10.1017/S0370164600023993](https://doi.org/10.1017/S0370164600023993) (1923). Publisher: Royal Society of Edinburgh Scotland Foundation.
38. Wright, S. EVOLUTION IN MENDELIAN POPULATIONS. *Genetics* **16**, 97–159, DOI: [10.1093/genetics/16.2.97](https://doi.org/10.1093/genetics/16.2.97) (1931).
39. Kimura, M. Diffusion Models in Population Genetics. *J. Appl. Probab.* **1**, 177–232, DOI: [10.2307/3211856](https://doi.org/10.2307/3211856) (1964). Publisher: Applied Probability Trust.
40. Marchi, N., Schlichta, F. & Excoffier, L. Demographic inference. *Curr. Biol.* **31**, R276–R279, DOI: [10.1016/j.cub.2021.01.053](https://doi.org/10.1016/j.cub.2021.01.053) (2021).
41. Desai, M. M. & Fisher, D. S. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759–98, DOI: [10.1534/genetics.106.067678](https://doi.org/10.1534/genetics.106.067678) (2007).
42. Osella, M., Nugent, E. & Cosentino Lagomarsino, M. Concerted control of Escherichia coli cell division. *Proc. Natl. Acad. Sci.* **111**, 3431–3435, DOI: [10.1073/pnas.1313715111](https://doi.org/10.1073/pnas.1313715111) (2014). Publisher: Proceedings of the National Academy of Sciences.
43. Taylor, D., Verdon, N., Lomax, P., Allen, R. J. & Titmuss, S. Tracking the stochastic growth of bacterial populations in microfluidic droplets. *Phys. Biol.* **19**, 026003, DOI: [10.1088/1478-3975/ac4c9b](https://doi.org/10.1088/1478-3975/ac4c9b) (2022).

- 741 44. D'Arrigo, M. *et al.* Indirect Measurement of the Lag Time Distribution of Single Cells of *Listeria innocua* in
742 Food. *Appl. Environ. Microbiol.* **72**, 2533–2538, DOI: [10.1128/AEM.72.4.2533-2538.2006](https://doi.org/10.1128/AEM.72.4.2533-2538.2006) (2006).
- 743 45. Gillespie, J. H. Natural Selection for within-Generation Variance in Offspring Number. *Genetics* **76**, 601–606
744 (1974).
- 745 46. Ferrare, J. T. & Good, B. H. Evolution of evolvability in rapidly adapting populations. *Nat. ecology & evolution* **8**,
746 2085–2096 (2024).
- 747 47. Ascensao, J. A. *et al.* Rediversification following ecotype isolation reveals hidden adaptive potential. *Curr. Biol.*
748 **34**, 855–867.e6, DOI: [10.1016/j.cub.2024.01.029](https://doi.org/10.1016/j.cub.2024.01.029) (2024).
- 749 48. Choi, K.-H. *et al.* A Tn7-based broad-range bacterial cloning and expression system. *Nat. Methods* **2**, 443–448,
750 DOI: [10.1038/nmeth765](https://doi.org/10.1038/nmeth765) (2005). Publisher: Nature Publishing Group.
- 751 49. Jeong, H. *et al.* Genome Sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394**, 644–652,
752 DOI: [10.1016/J.JMB.2009.09.052](https://doi.org/10.1016/J.JMB.2009.09.052) (2009).
- 753 50. Deatherage, D. E. & Barrick, J. E. Identification of Mutations in Laboratory-Evolved Microbes from Next-
754 Generation Sequencing Data Using breseq. 165–188, DOI: [10.1007/978-1-4939-0554-6_12](https://doi.org/10.1007/978-1-4939-0554-6_12) (2014).
- 755 51. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **1**, DOI:
756 [10.1038/s41586-018-0124-0](https://doi.org/10.1038/s41586-018-0124-0) (2018).
- 757 52. Wetmore, K. M. *et al.* Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly
758 bar-coded transposons. *mBio* **6**, 00306–15, DOI: [10.1128/mBio.00306-15](https://doi.org/10.1128/mBio.00306-15) (2015).
- 759 53. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using
760 NetworkX. In Varoquaux, G., Vaught, T. & Millman, J. (eds.) *Proceedings of the 7th Python in Science Conference*,
761 11–15 (Pasadena, CA USA, 2008).
- 762 54. Ba, A. N. N. *et al.* High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast.
763 *Nat.* **2019** 1–11, DOI: [10.1038/s41586-019-1749-3](https://doi.org/10.1038/s41586-019-1749-3) (2019).
- 764 55. Ascensao, J. A., Wetmore, K. M., Good, B. H., Arkin, A. P. & Hallatschek, O. Quantifying the local adaptive
765 landscape of a nascent bacterial community. *Nat. Commun.* **14**, 248, DOI: [10.1038/s41467-022-35677-5](https://doi.org/10.1038/s41467-022-35677-5) (2023).
766 Number: 1 Publisher: Nature Publishing Group.
- 767 56. Aitkin, M., Boys, R. J. & Chadwick, T. Bayesian point null hypothesis testing via the posterior likelihood ratio.
768 *Stat. Comput.* **15**, 217–230, DOI: [10.1007/s11222-005-1310-0](https://doi.org/10.1007/s11222-005-1310-0) (2005).
- 769 57. Smith, I. & Ferrari, A. Equivalence between the posterior distribution of the likelihood ratio and a p-value in an
770 invariant frame. *Bayesian Analysis* **9**, 939–962, DOI: [10.1214/14-BA877](https://doi.org/10.1214/14-BA877) (2014).
- 771 58. Stan Development Team. Stan modeling language users guide and reference manual (2023).
- 772 59. Miller, J. W. & Dunson, D. B. Robust Bayesian Inference via Coarsening. *J. Am. Stat. Assoc.*
773 **114**, 1113–1125, DOI: [10.1080/01621459.2018.1469995](https://doi.org/10.1080/01621459.2018.1469995) (2019). Publisher: Taylor & Francis _eprint:
774 <https://doi.org/10.1080/01621459.2018.1469995>.
- 775 60. Couce, A. *et al.* Changing fitness effects of mutations through long-term bacterial evolution. *Science* **383**,
776 eadd1417, DOI: [10.1126/science.add1417](https://doi.org/10.1126/science.add1417) (2024). Publisher: American Association for the Advancement of
777 Science.
- 778 61. Gumbel, E. J. Bivariate Exponential Distributions. *J. Am. Stat. Assoc.* **55**, 698–707, DOI: [10.2307/2281591](https://doi.org/10.2307/2281591) (1960).
779 Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

The evolution of genetic drift over 50,000 generations

Supplementary Information

Joao A Ascensao, QinQin Yu, Oskar Hallatschek

2025

Contents

S1 Supplemental Tables	3
S2 Supplementary Note: Relation to Wahl and Gerrish (2001)	4
S2.1 Notation and definitions	4
S2.1.1 Relationship between selection coefficients	5
S2.2 Effective population size in the Wahl-Gerrish Model	5
S2.3 Establishment rate	6
S3 Theory: Noisy Growth Traits	6
S3.1 Variation in lag times	6
S3.1.1 Exponential distribution of lag times	6
S3.1.2 Gamma distribution of lag times	7
S3.2 Variation in growth rates	8
S3.2.1 Exponential distribution of doubling times	8
S3.2.2 Gamma distribution of doubling times	8

List of Figures

S1	Variance in allele frequency trajectories increases over time	10
S2	Variance in allele frequency trajectories for 10k <i>S</i> and <i>L</i> clones	10
S3	Ara+2 50k barcode frequency trajectories	11
S4	Contributions to $\text{var}(\log \bar{N}_e)$	12
S5	CFU error measurements	13
S6	BarSeq technical replicates	14
S7	Simulation results for σ_d^2 inference	15
S8	Simulation results for $\bar{N}_b$ inference	16
S9	ROC curves for outlier detection	17
S10	Exponentially distributed doubling and lag times	18

List of Tables

S1	Tn7 barcoded time courses	3
S2	Table of genotyping oligonucleotides	4

S1 Supplemental Tables

Strain	Name	Line	Time point	Clone	Replicate 1	Replicate 2	Replicate 3	CFU plating media
REL606	R	all	0k		E1	E4	E5	LB
REL1159A	1159A	Ara+2	2k	A	E2	E4		LB
REL1159B	1159B	Ara+2	2k	B	E2	E4		LB
REL2174A	2174A	Ara+2	5k	A	E2	E4		LB
REL2174B	2174B	Ara+2	5k	B	E2	E4		LB
REL4531A	4531A	Ara+2	10k	A	E3	E4		LB
REL4531B	4531B	Ara+2	10k	B	E3	E4		LB
REL7184A	7184A	Ara+2	15k	A	E3	E4		LB
REL7184B	7184B	Ara+2	15k	B	E3	E4		LB
REL8601A	8601A	Ara+2	20k	A	E4	E6		DM2000
REL8601B	8601B	Ara+2	20k	B	E3	E6		DM2000
REL10407	10407	Ara+2	30k	A	E3	E4		DM2000
REL10408	10408	Ara+2	30k	B	E3	E4		DM2000
REL10950	10950	Ara+2	40k	A	E3	E5		DM2000
REL10951	10951	Ara+2	40k	B	E3	E5		DM2000
REL11342	11342	Ara+2	50k	A	E3	E5		DM2000
REL11343	11343	Ara+2	50k	B	E3	E5		DM2000
REL1165A	1165A	Ara-2	2k	A	E4	E6		LB
REL11829	11829	Ara-2	2k	B	E4	E6		LB
REL11556	28	Ara-2	6.5k	L	E2	E3		DM2000
REL11555	46	Ara-2	6.5k	S	E2	E3		DM2000
4537G	4537G	Ara-2	10k	L	E5	E6		DM2000
4537H	4537H	Ara-2	10k	S	E5	E6		DM2000
REL7178A	7178A	Ara-2	15k	S	E5	E6		DM2000
REL7178B	7178B	Ara-2	15k	L	E4	E6		DM2000
REL11830	11830	Ara-2	20k	S	E5	E6		DM2000
REL11831	11831	Ara-2	20k	L	E4	E6		DM2000
REL11832	11832	Ara-2	30k	S	E5	E6		DM2000
REL11833	11833	Ara-2	30k	L	E5	E6		DM2000
REL11035	11035	Ara-2	40k	L	E5	E6		DM2000
REL11036	11036	Ara-2	40k	S	E5	E6		DM2000
11319C	11319C	Ara-2	50k	L	E5	E6		DM2000
11319D	11319D	Ara-2	50k	S	E5	E6		DM2000

Table S1: Table showing when replicates of different Tn7 barcoded timecourses were run. Each set of experiments E1-6 was run separately, on different sets of days.

Oligo Name	Sequence	Purpose
ja35	GAAGAGGAT AAAACCGTGGA	Amplify arcA for genotyping
ja36	AGGTCAGGG ACTTTTGTAC	Amplify arcA for genotyping
ja40	TTCGAAGCG ACAGATGGC	Sanger sequence arcA for genotyping
ja37	CAGTTGTGACATA CAGCTAACGCT	Amplify aspS for genotyping
ja38	GCTCATGGGAGTT CACTCAGTTG	Amplify aspS for genotyping
ja42	ACTCTAACC ACGTCAACACC	Sanger sequence aspS for genotyping

Table S2: Table of genotyping oligonucleotides to distinguish *S* and *L* clones. [1]

S2 Supplementary Note: Relation to Wahl and Gerrish (2001)

This note clarifies the relationship between the parameterization used in our manuscript and the formulation of Wahl & Gerrish (2001) [2], who derived the fixation probability for beneficial mutations in periodically bottlenecked populations. We show that our per-cycle selection coefficient relates to the Wahl-Gerrish “selective advantage” (defined as the ratio of growth rates minus one) by a factor of $r\tau$, where r is the wild-type growth rate and τ is the time between bottlenecks. We also demonstrate that the rate of establishment of beneficial mutations in the Wahl-Gerrish model is equivalent to the classical formula $2sN_e\mu$, when parameters are appropriately interpreted.

S2.1 Notation and definitions

Wahl-Gerrish parameterization In Wahl & Gerrish (2001), populations undergo exponential growth between periodic bottlenecks. The key parameters are:

- N_0 : population size immediately after a bottleneck
- N_τ : population size immediately before the next bottleneck (i.e., at time τ)
- r : wild-type Malthusian growth rate (units of time^{-1})
- τ : time between bottlenecks
- $D = N_0/N_\tau = e^{-r\tau}$: dilution factor
- s_w : selective advantage, defined as the ratio of mutant to wild-type growth rates minus one, such that the mutant growth rate is $r(1 + s_w)$

With exponential growth, the population grows from N_0 to $N_\tau = N_0e^{r\tau}$ during each cycle, followed by dilution back to N_0 .

Our parameterization In our manuscript, we express quantities per growth cycle (dilution-to-dilution interval), treating the cycle as the natural unit of time:

- N_b : bottleneck population size ($\equiv N_0$)

- σ_d^2 : variance in descendant number per cycle
- $N_e = N_b/\sigma_d^2$: effective population size per cycle
- s : per-cycle selection coefficient (mean excess relative growth over one cycle)

S2.1.1 Relationship between selection coefficients

The Wahl-Gerrish selection coefficient Consistent with the standard parameterization in the LTEE literature, Wahl & Gerrish define the selective advantage s_w via the ratio of Malthusian fitnesses, subtracted by one,

$$s_w = \frac{N_{mut}(\tau)/N_{mut}(0)}{N_{wt}(\tau)/N_{wt}(0)} - 1, \quad (S1)$$

where $N_{mut}(t)$ and $N_{wt}(t)$ are the mutant and wild-type population sizes, respectively. In the Wahl-Gerrish model, the selective advantage solely depends on the growth rates,

$$s_w = \frac{r_{mut}}{r_{wt}} - 1, \quad (S2)$$

where r_{mut} and r_{wt} are the Malthusian growth rates of mutant and wild-type, respectively.

Per-cycle selection coefficient The per-cycle selection coefficient, which we denote s , measures the mean excess relative growth over one complete cycle, or the slope in the logit frequency of the mutant,

$$s = \text{logit} \left(\frac{f(\tau)}{1 - f(\tau)} \right) - \text{logit} \left(\frac{f(0)}{1 - f(0)} \right), \quad (S3)$$

$$= \log \left(\frac{N_{mut}(\tau)}{N_{mut}(0)} \right) - \log \left(\frac{N_{wt}(\tau)}{N_{wt}(0)} \right), \quad (S4)$$

where $f(t)$ is the mutant frequency.

For a mutant with growth rate $r_{mut} = r(1 + s_w)$, the per-cycle selective advantage is:

$$s = \tau(r_{mut} - r), \quad (S5)$$

$$= s_w r \tau. \quad (S6)$$

where $r\tau = \ln(N_\tau/N_0) = -\ln D$. For the LTEE with 1:100 daily dilutions, $r\tau = \ln(100) \approx 4.6$, corresponding to about 6.64 generations per cycle.

S2.2 Effective population size in the Wahl-Gerrish Model

Time-dependent fixation probability Wahl & Gerrish showed that a beneficial mutation (under weak selection, $s_w r \tau \ll 1$) arising at time t during the growth phase (where $t \in [0, \tau]$) has an approximate fixation probability of:

$$p_{fix}(t) \approx 2s_w r \tau \frac{N_0}{N(t)}. \quad (S7)$$

Using our definition of s ,

$$p_{fix}(t) \approx 2s \frac{N_0}{N(t)}. \quad (S8)$$

Effective population size Wahl & Gerrish derived an effective population size for periodically bottlenecked populations (in units number of generations):

$$N_e^{WG} = N_0 \cdot r\tau = N_0 \cdot (-\ln D) . \quad (\text{S9})$$

This effective population size captures the reduced efficacy of selection due to population bottlenecks. In our notation with per-cycle measurements, when the only source of descendant-number variance is the bottleneck itself (Poisson sampling with $\sigma_d^2 = 1$), our effective population size (in units of number of cycles) is simply

$$N_e = N_b = N_0 . \quad (\text{S10})$$

S2.3 Establishment rate

At time t during the growth phase, mutations arise at rate $\mu N(t)$ per unit time, where μ is the per-cell mutation rate. The instantaneous establishment rate is:

$$\text{establishment rate}(t) = \mu N(t) \cdot p_{fix}(t) . \quad (\text{S11})$$

Substituting Eq. (S8):

$$\text{establishment rate}(t) = \mu N(t) \cdot \frac{2sN_0}{N(t)} = 2sN_0\mu . \quad (\text{S12})$$

We see that the establishment rate is constant in time, independent of when during the cycle the mutation arises. This is because although mutations arising late in the cycle are more numerous (larger $N(t)$), they have proportionally lower fixation probability (smaller $N_0/N(t)$), and these effects exactly cancel. This expression is equivalent to the establishment rate in classical population genetics models [3].

S3 Theory: Noisy Growth Traits

S3.1 Variation in lag times

We consider a population of asexually reproducing microbes that is transferred to a new environment. After an initial lag time, individuals start growing exponentially at a growth rate r . The culture is allowed to grow for a total duration T .

S3.1.1 Exponential distribution of lag times

Suppose that the lag time for each individual, t_{lag} , is an exponential random variable,

$$P(t_{lag} < t) = 1 - e^{-t/\tau} . \quad (\text{S13})$$

The total population size of a clone that starts as a single individual at time 0 will be $N = \exp[r(T - t_{lag})]$ at time T . It follows that,

$$P(N > n) = P\left(t_{lag} < T - \frac{1}{r} \log(n)\right) \quad (\text{S14})$$

$$= 1 - e^{-\frac{T - \log(n)/r}{\tau}} \quad (\text{S15})$$

$$P(N = n) = -\partial_n P(N > n) \quad (\text{S16})$$

$$= \frac{1}{r\tau Q} \left(\frac{n}{Q}\right)^{\frac{1}{r\tau} - 1} . \quad (\text{S17})$$

We define $Q = e^{rT}$ as the maximum size of a clone (i.e. if it starts growing immediately; $t_{lag} = 0$). The expression is valid for $n < Q$ and assuming that $P(t_{lag} > T) \approx 0$. We see that the offspring number distribution $P(N = n)$ is a power-law, with an exponent that is always larger than -1 . For notational simplicity, we define a compound parameter $\gamma = 1/(r\tau)$. We can now calculate the variance of clone sizes,

$$\langle N \rangle = \frac{\gamma}{Q} \int_0^Q dn \left(\frac{n}{Q} \right)^{\gamma-1} n = \frac{\gamma}{\gamma+1} Q. \quad (\text{S18})$$

$$\langle N^2 \rangle = \frac{\gamma}{Q} \int_0^Q dn \left(\frac{n}{Q} \right)^{\gamma-1} n^2 = \frac{\gamma}{\gamma+2} Q^2. \quad (\text{S19})$$

$$\text{var } N = \left[\frac{\gamma}{\gamma+2} - \left(\frac{\gamma}{\gamma+1} \right)^2 \right] Q^2 \quad (\text{S20})$$

In a serial dilution regime, the population size at time T will be diluted down by a factor of $\langle N \rangle$, so that each clone will start again with one individual, on average. Thus, the effective additional descendant number variance across cycles is,

$$\sigma_d^2 = \frac{\text{var } N}{\langle N \rangle^2} = \frac{(\gamma+1)^2}{\gamma(\gamma+2)} - 1. \quad (\text{S21})$$

When $1/\gamma \gg 1$, we see that the descendant number variance increases linearly with the lag time mean/standard deviation, $\sigma_d^2 \propto \tau$.

S3.1.2 Gamma distribution of lag times

Following closely the derivation of the prior section, we now consider the case that the lag time for each individual, t_{lag} , is a gamma random variable,

$$P(t_{lag} < t) = 1 - \frac{\Gamma(\alpha, \beta t)}{\Gamma(\alpha)}. \quad (\text{S22})$$

Here $\Gamma(\cdot)$ is the gamma function, and $\Gamma(\cdot, \cdot)$ is the upper incomplete gamma function. In this parameterization, the mean of the gamma distribution is $\mu = \alpha/\beta$ and the variance is $v = \alpha/\beta^2$. As before, the total population size of a clone that starts as a single individual at time 0 will be $N = \exp[r(T - t_{lag})]$ at time T . It follows that,

$$P(N > n) = P\left(t_{lag} < T - \frac{1}{r} \log(n)\right) \quad (\text{S23})$$

$$= 1 - \frac{\Gamma(\alpha, \beta [T - \frac{1}{r} \log(n)])}{\Gamma(\alpha)} \quad (\text{S24})$$

$$P(N = n) = -\partial_n P(N > n) \quad (\text{S25})$$

$$= \frac{\beta}{r\Gamma(\alpha)} e^{-\beta T} x^{\frac{\beta}{r}-1} \left(\beta T - \frac{\beta}{r} \log(n) \right)^{\alpha-1}. \quad (\text{S26})$$

Again, we define $Q = e^{rT}$ as the maximum size of a clone (i.e. if it starts growing immediately; $t_{lag} = 0$). The expression is valid for $n < Q$ and assuming that $P(t_{lag} > T) \approx 0$. We calculate the moments as before,

$$\langle N \rangle = \int_0^Q dn P(N = n) n = Q \left(\frac{\beta}{\beta+r} \right)^\alpha. \quad (\text{S27})$$

$$\langle N^2 \rangle = \int_0^Q dn P(N = n) n^2 = Q^2 \left(\frac{\beta}{\beta+2r} \right)^\alpha. \quad (\text{S28})$$

In a serial dilution regime, the population size at time T will be diluted down by a factor of $\langle N \rangle$, so that each clone will start again with one individual, on average. Thus, the effective additional descendant number variance across cycles is,

$$\sigma_d^2 = \frac{\langle N^2 \rangle}{\langle N \rangle^2} - 1 = \left(1 + \frac{r^2}{\beta(\beta + 2r)} \right)^\alpha - 1. \quad (\text{S29})$$

S3.2 Variation in growth rates

We consider the case that the exponential growth rate is noisy. Again, we consider an asexual population of microbes that are growing in exponential phase for some time T . Using continuous time models, we take the time to split into two daughter cells as a random variable; we neglect mother-daughter correlations.

S3.2.1 Exponential distribution of doubling times

The simplest case of noisy growth rates is when doubling times are exponentially distributed with a mean parameter τ , as division is memoryless (i.e. a Poisson process). This problem can be formulated as a Galton-Watson process [4], which gives first and second moments of the population sizes,

$$\langle N \rangle = e^{T/\tau} \quad (\text{S30})$$

$$\text{var } N = e^{2T/\tau} - e^{T/\tau}. \quad (\text{S31})$$

As before, in a serial dilution regime, the population size at time T will be diluted down by a factor of $\langle N \rangle$, so that each clone will start again with one individual, on average. Thus, the effective additional descendant number variance across cycles is,

$$\sigma_d^2 = \frac{\text{var } N}{\langle N \rangle^2} = 1 - e^{-T/\tau}. \quad (\text{S32})$$

S3.2.2 Gamma distribution of doubling times

We now assume that doubling times are distributed according to a gamma distribution. We treat the process as a classical Bellman-Harris process, and obtain asymptotic solutions for the moments [5]. As the total time becomes large, $T \rightarrow \infty$, the mean number of individuals can be approximated by,

$$\langle N \rangle \approx n_1 e^{rT}. \quad (\text{S33})$$

Here, r is the Malthusian parameter, which is the solution to the integral,

$$2 \int_0^\infty e^{-rt} dG(t) = 1 \quad (\text{S34})$$

$$r = \beta(2^{1/\alpha} - 1). \quad (\text{S35})$$

The function $G(t)$ is the cumulative distribution function of the gamma distribution,

$$G(t) = 1 - \frac{\Gamma(\alpha, \beta t)}{\Gamma(\alpha)}. \quad (\text{S36})$$

The prefactor n_1 is given by the following integral,

$$n_1 = \left[4r \int_0^\infty t e^{-rt} dG(t) \right]^{-1} \quad (\text{S37})$$

$$= \frac{(\beta + r)^{\alpha+1}}{4\alpha r \beta^\alpha} . \quad (\text{S38})$$

Similarly, the second moment can be asymptotically approximated by,

$$\langle N^2 \rangle \approx n_2 e^{2rT} . \quad (\text{S39})$$

The prefactor n_2 is given by,

$$n_2 = \frac{2n_1^2 \int_0^\infty e^{-2rt} dG(t)}{1 - 2 \int_0^\infty e^{-2rt} dG(t)} \quad (\text{S40})$$

$$\int_0^\infty e^{-2rt} dG(t) = \frac{\beta^\alpha}{(2r + \beta)^\alpha} . \quad (\text{S41})$$

Finally, the effective additional descendant number variance is given by,

$$\sigma_d^2 = \frac{\langle N^2 \rangle}{\langle N \rangle^2} - 1 \approx \left[\frac{1}{2} \left(2^{1/\alpha+1} - 1 \right)^\alpha - 1 \right]^{-1} - 1 \quad (\text{S42})$$

$$\approx \frac{0.9609}{\alpha} . \quad (\text{S43})$$

We see that the effective descendant number variance does not depend on T or β . It only depends on α , which is related to the coefficient of variation of the gamma distribution, $\alpha = \mu^2/\sigma^2$. The effective descendant number variance is well approximated by a simple inverse relationship to α .

References

- [1] Joao A. Ascensao, Jonas Denk, Kristen Lok, QinQin Yu, Kelly M. Wetmore, and Oskar Hallatschek. Re-diversification following ecotype isolation reveals hidden adaptive potential. *Current Biology*, 34(4):855–867.e6, February 2024.
- [2] Lindi M. Wahl and Philip J. Gerrish. The Probability That Beneficial Mutations Are Lost in Populations with Periodic Bottlenecks. *Evolution*, 55(12):2606–2610, 2001. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0014-3820.2001.tb00772.x>.
- [3] Warren J. Ewens. *Mathematical Population Genetics 1: Theoretical Introduction*. Springer Science & Business Media, January 2004.
- [4] Patsy Haccou, Peter Jagers, and Vladimir A. Vatutin. *Branching Processes: Variation, Growth, and Extinction of Populations*. Cambridge University Press, May 2005. Google-Books-ID: Pff6J_VAboIC.
- [5] Richard Bellman and Theodore Harris. On Age-Dependent Binary Branching Processes. *The Annals of Mathematics*, 55(2):280, March 1952.

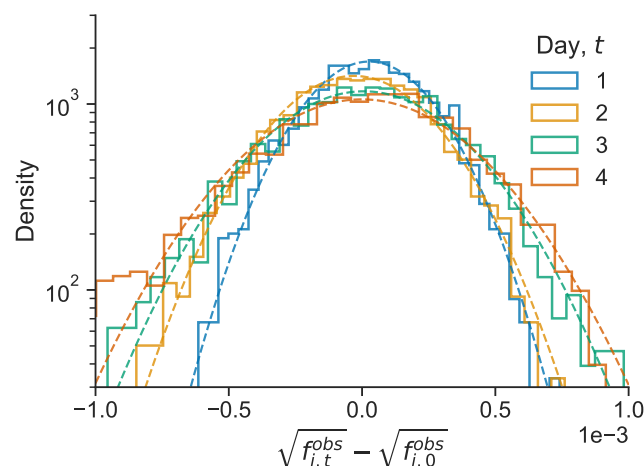


Figure S1: **Variance in allele frequency trajectories increases over time** (related to Figures 1D and 1E). Transition densities of barcode frequencies for the REL606 library, replicate 1. Here, we compute the transition densities in reference to the frequency at time point 0. We see that the variance increases over time, which is expected given that variance due to genetic drift accumulates over time. Dotted lines represent Gaussian fits to the empirical densities.

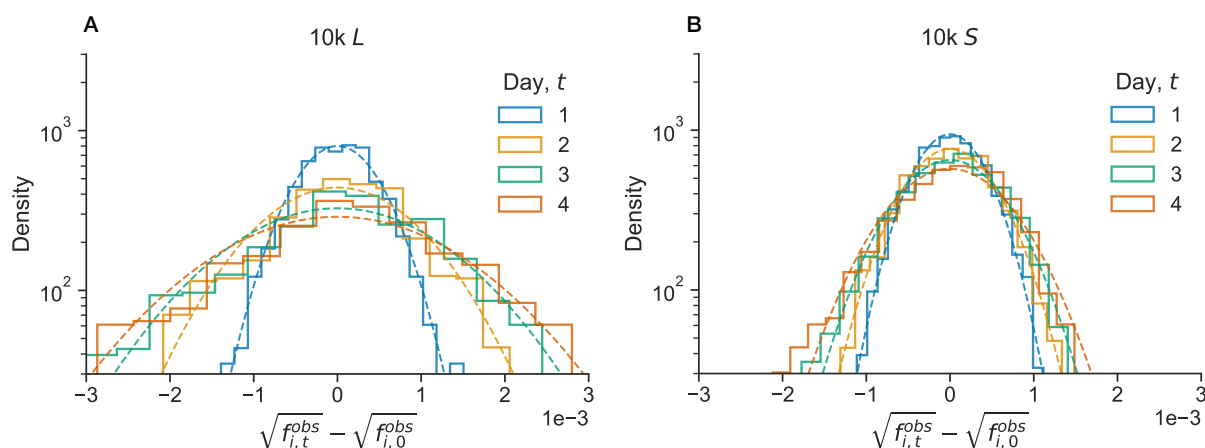


Figure S2: **Variance in allele frequency trajectories for 10k *S* and *L* clones** (related to Figures 1D and 2). Transition densities of barcode frequencies; dotted lines represent Gaussian fits to the empirical densities. Here, we compute the transition densities in reference to the frequency at time point 0. As suggested by the estimates of N_e , the neutral barcode trajectories of the *S* clone accumulate less variance than those of the *L* clone.

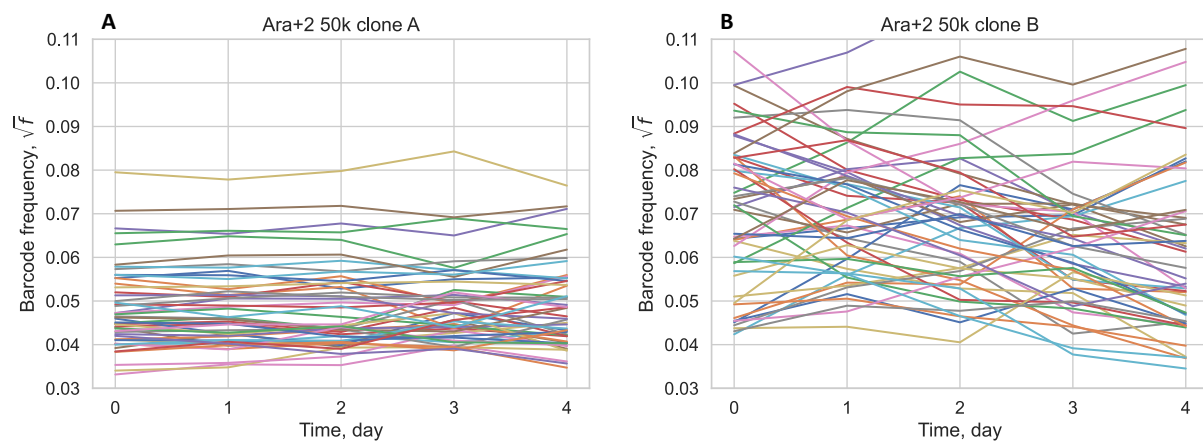


Figure S3: **Ara+2 50k barcode frequency trajectories.** We show example square-root transformed barcode frequency trajectories for clones (A) *A* and (B) *B*. We randomly sampled 50 barcodes from each clone, both from biological replicate 1. Clone *B* visibly has stronger fluctuations compared to clone *A*.

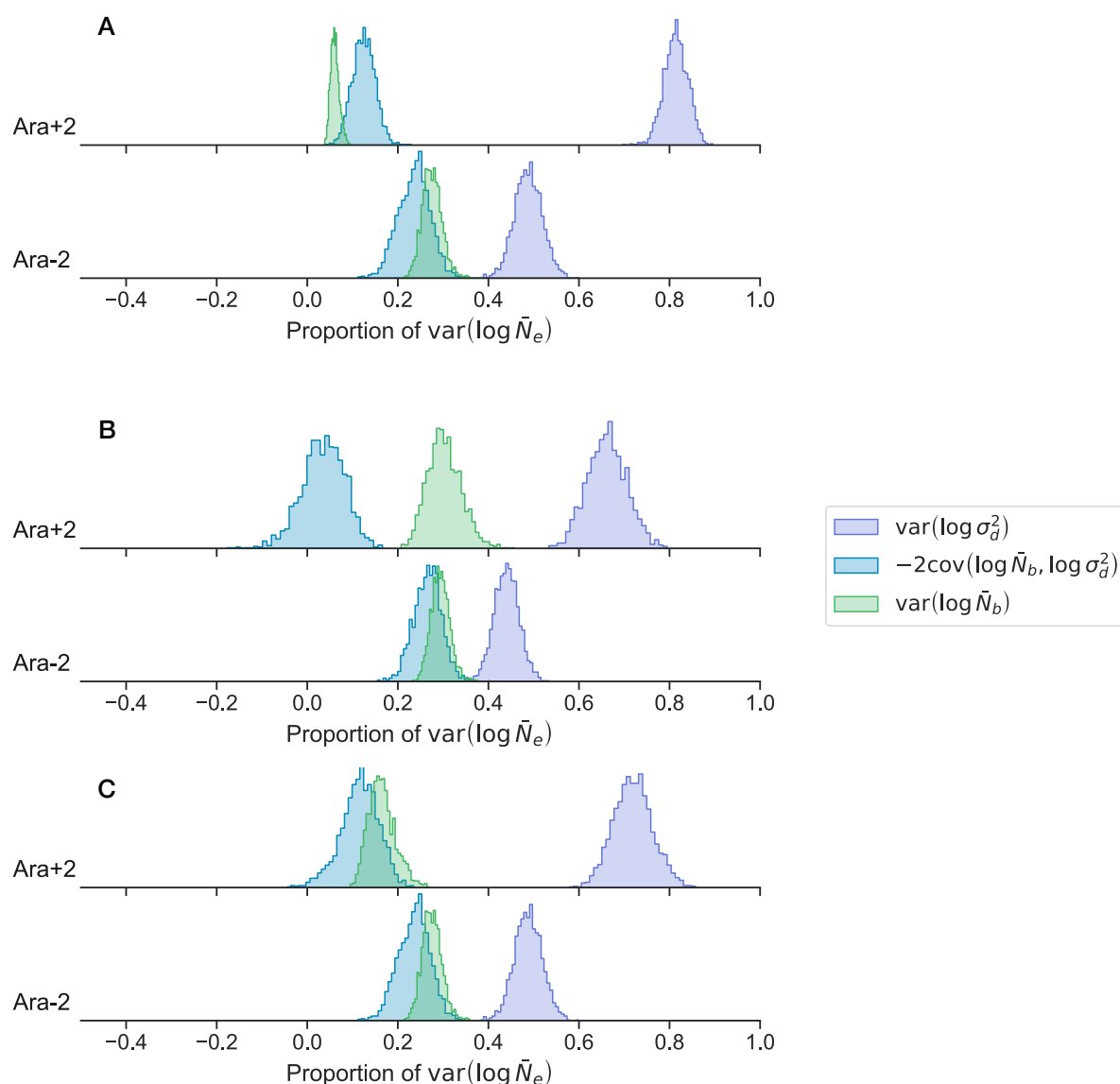


Figure S4: **Contributions to $\text{var}(\log \tilde{N}_e)$** . Same as Figure 4E, except **(A)** REL606 is excluded from the calculation, **(B)** the Ara+2 50k B clone is excluded (i.e. the outlier clone with the large σ_d^2 value), or **(C)** both REL606 and the Ara+2 50k B clone is excluded.

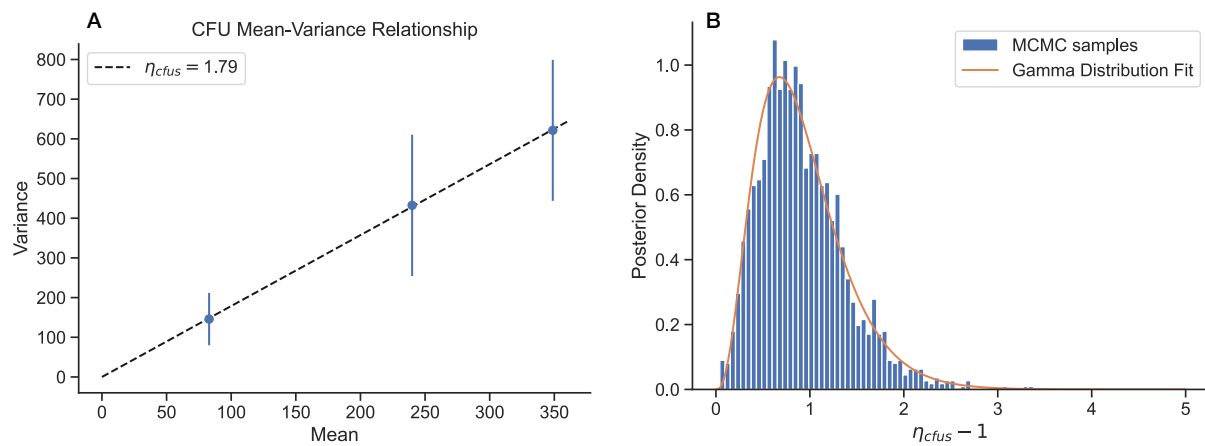


Figure S5: **Measurement of CFU errors.** (A) Using three different cultures of REL606 at different cell densities, we diluted and plated the cultures on agar plates as 16 independent technical replicates. We then computed the mean and variance across technical replicates for each culture. We see a linear mean-variance relationship. (B) Using a bayesian inference framework, we obtained an estimate for the marginal posterior of the overdispersion parameter η_{cfus} .

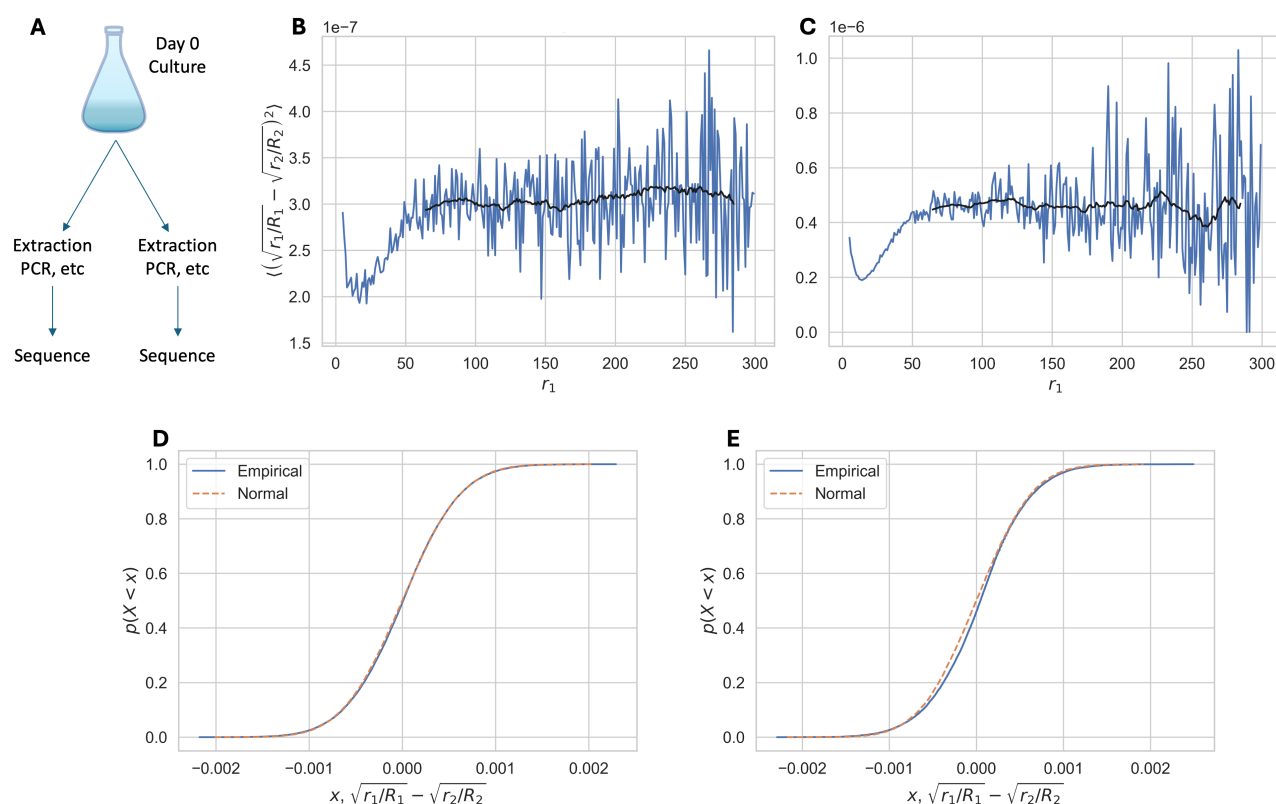


Figure S6: BarSeq technical replicates reveal patterns of measurement noise. (A) A single single culture of barcoded clones was split into two subcultures, and all sample processing steps (genomic DNA extraction, barcode amplification, sequencing) were done independently on the two cultures. (B-C) We quantified the average variance of square-root barcode frequencies, as a function of the read count of technical replicate 1, r_1 , for the (B) 6.5k *L* and (C) 6.5k *S* barcoded libraries. Blue lines represent raw data, black lines represent a rolling average. We see that the variance of the square-root transformed data is approximately stable for reads counts greater than about 50. (D-E) We found that the distribution of difference in frequencies between technical replicates was well approximated by a Gaussian distribution for both the (D) 6.5k *L* and (E) 6.5k *S* barcoded libraries.

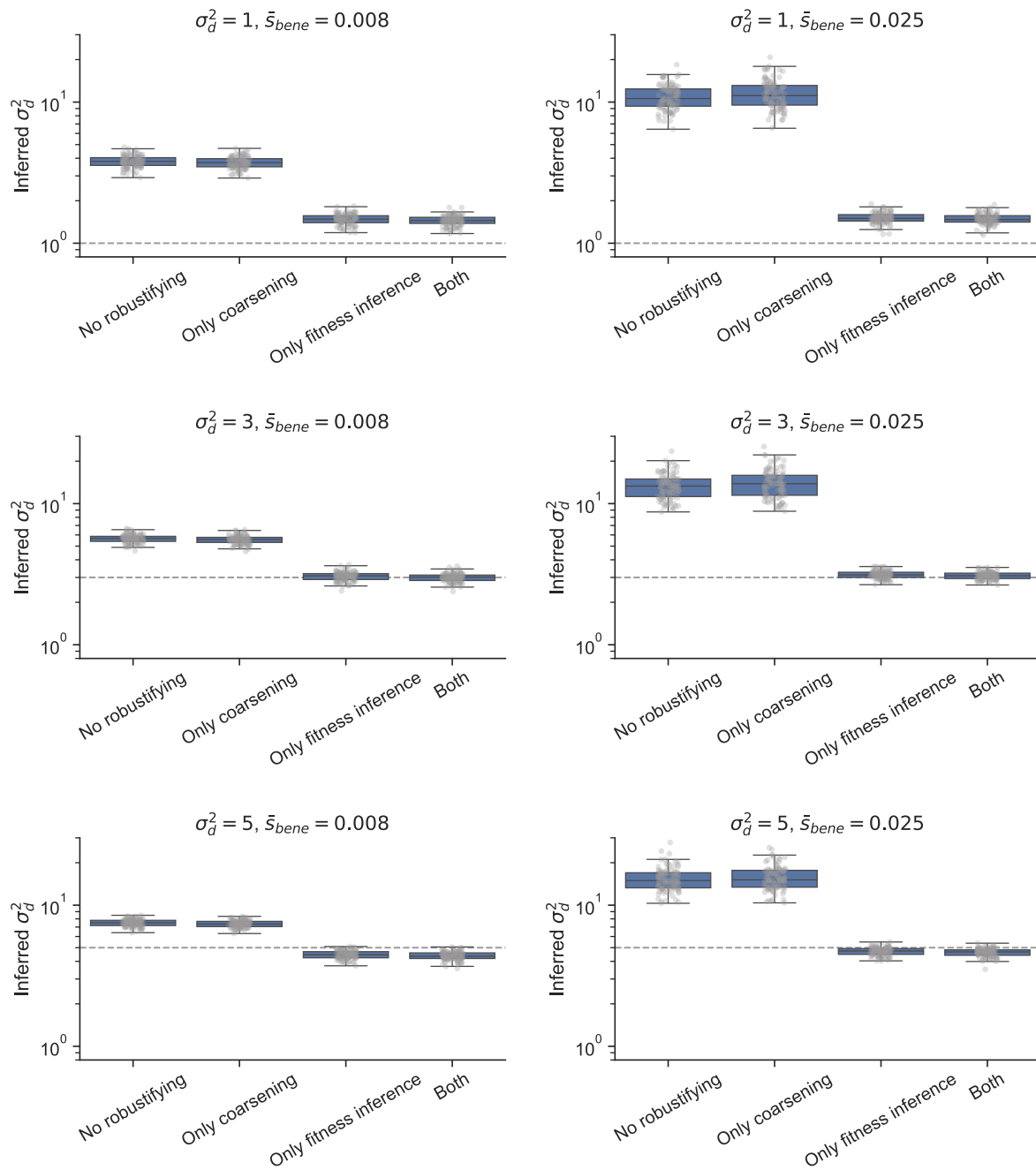


Figure S7: Simulation results for σ_d^2 inference.

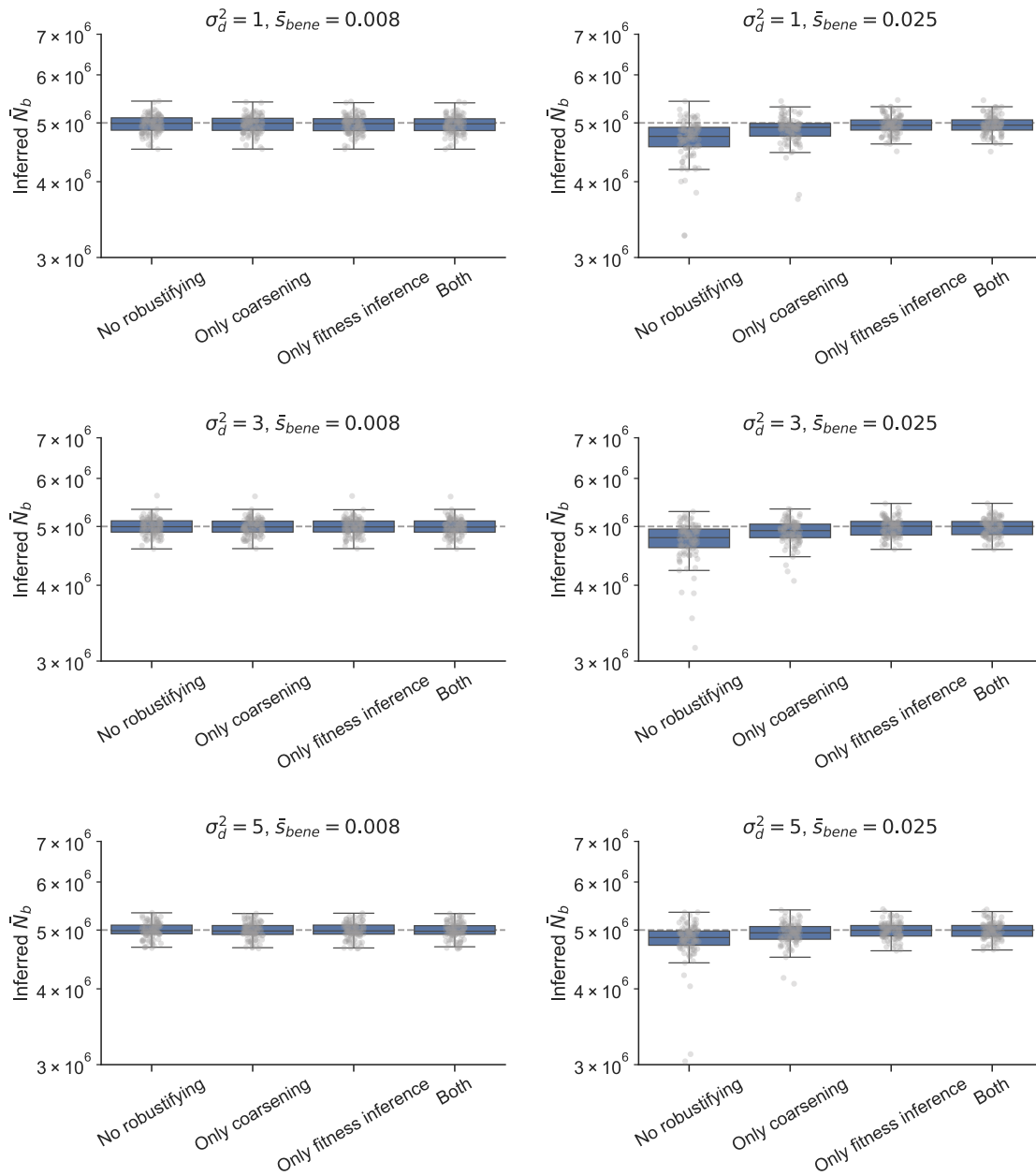


Figure S8: **Simulation results for $\bar{N}_b$ inference.**

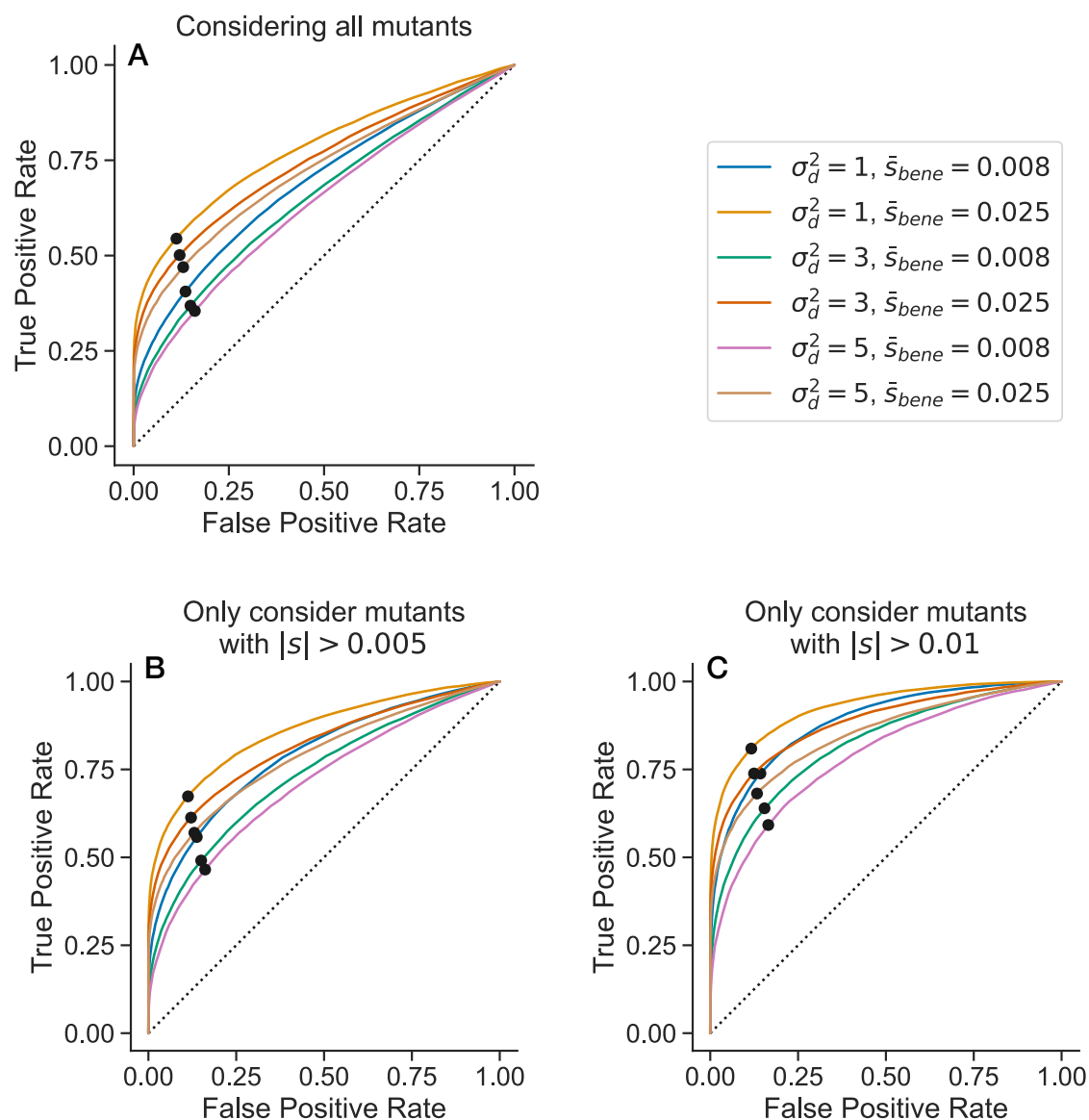


Figure S9: **Receiver operating characteristic (ROC) curves for outlier detection.** (A) Across all of our simulated data sets, as we vary the threshold p value to call a lineage as non-neutral, we compute the true and false positive rate that the lineage is actually non-neutral. We perform the same computation, but testing our ability to distinguish larger fitness effect mutants from neutrality, considering (B) $|s| > 0.005$ or (C) $|s| > 0.01$. Black dots represent the point where $p = 0.05$.

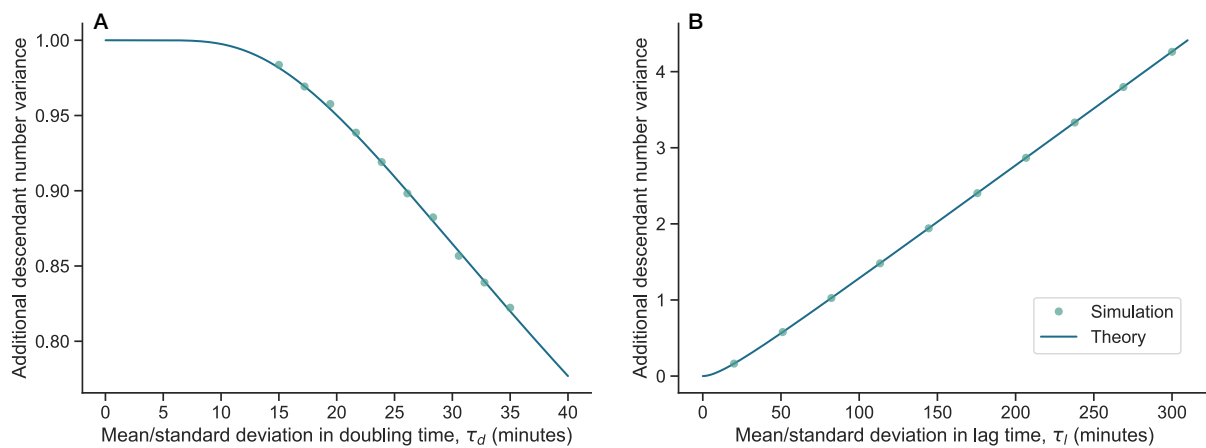


Figure S10: **Physiological causes of changes to descendant number variance.** Related to Figure 5A-B. We treat both (A) doubling times and (B) lag times as exponentially distributed random variables. Solid lines represent analytical results presented in section S3, and circles represent simulation results.