AN ANALYSIS OF FITNESS IN LONG-TERM ASEXUAL EVOLUTION EXPERIMENTS

By

Michael J Wiser

A DISSERTATION

Submitted to
Michigan State University
In partial fulfillment of the requirements
for the degree of

Zoology- Doctor of Philosophy
Ecology, Evolutionary Biology and Behavior- Dual Major

2015

ABSTRACT

AN ANALYSIS OF FITNESS IN LONG-TERM ASEXUAL EVOLUTION EXPERIMENTS

By

Michael J Wiser

Evolution is the central unifying concept of modern biology. Yet it can be hard to study in natural system, as it unfolds across generations. Experimental evolution allows us to ask questions about the process of evolution itself: How repeatable is the evolutionary process? How predictable is it? How general are the results? To address these questions, my collaborators and I carried out experiments both within the Long-Term Evolution Experiment (LTEE) in the bacteria *Escherichia coli*, and the digital evolution software platform Avida.

In Chapter 1, I focused on methods. Previous research in the LTEE has relied on one particular way of measuring fitness, which we know becomes less precise as fitness differentials increase. I therefore decided to test whether two alternate ways of measuring fitness would improve precision, using one focal population. I found that all three methods yielded similar results in both fitness and coefficient of variation, and thus we should retain the traditional method.

In Chapter 2, I turned to measuring fitness in each of the populations. Previous work had considered fitness to change as a hyperbola. A hyperbolic function is bounded, and predicts that fitness will asymptotically approach a defined upper bound; however, we knew that fitness in these populations routinely exceeded the asymptotic limit calculated from a hyperbola fit to the earlier data. I instead used to a power law, a mathematical function that does not

have an upper bound. I found that this function substantially better describes fitness in this system, both among the whole set of populations, and in most of the individual populations. I also found that the power law models fit on just early subsets of the data accurately predict fitness far into the future. This implies that populations, even after 50,000 generations of evolution in consistent environment, are so far from the tops of fitness peaks that we cannot detect evidence of those peaks.

In Chapter 3, I examined to how variance in fitness changes over long time scales. The among-population variance over time provides us information about the adaptive landscape on which the populations have been evolving. I found that among-population variance remains significant. Further, competitions between evolved pairs of populations reveal additional details about fitness trajectories than can be seen from competitions against the ancestor. These results demonstrate that our populations have been evolving on a complex adaptive landscape.

In Chapter 4, I examined whether the patterns found in Chapter 2 apply to a very different evolutionary system, Avida. This system incorporates many similar evolutionary pressures as the LTEE, but without the details of cellular biology that underlie nearly all organic life. I find that in both the most complex and simplest environments in Avida, fitness also follows the same power law dynamics as seen in the LTEE. This implies that power law dynamics may be a general feature of evolving systems, and not dependent on the specific details of the system being studied.

ACKNOWLEDGEMENTS

In my time at Michigan State University, I've been fortunate to receive the help and support of a wide range of individuals. I cannot take the space to acknowledge everyone, but there are some who need to be specifically mentioned.

My advisor, Richard Lenski, has provided a superb environment in which to be a graduate student. He has allowed me freedom to pursue the projects which interested me, while simultaneously offering ideas of projects related to the main work in the laboratory whenever my other projects ran into methodological difficulties. His laboratory funding has allowed me to pursue projects merely because I found them interesting, and not because he already had a grant specifically tied to that work. He has challenged me to become a better writer, to improve my statistical skills, to place my work in the context of the field (forcing me to at least occasionally see the forest among the trees), and to work independent of direct oversight.

My committee members, past and present, have each brought their own additions to my work. Charles Ofria, who was all but officially a second advisor, taught me how to explain my work to a broader audience. Thomas Schmidt kept me grounded in the molecular details of my system. Andrew McAdam taught me to look beyond the mechanics of a statistical test to what the fundamental point of the comparison was. Ian Dworkin forced me to be able to defend my mathematical choices. Arend Hintze provided a useful sanity check when I was

trying to write up what felt like disparate parts of my work. Without each of them, my dissertation would have suffered.

Support staff are essential to any well-functioning enterprise. A few of them I need to call out for particular recognition. Neerja Hajela, the lab manager who has kept me supplied with glassware for the thousands of competitions assays contained in these pages, and who therefore freed up substantial time for me to have something resembling a life while a graduate student. Brian Baer, the computer manager who took care of many technological issues along the way. Connie James, the administrative assistant in the BEACON center, without whom the center would grind to a halt. Darcie Zubeck, the financial officer for the BEACON Center, who has been almost unbelievably kind in helping me with reimbursement protocols.

Within both the Lenski and Ofria labs, I've overlapped with dozens of individuals. Here I list just the most notable of the interactions with them. Noah Ribeck, a postdoc with whom I collaborated on most of my fitness trajectory work in the LTEE. David Bryson, a developer and (former) graduate student with whom I collaborated on all of the Avida work contained in these pages. Jeffrey Barrick, a postdoc who started on the same day in the lab as I did, and who acted as a pseudo-advisor for several years. Jeffrey Morris, a postdoc who started later, but with whom I collaborated on multiple projects not in this dissertation. Caroline Turner, Rohan Maddamsetti, and Alita Burmeister, all fellow Lenski lab students without whose help in our laboratory writing group I may still not have finished writing this dissertation. Emily Dolson and Anya

Vostinar, graduate students in the Ofria lab whose help was instrumental in repeating some of the Avida analyses. Bess Walker, a former student in the Ofria lab, with whom I discussed most of my work over many years; Justin Meyer, a former student in the Lenski lab, for the same. Jeffrey Barrick, Jeffrey Morris, Rohan Maddamsetti, Caroline Turner, Emily Dolson, Brian Connelly, Luis Zaman, David Knoester, Rosangela Canino-Koning, Daniel Mitchell, and Neem Serra, all individuals with whom I collaborated with in my time at Michigan State University on projects not in this dissertation.

Beyond the university, there is also the support of family. I thank my brother, Matthew Wiser, for taking a surprisingly long time before we achieved the point of mutual incomprehensibility in the specifics of our research. And I thank my father, James Wiser, who never objected to the fact that his science-minded son had no interest in becoming a physician.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

ASR: Altered Starting Ratio

BIC: Bayesian Information Criteria

DCC: Different Common Competitor

E. coli: Escherichia coli

LTEE: Long-Term Evolution Experiment

CHAPTER 1: A COMPARISON OF METHODS TO MEASURE FITNESS IN *ESCHERICHIA COLI*

Authors: Michael J. Wiser and Richard E. Lenski

Abstract:

In order to characterize the dynamics of adaptation, it is important to be able to quantify how a population's mean fitness changes over time. Such measurements are especially important in experimental studies of evolution using microbes. The Long-Term Evolution Experiment (LTEE) with *Escherichia coli* provides one such system in which mean fitness has been measured by competing derived and ancestral populations. The traditional method used to measure fitness in the LTEE and many similar experiments, though, is subject to a potential limitation. As the relative fitness of the two competitors diverges, the measurement error increases because the less-fit population becomes increasingly small and cannot be enumerated as precisely. Here, we present and employ two alternatives to the traditional method. One is based on reducing the fitness differential between the competitors by using a common reference competitor from an intermediate generation that has intermediate fitness; the other alternative increases the initial population size of the less-fit, ancestral competitor. We performed a total of 480 competitions to compare the statistical properties of estimates obtained using these alternative methods with those obtained using the traditional method for samples taken over 50,000 generations from one of the LTEE populations.

On balance, neither alternative method yielded measurements that were more precise than the traditional method.

Introduction:

The concept of fitness is central to evolutionary biology. Genotypes with higher fitness will tend to produce more offspring and thereby increase in frequency over time compared to their less-fit competitors. Fitness, however, is often difficult to measure, especially for long-lived organisms. Unlike traits such as color, fitness cannot be observed at a single point in time, but instead it must be measured and integrated across the lifespan of the individuals. Thus, researchers typically measure fitness components – such as the number of seeds produced or offspring fledged – and use them as proxies for fitness.

These limitations can be overcome in experimental evolution studies using microorganisms. Microbes typically have rapid generations and require little space, making them attractive for laboratory-based studies. Replicate populations founded from a common ancestor allow researchers to examine the repeatability of evolutionary changes. Environments can be controlled, reducing uninformative variation between samples or populations and allowing precise manipulations of conditions of interest. Also, one can often freeze microbial populations at multiple points along an evolutionary trajectory and revive them later, allowing direct comparisons between ancestral and derived populations (*1*, *2*). Owing to these advantages, evolution experiments with microbes are becoming increasingly common (*3*–*5*). Thus, it is important to be able to

accurately quantify fitness in these experiments, in order to understand the evolutionary dynamics at work.

One commonly employed method of quantifying microbial fitness is to calculate the maximum growth rate (V_{max}) of a culture growing on its own (*6*–*10*), usually by measuring the optical density of the culture over time. These measurements have the advantages of being simple and fast; a spectrophotometer can measure many samples in a multi-well plate in quick succession, and systems can be programmed to take measurements over the full growth cycle of a culture. However, maximum growth rate is typically only one component of fitness even in the simplest systems (*11*), and hence it provides, at best, only a proxy for fitness.

A second type of fitness measurement comes from studies where microbes are adapting to stressful compounds, such as antibiotics. In these situations, researchers typically quantify the Minimum Inhibitory Concentration (MIC) of the compound, and those organisms with higher MICs are considered to be more fit in environments that contain that compound, as it takes more of the substance to inhibit their growth (*12*, *13*).

A third approach for quantifying fitness in microbial systems—and the approach that most closely corresponds to the meaning of fitness in evolutionary theory—uses a competition assay. The basic approach is to compete one strain or population against another and directly measure their relative contributions to future generations. This approach typically produces a measure of relative, rather than absolute, fitness. Relative fitness is more important than absolute fitness when considering the evolutionary fate of a particular genotype, provided that absolute fitness is high enough to prevent extinction of the entire population (*14*, *15*). Competitive fitness assays, by

measuring the net growth of two different populations, incorporate and integrate differences across the full culture cycle, which may include such fitness components as lag times, exponential growth rates, and stationary phase dynamics in batch culture (*11*, *16*).

Despite their relevance to evolutionary theory, competitive fitness assays sometimes have practical limitations. In particular, and the focus of our paper, these measurements are more precise when the two competitors have similar fitness than when one is substantially more fit than the other. When one competitor is markedly less fit, its abundance will decrease over the course of the competition assay, potentially reaching values low enough that measurement error has a large impact. Thus, as the duration of an evolution experiment increases, and the fitness of the evolved organisms increases relative to the ancestral competitor, the measurement error also tends to increase, as we will show in this study.

We used a population from the Long-Term Evolution Experiment (LTEE) with *Escherichia coli* to investigate whether changes in the methods of performing competition assays – changes meant to reduce the discrepancy in the final abundances of the competitors – would yield more precise fitness measurements. The LTEE has been described in detail elsewhere (*1*, *17*–*19*), and a brief summary is provided in the Materials and Methods section below. Previous work in this system has established that changes in V_{max} explained much, but not all, of the improvement in relative fitness in this system, at least in the early generations (*11*).

Materials and Methods:

Experimental conditions:

The LTEE is an ongoing experiment that began in 1988, and which has now surpassed 50,000 bacterial generations. The experiment uses a Davis Minimal salts medium with 25 μg/mL glucose (DM25), which supports densities of ~3-5 x 10^7 bacteria per mL. Each population is maintained in 10 mL of DM25 in a 50-mL glass Erlenmeyer flask incubated at 37C and shaken at 120 rpm. Every day, each population is diluted 1:100 into fresh media. This dilution sets the number of generations, as the regrowth up to the carrying capacity allows $\log_2 100 \approx 6.64$ cell divisions per day.

Bacterial strains:

The LTEE has 12 populations of *E. coli* (*1*). Six populations were founded by the strain REL606 (*20*) and six by the strain REL607. REL606 is unable to grow on the sugar arabinose (Ara^-); REL607 is an Ara^+ mutant derived from REL606. The DM25 medium does not contain arabinose, and the arabinose-utilization marker is selectively neutral in the LTEE environment [20]. In this study, we use both ancestral strains as well as samples taken from one population, called Ara-1, at generations 500, 1000, 1500, 2000, 5000, 10,000, 15,000, 20,000, 25,000, 30,000, 35,000, 40,000, 45,000, and 50,000. We also use another strain REL11351, which is an Ara^+ mutant of a clone isolated from the 5000-generation sample of population Ara-1.

Fitness measurements:

We quantify fitness in this system as the ratio of the realized growth rates of two populations while they compete for resources in the same flask and under the same environmental conditions used in the LTEE. This calculation is identical to the ratio of the number of doublings achieved by the two competitors. In all cases, we compete samples from the Ara-1 population (including the ancestor REL606) against an Ara^+ competitor (either REL607 or REL11351). We distinguish the two competitors on the basis of their arabinose-utilization phenotypes; Ara^- and Ara^+ cells produce red and white colonies, respectively, on Tetrazolium Arabinose (TA) agar plates (*1*, *21*).

We employ three different methods for measuring fitness in this study. For all three methods, we begin by removing aliquots of the competitors from the vials in which they are stored at –80C into separate flasks containing Luria-Bertani (LB) broth. The cultures grow overnight at 37C and reach stationary phase. We then dilute each culture 100-fold into 0.86% (w/v) saline solution and transfer 100 μL into a flask containing 9.9 mL of DM25. These cultures grow for 24 h under the same conditions as the LTEE, so that all competitors are acclimated to this environment. We then jointly inoculate 100 μL in total of the Ara-1 population sample and the Ara^+ competitor into 9.9 mL of DM25. We immediately take an initial 100-μL sample of this mixture, dilute it in saline solution, and spread the cells onto a TA plate. The competition mixture is then incubated in the same conditions as the LTEE for 24 h, at which point we take a final 100-μL sample, dilute it, and spread the cells onto a TA plate. We count each competitor on the TA plates, and multiply the numbers by the appropriate dilution factor to determine their initial and final population sizes. We calculate fitness as

$$w = \frac{\ln\left(\frac{A_f}{A_i}\right)}{\ln\left(\frac{B_f}{B_i}\right)}$$

where *w* is fitness, *A* and *B* are the population sizes of the two competitors, subscripts *i* and *f* indicate the initial and final time points in the assay; here, ln refers to the natural logarithm in order to reflect population growth, although the ratio used to express fitness is insensitive to the choice of base used.

For the Traditional method, we measure the relative fitness of the evolved population samples against the Ara^+ ancestor, REL607. We inoculate the competition flasks with 50 μL (an equal volumetric ratio) of each competitor. This method has been used extensively in evaluating fitness in the LTEE (*1*, *2*).

The Altered Starting Ratio (ASR) method also uses the ancestral Ara^+ strain as the common competitor. However, we inoculate the competition flasks with 20 μL of the evolved population and 80 μL of the ancestral population, leading to an initial 1:4 volumetric ratio. This difference in the starting ratio increases the population size of the ancestor at the end of the competition assay, which reduces the problem of small numbers when the ancestor is much less fit than the evolved population. The initial ratio is not so extreme, however, that it is difficult to enumerate the evolved population at the start of the competition assay. We attempted to keep total plate counts around a few hundred colonies, with at least 20 of the minority competitor, to reliably estimate population densities (*22*), and we chose this initial ratio with that objective in mind. It seemed particularly important to increase the final count of the ancestral population in

the context of our fitness measurements; smaller numbers are subject to increased sampling error, and the realized growth rate of the ancestor is the denominator when calculating the relative fitness of the evolved population, which can magnify the measurement error. More extreme ratios have been used in some experiments testing invasion when rare (*23*), but these ratios would result in minority populations of fewer than 20 colonies per plate; therefore, they were not tested in this study. It is also important to note that we test different ratios of culture volume, not specifically of different numbers of starting cells per se; differences in carrying capacity between the ancestral and evolved bacteria (*11*, *17*) and stochastic sampling effects will prevent the initial ratio of cell numbers from precisely matching these volumetric ratios.

Using the Different Common Competitor (DCC) method, we compete the evolved population samples against the marked clone from generation 5,000, rather than against the marked ancestor. We chose a 5,000-generation clone because its fitness was near the geometric mean of the expected fitness values spanning generations 0 to 50,000, and thus it might reduce the overall disparity in population counts across the full time series being considered. We inoculate the competitions with equal volumes (50 μL each) of the Ara-1 population sample and reference competitor. We considered that this method might increase the precision of our fitness measurements because the ratios used in the fitness calculation tend to be more precise as they approach 1.

We selected 15 time points from the focal population Ara-1 to evaluate these three methods: generations 0, 500, 1,000, 1,500, 2,000, 5,000, 10,000, 15,000, 20,000, 25,000, 30,000, 35,000, 40,000, 45,000, and 50,000. We ran competitions as complete blocks; each block included one competition for each time point using each method,

plus an additional competition (see below) used as a scaling factor to compare the methods. We performed a total of 10 replicate blocks, and so there were a total of 450 competition assays to measure fitness (3 methods x 15 time points x 10 blocks) plus an additional 30 assays to generate the scaling factors.

A scaling factor was necessary for comparing the DCC method with the Traditional and ASR methods, because the DCC method measured fitness relative to a different competitor than the ancestor used for the other two methods. To calculate this scaling factor, we performed an additional competition between the Ara^- ancestor (REL606) and the Ara^+ reference competitor (either REL607 or REL11351) for each method in every block. We then divided the fitness values from all of the competition assays for a given method and block by the fitness value that served as the scaling factor. We did not otherwise include the scaling-factor competitions in our data analysis. We applied the same procedure to all three methods to ensure consistency, although adjusting for the scaling factor was not otherwise required for the Traditional and ASR methods.

The data and analysis scripts are available at the Dryad Digital Depository (doi: http://dx.doi.org/10.5061/dryad.4875k). The data obtained using the Traditional method previously appeared in (*19*); the data for the ASR and DCC methods, as well as all of the analyses in this article, are new .

Statistical methods:

We performed statistical analyses in R version 2.14.1. We fit the fitness trajectories using nonlinear least-squares regression, as implemented with the nls()

function. We performed ANOVAs using the aov() function. For the single-generation ANOVAs, Method was a fixed factor and Block was a random factor. For the combined ANOVA, Generation was included as a fixed factor.

Bootstrapping:

We employed a bootstrap procedure to compare the differences between the coefficients of variation in our three methods to a null distribution. We sampled the total dataset with replacement, to produce 3 datasets of equal size, each containing 10 measurements at each generation. We then fit a linear regression of the coefficient of variation against time (i.e., generation) to each of the 3 datasets. We then summed the squares of the differences between each pairwise combination of the 3 linear regressions over all 15 time points when fitness was measured. We repeated this entire procedure 1,000,000 times, and we compared the observed sum of the squared differences to this distribution.

Results and Discussion:

There are two fundamental ways in which these different methods could produce meaningfully different results. One way is that different methods could produce significantly different fitness estimates. In that case, we would need additional information or another criterion to determine which method was superior. The other way is that different methods could have different levels of precision; that is, one method may have significantly less variation in measured values across replicate

assays than another. In this case, the method with the greater precision would clearly be preferred.

Figure 1.1 shows the results of our fitness assays for all three methods, with trajectories fit to the data obtained using each method. These trajectories are in the form of an Offset Power Law:

$w = (bT + 1)^a$,

where w is fitness, T is time in generations, and *a* and *b* are model parameters, as derived in [2]. All three methods produce virtually identical fitness trajectories. S1 Table shows the results of ANOVAs performed at each generation to test for variation among the three methods in the mean fitness values they produce; the effect of Method was not significant in any of the 15 tests, even without accounting for multiple tests. From these results, we conclude that the three methods do not produce meaningfully different estimates of mean fitness.

Next, we calculated the coefficient of variation (i.e., the standard deviation divided by the mean) for each method at each time point to determine whether they differed in their precision. We then constructed a linear model of the coefficient of variation as a response to time (i.e., generation) and method. Figure 1.2 shows the data and linear model fit to the coefficients of variation for all three methods. Table 1.1 presents the ANOVA table for this model. There is a highly significant tendency for the coefficient of variation to increase in later generations, as the evolving bacteria become progressively more fit, as discussed in the Introduction. However, the effect of Method was not significant as a predictor of the coefficient of variation, although a p-value of

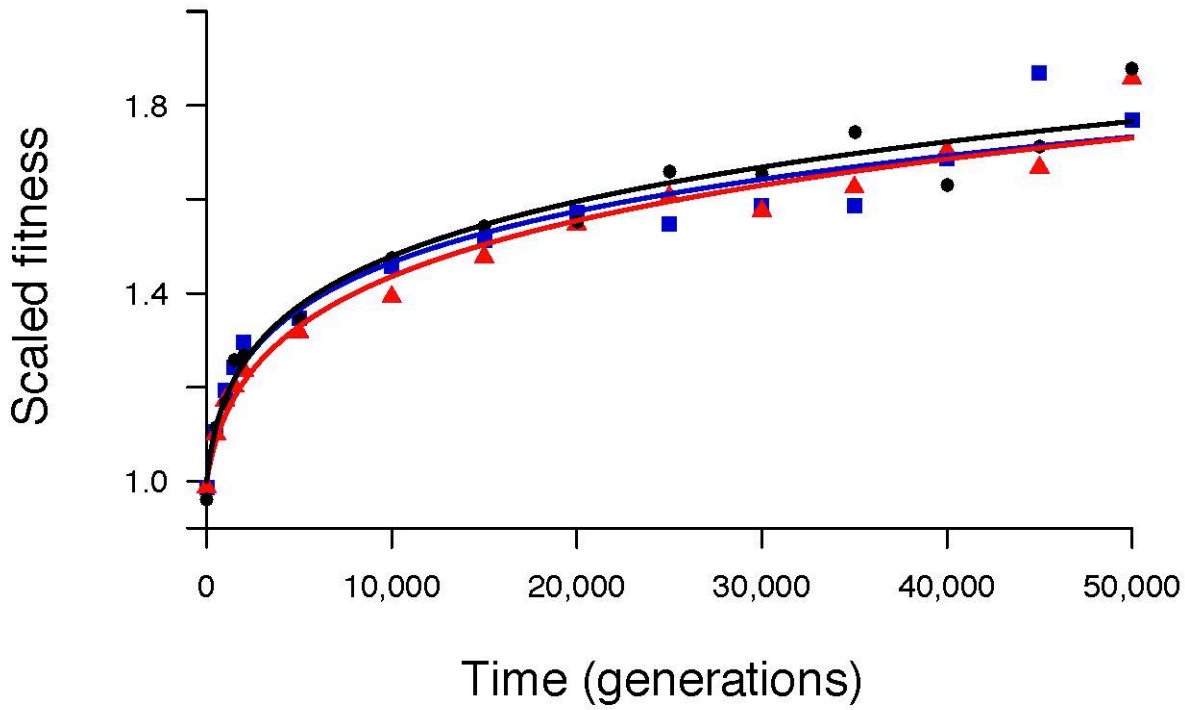

Figure 1.1: **Fitness trajectories over time**. Fitness trajectories for each method, shown separately, have the form $w = (bT + 1)^a$, where w is fitness, T is time in generations, and *a* and *b* are model parameters. Black circles and curve show the Traditional method; blue squares and curve show the ASR method; red triangles and curve show the DCC method.

0.0762 is suggestive. On inspection of the data (Figure 1.2), it is clear that any difference between the methods is driven by the ASR method having a higher coefficient of variation – and thus lower precision – in early generations. Indeed, when we removed the ASR method from the analysis and performed an ANOVA on the remaining data, there was no suggestion of any difference between the Traditional and DCC methods (Table 1.2, p = 0.8802).

	df	SS	MS	F	p
Time	1	0.03672	0.03672	69.664	<0.0001
Method	2	0.00289	0.00145	2.743	0.0762
Residuals	41	0.21610	0.00053		

Table 1.1: **ANOVA on the coefficient of variation across time and comparing the three methods used to estimate fitness.**

	df	SS	MS	F	p
Time	1	0.03068	0.03068	70.035	<0.0001
Method	1	0.00001	0.00001	0.023	0.8802
Residuals	27	0.01183	0.00044		

Table 1.2: **ANOVA on the coefficient of variation across time and comparing the Traditional and DCC methods.**

We can also express the differences between these methods as follows. The regression line for the coefficient of variation based on the ASR method is always higher than at least one of the other two methods (Figure 1.2), and therefore it is never the best

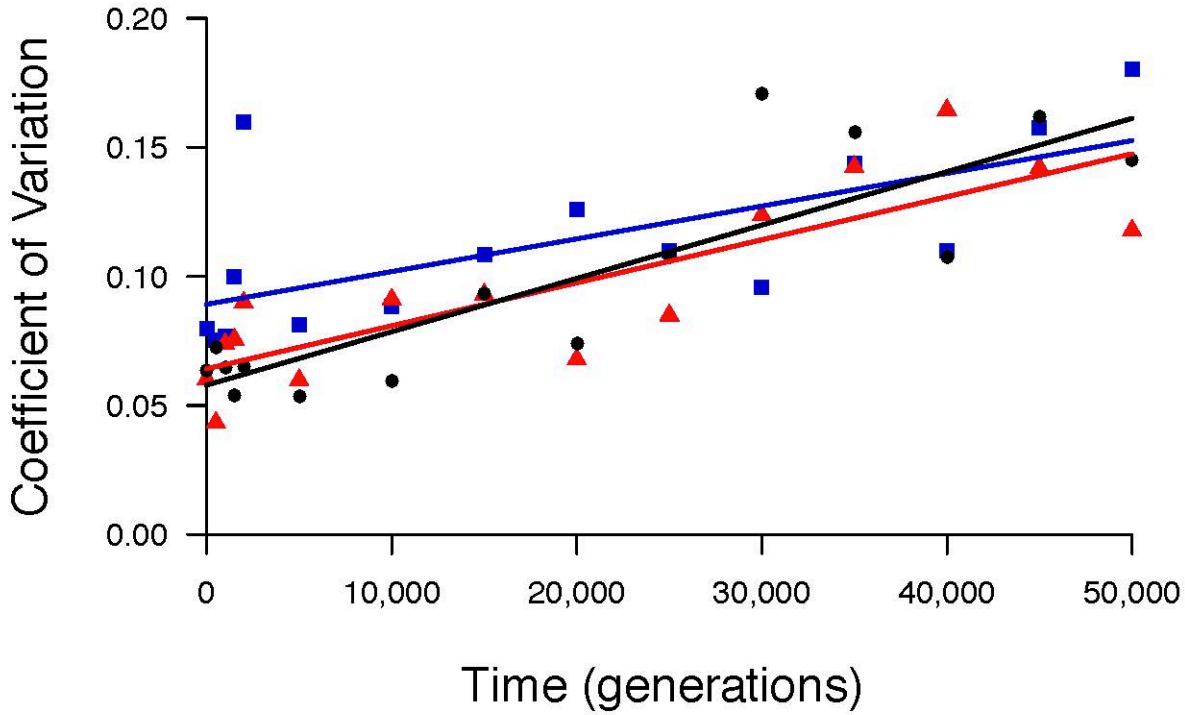

Figure 1.2: **Coefficient of variation over time**. Lines are linear regressions on the relevant data. Black circles and line show the Traditional method; blue squares and line show the ASR method; red triangles and line show the DCC method. Figure S1.1 shows the confidence bands associated with each regression line.

method, at least for the system and generations analyzed here. By contrast, the Traditional and DCC methods yield coefficients of variation, as inferred from the regression lines, that are very similar and always within the 95% confidence interval of one another (Figure S1.1). Which of these two methods gave a lower point estimate of the coefficient of variation varied over time, but the difference was not significant (Table 1.2).

An alternative way to assess whether the differences in the coefficient of variation between the methods are statistically significant involves bootstrapping the data, as detailed in the Methods section. Figure 1.3 shows that the observed differences in the coefficient of variation among the three methods are no greater than would be expected by chance if there were no differences among the methods.

Over the range of fitness changes that we observed in the LTEE (i.e., from 1 to ~1.8), neither alternative method for assaying fitness (ASR or DCC) outperformed the Traditional method. Given its extensive prior use in this study system [1,2,17], we therefore prefer to use the Traditional method for fitness competitions that span this range. It is important to note, however, that the ASR or the DCC method might turn out to have higher precision in systems that exhibit larger fitness changes than the system studied here, as suggested by the regression lines in Figure 1.2. The LTEE has, to our knowledge, run for many more generations than any other evolution experiment, but the extent of fitness improvements has been less than that seen in some other shorter-duration experiments. The relatively limited fitness gains that have occurred during the LTEE reflect the fact that the experimental environment is quite benign; also, the ancestor of the LTEE had been studied by microbiologists for many decade (*24*) and

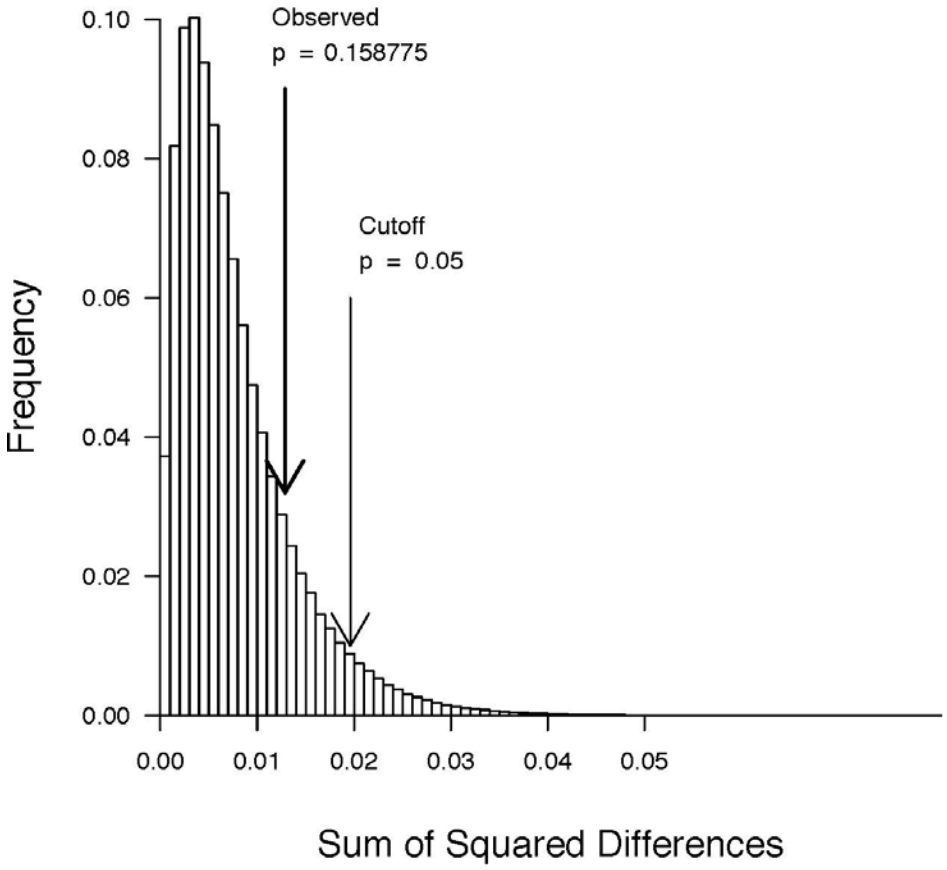

Figure 1.3: **Histogram of bootstrap analysis**. Histogram showing the distribution for the bootstrapped sums of squared differences in the coefficient of variation for 3 arbitrary groupings of the combined data. The dark arrow indicates the difference for the actual grouping of the 3 methods employed. The light arrow shows the most extreme 5% of the sums of the squared differences.

was thus probably already well-adapted to general laboratory conditions. Other experiments conducted for fewer generations, but performed under more stressful conditions or founded by less-fit ancestors, might reach fitness differences where these or other alternative methods would be helpful. Table 3 summarizes the duration and range of fitness improvements reported in a number of other evolution experiments that used a variety of microorganisms including bacteria, fungi, and viruses (see also Table 2.3 in (*25*)). We have included values for both relative fitness, W_f / W_i, and the difference between final and initial fitness values, $W_f - W_i$, when the latter was reported in the paper cited. The value of $W_f - W_i$ necessarily depends on the time frame of the experiment, whereas W_f / W_i is a dimensionless number and thus readily compared across experiments.

Reference	Organism	Generations	W_f / W_i	$W_f - W_i$
This study	*E. coli*	50,000	1.88	3.5 / day
(*26*)	*E. coli* at 32C	2,000	1.10	
	E. coli at 42C	2,000	1.19	
(*27*)*	*E. coli*	1,100	1.98	0.23 / h
(*28*)**	*Saccharomyces cerevisiae*	300	1.80	
(*29*)	*Aspergillus nidulans*	800	1.48	
(*30*)	phage Φ6 with bottleneck = 10	100	1.26	
	phage Φ6 with bottleneck = 1,000	40	2.03	
(*31*)	phage G4	180	1.18	3.8 / h
	phage ID2	600	2.55	13.5 / h

* Mean calculated from four replicate populations
** Value estimated from figure
W_f is the fitness at the end of the evolution experiment.
W_i is the fitness at the start of the evolution experiment.

Table 1.3: **Selected evolution experiments**

Conclusions:

We performed 480 assays to compare three different methods for estimating the relative fitness of bacterial competitors. The three methods generated results that were not meaningfully or significantly different in terms of either their mean values or dispersion. The only suggestion of a meaningful difference was that the ASR method appeared worse than the other two methods in the early generations, when the fitness gains of the evolved bacteria were still fairly small. Therefore, we see no compelling reason to adopt one of the alternatives to the Traditional method when analyzing systems that have achieved fitness gains less than or similar to those measured in the LTEE over its first 50,000 generations. When expected relative fitness values are much greater than 1.8, or when fitness differences are compounded for more generations, researchers may need to consider using one of these or other alternative methods.

Acknowledgments:

We thank Caroline Turner, Amy Lark, Rohan Maddamsetti, and Christopher Strelioff for helpful discussions during manuscript preparation, two reviewers for constructive suggestions, and Neerja Hajela for technical assistance.

APPENDIX

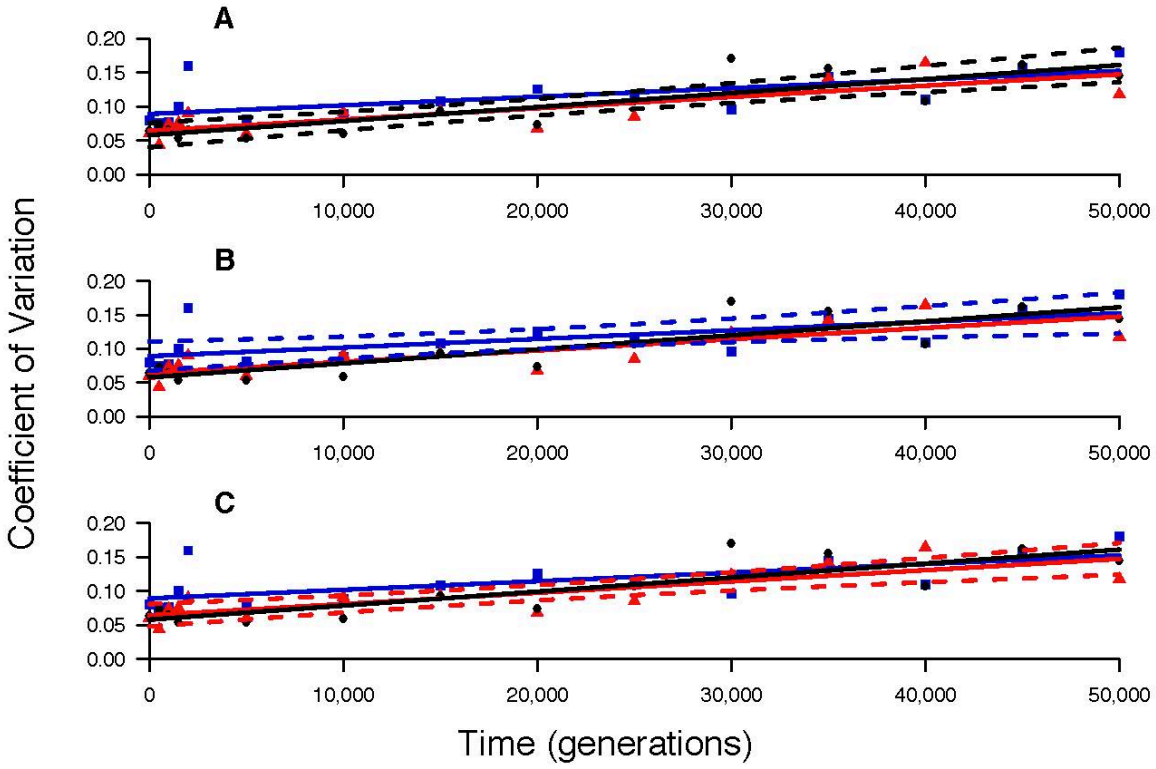

Figure S1.1: **Temporal trends in coefficient of variation**. Temporal trends in the coefficient of variation across replicate assays for the three different methods used to measure fitness. Black circles show the Traditional method; blue squares show the ASR method; red triangles show the DCC method. The solid colored lines show the linear regressions based on the corresponding data. The dashed colored curves show the 95% confidence bands for the regressions for the three methods: A) Traditional, B) ASR, and C) DCC. The points and regression lines are the same across all three panels, but the confidence bands are shown separately for clarity.

		df	SS	MS	F	p
Generation 0	Method	2	0.00452	0.00226	0.488	0.6219
	Block	9	0.03757	0.00417	0.901	0.5441
	Residuals	18	0.08338	0.00463		
Generation 500	Method	2	0.00097	0.00049	0.090	0.9147
	Block	9	0.04394	0.00488	0.900	0.5454
	Residuals	18	0.09770	0.00543		
Generation 1,000	Method	2	0.00377	0.00188	0.254	0.7787
	Block	9	0.06030	0.00670	0.902	0.5437
	Residuals	18	0.13375	0.00743		
Generation 1,500	Method	2	0.01654	0.00827	0.931	0.4123
	Block	9	0.09422	0.01047	1.179	0.3646
	Residuals	18	0.15989	0.00888		
Generation 2,000	Method	2	0.01747	0.00874	0.550	0.5863
	Block	9	0.27162	0.03018	1.900	0.1178
	Residuals	18	0.28588	0.01588		
Generation 5,000	Method	2	0.00561	0.00280	0.326	0.7258
	Block	9	0.05552	0.00617	0.718	0.6866
	Residuals	18	0.15460	0.00859		
Generation 10,000	Method	2	0.03815	0.01908	1.881	0.1812
	Block	9	0.18114	0.02013	1.984	0.1032
	Residuals	18	0.18257	0.01014		
Generation 15,000	Method	2	0.02064	0.01032	0.368	0.6973
	Block	9	0.09212	0.01024	0.365	0.9374
	Residuals	18	0.50501	0.02806		
Generation 20,000	Method	2	0.00341	0.00170	0.122	0.8858
	Block	9	0.31731	0.35257	2.527	0.0450
	Residuals	18	0.25116	0.01395		
Generation 25,000	Method	2	0.06087	0.03043	1.286	0.3006
	Block	9	0.28899	0.03211	1.357	0.2772
	Residuals	18	0.42593	0.02366		
Generation 30,000	Method	2	0.03426	0.01713	0.595	0.5623
	Block	9	0.74363	0.08263	2.868	0.0273
	Residuals	18	0.51864	0.02881		
Generation 35,000	Method	2	0.13333	0.06667	1.163	0.3349
	Block	9	0.58220	0.06469	1.129	0.3929
	Residuals	18	1.03167	0.05732		
Generation 40,000	Method	2	0.02934	0.01467	0.392	0.6813
	Block	9	0.61796	0.06866	1.835	0.1306
	Residuals	18	0.67359	0.03742		

Table S1.1: **ANOVAs of fitness for three methods, by generation**. Analyses of variance of measured fitness values for the three methods, analyzed separately for the various generations examined.

Table S1.1 (cont'd)

		df	SS	MS	F	p
Generation 45,000	Method	2	0.22400	0.11200	2.113	0.1499
	Block	9	1.02081	0.11342	2.140	0.0811
	Residuals	18	0.95425	0.05301		
Generation 50,000	Method	2	0.06732	0.03366	0.057	0.5764
	Block	9	0.94686	0.10521	1.776	1.4320
	Residuals	18	1.06627	0.05924		

This chapter was originally published as:

Wiser MJ, Lenski RE (2015) A Comparison of Methods to Measure Fitness in *Escherichia coli*. PLoS ONE 10(5): e0126210. doi: 10.1371/journal.pone.0126210

REFERENCES

REFERENCES

1. R. E. Lenski, M. R. Rose, S. C. Simpson, S. C. Tadler, Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138**, 1315–1341 (1991).

2. M. J. Wiser, N. Ribeck, R. E. Lenski, Long-term dynamics of adaptation in asexual populations. *Science*. **342**, 1364–1367 (2013).

3. S. F. Elena, R. E. Lenski, Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet*. **4**, 457–469 (2003).

4. T. J. Kawecki *et al.*, Experimental evolution. *Trends Ecol. Evol.* **27**, 547–560 (2012).

5. J. E. Barrick, R. E. Lenski, Genome dynamics during experimental evolution. *Nat Rev Genet*. **14**, 827–839 (2013).

6. W. Paulander, S. Maisnier-Patin, D. I. Andersson, Multiple mechanisms to ameliorate the fitness burden of mupirocin resistance in Salmonella typhimurium. *Mol. Microbiol.* **64**, 1038–1048 (2007).

7. A. I. Nilsson *et al.*, Reducing the fitness cost of antibiotic resistance by amplification of initiator tRNA genes. *Proc. Natl. Acad. Sci.* **103**, 6976–6981 (2006).

8. L. Sandegren, A. Lindqvist, G. Kahlmeter, D. I. Andersson, Nitrofurantoin resistance mechanism and fitness cost in Escherichia coli. *J. Antimicrob. Chemother.* **62**, 495–503 (2008).

9. D. I. Andersson, D. Hughes, Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat Rev Micro*. **8**, 260–271 (2010).

10. K. Walkiewicz *et al.*, Small changes in enzyme function can lead to surprisingly large fitness effects during adaptive evolution of antibiotic resistance. *Proc. Natl. Acad. Sci.* **109**, 21408–21413 (2012).

11. F. Vasi, M. Travisano, R. E. Lenski, Long-term experimental evolution in Escherichia coli. II. Changes in life-history traits during adaptation to a seasonal environment. *Am. Nat.* **144**, 432–456 (1994).

12. D. M. Weinreich, N. F. Delaney, M. A. DePristo, D. L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*. **312**, 111–114 (2006).

13. A. Ripoll *et al.*, In vitro selection of variants resistant to β-lactams plus β-lactamase inhibitors in CTX-M β-lactamases: Predicting the in vivo scenario? *Antimicrob. Agents Chemother.* **55**, 4530–4536 (2011).

14. G. Bell, Evolutionary rescue and the limits of adaptation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368** (2012), doi:10.1098/rstb.2012.0080.

15. A. F. Bennett, R. E. Lenski, Evolutionary adaptation to temperature II. Thermal niches of experimental lines of Escherichia coli. *Evolution*. **47**, 1–12 (1993).

16. S. C. Sleight, R. E. Lenski, Evolutionary adaptation to freeze-thaw-growth cycles in Escherichia coli. *Physiol. Biochem. Zool.* **80**, 370–385 (2007).

17. R. E. Lenski, M. Travisano, Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci.* **91**, 6808–6814 (1994).

18. J. E. Barrick *et al.*, Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*. **461**, 1243–1247 (2009).

19. S. Wielgoss *et al.*, Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl. Acad. Sci.* **110**, 222–227 (2013).

20. F. W. Studier, P. Daegelen, R. E. Lenski, S. Maslov, J. F. Kim, Understanding the differences between genome sequences of Escherichia coli B strains REL606 and BL21(DE3) and comparison of the E. coli B and K-12 genomes. *J. Mol. Biol.* **394**, 653–680 (2009).

21. R. E. Lenski, Experimental studies of pleiotropy and epistasis in Escherichia coli. I. Variation in competitive fitness among mutants resistant to virus T4. *Evolution*. **42**, 425–432 (1988).

22. R. S. Breed, W. D. Dotterrer, The number of colonies allowable on satisfactory agar plates. *J. Bacteriol.* **1**, 321–331 (1916).

23. R. F. Inglis, S. West, A. Buckling, An experimental study of strong reciprocity in bacteria. *Biol. Lett.* **10** (2014), doi:10.1098/rsbl.2013.1069.

24. P. Daegelen, F. W. Studier, R. E. Lenski, S. Cure, J. F. Kim, Tracing ancestors and relatives of Escherichia coli B, and the derivation of B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394**, 634–643 (2009).

25. R. Kassen, *Experimental evolution and the nature of biodiversity* (Roberts and Company Publishers, Inc, Greenwood Village, Colorado, 2014).

26. A. F. Bennett, R. E. Lenski, J. E. Mittler, Evolutionary adaptation to temperature. I. Fitness responses of Escherichia coli to changes in its thermal environment. *Evolution*. **46**, 16–30 (1992).

27. T. Conrad *et al.*, Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol.* **10**, R118 (2009).

28. M. R. Goddard, H. C. J. Godfray, A. Burt, Sex increases the efficacy of natural selection in experimental yeast populations. *Nature*. **434**, 636–640 (2005).

29. S. E. Schoustra, T. Bataillon, D. R. Gifford, R. Kassen, The properties of adaptive walks in evolving populations of fungus. *PLoS Biol*. **7**, e1000250 (2009).

30. C. L. Burch, L. Chao, Evolution by small steps and rugged landscapes in the RNA virus ϕ6. *Genetics*. **151**, 921–927 (1999).

31. D. R. Rokyta, Z. Abdo, H. A. Wichman, The genetics of adaptation for eight microvirid bacteriophages. *J. Mol. Evol.* **69**, 229–239 (2009).

CHAPTER 2: LONG-TERM DYNAMICS OF ADAPTATION IN ASEXUAL POPULATIONS

Authors: Michael J. Wiser, Noah Ribeck, and Richard E. Lenski

Abstract:

Experimental studies of evolution have increased greatly in number in recent years, stimulated by the growing power of genomic tools. However, organismal fitness remains the ultimate metric for interpreting these experiments, and the dynamics of fitness remain poorly understood over long time scales. Here, we examine fitness trajectories for 12 *Escherichia coli* populations during 50,000 generations. Mean fitness appears to increase without bound, consistent with a power law. We also derive this power-law relation theoretically by incorporating clonal interference and diminishing-returns epistasis into a dynamical model of changes in mean fitness over time.

Main Text:

The dynamics of evolving populations are often discussed in terms of movement on an adaptive landscape, where peaks and valleys are states of high and low fitness, respectively. There is considerable interest in the structure of these landscapes (*1*–*7*). Recent decades have seen tremendous growth in experiments using microbes to address fundamental questions about evolution (*8*), but most have been short in duration. The Long-Term Evolution Experiment

(LTEE) with *Escherichia coli* provides the opportunity to characterize the dynamics of adaptive evolution over long periods under constant conditions (*1*, *9*, *10*). Twelve populations were founded from a common ancestor in 1988 and have been evolving for >50,000 generations, with samples frozen every 500 generations. The frozen bacteria remain viable, and we use this "fossil record" to assess whether fitness continues to increase and to characterize mean fitness trajectories (see Appendix: Material and Methods).

We first performed 108 competitions, in the same conditions as the LTEE, between samples from nine populations at 40,000 and 50,000 generations against marked 40,000-generation clones (see Appendix: Material and Methods). Three populations were excluded for technical reasons (see Appendix: Material and Methods). Fitness was quantified as the dimensionless ratio of the competitors' realized growth rates. Most populations experienced significant improvement (Figure 2.1A), and the grand mean fitness increased by 3.0% (Figure 2.1B).

To examine the shape of the fitness trajectory, we competed samples from all 12 populations and up to 41 time points against the ancestor (see Appendix: Material and Methods). We compared the fit of two alternative models with the fitness trajectories. The hyperbolic model describes a decelerating trajectory with an asymptote. The power law also decelerates (provided the exponent is <1), but fitness has no upper limit.

Hyperbolic model

$$\overline{\omega} = 1 + at/(t+b)$$

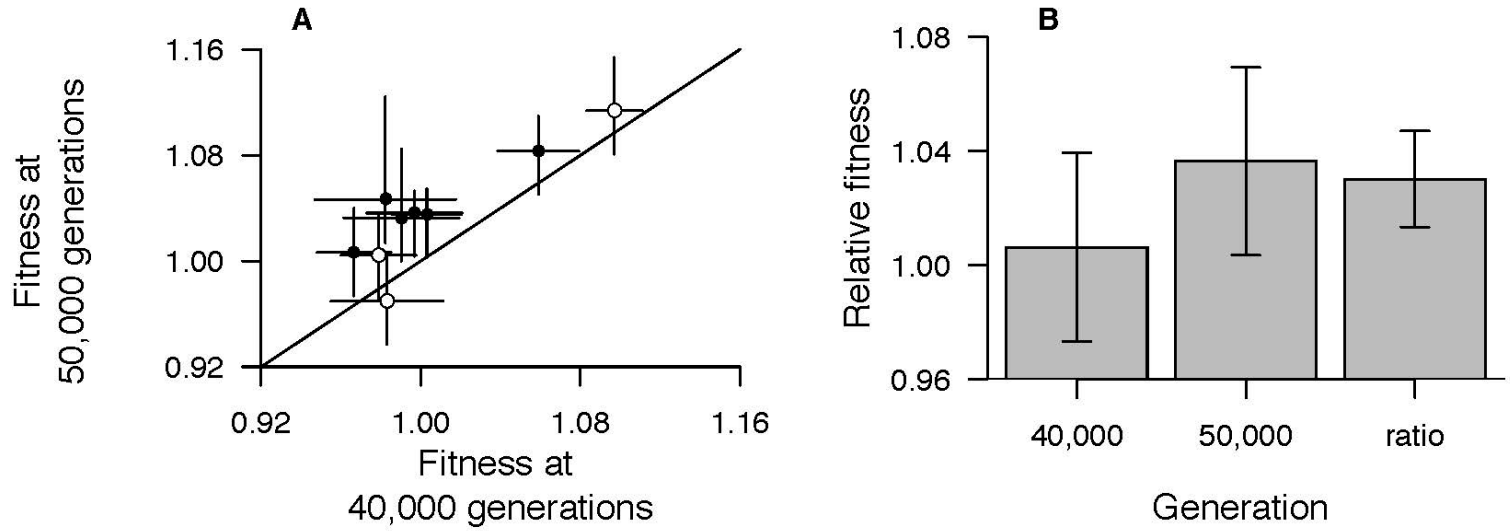

Figure 2.1: **Fitness changes in nine *E. coli* populations between 40,000 and 50,000 generations**. (**A**) Filled symbols: six populations whose improvement was significant ($P < 0.05$); open symbols: three populations without significant improvement. (**B**) Grand-mean fitness at 40,000 and 50,000 generations relative to 40,000-generation competitor and the ratio of means showing overall gain. Error bars are 95% confidence limits based on replicate assays (A) or populations (B).

Power law

$$\overline{\omega} = (bt + 1)^a$$

Mean fitness is $\overline{\omega}$, time in generations is *t*, and each model has two parameters, *a* and *b*. Both models are constrained such that the ancestral fitness is 1, hence the offset of +1 in the power law. The hyperbolic model was fit to the first 10,000 generations of the LTEE (*9*), but others suggested an alternative nonasymptotic trajectory (*11*). The grand mean fitness values and the trajectory for each model are shown in Figure 2.2A and the individual populations in Figure S2.1. Both models fit the data very well; the correlation coefficients for the grand means and model trajectories are 0.969 and 0.986 for the hyperbolic and power-law models, respectively. When Bayesian information criterion scores (see Appendix: Material and Methods) are used, the power law outperforms the hyperbolic model with a posterior odds ratio of ~30 million (Table S2.1). The superior performance of the power law also holds when populations are excluded because of incomplete time series or evolved hypermutability (Table S2.1). The power law provides a better fit to the grand-mean fitness than the hyperbolic model in early, middle, and late generations (Figure S2.2). The power law is supported (odds ratios >10) in six individual populations, whereas none supports the hyperbolic model to that degree (Table S2.2). The power law also predicts fitness gains more accurately than the hyperbolic model. When fit to data for the first 20,000 generations only, the hyperbolic model badly underestimates later measurements, whereas the power-law trajectory predicts them accurately (Figure 2.2B and Figure S2.3).

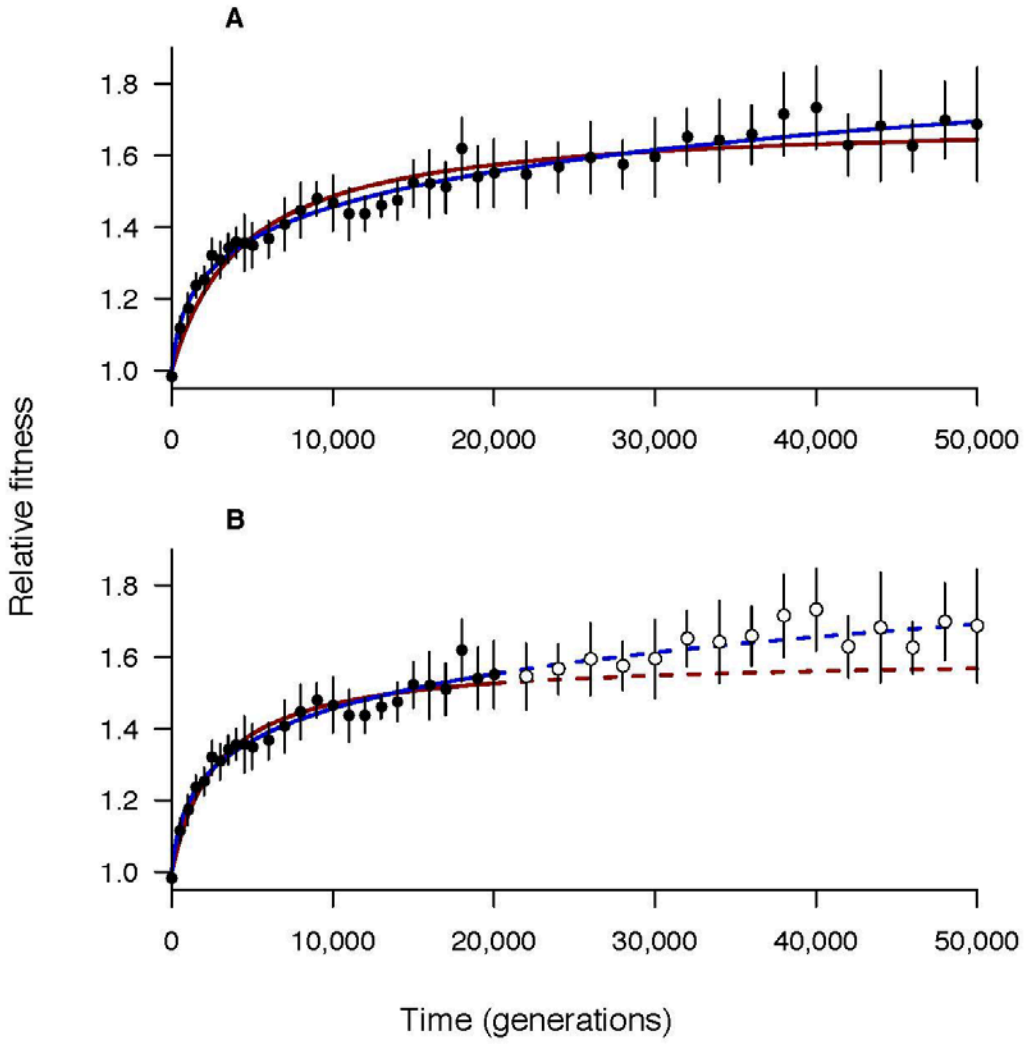

Figure 2.2: **Comparison of hyperbolic and power-law models**. (**A**) Hyperbolic (red) and power-law (blue) models fit to the set of mean fitness values (black symbols) from all 12 populations. (**B**) Fit of hyperbolic (solid red) and power-law (solid blue) models to data from first 20,000 generations only (filled symbols), with model predictions (dashed lines) and later data (open symbols). Error bars are 95% confidence limits based on the replicate populations.

The power law describes the fitness trajectories well, but it is not explanatory. We have derived a dynamical model of asexual populations with clonal interference and diminishing-returns epistasis, which generates mean-fitness trajectories that agree well with the experimental data. Clonal interference refers to competition among organisms with different beneficial mutations, which impedes their spread in asexual populations (*12–15*). Diminishing-returns epistasis occurs when the marginal improvement from a beneficial mutation declines with increasing fitness (*5*, *6*). We outline key points of the model below (see Appendix: Material and Methods).

We used a coarse-grained approach that describes the magnitudes and time scales of fixation events (*12*). Beneficial mutations of advantage s are exponentially distributed with probability density $\alpha e^{-\alpha s}$, where $1/\alpha$ is the mean advantage. This distribution is for mathematical convenience; the theory of clonal interference is robust to the form of the distribution (*12*). We assume that deleterious mutations do not appreciably affect the dynamics; deleterious mutations occur at a higher rate than beneficial mutations, but the resulting load is very small relative to the fitness increase measured over the course of the LTEE (*16*).

We assume the distribution of available benefits declines after a mutation with advantage $\langle s \rangle$ fixes, such that α increases by a factor linearly related to $\langle s \rangle$:

where $g > 0$ is the diminishing-returns parameter, $\langle s_n \rangle$ is beneficial effect of the nth fixed mutation, and α_n is α after n fixations. Then, the mean fitness of an asexual population adapting to a constant environment is approximated by (see

Appendix: Material and Methods): where $\langle s_1 \rangle$ and $\langle t_1 \rangle$ are the beneficial effect and fixation time, respectively, for the first fixed mutation.

Comparing this formula with the power law, $g = 1/2a$. The value of g estimated for the six populations that retained the low ancestral mutation rate throughout 50,000 generations is 6.0 (95% confidence interval 5.3 to 6.9). In the LTEE, the beneficial effect of the first fixation, $\langle s_1 \rangle$, is typically ~0.1 (*1*, *9*, *10*). It follows that the distribution of beneficial effects immediately after the first fixation is shifted such that the mean advantage is $1/(1+g\langle s_1 \rangle) \approx 63\%$ of its initial value (see Appendix: Material and Methods). This estimate of g also accords well with epistasis observed for early mutations in one of the populations (Figure S2.4). In principle, g might vary among populations if some fixed mutations lead to regions of the fitness landscape with different epistatic tendencies (*17*). However, an analysis of variance shows no significant heterogeneity in g among the six populations that maintained the ancestral mutation rate ($p = 0.3478$) (Table S2.3). The g values tend to be lower for several populations that evolved hypermutability (Table S2.4). However, these fits are confounded by the change in mutation rate; we show below that it is not necessary to invoke a difference in diminishing-returns epistasis between the hypermutable populations and those that retained the low ancestral mutation rate.

Diminishing-returns epistasis generates the power-law dynamics through the relation between a and g. Clonal interference affects the dynamics through the parameter b, which depends on $\langle s_1 \rangle$ and $\langle t_1 \rangle$, which in turn are functions of

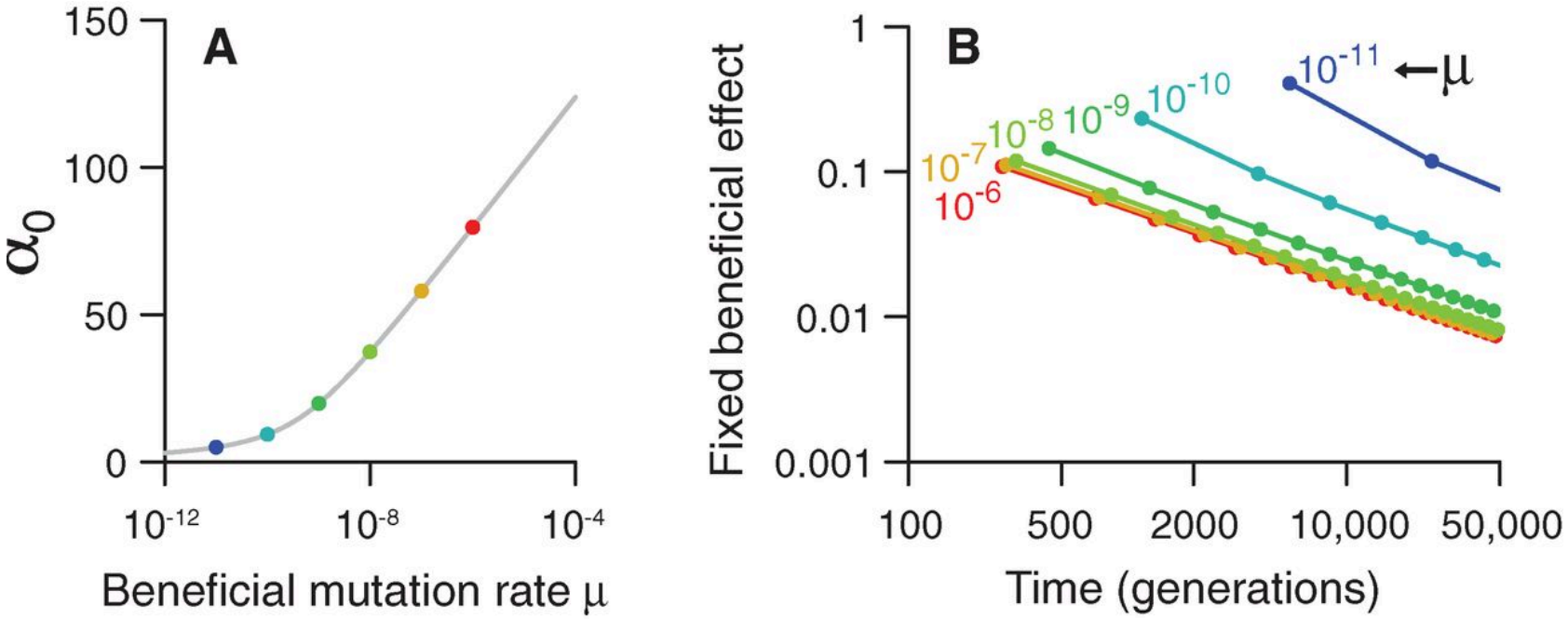

Figure 2.3: **Theoretical model generating power-law dynamics**. (**A**) Parameter pairs for μ and α_0 that match best fit of power law to fitness trajectories for populations that retained ancestral mutation rate for 50,000 generations. (**B**) Expected times and beneficial effects of successive fixations for different pairs that match the best fit. The α_0 values corresponding to each μ are shown in (A). In both panels $g = 6.0$, and $N = 3.3 \times 10^7$.

the population size *N*, beneficial mutation rate μ, and initial mean beneficial effect $1/\alpha_0$ (see Appendix: Material and Methods). For the LTEE, $N = 3.3 \times 10^7$, which takes into account the daily dilutions and regrowth (*1*). However, μ and α_0 are unknown. Pairs of values that all match the best fit to the populations that retained the low mutation rate are shown in Figure 2.3A. The expected values for beneficial effects and fixation times across a range of pairs are shown in Figure 3.3B. The dynamics are similar among pairs with high beneficial mutation rates ($\mu > 10^{-8}$), giving $\langle s_1 \rangle \approx 0.1$ and $\langle t_1 \rangle \approx 300$ generations for the first fixation, which agree well with observations from the LTEE (*1*, *9*, *10*). At lower values of μ, adaptation becomes limited by the supply of beneficial mutations, and fixation times are inconsistent with the LTEE. This model also predicts that the rate of adaptation decelerates more sharply than the rate of genomic evolution (Figure S2.5), which is qualitatively consistent with observations (*10*) (see Appendix: Material and Methods). The model assumes that individual beneficial mutations sweep sequentially, although "cohorts" of beneficial mutations may co-occur, especially at high μ (*14*, *15*, *18*) (see Appendix: Material and Methods). However, the inferred role of diminishing returns in generating population mean-fitness dynamics is unaffected by this complication, because the power-law exponent is independent of μ. Moreover, we have verified by numerical simulations that co-occurring beneficial mutations have no appreciable affect on long-term fitness trajectories over the range of parameters considered here (Figure S2.6).

Six populations evolved hypermutator phenotypes that increased their point-mutation rates by ~100-fold (see Appendix: Material and Methods). Three

of them became hypermutable early in the LTEE (between ~2500 and ~8500 generations) and had measurable fitness trajectories through at least 30,000 generations (Table S2.2). Our model predicts these populations should adapt faster than those that retained the ancestral mutation rate. We pooled the data from these early hypermutators and confirmed that their composite fitness trajectory was substantially higher than that of the populations with the low mutation rate (Figure 2.4). If the hypermutators' beneficial mutation rate also increased by ~100-fold, the difference in trajectories is best fit by an ancestral rate $\mu = 1.7 \times 10^{-6}$ (95% confidence interval 2.5×10^{-7} to 6.1×10^{-5}), although higher values cannot be ruled out (see Appendix: Material and Methods). Note that this fit was obtained by using the same initial distribution of fitness effects, α_0, and epistasis parameter, *g*, for the hypermutators and the populations that retained the ancestral mutation rate.

Both our empirical and theoretical analyses imply that adaptation can continue for a long time for asexual organisms, even in a constant environment. The 50,000 generations studied here occurred in one scientist's laboratory in ~21 years. Now imagine that the experiment continues for 50,000 generations of scientists, each overseeing 50,000 bacterial generations, for 2.5 billion generations total. At that time, the predicted fitness relative to the ancestor is ~4.7 based on the power-law parameters estimated from all 12 populations (Table S2.4). The ancestor's doubling time in the glucose-limited minimal

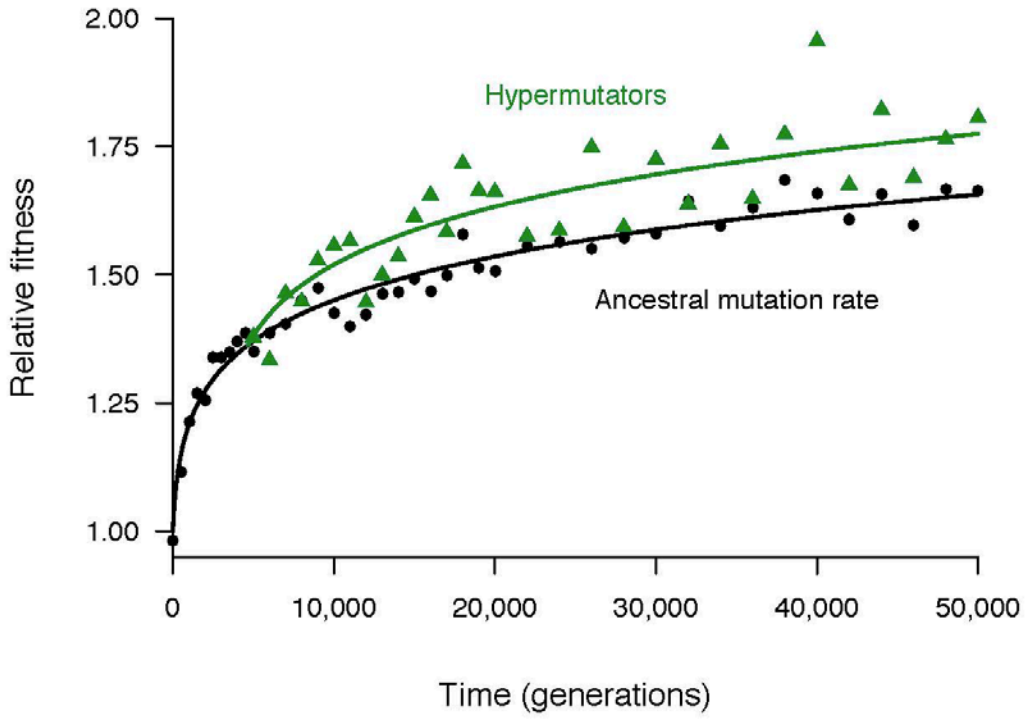

Figure 2.4: **Effect of hypermutability on observed and predicted fitness trajectories**. Black circles: mean fitness of six populations that retained low ancestral mutation rate. Green triangles: mean fitness of three populations that evolved hypermutability early in the LTEE, including one with measurable values through 30,000 generations only. The hypermutators have higher mean fitness at 28 of 31 time points from 5000 to 50,000 generations. Black curve: Predicted trajectory of dynamic model with $\mu = 1.7 \times 10^{-6}$, $\alpha_0 = 85$, $g = 6.0$, and $N = 3.3 \times 10^7$. Green curve: Predicted trajectory with μ increased 100-fold starting at 4667 generations and all other

medium of the LTEE was ~55 min, and its growth commenced after a lag phase of ~90 min (*19*). If the bacteria eliminate the lag, a fitness of 4.7 implies a doubling time of ~23 min (Figure S2.7). Although that is fast for a minimal medium where cells must synthesize most constituents, it is slower than the 10 min that some species can achieve in nutrient-rich media (*20*). At some distant time, biophysical constraints may come into play, but the power-law fit to the LTEE does not predict implausible growth rates even far into the future. Also, some equilibrium might eventually be reached between the fitness-increasing effects of beneficial mutations and fitness-reducing effects of deleterious mutations (*21*), although it is impossible to predict when for realistic scenarios with heterogeneous selection coefficients, compensatory mutations, reversions, and changing mutation rates.

Fitness may continue to increase because even very small advantages become important over very long time scales in large populations. Consider a mutation with an advantage $s = 10^{-6}$. The probability that this mutation escapes drift loss is ~4s for asexual binary fission (*12*), so it would typically have to occur 2.5×10^5 times before finally taking hold. Given a mutation rate of 10^{-10} per base pair per generation (*22*) and effective population size of ~3.3×10^7, it would require ~10^8 generations for that mutation to escape drift and millions more to fix. Also, pleiotropy and epistasis might allow a sustained supply of advantageous mutations, because many net-beneficial mutations have maladaptive side effects that create opportunities for compensatory mutations to ameliorate those effects.

The LTEE uses a simple, constant environment to minimize complications and thus illuminates the fundamental dynamics of adaptation by natural selection in asexual populations. The medium has one limiting resource and supports low population densities (for bacteria) to minimize the potential for cross-feeding on, or inhibition by, secreted by-products. Frequency-dependent interactions are weak in most populations, although stronger in some others (*23*). Also, such interactions should favor organisms that are more fit than their immediate predecessor, but they are not expected to amplify gains relative to a distant ancestor, as fitness was measured here. In fact, such interactions may cause fitness to fall relative to a distant ancestor (*24*). In any case, small-effect beneficial mutations should allow fitness to increase far into the future.

At present, the evidence that fitness can increase for tens of thousands of generations in a constant environment is limited to the LTEE, but these findings have broader implications for understanding evolutionary dynamics and the structure of fitness landscapes. It might be worthwhile to examine fitness trajectories from other evolution experiments in light of our results, although data from short-term experiments may not suffice to discriminate between asymptotic and nonasymptotic trajectories. We hope other teams will perform long experiments similar to the LTEE and that theoreticians will refine our models as appropriate.

APPENDIX

Materials and Methods:

Evolution experiment:

The long- term evolution experiment (LTEE) began in 1988, and it has continued (with occasional interruptions) since then (*1*). Six populations were founded from each of two variants of the same ancestral strain of *Escherichia coli* B (*25*). One ancestral variant, REL607, is able to grow on arabinose (Ara^+) while the other, REL606, cannot (Ara^-). The 12 populations are called Ara-1 to Ara-6 and Ara+1 to Ara+6. They are maintained by daily serial transfer in 10 mL of Davis minimal medium supplemented with limiting glucose at 25 µg/mL (DM25). The cultures are held in 50-mL Erlenmeyer flasks and incubated with shaking at 120 rpm and 37 °C. These conditions support a stationary-phase cell density of ~5×10^7per mL for the ancestral strain (*1*); the evolved populations tend to produce somewhat fewer and larger cells (*19*). The 1:100 dilution and re-growth allow $\log_2 100 \approx 6.64$ generations per day. The effective population size is ~3.3×10^7, which takes into account both the population bottleneck and re-growth (*1*). Every 75 days (500 generations), after the populations have been transferred to fresh medium, glycerol is added to the remaining culture, the material is split between two vials and stored frozen at –80°C. The bacteria remain viable and can be revived for later study; the freezer samples thus provide a living fossil record.

Populations with truncated fitness data:

We obtained complete fitness trajectories for nine of the populations. However, the trajectories for three populations were truncated, even though the populations

themselves continued to evolve for the full 50,000 generations. Populations Ara+6 and Ara-2 no longer produced reliable colonies on the agar plates used to enumerate competitors in the fitness assays after 4000 and 30,000 generations, respectively. Ara-3 evolved the ability to use the citrate in the DM25 medium, which led to a greatly increased cell density (*26*) and other complications for assessing fitness, and therefore its fitness was only measured through 32,000 generations. The same three populations were also excluded from the assays comparing fitness levels at 40,000 and 50,000 generations.

General procedures for fitness assays:

Fitness is measured by mixing two bacterial strains or populations and assessing their relative growth rates during head-to-head competition. In this study, all competitions were performed in the same DM25 medium and other culture conditions as used in the LTEE. The competitors were distinguished on the basis of an arabinose-utilization marker, which is selectively neutral under these conditions (*1*); Ara^- and Ara^+ cells form red and white colonies, respectively, on tetrazolium-arabinose (TA) agar plates. To begin, samples of the population of interest and the reciprocally marked reference competitor were taken from the freezer, transferred into 10 mL of Luria Broth (LB), and grown overnight at 37 °C. These cultures were then diluted 100-fold in saline solution, and 100 µL of the dilutions were inoculated separately into 9.9 mL of DM25. These cultures were incubated for 24 h under the same conditions as the LTEE, such that each competitor was comparably acclimated to those conditions. To

start the actual competition, 50 µL from each acclimation culture were inoculated into 9.9 mL of DM25 and mixed together. An initial 100-µL sample was taken immediately after mixing, then diluted and spread on a TA plate to enumerate the initial density of each competitor. The competition culture was then incubated under the same conditions as the LTEE. For assays used to obtain the fitness trajectories, a final 100-µL sample was taken after 24 h, diluted, and plated on TA agar. For assays comparing fitness levels between 40,000 and 50,000 generations, the competitions were propagated through daily 100-fold dilutions until, after three days, a sample was taken to enumerate the final density of each competitor. In each case, relative fitness was calculated as the ratio of the realized Malthusian parameters of the two competitors over the course of the competition (*1*). For the one-day assays, fitness was calculated simply as

$$w = \ln(A_f / A_i) / \ln(B_f / B_i) ,$$

where *A* and *B* are the respective densities of the evolved population and reference competitor, and subscripts *i* and *f* indicate initial and final densities, respectively. For the three-day assays, the final densities were both multiplied by 10,000 to account for the two additional cycles of dilution and re-growth. In either case, this metric encompasses any and all differences between the competitors in their lag, growth, and stationary phases over the same serial-transfer cycle as used in the LTEE itself (*1*, *19*).

Specific procedures for comparing fitness levels between 40,000 and 50,000 generations:

In general, the statistical error associated with competition assays becomes larger as the difference in fitness increases, because the losing competitor becomes increasingly rare and its abundance less certain in the final sample. Given the small fitness changes expected in later generations, we decided to compete the 40,000- and 50,000-generation populations against a late-generation competitor rather than the ancestor in order to reduce the fitness differential and thereby improve statistical power. To that end, we used a clone, REL10948, sampled from population Ara-5 at 40,000 generations, and we isolated an Ara+ mutant of that clone, REL11638, by plating millions of cells on a minimal medium supplemented with arabinose. Competition assays confirmed that the marker was selectively neutral on this background under the conditions of the LTEE. Samples from the nine populations (excluding the three with truncated fitness trajectories) at generations 40,000 and 50,000 were competed against the reciprocally marked reference clone for three days. Each pairwise competition was replicated six-fold in a complete-block design.

Specific procedures for obtaining fitness trajectories through 50,000 generations:

To ensure uniformity of procedures, all of the data used to characterize the long-term fitness trajectories were based on one-day competitions against the reciprocally marked ancestral clone, either REL606 (Ara^{-}) or REL607 (Ara^{+}). All of the generational samples from a given population were simultaneously

placed in competitions, and each complete time-series was replicated twice at different times (with a few missing values caused by procedural errors). The fitness trajectory for each individual population was fit using the replicate values at each time point. The trajectory for the grand-mean fitness of the ensemble of populations was fit using the average of the replicate values for each population at each time point.

Statistical analyses:

All of the statistical analyses of experimental data were performed using the R software suite (version 2.14.1). The hyperbolic and power law models were fit to the fitness trajectories using the nls function in R. Both models have two parameters, and they are not nested, so they cannot be compared using likelihood-ratio or F tests. Instead, we compare them using Bayesian Information Criterion (BIC) scores (*27*). To calculate the 95% confidence interval for the diminishing-returns parameter g for populations that retained the low mutation rate throughout, we first calculated the confidence interval for a using the estimates from the six corresponding populations. The endpoints of that interval were then transformed to g values based on the relationship $g = 1 / (2a)$. As a consequence, the interval for g is asymmetric around the point estimate. The datasets and analysis scripts have been deposited in the Dryad database.

Derivation of theory:

The derivation of the theory that generates the power-law dynamics was checked by obtaining numerical solutions using Wolfram Mathematica (version 8.0). The script has been deposited in the Dryad database. To examine the possible effects of co-occurring beneficial mutations, we used LabVIEW 2010 (version 10.0.1) to simulate the dynamics of mean fitness and fixed beneficial mutations; these dynamics were then compared to predictions from our theory, which does not consider multiple co-occurring mutations. In particular, adaptation was simulated using a Wright-Fisher model with discrete generations. Distinct genotypes were tracked along with their corresponding frequencies, fitnesses, and α values. Binary fission was simulated by updating each genotype's population size in the following generation by drawing from a binomial distribution with 2*x* trials and (1/2)($w / \overline{w}$) success probability in each trial, where *x* and w are the genotype's population size and fitness, respectively, and $\overline{w}$ is the mean fitness of the entire population. Each generation, a number of beneficial mutations (drawn from Poisson distribution with mean Nμ, where N is the total population size and μ is the beneficial mutation rate) were assigned randomly to genotypes (with probability weighted by *x*), with the mutant designated as a new genotype with new fitness $w_{new} = w(1+s)$, where s is drawn from an exponential distribution $\alpha e^{-\alpha s}$, and new αvalue $\alpha_{new} = \alpha(1+gs)$

The full derivation of dynamical model of long-term fitness trajectory that incorporates clonal interference and diminishing-returns epistasis can be found in www.sciencemag.org/content/342/6164/1364/suppl/DC1, the Supplementary

Materials for this paper. We have omitted this for the purposes of this dissertation, as the model was derived by Noah Ribeck.

For trajectories with $\mu > 10^{-8}$ that match the best fit to the populations that retained the ancestral mutation rate, the model predicts ~13 fixation events of beneficial mutations over the course of 20,000 generations. However, 45-50 mutations were discovered by sequencing the genomes of clones from two LTEE populations that were not hypermutators at that time (*10*, *28*). Some of the discrepancy may reflect neutral mutations that hitchhiked along with beneficial ones. However, this explanation is insufficient given the paucity of synonymous substitutions (*22*), the prevalence of parallel changes across replicate populations (*10*), and the results of competitions between isogenic strains (*10*).

Another factor that could contribute to this discrepancy is the sequential, one-at-a-time fixations of beneficial mutations assumed by our model of clonal interference. That is, a single fixation event may sometimes involve multiple beneficial mutations. At high $N\mu$, "cohorts" of multiple beneficial mutations can co-occur in the same lineage before one of them fixes and, in some cases, they may alleviate the effect of clonal interference on the rate of adaptation (*18*, *29*). Some theoretical work has examined the effect of co-occurring beneficial mutations on the rate of adaptation (*18*, *29*–*32*), but its direct application here is prevented by the pervasive epistasis in our model. At intermediate $N\mu$, selective sweeps are sometimes caused by a single large-effect beneficial "driver" mutation accompanied by a weakly beneficial passenger that hardly affects the

dynamics of adaptation (*29*, *30*). Indeed, such weakly beneficial passenger mutations have been observed in the LTEE populations (*33*, *34*).

To test whether our theoretical model is accurate, despite ignoring cohorts of beneficial mutations, we ran individual-based simulations of asexual populations for a range of μ values, each with the corresponding value of α_0 such that the simulation matches the best fit to the fitness trajectories for the populations that retained the low mutation rate throughout the LTEE (Figure 2.3A). These simulations show that fitness trajectories are consistent across a wide range of (μ, α_0) values, and they closely match the theoretical fitness trajectory that assumes one-at-a-time fixations (Figure S2.6). Thus, our theoretical model with its simplifying assumptions does well with respect to the fitness trajectory. With respect to genomic evolution, the individual-based simulations show a number of fixed beneficial mutations that is slightly higher than the theoretical values for $\mu > 10^{-7}$, with the discrepancy increasing with higher μ (Figure S2.6). Taken together, these observations are consistent with the intermediate Nμ regime, where weakly beneficial passengers occasionally fix along with highly beneficial drivers but do not appreciably affect the rate of adaptation. The pervasive diminishing-returns epistasis inherent to our model likely reduces the effect of weakly beneficial passenger mutations relative to previous theory that does not include this epistasis.

From our analysis of the effect of hypermutators on fitness trajectories (Figure 2.4), we estimated the ancestral rate of beneficial mutations to be 1.7×10^{-6}. At this rate, however, the simulations predict only ~14 beneficial mutations

to fix by 20,000 generations. Therefore, weakly beneficial passenger mutations—at least those that occur at typical point-mutation rates—cannot account for the discrepancy between the observed number of mutations in the LTEE and that predicted by both theory and simulations. Instead, we suspect that certain types of insertion and deletion mutations that occur at much higher rates than point mutations (*33*–*35*)—in particular those that are neutral or nearly neutral—might help to explain why the rate of genomic evolution exceeds the number of beneficial fixation events to the extent that it does. In that respect, it is noteworthy that the two weakly beneficial mutations that fixed early in the most intensively studied LTEE population, Ara-1, were non-point mutations of types known to occur at unusually high rates (*33*–*35*). More generally, 16 of the 45 mutations in a 20,000-generation clone from that population were non-point mutations (*10*), which potentially reduces by about half the discrepancy between the observed number of mutations and the number predicted by theory and simulations.

For the number of observed fixed mutations to be in close agreement with the simulations would require a higher ancestral beneficial mutation rate than we have estimated here (Figure S2.6). In fact, we cannot rule out this possibility. For simulations with $\mu = 10^{-4}$, the fitness trajectory is slightly higher than the theory predicts, indicating entry into the high $N\mu$ regime, where co-occurring beneficial mutations alleviate the inhibitory effect of clonal interference on the rate of adaptation. If the hypermutators have entered this regime, then our theory would underestimate the fitness trajectory, and our derived estimate of the beneficial

mutation rate would be too low. We therefore interpret our estimate of $\mu = 1.7 \times 10^{-6}$ for the ancestral beneficial mutation rate to be a lower bound on the actual value.

Reflecting these complications and uncertainties, our dynamical model cannot predict the overall rate of genomic evolution. However, we can use the model's predicted rate of fixation events as a proxy for the overall rate. Figure S2.5 shows the predicted fixation trajectory and the corresponding mean fitness trajectory that fits the LTEE data for the populations that maintained the low ancestral mutation rate. Both the rate of fitness improvement and the rate of fixation events decline over time; however, the deceleration in the rate of fixation events is much less pronounced, giving the appearance of relative constancy. It has also been shown elsewhere, using another theoretical framework, that evolution on fitness landscapes with antagonistic (e.g., diminishing-returns) epistasis can produce nearly linear fixation trajectories (*4*). In any case, the difference in the relative curvature of the trajectories for mean fitness and genomic evolution observed in the LTEE (*10*) is consistent with our model.

Parameterization of diminishing-returns epistasis fits well with other data from the LTEE. Khan et al. (*6*) constructed the 32 possible combinations of the first five mutations that fixed in the Ara-1 population. The fitness of each construct was then measured against the ancestor, providing estimates of the marginal effect of each mutation on backgrounds of varying fitness. Three of the mutations—those affecting the *topA*, *spoT*, and *glmUS* genes—exhibited significant diminishing-returns epistasis, i.e., they had smaller beneficial effects in

higher fitness backgrounds. Of the other two, one was nearly neutral and showed no significant trend, and one exhibited positive epistasis. Here, we compare the data for the three mutations with diminishing-returns epistasis to the best-fit parameter g obtained from our theoretical model of long-term adaptation. From our model, we expect the effects of beneficial mutations to scale as:

$\alpha(\bar{w}) = \alpha_0 e^{g \ln \bar{w}}$, or $s(\bar{w}) = s_0 e^{-g \ln \bar{w}}$, or equivalently:

$$\frac{s(\bar{w})}{s_0} = \bar{w}^{-g}.$$

Figure S2.4 shows fits of this equation to these independently measured data, which appear consistent with the general form of diminishing-returns epistasis assumed in our theoretical model.

Khan et al. (*6*) concluded there is a tendency toward diminishing-returns epistasis among beneficial mutations. However, the magnitude of that epistasis seems to vary even among the mutations that clearly show diminishing returns, as evidenced by the best-fit g values of 3.1, 2.9, and 7.2 for the mutations affecting *topA*, *spoT*, and *glmUS*, respectively. In comparison, the dashed curve in Figure S2.4 corresponds to g = 6.0, which derives from 50,000 generations of fitness measurements for all six populations that maintained the low ancestral mutation rate throughout 50,000 generations. This value is our best estimate of the mean strength of diminishing-returns epistasis for the LTEE as a whole.

As a technical aside, we note that our model is meant for use in the long-time limit. However, by evaluating $c = \alpha_0$ here, it is also reasonably accurate for

values of $\overline{w}$ near 1. For the values of s/s_0 shown in Figure S2.4, this approximation is accurate to within ~10% at w = 1.3.

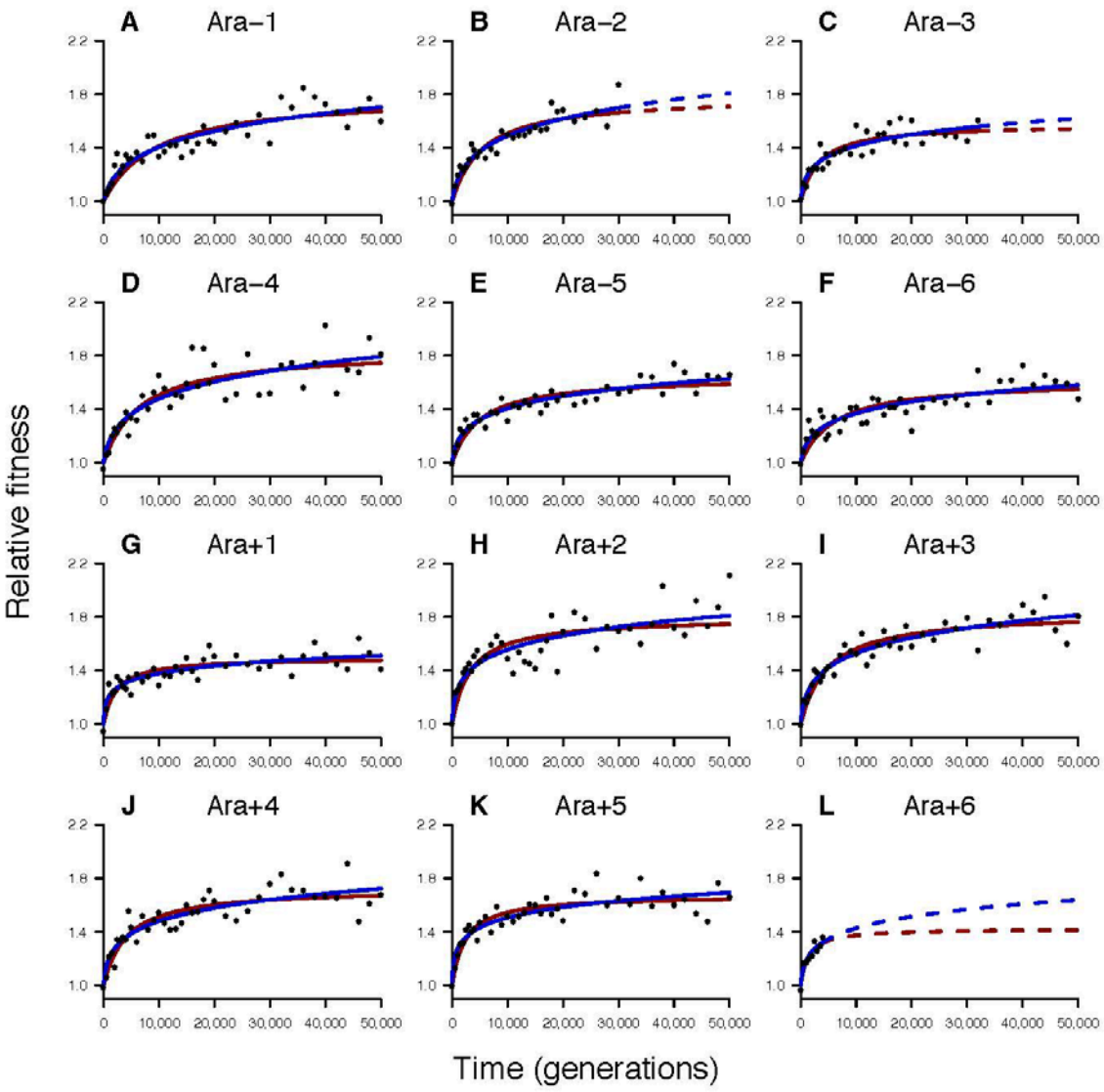

Figure S2.1: **Comparison of the fit of the hyperbolic (red) and power-law (blue) models to the fitness trajectories for the 12 individual *Escherichia coli* populations**. (A) Ara-1. (B) Ara-2. (C) Ara-3. (D) Ara-4. (E) Ara-5. (F) Ara-6. (G) Ara+1. (H) Ara+2. (I) Ara+3. (J) Ara+4. (K) Ara+5. (L) Ara+6. Three trajectories are truncated because of difficulties in measuring fitness that arose in those populations, as explained in the Materials and Methods.

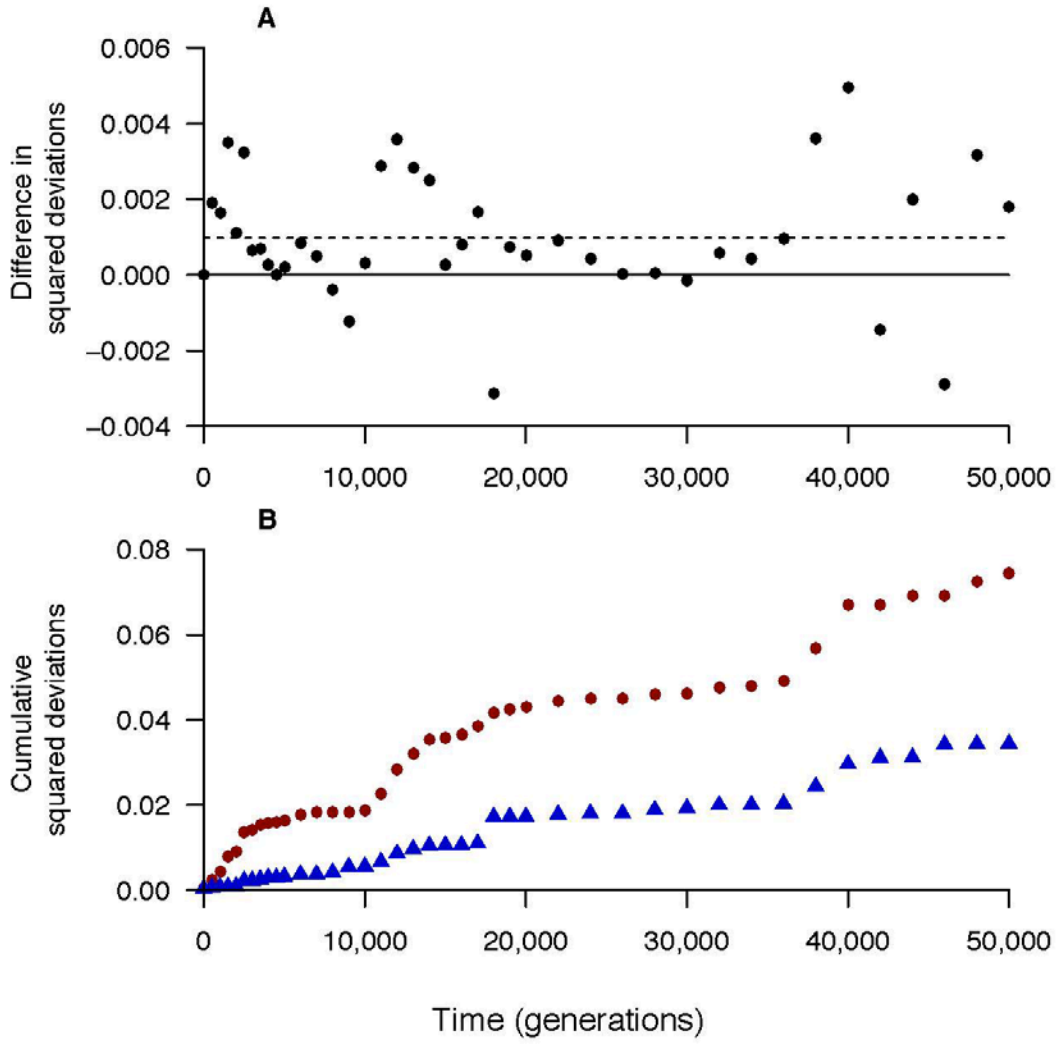

Figure S2.2: **Comparison of hyperbolic and power-law models in terms of squared deviations between their fit trajectories and measured grand-mean fitness values over time**. (**A**) Difference in squared deviations between the two models; positive values indicate the power law provides a better fit. The dashed line shows the average difference in squared deviations over 50,000 generations. (**B**) Cumulative squared deviations between the hyperbolic (red circles) and power-law (blue triangles) models and the measured values over time.

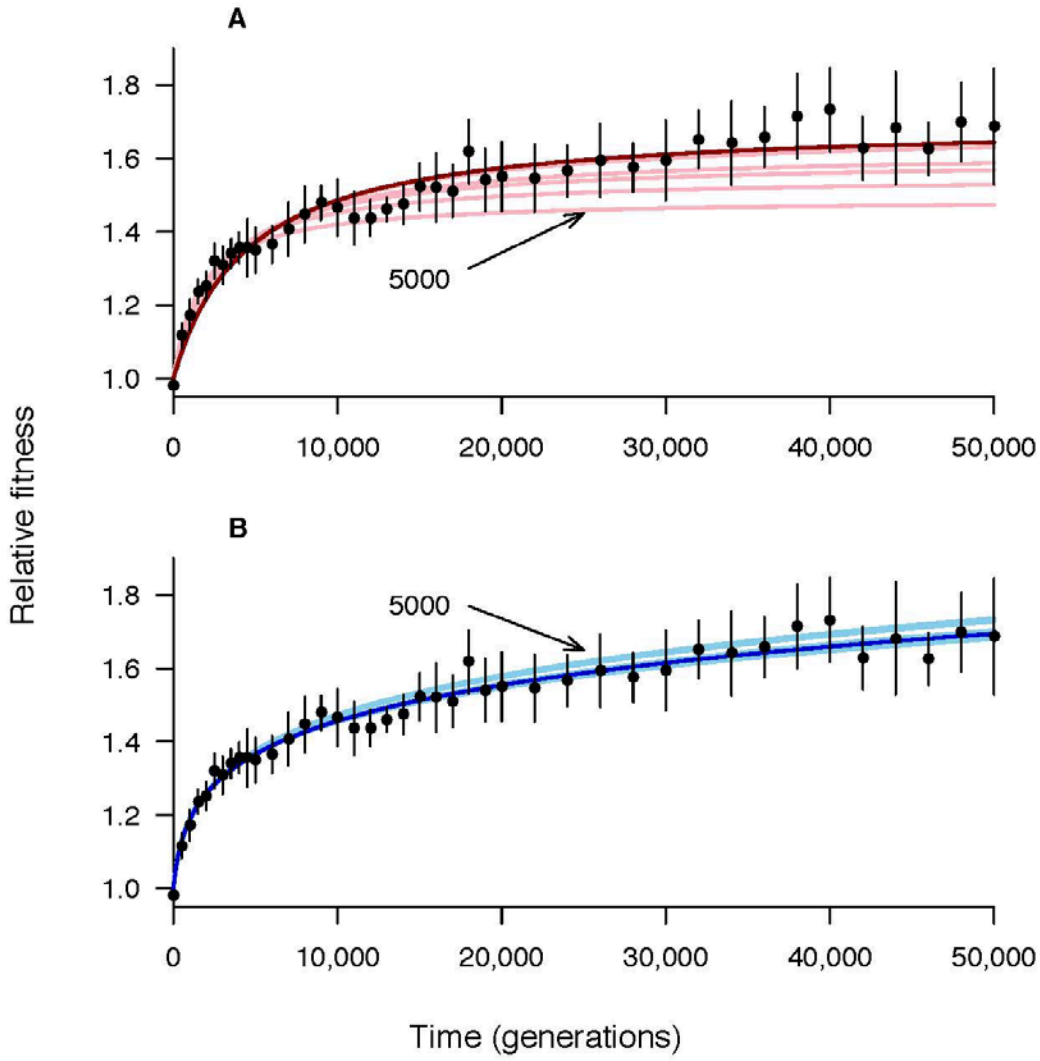

Figure S2.3: **Comparison of hyperbolic and power-law models in their ability to predict future fitness values from temporally truncated datasets**. (**A**) Fit of the hyperbolic model to all 12 populations using data from several subsets of generations (light red) or from all 50,000 generations (dark red). The subsets, from bottom to top, include data through 5000, 10,000, 20,000, 30,000, and 40,000 generations. The underestimation of the later values becomes progressively more severe as the data are truncated at earlier time points. (**B**) Fit of the power-law model to all 12 populations using data from several subsets of generations (light blue) or from all 50,000 generations (dark blue). The subsets include data through 5000, 10,000, 20,000, 30,000, and 40,000 generations, and all are very close to the trajectory fit to the complete 50,000 generations. Error bars are 95% confidence limits based on replicate populations.

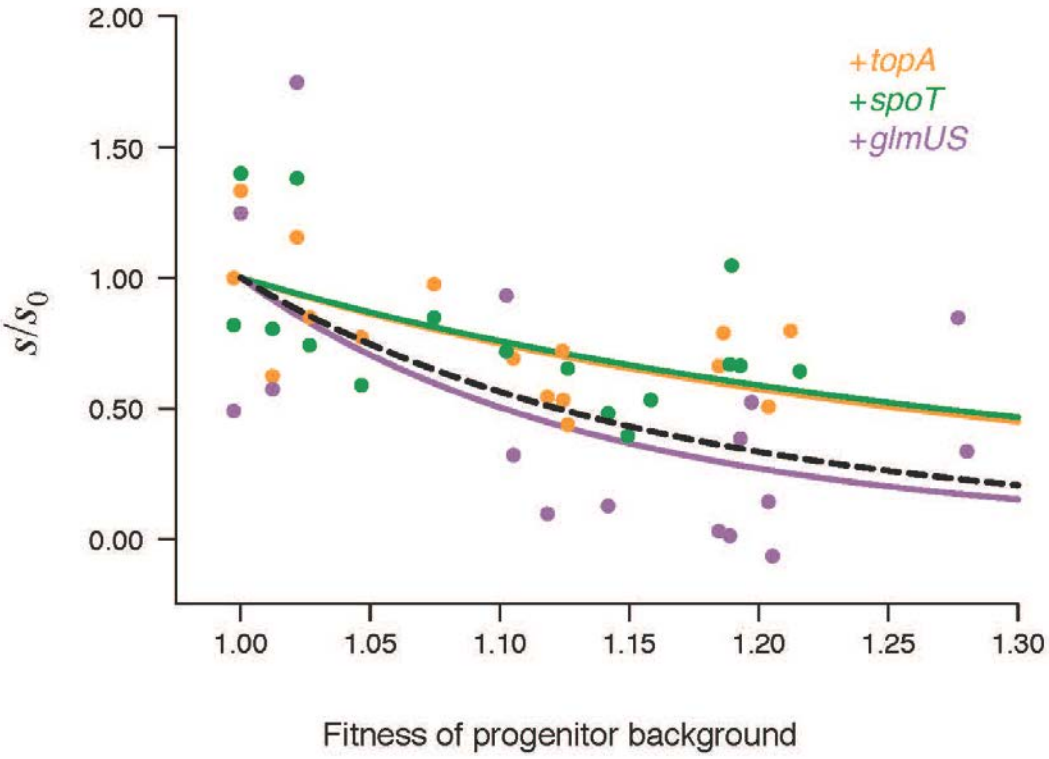

Figure S2.4: **Parameterization of diminishing-returns epistasis based on the fit of the dynamic model to the fitness trajectories accords well with independent data on the form and strength of epistasis from the LTEE**. Each set of points shows the beneficial effect of adding an individual mutation to different progenitor backgrounds of varying fitness, as measured by Khan et al. (*6*) using the first several mutations that fixed in the Ara-1 population. The solid colored curves are fits to the parameterization of diminishing-returns epistasis used in our theoretical model, giving g values of 3.1 for the addition of a beneficial mutation in *topA*, 2.9 for *spoT*, and 7.2 for *glmUS*. The black dashed curve corresponds to g = 6.0, the value that provides the best fit of the power-law model to the fitness trajectories of the populations that retained the low ancestral mutation rate throughout the 50,000 generations.

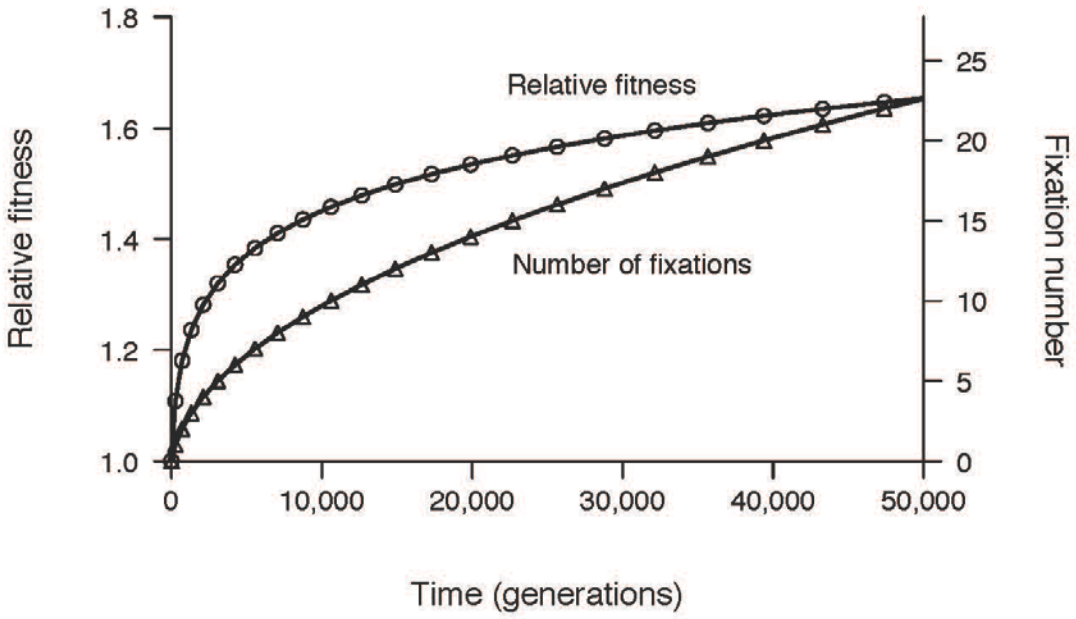

Figure S2.5: **Predicted number of beneficial fixation events in relation to the fitness trajectory, based on the theoretical model with clonal interference and diminishing-returns epistasis**. The expected fitness trajectory and number of fixation events are shown for the follow set of parameters: $\mu = 1.7 \times 10^{-6}$, $\alpha_0 = 85$, $g = 6.0$, and $N = 3.3 \times 10^{7}$.

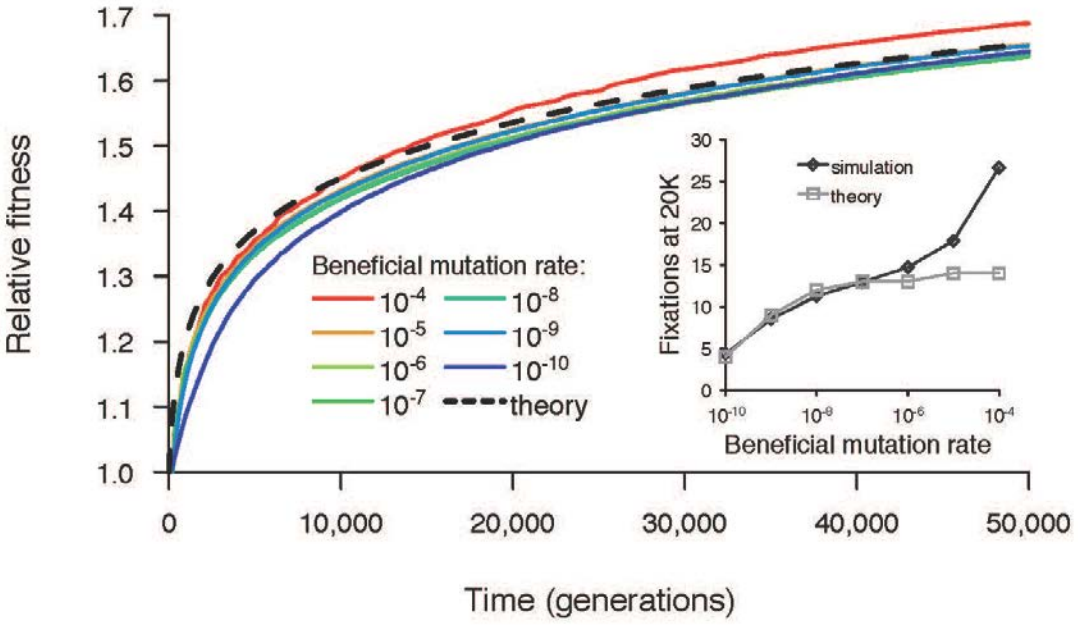

Figure S2.6: **Numerical simulations of fitness trajectories show good agreement with the theory over a wide range of the beneficial mutation rate** μ. Simulations of the theory of clonal interference with diminishing-returns epistasis are shown, with different colors representing different pairs of the parameters μ and α_0(each curve is labeled by its μ) that equivalently give the best fit to the set of the populations that maintained the ancestral mutation rate for 50,000 generations (Figure 2.3A). The dashed line represents the theoretical fitness trajectory for this family of parameters. Deviations from the theory at early times are small for simulations with $\mu \geq 10^{-9}$; deviations at lower $\mu$are caused by the small $\langle s \rangle$approximation. Deviations from the theory at later times are negligible for $\mu \leq 10^{-5}$; deviations at higher μresult from co-occurring beneficial mutations. All simulations were run with $g = 6.0$and $N = 3.3 \times 10^7$. Curves are based on the mean fitness from multiple runs, with 3 runs for $\mu = 10^{-4}$, 50 each for $\mu = 10^{-5}$ and 10^{-6}, 200 for $\mu = 10^{-7}$, 500 for $\mu = 10^{-8}$, and 2000 each for $\mu = 10^{-9}$and 10^{-10}. The inset panel shows the mean number of beneficial mutations fixed in each set of simulations at 20,000 generations, compared to the number predicted by the theory, which does not account for co-occurring beneficial mutations.

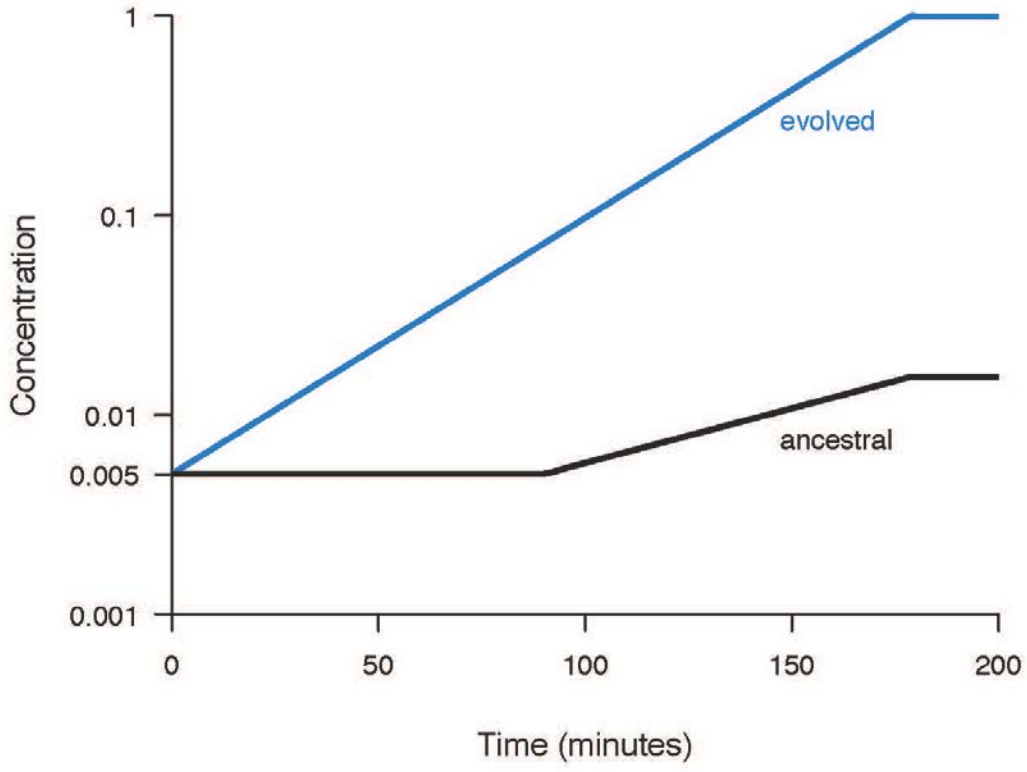

Figure S2.7: **Hypothetical growth kinetics of evolved (blue) and ancestral (black) competitors that would produce a relative fitness of ~4.7**. The LTEE ancestral strain grows with a doubling time of 55 minutes, following a lag phase of 90 minutes. The hypothetical evolved population has a doubling time of ~23 minutes without any lag.

	BIC Hyperbolic	BIC Power Law	BIC Difference	Odds Ratio
(a) All 12 populations and all time points	-649.876	-684.346	34.470	3.056 * 10^7
(b) Excluding 3 populations with incomplete trajectories	-505.915	-534.278	28.362	1.442 * 10^6
(c) Excluding 6 populations that evolved hypermutability	-337.257	-359.679	22.422	7.394*10^4

Table S2.1: **Differences in Bayesian Information Criteria (BIC) scores between hyperbolic and power-law model trajectories fit to the measured fitness values.** Contrasts are based on: (a) the full dataset including all 12 populations and all time points available for each population; (b) the dataset excluding 3 populations with incomplete fitness trajectories; and (c) the dataset excluding 6 populations that evolved hypermutability. A BIC difference >10 is considered to provide very strong support for one model over another (*27*), which can also be expressed as a posterior odds ratio.

Population	BIC Hyperbolic	BIC Power Law	BIC Difference	Odds Ratio	Complete?	Hyper-mutator?
Ara - 1	-80.290	-85.544	5.253	13.826	Yes	Yes (10) ~27,000 gens.
Ara - 2	-97.283	-104.091	6.808	30.080	No 30,000 gens.	Yes (36) ~2500 gens.
Ara - 3	-95.662	-94.666	-0.996	0.608	No 32,000 gens.	Yes (28) ~35,000 gens.
Ara - 4	-62.844	-60.942	-1.902	0.386	Yes	Yes (36) ~8500 gens.
Ara - 5	-167.638	-182.031	14.393	1334.878	Yes	No
Ara - 6	-58.165	-63.640	5.475	15.446	Yes	No
Ara + 1	-138.723	-144.565	5.842	18.558	Yes	No
Ara + 2	-33.378	-42.974	9.597	121.311	Yes	No
Ara + 3	-61.268	-63.430	2.162	2.948	Yes	Yes (36) ~3000 gens.
Ara + 4	-97.446	-98.089	0.643	1.379	Yes	No
Ara + 5	-108.132	-106.181	-1.951	0.377	Yes	No
Ara + 6	-30.308	-30.659	0.351	1.192	No 4000 gens.	Yes ~5000 gens.

Table S2.2: **Differences in BIC scores between the hyperbolic and power-law trajectories fit to the measured fitness values for 12 individual *E. coli* populations**. The column labeled "Complete?" indicates whether the population's fitness trajectory extended the full 50,000 generations ("Yes") or was terminated at an earlier generation ("No" followed by the last generation with fitness data in Figure S2.1). The column labeled "Hypermutator?" indicates whether the population evolved hypermutability ("Yes" followed by the approximate generation when the hypermutable genotype become the majority (*10*, *28*, *36*)) or retained the low ancestral mutation rate throughout ("No"). An odds ratio >1 or <1 indicates support for the power law or the hyperbolic model, respectively.

	Degrees of Freedom	Sums of Squares	Mean Square	F	p
Population	5	17.624	3.5247	1.3853	0.3478
Residuals	6	15.266	2.5443		

Table S2.3: **Analysis of variation to test for heterogenetic ln g values among the six populations that maintained the low ancestral mutation rate throughout the LTEE**. Each fitness trajectory was replicated twice, giving two estimates of the power-law exponent a and, using the relationship $g = 1/(2a)$, two estimates of the diminishing-returns parameter g.

Population	a	b	Complete?	Hypermutator?	g
Ara - 1	0.1239	0.001461	Yes	~27,000 gens.	[4.035]
Ara - 2	0.1235	0.002406	30,000 gens.	~2500 gens.	[4.048]
Ara - 3	0.0842	0.006105	32,000 gens.	~35,000 gens.	5.941
Ara - 4	0.1230	0.002297	Yes	~8500 gens.	[4.066]
Ara - 5	0.0923	0.003913	Yes	No	5.418
Ara - 6	0.0903	0.003147	Yes	No	5.534
Ara + 1	0.0576	0.026337	Yes	No	8.682
Ara + 2	0.0944	0.010771	Yes	No	5.295
Ara + 3	0.1074	0.005101	Yes	~3000 gens.	[4.653]
Ara + 4	0.0951	0.006051	Yes	No	5.256
Ara + 5	0.0730	0.027248	Yes	No	6.848
Ara + 6	0.1159	0.003517	4000 gens.	~5000 gens.	4.314
All 12 populations	0.0950	0.005149	n/a	n/a	5.265
Six populations with complete trajectories and low ancestral mutation rate throughout	0.0830	0.008706	n/a	n/a	6.022

Table S2.4: **Parameter estimates for the power-law model fit to each individual population's measured fitness values**. Parameter estimates for the exponent a and scaling factor b are also shown for the model fit to the set of all 12 populations and to the subset of six populations with complete trajectories that maintained the low ancestral mutationr ate throughout the 50,000 generations. The column labeled "Complete?" indicates whether a population's fitness trajectory extended for the full 50,000 generations "Yes") or was terminated, shown as the last generation with fitness data. The column labeled "Hypermutator?" indicates whether the population evolved hypermutability, shown by the approximate generation when the hypermutable genotype became the majority, or retained the ancestral mutation rate throughout ("No"). The column labeled g shows the estimate of the diminishing-returns epistasis parameter in the dynamical model, calculated using $g = 1/(2a)$. However, for populations that evolved hypermutability, the change in mutation rate complicates the fit of the power law and the estimation of those parameters. As shown in Figure 2.4, the change in mutation rate can explain the difference in trajectories between the hypermutator populations and those that retained the ancestral mutation rate, without any change in the diminishing-returns parameter g. For that reason, the estimates of g are inaccurate for the populations that evolved hypermutability before their trajectories were terminated (g values shown in brackets).

From M. J. Wiser, N. Ribeck, R. E. Lenski, Long-term dynamics of adaptation in asexual populations. *Science*. **342**, 1364–1367 (2013). Reprinted with permission from AAAS.

REFERENCES

REFERENCES

1. R. E. Lenski, M. R. Rose, S. C. Simpson, S. C. Tadler, Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138**, 1315–1341 (1991).

2. C. L. Burch, L. Chao, Evolution by small steps and rugged landscapes in the RNA virus ϕ6. *Genetics*. **151**, 921–927 (1999).

3. D. M. Weinreich, N. F. Delaney, M. A. DePristo, D. L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*. **312**, 111–114 (2006).

4. S. Kryazhimskiy, G. Tkačik, J. B. Plotkin, The dynamics of adaptation on correlated fitness landscapes. *Proc. Natl. Acad. Sci.* **106**, 18638–18643 (2009).

5. H.-H. Chou, H.-C. Chiu, N. F. Delaney, D. Segrè, C. J. Marx, Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science*. **332**, 1190–1192 (2011).

6. A. I. Khan, D. M. Dinh, D. Schneider, R. E. Lenski, T. F. Cooper, Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. *Science*. **332**, 1193–1196 (2011).

7. I. G. Szendro, M. Schenk, J. Franke, J. Krug, J. A. G. M. de Visser, Quantitative analyses of empirical fitness landscapes. *J. Stat. Mech. Theory Exp.* **2013**, P01005 (2013).

8. T. J. Kawecki *et al.*, Experimental evolution. *Trends Ecol. Evol.* **27**, 547–560 (2012).

9. R. E. Lenski, M. Travisano, Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci.* **91**, 6808–6814 (1994).

10. J. E. Barrick *et al.*, Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*. **461**, 1243–1247 (2009).

11. P. Sibani, M. Brandt, P. Alstrøm, Evolution and Extinction Dynamics in Rugged Fitness Landscapes. *Int. J. Mod. Phys. B*. **12**, 361–391 (1998).

12. P. Gerrish, R. Lenski, The fate of competing beneficial mutations in an asexual population. *Genetica*. **102-103**, 127–144 (1998).

13. M. Hegreness, N. Shoresh, D. Hartl, R. Kishony, An Equivalence Principle for the Incorporation of Favorable Mutations in Asexual Populations. *Science*. **311**, 1615–1617 (2006).

14. S.-C. Park, J. Krug, Clonal interference in large populations. *Proc. Natl. Acad. Sci.* **104**, 18135–18140 (2007).

15. G. I. Lang *et al.*, Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*. **500**, 571–574 (2013).

16. S. Wielgoss *et al.*, Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl. Acad. Sci.* **110**, 222–227 (2013).

17. R. J. Woods *et al.*, Second-Order Selection for Evolvability in a Large Escherichia coli Population. *Science*. **331**, 1433–1436 (2011).

18. M. M. Desai, D. S. Fisher, A. W. Murray, The Speed of Evolution and Maintenance of Variation in Asexual Populations. *Curr. Biol.* **17**, 385–394 (2007).

19. F. Vasi, M. Travisano, R. E. Lenski, Long-term experimental evolution in Escherichia coli. II. Changes in life-history traits during adaptation to a seasonal environment. *Am. Nat.* **144**, 432–456 (1994).

20. R. G. Eagon, Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes. *J. Bacteriol.* **83**, 736–737 (1962).

21. S. Goyal *et al.*, Dynamic Mutation–Selection Balance as an Evolutionary Attractor. *Genetics*. **191**, 1309–1319 (2012).

22. S. Wielgoss *et al.*, Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli. *G3 Genes Genomes Genet.* **1**, 183–186 (2011).

23. S. F. Elena, R. E. Lenski, Long-Term Experimental Evolution in Escherichia coli. VII. Mechanisms Maintaining Genetic Variability Within Populations. *Evolution*. **51**, 1058–1067 (1997).

24. C. E. Paquin, J. Adams, Relative fitness can decrease in evolving asexual populations of S. cerevisiae. *Nature*. **306**, 368–371 (1983).

25. P. Daegelen, F. W. Studier, R. E. Lenski, S. Cure, J. F. Kim, Tracing ancestors and relatives of Escherichia coli B, and the derivation of B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394**, 634–643 (2009).

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

27. A. E. Raftery, Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–164 (1995).

28. Z. D. Blount, J. E. Barrick, C. J. Davidson, R. E. Lenski, Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*. **489**, 513–518 (2012).

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

30. S. Schiffels, G. J. Szöllősi, V. Mustonen, M. Lässig, Emergent Neutrality in Adaptive Asexual Evolution. *Genetics*. **189**, 1361–1375 (2011).

31. Y. Kim, H. A. Orr, Adaptation in Sexuals vs. Asexuals: Clonal Interference and the Fisher-Muller Model. *Genetics*. **171**, 1377–1386 (2005).

32. P. D. Sniegowski, P. J. Gerrish, Beneficial mutations and the dynamics of adaptation in asexual populations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 1255–1263 (2010).

33. V. S. Cooper, D. Schneider, M. Blot, R. E. Lenski, Mechanisms Causing Rapid and Parallel Losses of Ribose Catabolism in Evolving Populations of Escherichia coli B. *J. Bacteriol.* **183**, 2834–2841 (2001).

34. M. T. Stanek, T. F. Cooper, R. E. Lenski, Identification and dynamics of a beneficial mutation in a long-term evolution experiment with Escherichia coli. *BMC Evol. Biol.* **9**, 1–13 (2009).

35. E. R. Moxon, P. B. Rainey, M. A. Nowak, R. E. Lenski, Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).

36. P. D. Sniegowski, P. J. Gerrish, R. E. Lenski, Evolution of high mutation rates in experimental populations of E. coli. *Nature*. **387**, 703–705 (1997).

CHAPTER 3: PERSISTENT AMONG-POPULATION VARIANCE IN FITNESS IN A LONG-TERM EVOLUTION EXPERIMENT WITH *ESCHERICHIA COLI*

Authors: Michael J. Wiser and Richard E. Lenski

Abstract:

Adaptive landscapes for real populations are difficult to characterize both qualitatively and quantitatively, in part because individual natural populations often occupy only one small region of any given landscape. However, variation in fitness across independent experimental populations can provide insight about the adaptive landscapes on which they evolve. Previous research has addressed how mean fitness changes in populations over time, but there has been much less work on how variation among populations changes over time. Here, we investigate populations from a long-term evolution experiment (LTEE) in *Escherichia* coli that evolved for 50,000 generations. We look collectively at the populations to measure the trajectory of the among-population variance in fitness over that time. We further measure the relative fitness of pairs of evolving populations, and compare these measurements to predictions based on each individual population's fitness relative to the ancestor. We find persistent among-population variance in fitness, providing evidence that the populations have not converged – and probably are not converging – to the same fitness level in the adaptive landscape. Our data indicate a rich and complex adaptive landscape even in a simple and nearly constant physical environment.

Introduction:

In order to understand how evolution will unfold over long time periods, it is critical to understand how variance within the population changes over time. If variance within a population remains substantial, there will be sufficient differences among individuals for natural selection to operate. However, if variance diminishes to negligible levels, the rate of adaptation will slow dramatically, and perhaps even come to a stop.

As a thought experiment, consider a hypothetical population with a finite number of possible beneficial mutations. Without some sort of change to the environment causing additional mutations to be beneficial, the population will inevitably reach a point of having incorporated the full set of beneficial mutations. At this point, adaptation would essentially stop, except for second-order effects such as the population moving to regions of the landscape that limit the deleterious impact of new mutations (*1*). In such a case, the process of evolution is limited by the availability of beneficial mutations; once those mutations are incorporated, the stock of potentially adaptive mutations has been exhausted, and the population stagnates.

In real populations, there are at least two ways in which the supply of beneficial mutations can be refreshed. One is epistasis: that is, some mutations may not be beneficial at the moment, but would be beneficial on a different genetic background (*2*). Every step along an adaptive trajectory thus brings an individual to a new area of the adaptive landscape. While the number of

mutations that are beneficial at any one point is finite, that does not necessarily mean that there are a finite total number of potentially adaptive mutations; in fact, because genome length is variable, we cannot *a priori* list all possible genotypes. A second possibility is that the environment could change. For example, the availability of a new resource may favor mutations allowing use of that resource, while absence of the resource prevents those mutations from being beneficial (*3*, *4*). Changes in the environment can also involve biotic interactions; changes in predator, prey, competitor, or mutualist populations can all alter what mutations will be favored within a given population (*5*, *6*).

Looking beyond any single population, the variance among populations provides insight into the topology of the underlying fitness landscape. While each population in nature typically experiences a different environment, and hence evolves on a different fitness landscape, theory and experiments allow us to consider the case of initially identical populations evolving under identical conditions. In the theoretical case described above, in which each individual population has exhausted the within-population variance, there are still two possible outcomes for the among-population variance. First, if there is only a single accessible fitness peak (a smooth landscape), then the among-population variance should eventually drop to zero because all of the populations approach the same equilibrium mean fitness. Second, if multiple peaks exist and are accessible (a rugged landscape), then different populations may reach different fitness peaks. Because populations can become stuck at the sub-optimal peaks in a rugged landscape, evolutionary outcomes will be more variable in rugged

adaptive landscapes than in smooth ones (*7*), and the among-population variance in fitness may remain positive indefinitely.

From previous work (*8*), we already know that 50,000 generations has not been enough time for populations to reach fitness peaks in the LTEE. Instead, the grand mean fitness is well described by a power law, of the form

$$w = (bT + 1)^a$$

where w is fitness, T is time in generations, and a and b are model parameters. Because the power law does not have an asymptotic limit – fitness keeps increasing indefinitely in this model – it implies that the evolving populations are so far from the top of whatever peaks might exist that it is not useful to think of them reaching these peaks over 50,000 generations, or even over much longer timescales. Therefore, we do not expect the among-population variance in fitness to decline to zero in this time frame. However, we might be able to use estimates of the among-population variance in fitness, and especially its trajectory over time, to infer more information about how the set of population fitness trajectories map onto the adaptive landscape. If the among-population variance remained zero (i.e., its initial state given that the populations all started from the same ancestor) throughout the evolution experiment, then this would imply that either i) the populations are all following the same adaptive path, or ii) the populations are on different paths that nonetheless map onto regions of the adaptive landscape with parallel slopes. Conversely, if the among-population variance continues to exceed zero indefinitely, then this implies that either i) the different populations are following different paths, or ii) the timing of the

appearance of equivalent beneficial mutations varies enough to sustain the among-population variance in fitness.

Meanings of changes in variance:

To better understand evolutionary dynamics in this system, we examine how the variance in fitness changes over evolutionary time. By definition, the populations in the LTEE have no variance in fitness at generation 0, as all populations have the same fitness. Previous work showed that among-population variance in fitness increased over the first 10,000 generations in this experiment. Lenski and Travisano (1994) suggested that this among-population variance in fitness may have leveled off, but did not explicitly test whether an asymptotic model provided a better fit to the data than did an unbounded model (*9*). There are three hypothetical possibilities of how variance in fitness changes after these first 10,000 generations: continued increase, constancy, or decrease after the initial increase.

One possibility is that among-population variance in fitness continues to increase across the 50,000 generations of data. This possibility is most consistent with populations continuing to explore different areas of the fitness landscape, or else exploring the same peak but at very different rates by climbing faces with different slopes. Because our previous work showed no evidence for populations reaching fitness peaks (*8*), we hypothesize that this is the most likely scenario.

Another possibility is that among-population variance in fitness could increase for some length of time before reaching a plateau. This possibility is most consistent with a scenario in which different populations explore different peaks in the adaptive landscape, eventually reaching peaks of different heights. Because each population would reach its own fitness maximum, variance in fitness would stop increasing once all of the populations reached their fitness peaks. Alternately, different populations could reach regions of the fitness landscape where they experience the same slopes as each other, leading to a consistent variance in fitness. Because we saw no evidence of populations reaching fitness peaks in previous work (*8*), and because truly parallel slopes are mathematically unlikely, we do not expect this result to occur.

A third possibility is that after an initial increase, among-population variance in fitness could decrease. This possibility is most consistent with different populations exploring different routes to the same fitness peak – or different peaks of the same fitness – and then converging together at the peak(s). Again, because our previous work showed no indication of populations reaching fitness peaks, we do not expect this scenario to occur.

Study System:

The Long-Term Evolution Experiment (LTEE) is an ongoing evolution experiment, using populations of the bacteria *E. coli*. This experiment has been described in detail in previous chapters, but a brief summary follows. The experiment consists of twelve populations of *E. coli*, each descended from a

common ancestor. Six of the populations are Ara^+, capable of growing on the sugar arabinose as their sole carbon source; the other six populations are Ara^-, unable to grow on arabinose as a sole carbon source. Each population exists within a separate 50 mL Erlenmeyer flask, with the cells growing in a growth medium of Davis Minimal salts supplemented with 25 mg/L glucose (DM 25). Each day, a member of the research team transfers 0.1 mL of the previous day's culture into 9.9 mL of fresh DM 25, repeating separately for each of the twelve populations, and places the new cultures in a shaking incubator at 37 ^oC and 120 rpm. All populations grow rapidly enough to exhaust the available glucose prior to the next transfer. Every 75 days, corresponding to every 500 generations, frozen samples are made from each of the LTEE populations, and stored at –80 ^oC.

Previous work:

Lenski and Travisano (1994) previously showed that in the LTEE, among-population variance in fitness increased during the first 10,000 generations of the experiment (*9*). By 10,000 generations, they calculate an among population standard deviation in fitness of between 0.04 and 0.05. Interestingly, though they fit a curve to the among-population standard deviation in fitness as a form of a hyperbola, they did not test whether the hyperbola is a better fit than a linear regression. This data, Figure 7 within the original paper, is reproduced here as Figure 3.1.

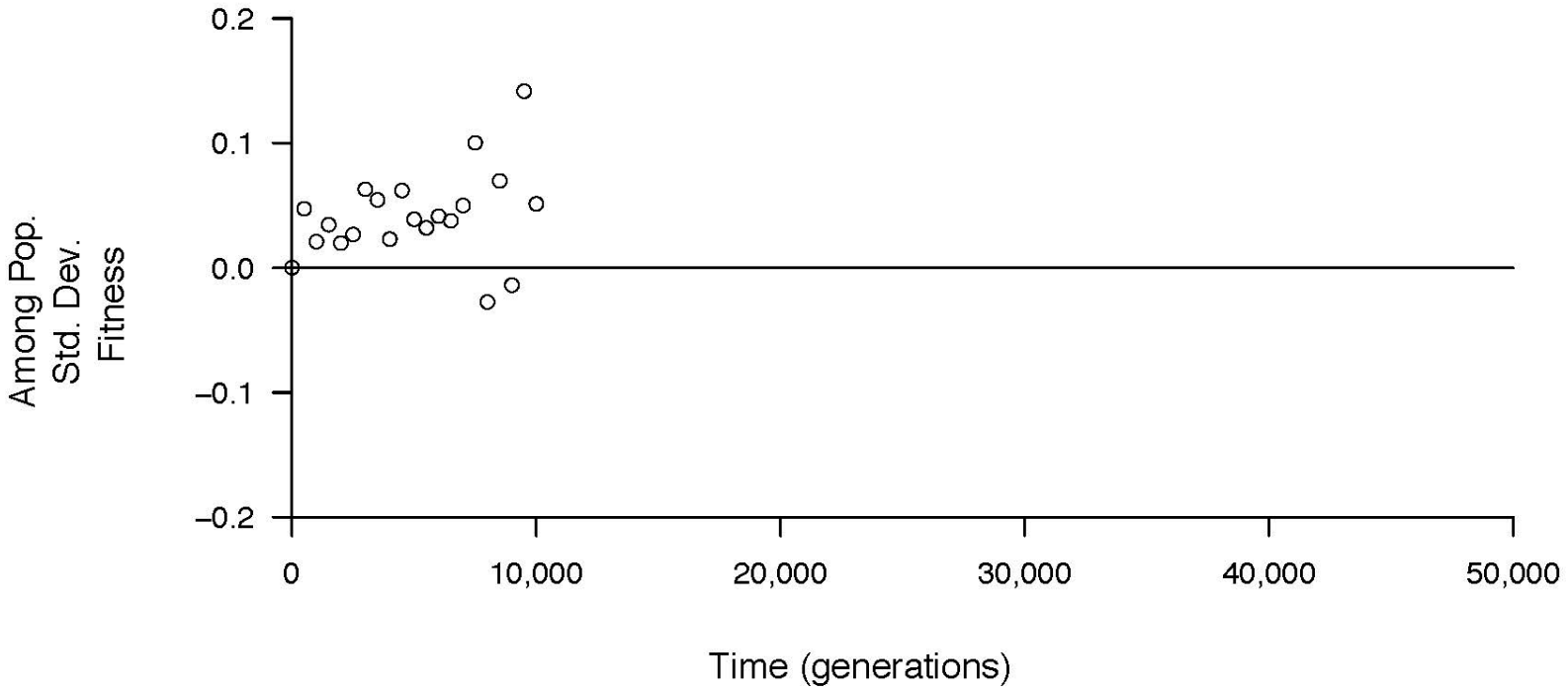

Figure 3.1: **Among-population standard deviation in fitness over the first 10,000 generations across all populations in the LTEE**. Data from Lenski and Travisano (1994).

This figure has been replicated from the summary data available from Lenski's website (*10*). Because only summary data is available, not all of the analyses we will be performing on our other data are applicable to this data set. However, it is clear that there is appreciable among population variance in fitness during the first 10,000 generations of this experiment.

Fitness assays:

We performed fitness assays much as discussed previously (Chapters 1 and 2). Each of these assays has one, but only one, of two differences from the Traditional method outlined in Chapter 1. One, in almost all cases, we performed competitions over the course of three days (roughly 20 generations) rather than one day (roughly 6.67 generations). These additional generations allow greater precision in fitness, but require that the two competing populations have fitness differences no more than approximately 10%. In five individual measurements of 592 three day competitions, the plate from day 3 was uncountable due to error; in these cases, we used the counts and dilution factor from day 2 instead, making these two day (roughly 13.33 generation) competitions. In three additional individual measurements, the plate from day 0 was uncountable; we excluded these three measurements.

Statistical methods:

To calculate the among-population variance in fitness, we followed the procedure outlined in Sokal and Rohlf (1995) (*11*). We treated each time point

separately. Within each time point, we performed an ANOVA, with Population as a random effect. From these ANOVAs, we subtracted the mean square error term from the mean square population term. We divided this difference by the number of replicate blocks. This produces an estimate of the among-population variance. To obtain an estimate of the among-population standard deviation, we first preserved the sign of the variance estimate, and then calculated the square root of the absolute value of the variance estimate. Because the among-population standard deviation in fitness is simply the square root of the among-population variance in fitness, broad-scale patterns (i.e. increases, decreases, or consistency) will be consistent across the two calculations.

All statistical analyses were performed in R version 3.0.2 (*12*). The local smoothing function was fit with the loess() command.

Results and Discussion:

In Figure 3.1, we have replicated a figure from a previous study that examined among-population standard deviation in fitness over the first 10,000 generations of the experiment. Before we look at additional data, it is worth taking a moment to interpret these findings.

One immediate point to notice in Figure 3.1 is that the estimate of the among-population standard deviation varies from one measured time point to another. Part of this difference likely reflects real changes in the degree to which different populations have achieved different levels of fitness over time. Part of the difference, however, is due to measurement error. Indeed, this measurement

error can clearly be seen in the estimates for generations 8,000 and 9,000, when the estimated among-population standard deviation in fitness is negative. This negative result is directly due to measurement error – when the ANOVA mean square error term is larger than the mean square population term, the estimate will be a negative number. The magnitude of these negative numbers, though, can give us an indication of the size of this measurement error. The fact that the majority of the positive estimates of among-population standard deviation in fitness (14 out of 18) are greater than the largest of the negative estimates is a strong point in favor of the among population standard deviation being appreciably greater than 0. Therefore, these populations are achieving different fitness values.

As an initial look at how among-population variance has changed in the first 50,000 generations of the LTEE, we calculated the among population standard deviation in fitness from the data set (*13*) that formed the basis of Chapter 2. That data is presented in Figure 3.2.

In Figure 3.2, we can see that among-population standard deviation in fitness has remained mostly positive across the first 50,000 generations of evolution in the LTEE. However, the relative number of negative estimates has also increased later on in the experiment; half of the estimates after 30,000 generations are negative. There are several possible causes of this.

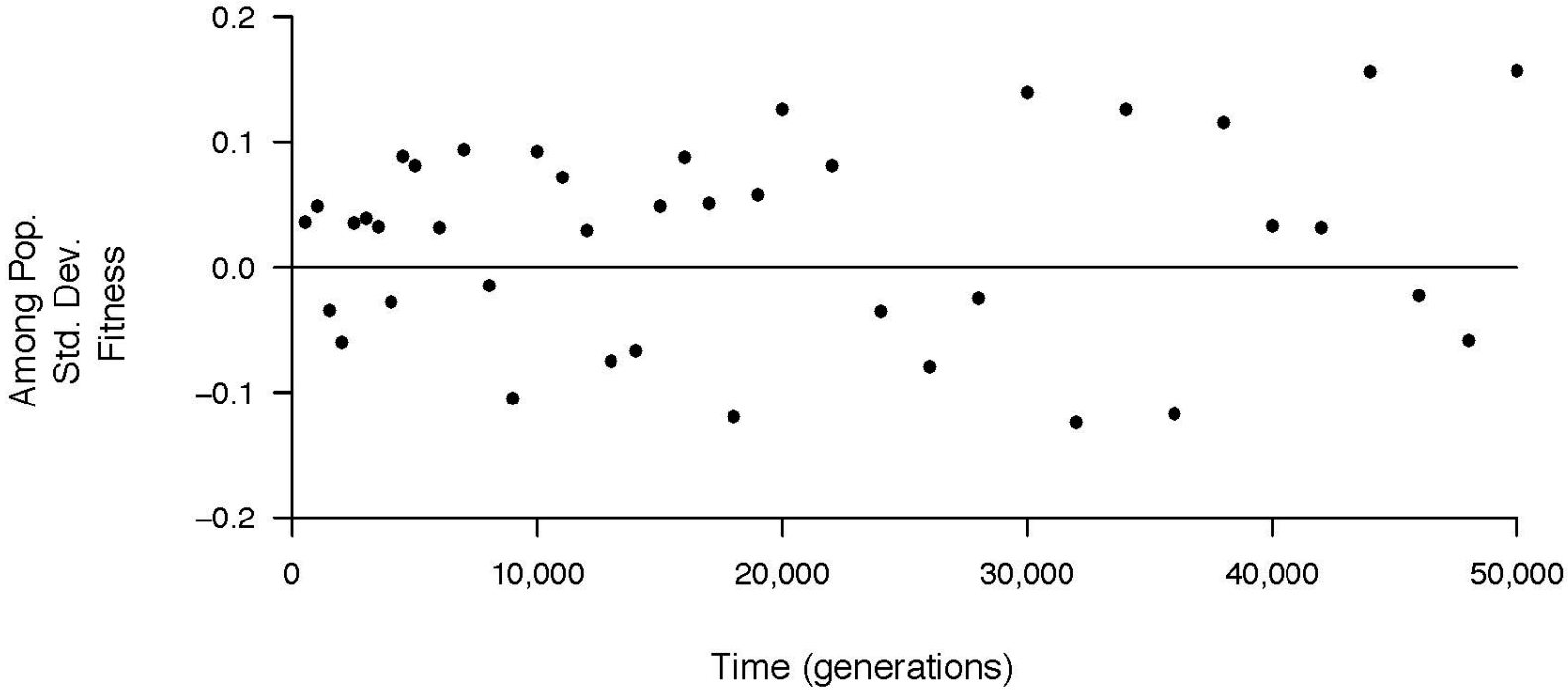

Figure 3.2: **Among-population standard deviation in fitness calculated across all populations in the LTEE**. Data from Wiser et al (2013).

LTEE became hypermutators during the time period studied here (*8*, *14*, *15*), with an additional one becoming a hypermutator after it was already excluded (*16*). These hypermutator populations will therefore have more diverse populations, and consequently greater measurement error in population fitness.

Because our previous results have already shown that increases in the mutation rate of a population lead to increases in fitness (*8*), we would expect that including populations that have become hypermutators at different times would increase the among-population variance in fitness. We therefore choose to restrict our analysis to just the six populations that maintained the ancestral mutation rate in order to make this a more conservative test. Figure 3.3 shows the among-population standard deviation in fitness in the previous data, using only the six populations that maintained the ancestral mutation rate through all 50,000 generations. In these populations, we find that among-population variance in fitness is positive in 28 of the 40 generations after generation 0, a significant result (binomial test, one-tailed $p = 0.008295$).

To address lack of precision cause by low degrees of replication, we generated new data using a smaller number of time points, but a greater degree of replication at each time point – five replicate fitness measurements from each population at each time point, rather than two. For this data set, data from each time point was collected separately, and thus there cannot be a Block effect in the relevant ANOVA, as all replicates of a given population from a given generation were conducted simultaneously. In Figure 3.4 we present this data from just the six populations that did not become hypermutators:

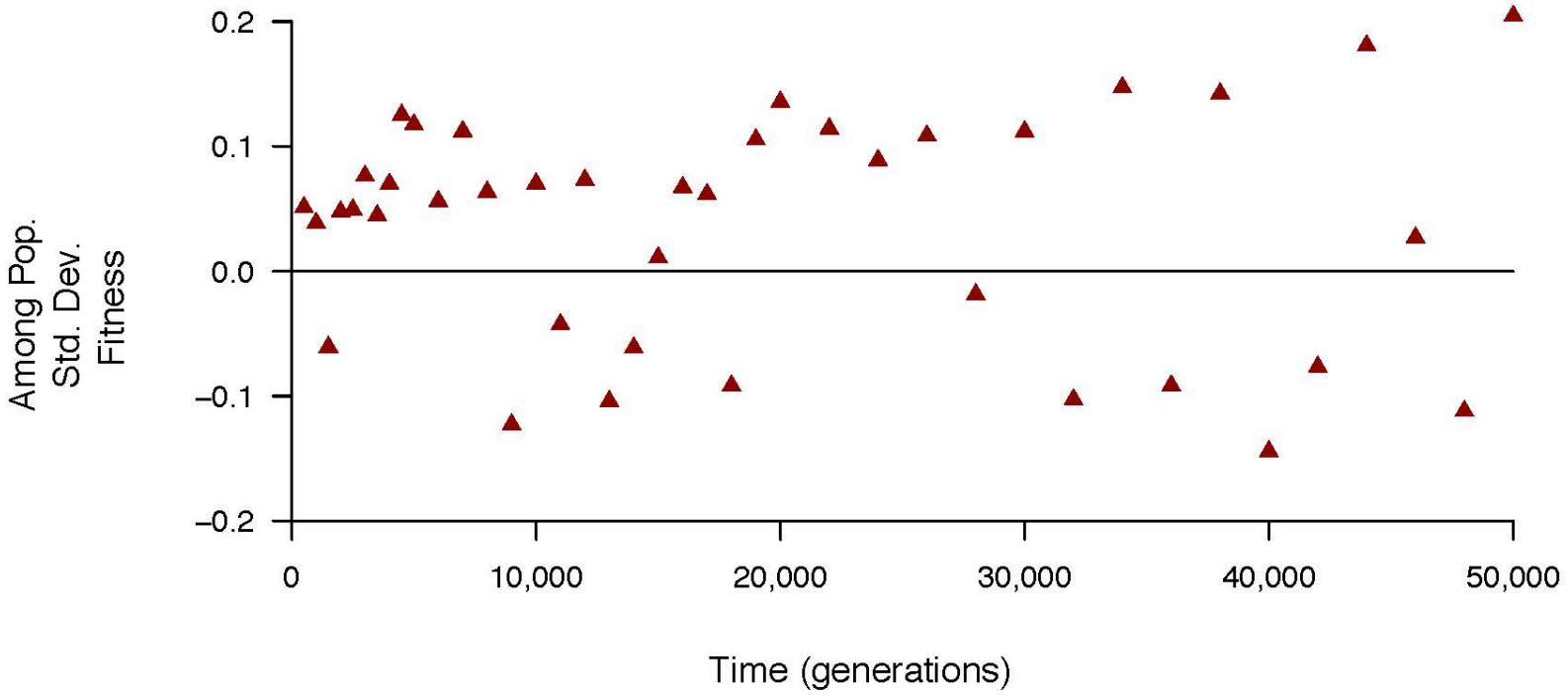

Figure 3.3: **Among-population standard deviation in fitness calculated across LTEE populations that did not become hypermutators**. Data from Wiser et al (2013).

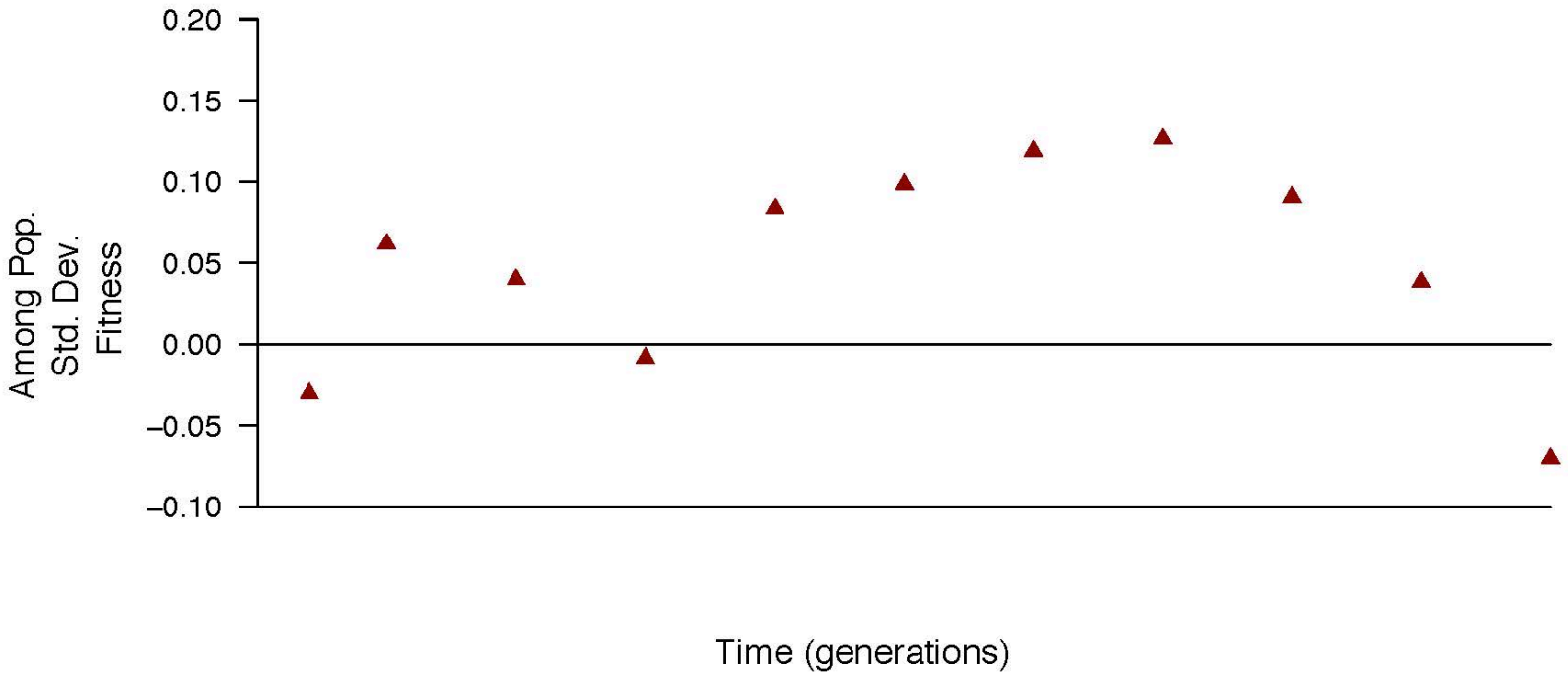

Figure 3.4: **Among-population standard deviation in fitness, calculated across LTEE populations that did not become hypermutators**. Data are new to this study.

As we can see from Figure 3.4, increasing the replication level on measurements decreases the relative frequency of negative estimates for the among-population standard deviation in fitness. We find that eight of the eleven time points produce positive among-population standard deviation estimates. This is not statistically significant (binomial test, one-tailed p=0.1133), although this test is very conservative and suffers from a low statistical power. Although it would be tempting to interpret the variance as declining in the latest generations, we should be cautious about not over-interpreting the data. Much of the apparent decline is driven by the negative estimate at 50,000 generations. Overall, there is substantial agreement between our data sets on the among-population standard deviation in fitness early in the experiment, with decreasing precision in these measurements as populations deviate further from the ancestor.

Komologrov-Smirnov tests:

Although many of the individual time points considered are not, themselves, statistically significant, we can still look for statistical significance in the data set as a whole. Each individual among-population variance in fitness has an associated significance value, because the among-population variance is calculated from an ANOVA table. Under a null distribution, we would expect the cumulative relative frequency of p-values to be equal to that p-value; in other words, 30% of the p-values would be 0.3 or less, 65% of the p-values would be 0.65 or less, etc. The Kolmogorov-Smirnov test allows us to compare the distribution of p-values from our series of ANOVAs to a null distribution and

determine whether we have an excess of small p-values; that is, whether our results as a whole are more significant than expected by chance. We have chosen to use a Kolmogorov-Smirnov test, rather than calculating a False Discovery Rate, as we are interested in whether there is overall evidence of a significant among-population variance in fitness within this data, and are not particularly interested in determining how many of the significant values are likely to only appear significant due to chance.

Figure 3.5 shows the cumulative relative frequency of p-values for our combined data set, considering only the six populations that maintained the ancestral mutation rate. Many more of our p-values are at the small end of the distribution – particularly under 0.2 – than would be expected by chance. This is a highly significant result (Kolmogorov-Smirnov test, 2-tailed, D=0.3024, p=0.0001239. From this, we can see that although individual time points often do not show statistically significant among-population variance in fitness, the data set as a whole does.

The same basic pattern holds for both of the two data sets considered separately. Figure 3.6 shows the cumulative frequency of p-values for just the Wiser et al (2013) data set, again considering only the populations that maintained the ancestral mutation rate. These data are highly significant (Kolmogorov-Smirnov test, 2-tailed, D=0.2689, p=0.004812). We see the same pattern in Figure 3.7 for the data set of higher replication but fewer time points. These data are highly significant (Kolmogorov-Smirnov test, 2-tailed, D=0.514, p=0.003187).

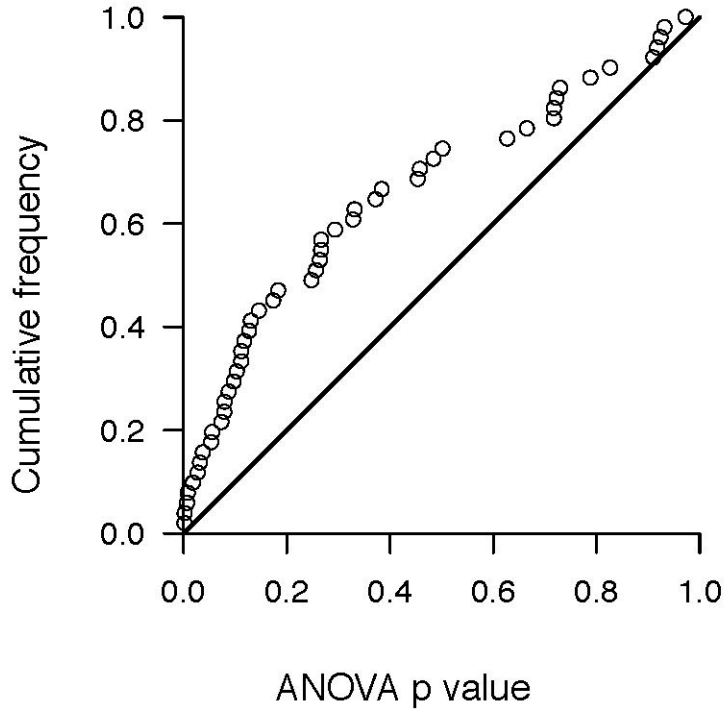

Figure 3.5: **Cumulative frequency of p values among ANOVAs used to calculate among-population variance in fitness**. Points represent empirical data; the solid line at y = x shows the null expectation. Data is combined from Wiser et al (2013) and new data for this study; populations that became hypermutators were excluded.

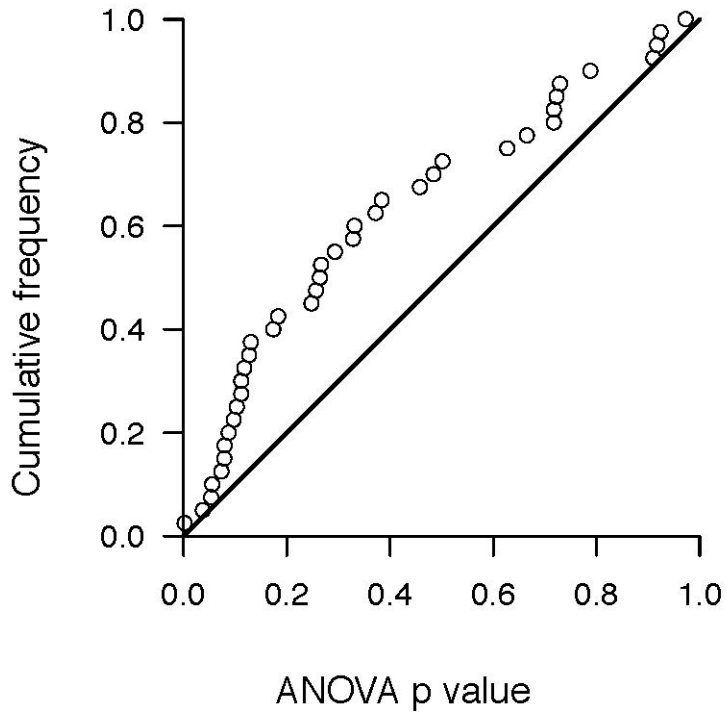

Figure 3.6: **Cumulative frequency of p values among ANOVAs used to calculate among-population variance in fitness**. Points represent empirical data; the solid line at y=x shows the null expectation. Data from Wiser et al (2013).

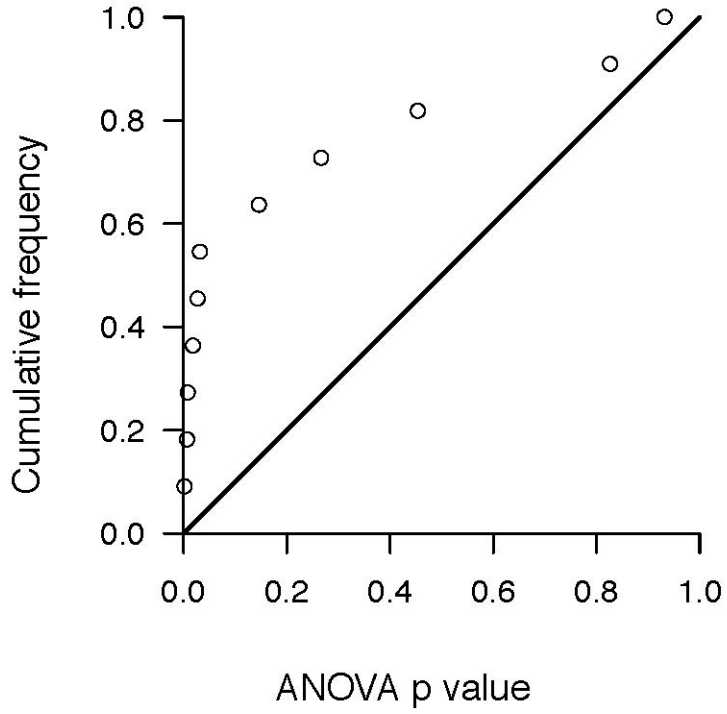

Figure 3.7: **Cumulative frequency of p values among ANOVAs used to calculate among-population variance in fitness**. Points represent empirical data; the solid line at y=x shows the null expectation. Data are new for this study.

Given that our data is highly significant in both individual data sets considered separately, as well as in our combined data considered as a whole, we conclude that there is significant evidence of among-population variance in fitness within our data. This further strengthens our conclusion that our populations are not converging at the top of a single peak in the adaptive landscape.

Using population pairs to examine finer scale differences:

The preceding analyses show that there is a substantial among-population variance in fitness across the first 50,000 generations of the LTEE. This variance increases rapidly from 0 at the start of the experiment to significant levels within the first few thousand generations, and remains positive thereafter. These analyses lack sufficient statistical power to state with confidence whether this variance continues to increase or remains at a constant level. However, examining the among-population variance across a range of populations is not the only way to look at differences that have evolved across different populations.

From our previous work, we have already established fitness trajectories for individual populations within the LTEE (*8*). We also know that our fitness assays are most precise when our two competitors have similar fitnesses (*17*). This poses a potential problem for accurately determining differences in fitness of two evolved populations late in the experiment: each one is being compared to a common ancestor, and thus each has an increasing measurement error. Further, comparing two different populations to a common competitor to

determine which one is competitively superior assumes complete transitivity in fitness; if A > B, and B > C, it assumes A > C, which may or may not be the case.

One obvious way to overcome these limitations is to compete different evolved populations against each other. Competing populations directly against each other, instead of competing each against a common competitor, avoid the issue of error propagation from non-transitivity. We would further expect populations that have been evolving in the lab for the same number of generations to have fitness values closer to each other than they would to their common ancestor, inherently reducing the impact of measurement error. Additionally, two populations of similar fitness can be competed against each over a larger number of generations, which increases the precision of our fitness measurements.

We therefore chose to compete pairs of evolved populations from each of the time points involved in the individual population fitness trajectories. In order to make our pairings as independent as possible, we assigned each population to only a single pairing. By making the population pairings independent, we reduce the capacity for one population to have excessive influence on our findings – similar patterns would be caused by similar evolutionary patterns, rather than the effect of a single population on multiple different pairings.

For each of our pairings, we need both an Ara^+ and an Ara^- population. Because our populations are labeled as Ara-1, Ara-2,…Ara-6, Ara+1, Ara+2…Ara+6, the simplest approach is to compete each Ara^+ population against the equivalently numbered Ara^- population. This is not strictly necessary

– there is nothing in particular linking Ara+1 to Ara-1 more than to Ara-5. Three of our populations become difficult to work with in later time points: Ara-2 and Ara+6 stop growing reliably on the TA medium and Ara-3 has a substantial population increase due to its ability to metabolize citrate in the presence of oxygen in later generations (*3*). We competed Ara-2 against Ara+2 for the first 30,000 generations, the time period in which Ara-2 grows reliably on TA plates. Similarly, we only consider competitions of Ara-3 against Ara+3 for the first 32,000 generations of the experiment, as this is before the Cit^+ population expansion. We conducted competitions over the course of 3 days (roughly 20 generations), with two replicate measurements at each generation. We structured the replicates such that one measurement for each generation collectively formed a block, and we repeated this block a second time.

For the pairing of Ara+1 v Ara-1, we chose to expand the number of generations under consideration. In the other pairings we looked at as many of the 40 distinct time points as possible that were used to establish individual population fitness trajectories (see Chapter 2). For this pairing, we looked at each 500 generation interval across the first 50,000 generations. Because 101 unique time points were too many to include in a single block, we had to split the time points into two separate collections. We chose to do so in the form of one set of every generation evenly divisible by 1,000, and a second set of those generations ending in 500. This interweaving, as opposed to splitting populations between early and late generations, reduces the likelihood of a

systematic temporal difference between blocks having a significant effect on the pattern of fitness change.

Figure 3.8 shows the fitness of population Ara-1 relative to population Ara+1 for the first 50,000 generations of the LTEE. All of the graphs in this section follow the same basic format. The light colored, open symbols represent each individual measurement of fitness. The dark, filled symbols represent the average at each individual generation. The line is a local smoothing function, finding the average trend through nearby points, without imposing a specific mathematical relationship across the data set as a whole. From this figure, we can see that Ara-1 quickly gained a lead over Ara+1 in fitness, rising to about a 5% advantage by 25,000 generations and maintaining that lead through 50,000 generations.

It is not surprising that Ara+1 is lagging in fitness – previous findings (Chapter 2) showed that it has a substantially lower fitness than all other populations – but a comparison directly against another population allows us to glean additional information. For one, the precision in these measurements is substantially greater than what we were able to obtain by competing the evolved populations against their ancestor. Figure 3.9 places the data from Figure 3.8 in context with our expectations. The light-colored, solid line is the ratio of the fitness of Ara-1 to Ara+1 from the curves fit in Chapter 2; the dashed lines to either side show the 95% confidence interval of this expectation, obtained through bootstrapping the data.

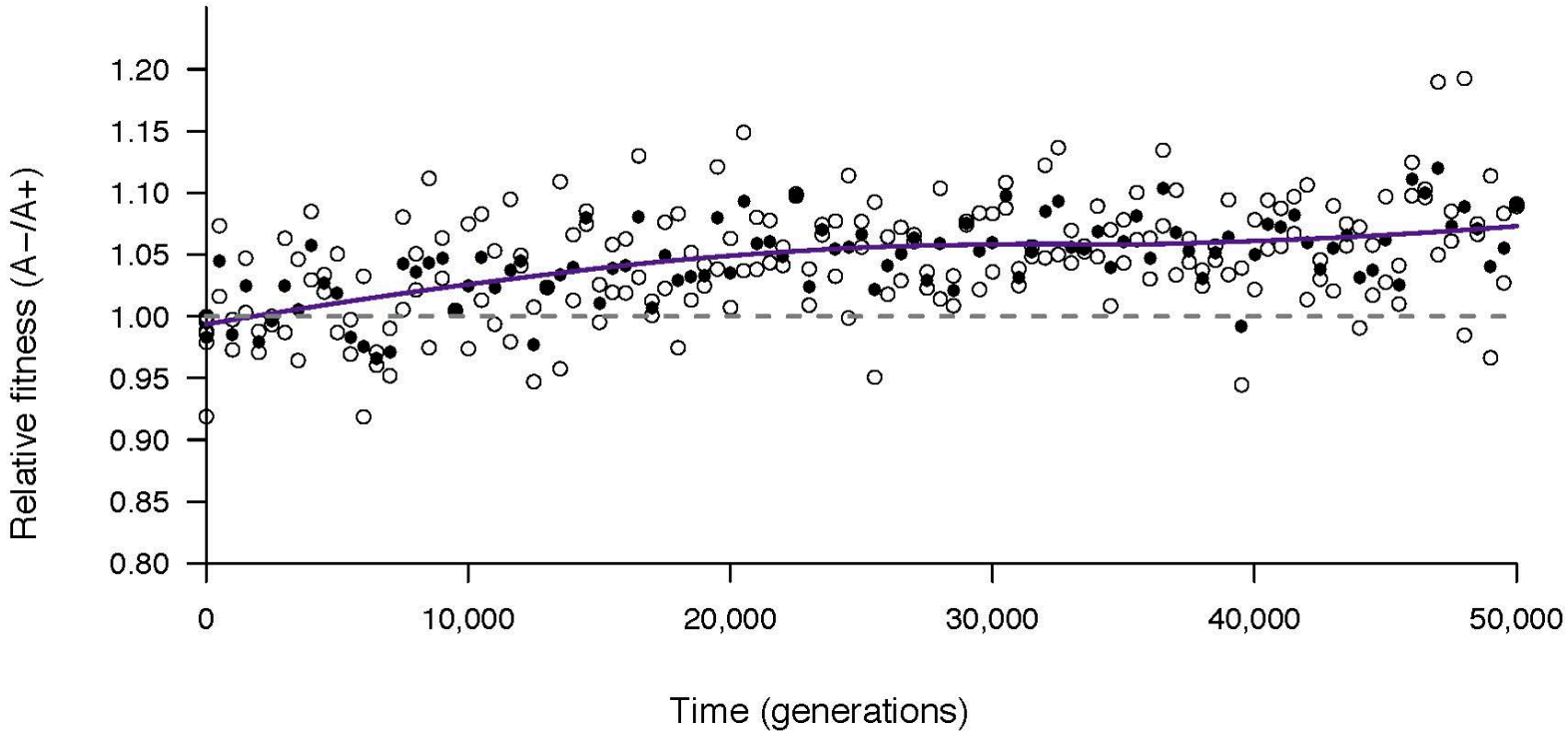

Figure 3.8: **Ara-1 v Ara+1**. Open symbols show each measured value of relative fitness from head-to-head competitions. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness. The solid purple curve is a local smoothing function showing the general trend of the data.

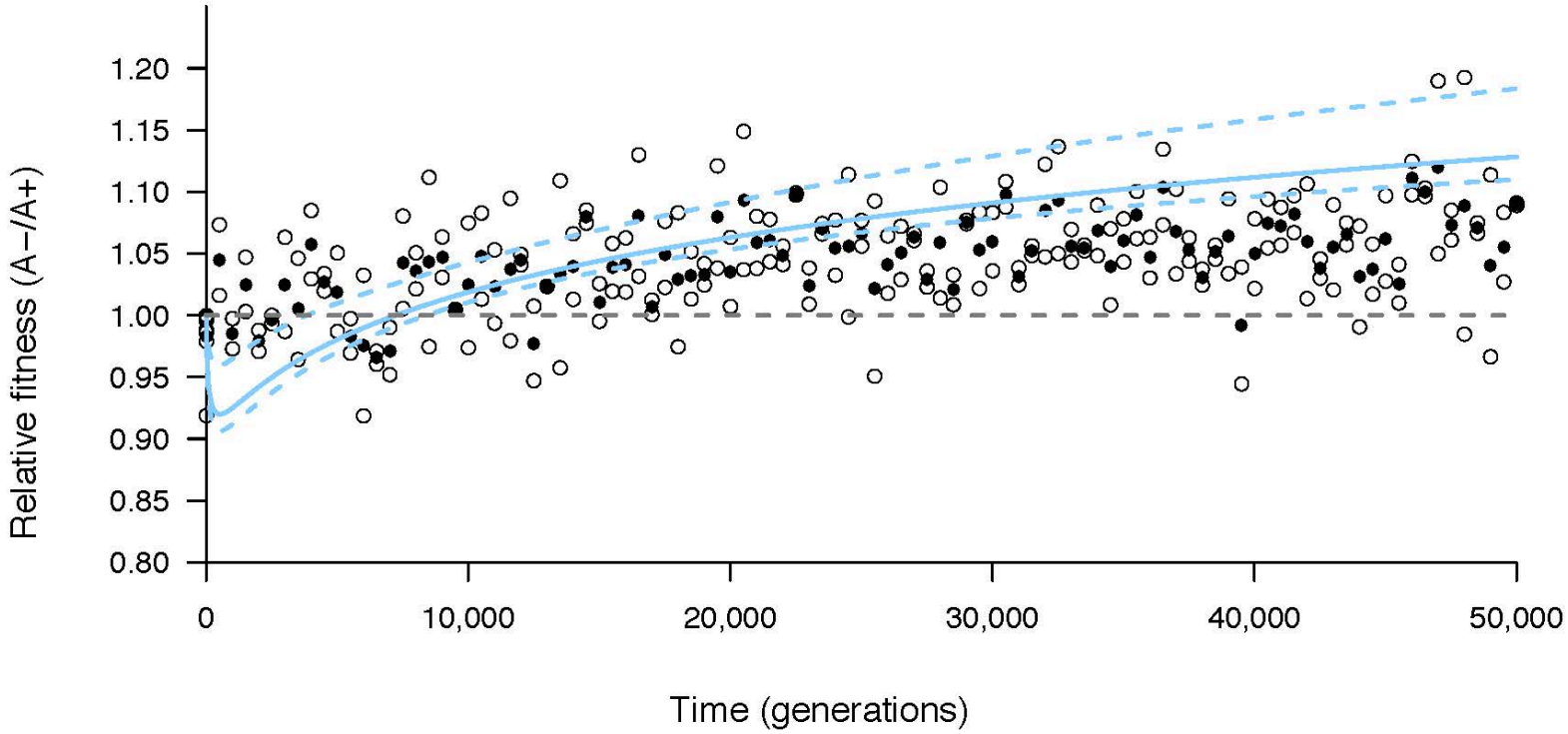

Figure 3.9: **Ara-1 v Ara+1**. The solid light blue line shows the mean fitness of Ara-1 relative to Ara+1, based on 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show the corresponding, non-parametric 95% confidence. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness.

As we can see, the actual data differ from the expectations in a number of ways. The expectations suggest Ara+1 would have a sizeable advantage at the earliest time points, while the empirical data show much less, or possibly no advantage for Ara+1 even at the earliest time points. This is influenced by constraints on model fits. All of the populations start out with the same fitness at generation 0, and all are fit to the same mathematical function, though with different parameter values. In order for a population to have a steeper increase in fitness later in the experiment, it must by necessity have a shallower increase in fitness early in the experiment. Therefore, the fact that Ara+1 shows a slow rate of increase in fitness in later time points requires it to have a relatively rapid increase in fitness early, which causes the prediction that it would have a higher fitness than anything it competed against at these early time points. The fact that our expectations are calculated from a ratio of two smooth power law curves means that we can only predict zero, one, or two changes in which population is gaining fitness more rapidly in any particular pairing. Direct measurements of pairs of evolved populations could have many more changes in which population is gaining fitness more rapidly. The expectations also suggest that Ara-1 would end up with a larger advantage over Ara+1 than it achieved, with a mean advantage in the 40,000+ generation range that is more consistent with the highest individual measurements than the means at each generation. Further, the expectations show a widening uncertainty as time progresses, while the visible spread in measurements of fitness differential between the two populations remain roughly constant.

It is also noteworthy that the fitness differential between these populations is as low as it is. Even Ara+1, the population that has seen the least gain in fitness by 50,000 generations, is roughly 40% more fit than its generation 0 ancestor. That the fitness differential between the populations is only ~5% means that there has been a striking degree of parallelism in fitness changes across replicate populations. Figure 3.10 shows this in stark contrast. The two individual population fitness trajectories each increase markedly from the ancestor, while staying relatively close to each other. As a consequence, both the expectation for, and the measured values of, their fitness relative to each other remains much closer to 1. If anything, the measured population pair relative fitness is closer to 1 than the expectation is, suggesting that these populations have more similar fitness trajectories than each would appear from the individual trajectories against the ancestor.

The pairing of populations Ara-4 and Ara+4, shown in Figure 3.11, displays a somewhat different pattern than Ara-1 and Ara+1. In this pairing, Ara+4 has a notable early lead of roughly 5% by 5,000 generation, but it is only temporary. Ara-4 catches up by generation 10,000, and then takes a lead of its own of roughly 2-3% for the next several tens of thousands of generations. From Figure 3.12, we can see that this pairing behaves largely as expected, with the majority of measured relative fitness points falling within the confidence interval of the expectations. This is notable, because as we can see from Figure 3.13, populations Ara-4 and Ara+4 have individual population fitness trajectories that are much more similar to each other than Ara-1 and Ara+1 do. Yet despite these

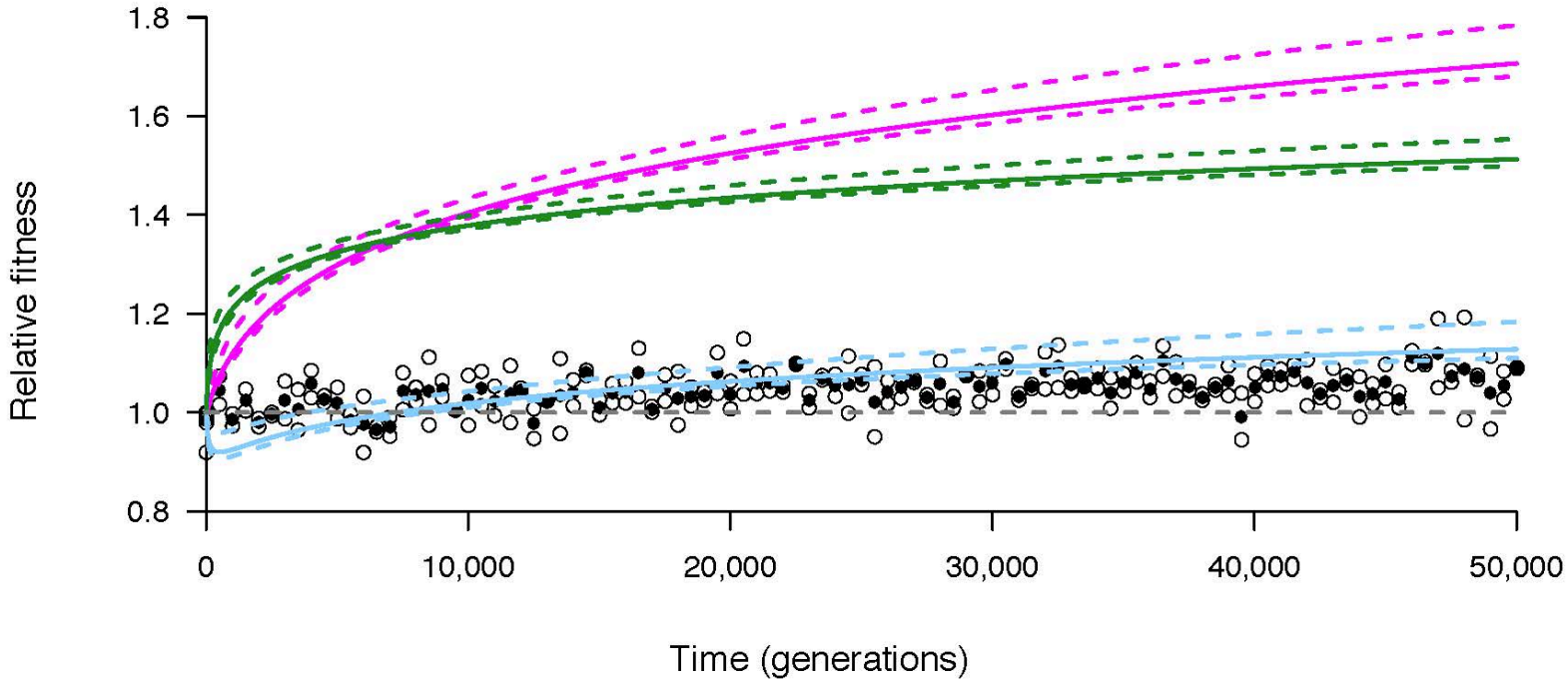

Figure 3.10: **Ara-1 v Ara+1**. The solid magenta curve shows the mean Power Law fitness trajectory for population Ara-1. The solid green curve shows the mean Power Law fitness trajectory for population Ara+1. The solid light blue curve shows the mean fitness of Ara-1 relative to Ara+1. Each of these lines is the mean across 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show corresponding non-parametric 95% confidence intervals around the solid curves. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point.

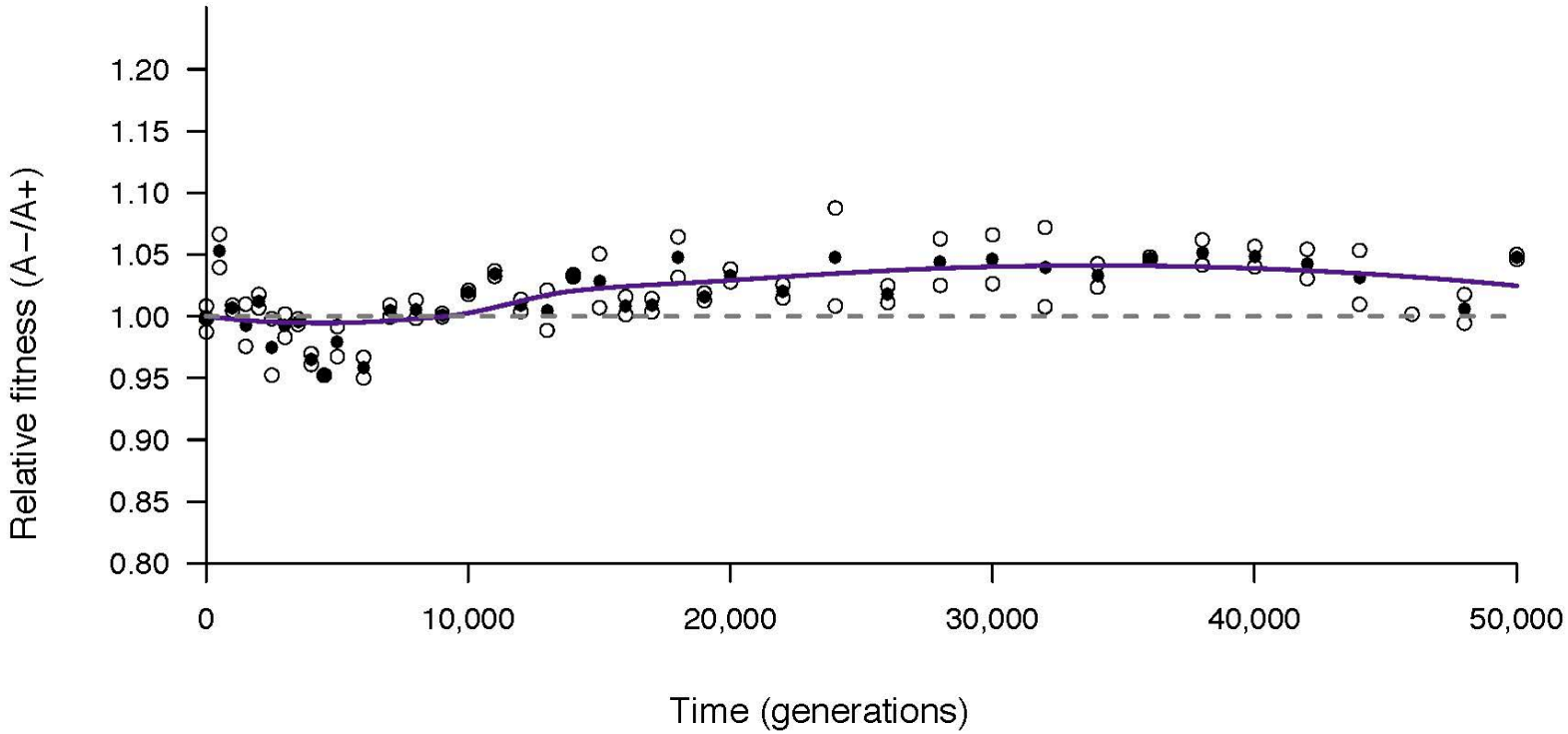

Figure 3.11: **Ara-4 v Ara+4**. Open symbols show each measured value of relative fitness from head-to-head competitions. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness. The solid purple curve is a local smoothing function showing the general trend of the data.

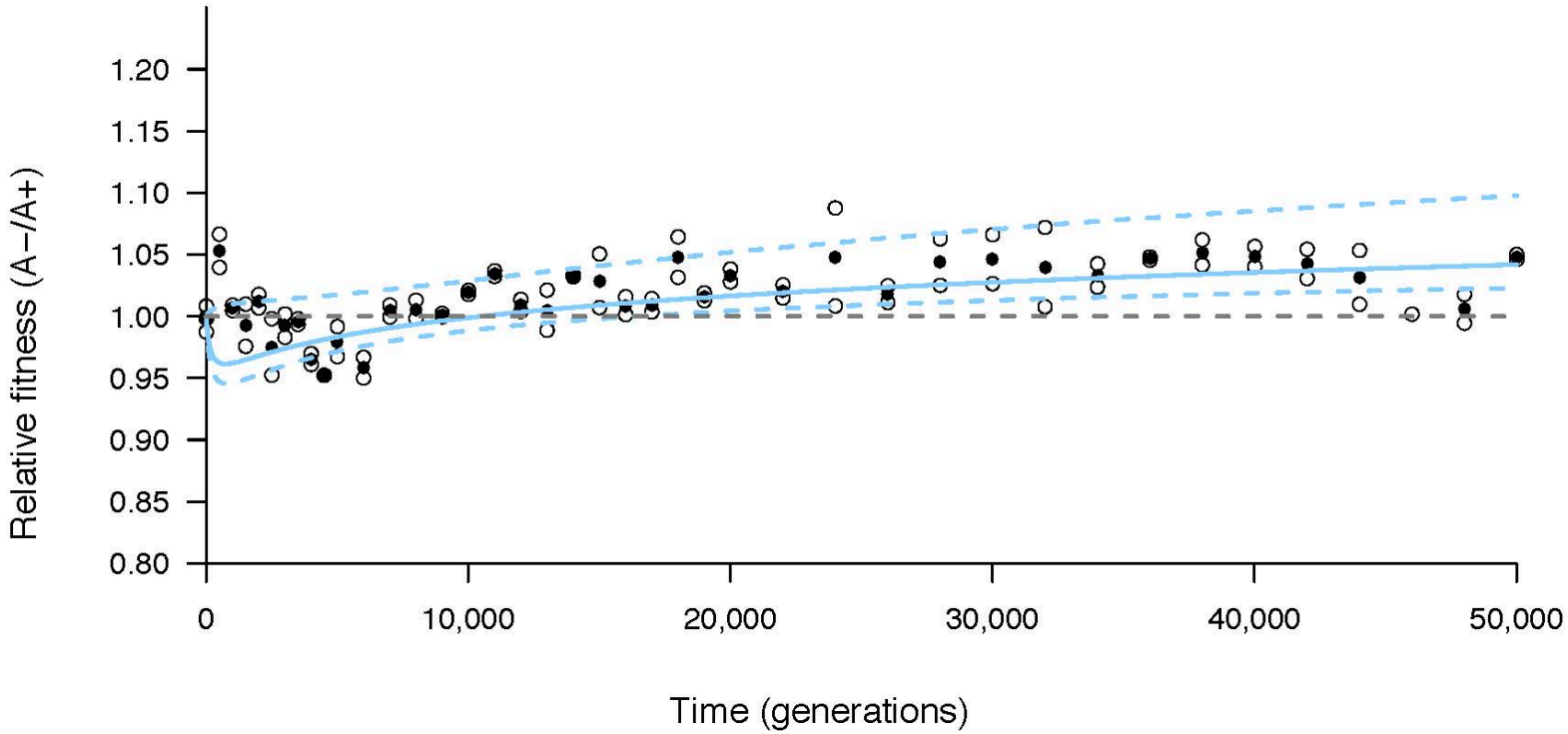

Figure 3.12: **Ara-4 v Ara+4**. The solid light blue line shows the mean fitness of Ara-4 relative to Ara+4, based on 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show the corresponding, non-parametric 95% confidence. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness.

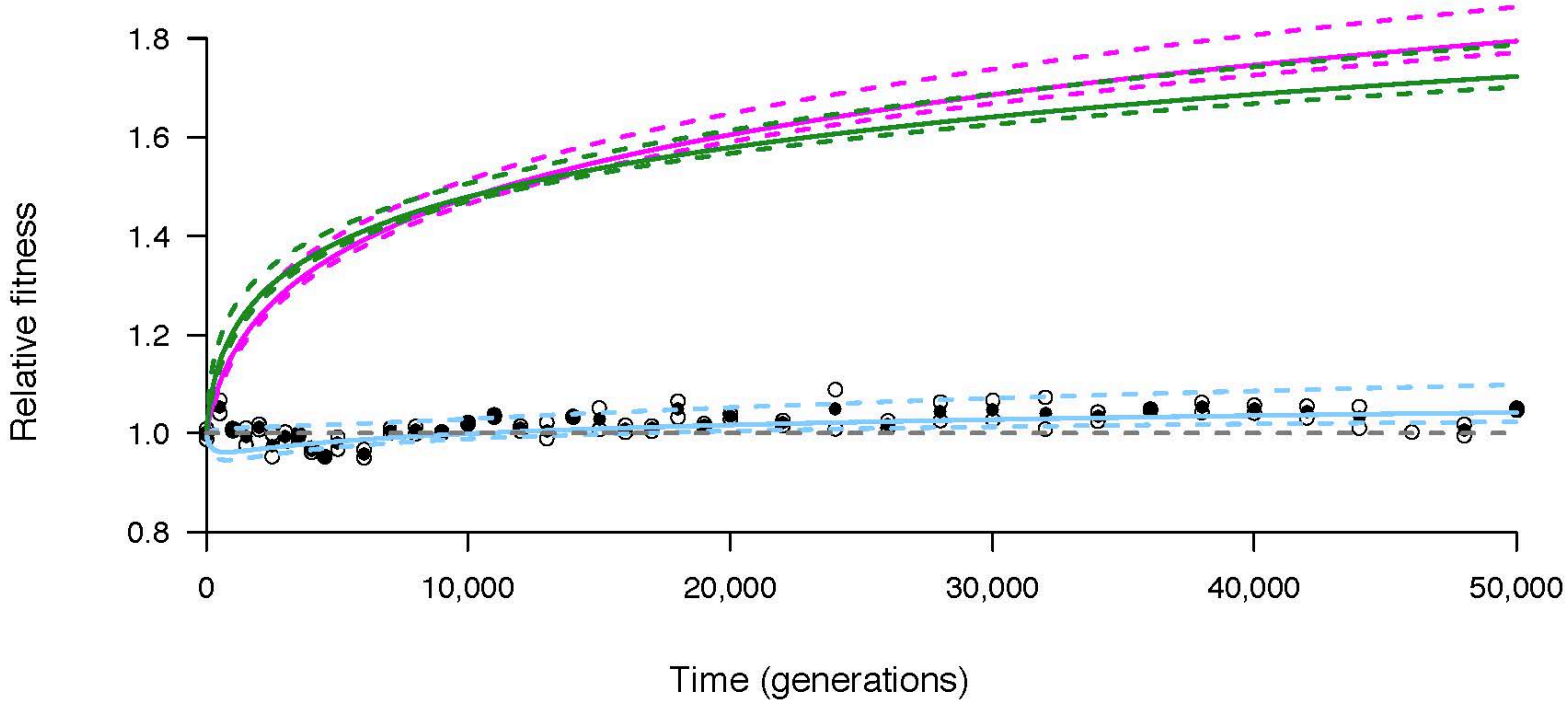

Figure 3.13: **Ara-4 v Ara+4**. The solid magenta curve shows the mean Power Law fitness trajectory for population Ara-4. The solid green curve shows the mean Power Law fitness trajectory for population Ara+4. The solid light blue curve shows the mean fitness of Ara-4 relative to Ara+4. Each of these lines is the mean across 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show corresponding non-parametric 95% confidence intervals around the solid curves. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point.

small differences between populations, compared to both of their substantial differences from the ancestor, we observe essentially the expected pattern in the empirical data.

Figure 3.14 shows the pairing of populations Ara-5 and Ara+5. Much more so than the previous pairings, this one shows marked change over time. For the first 10- to 15,000 generations, Ara+5 has a substantial and widening advantage over Ara-5, reaching roughly a 5% advantage around generation 15,000. At this point, however, Ara-5 begins to rise in relative fitness, reaching roughly equal fitness to Ara+5 by approximately generation 35,000, and subsequently surpassing Ara-5, reaching a roughly 3% advantage over Ara+5 by 50,000 generations. From Figure 3.15, we can see that this broad-strokes pattern is very similar to what we expect – an initial lead for population Ara+5, followed by population Ara-5 catching up – but the measured fitness difference between the two populations is typically tilted more in favor of population Ara-5 than expected. As we can see in Figure 3.16, in the latest generations the two populations are expected to have such similar fitnesses that the confidence intervals overlap, though with population Ara+5 having the higher mean estimate. However, the direct competition data show population Ara-5 having a slightly higher mean estimate for fitness.

The pattern for the pairing of Ara-5 and Ara+5 is particularly striking, because it demonstrates how different populations can reach very different local regions of the adaptive landscape. In Chapter 2, we saw that most of the population in the LTEE have fitness trajectories that follow power laws, including

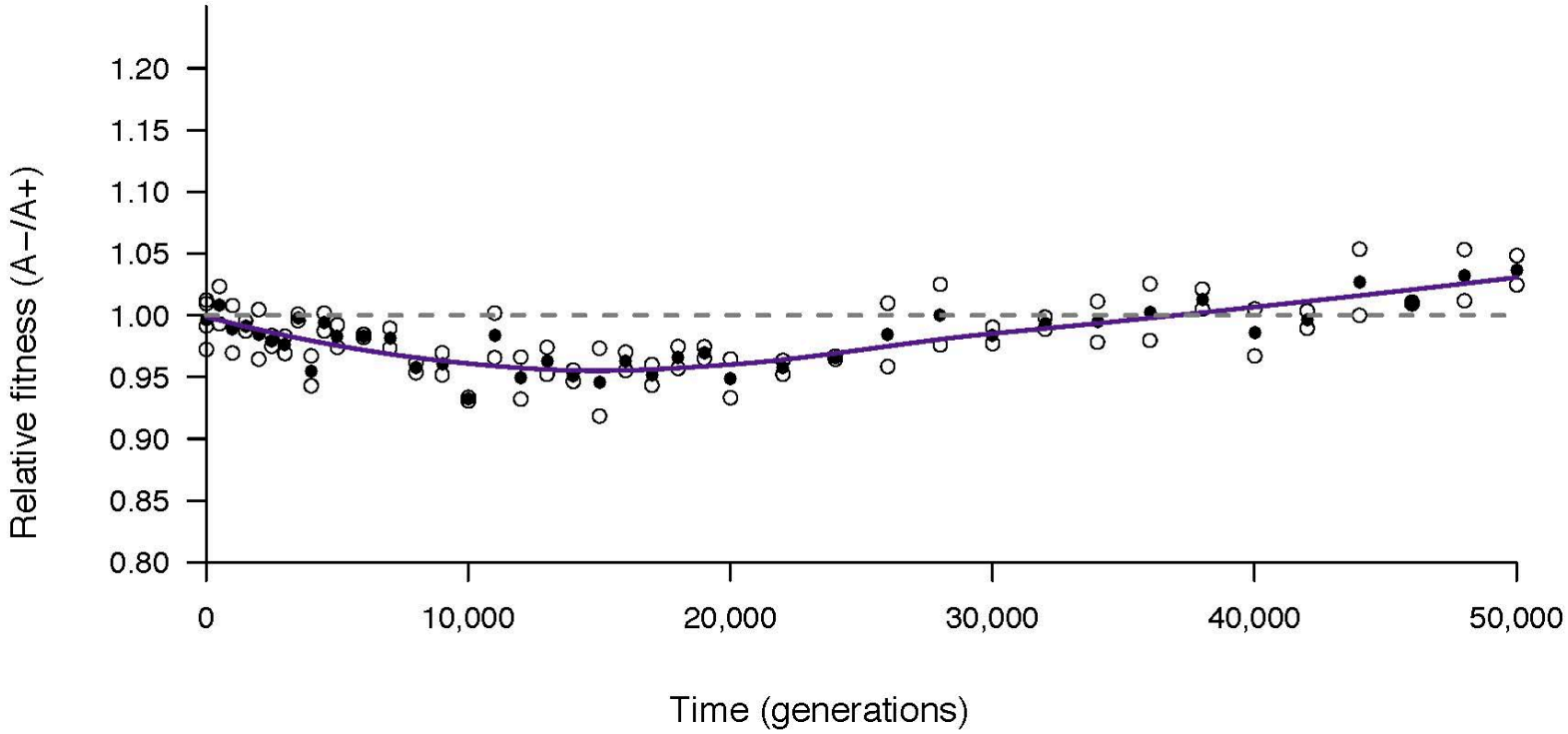

Figure 3.14: **Ara-5 v Ara+5**. Open symbols show each measured value of relative fitness from head-to-head competitions. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness. The solid purple curve is a local smoothing function showing the general trend of the data.

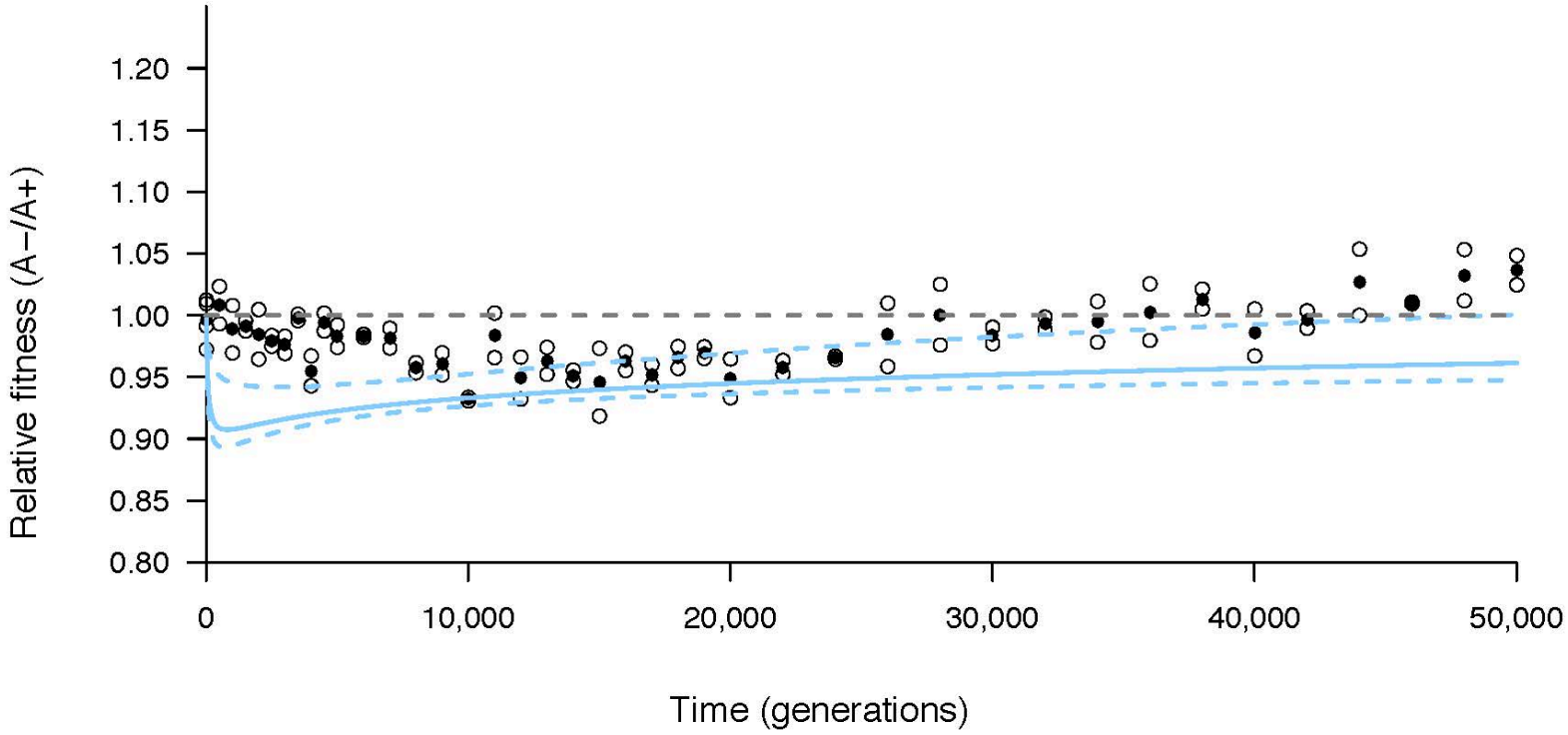

Figure 3.15: **Ara-5 v Ara+5**. The solid light blue line shows the mean fitness of Ara-5 relative to Ara+5, based on 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show the corresponding, non-parametric 95% confidence. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness.

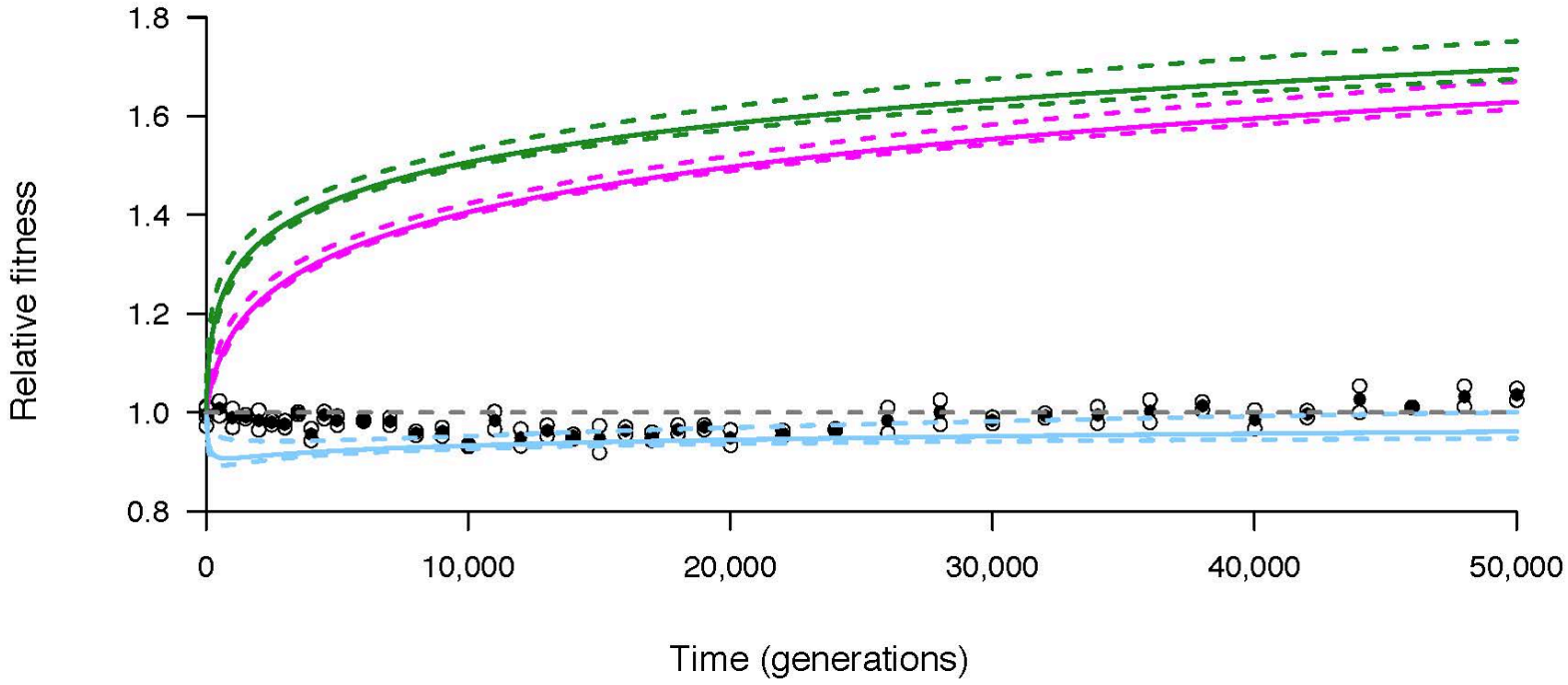

Figure 3.16: **Ara-5 v Ara+5**. The solid magenta curve shows the mean Power Law fitness trajectory for population Ara-5. The solid green curve shows the mean Power Law fitness trajectory for population Ara+5. The solid light blue curve shows the mean fitness of Ara-5 relative to Ara+5. Each of these lines is the mean across 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show corresponding non-parametric 95% confidence intervals around the solid curves. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point.

both Ara-5 and Ara+5. Different parameter values within the power law can lead to different populations improving at different rates at various points in their evolution, but each individual population would be expected to follow a relatively simple trajectory in fitness over time. However, measuring the populations directly against each other can show cases like this pairing, where one of the two gets an early lead, but that lead is subsequently lost as the initially-trailing population catches up and later surpasses the one with the faster start.

Figure 3.17 shows the pairing of Ara-2 and Ara+2. This pairing only extends through the first 30,000 generations, before population Ara-2 no longer grows reliably on TA plates. In this pair, Ara-2 takes a rapid early lead, climbing to approximately a 7-8% fitness advantage by 5,000 generations. This trend then reverses, with the two populations reaching approximately equal fitness by generations 15,000. Subsequently, population Ara-2 regains a lead, reaching a roughly 5% fitness advantage over Ara+2 by generation 30,000. Interestingly, this is not even close to the pattern we expected, as shown in Figure 3.18. Our expectation is for an initial advantage in population Ara+2, gradually shrinking or even disappearing by 50,000 generations. This pairing shows an unusually wide confidence interval in its expectation. This is likely influenced by how much of their individual fitness trajectories overlap; as we can see in Figure 3.19, the confidence intervals of Ara-2 and Ara+2's individual fitness trajectories overlap by 15,000 generations, and the mean values lie within each other's confidence intervals by 25,000 generations.

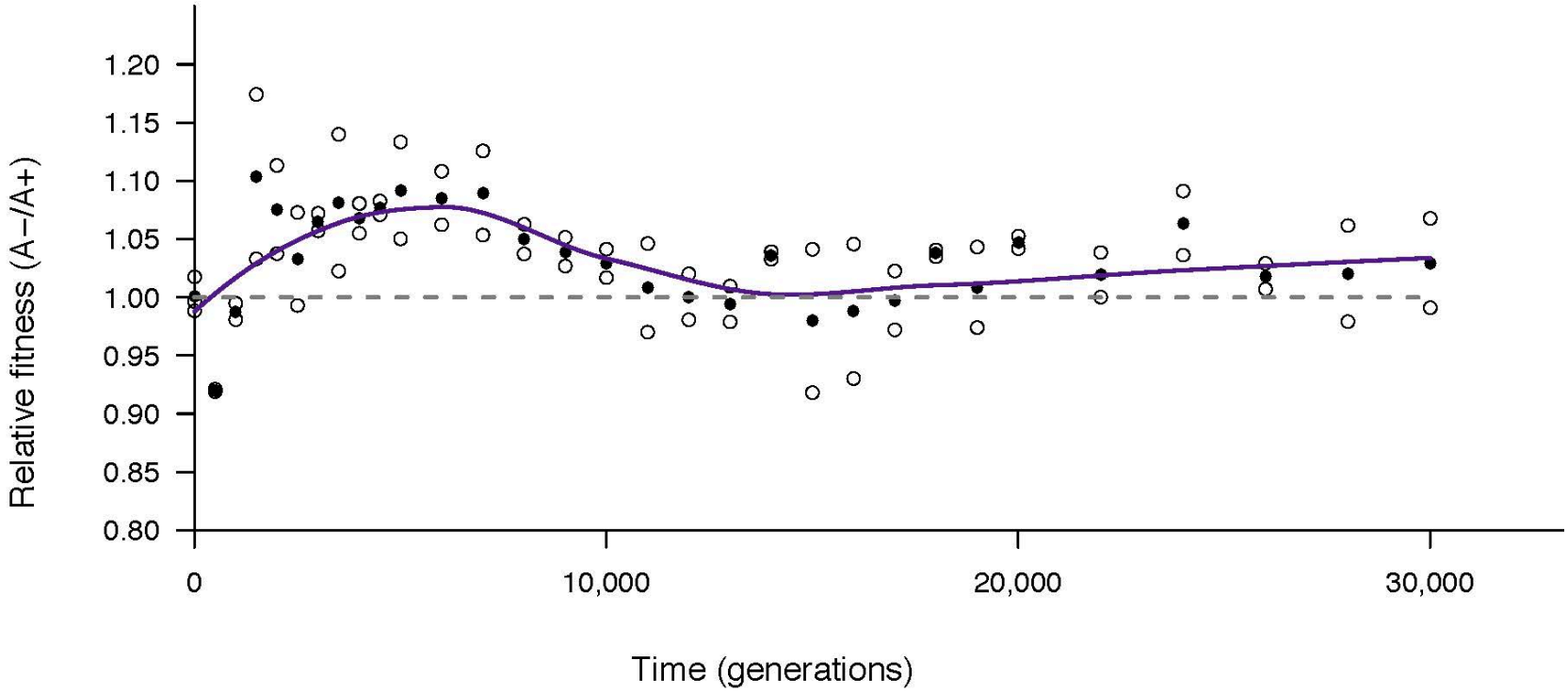

Figure 3.17: **Ara-2 v Ara+2**. Open symbols show each measured value of relative fitness from head-to-head competitions. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness. The solid purple curve is a local smoothing function showing the general trend of the data.

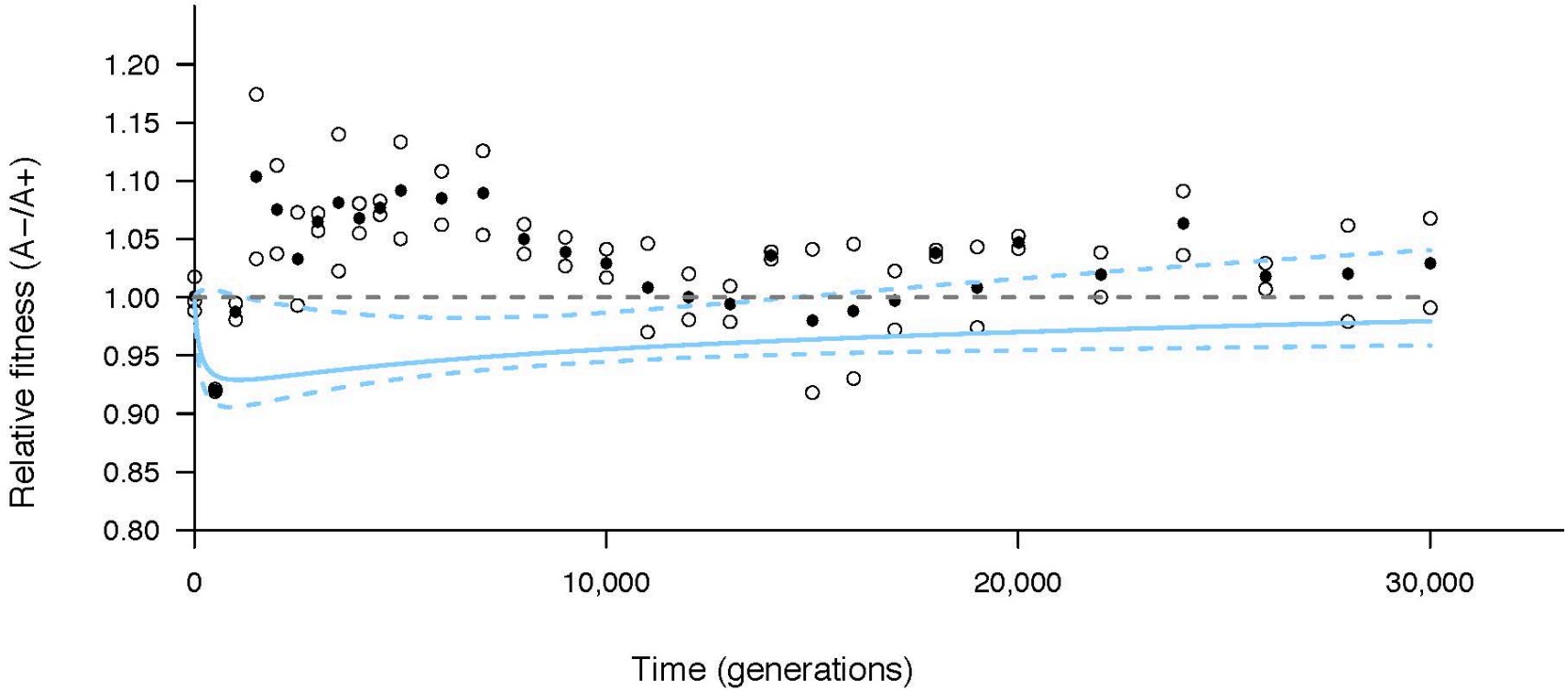

Figure 3.18: **Ara-2 v Ara+2**. The solid light blue line shows the mean fitness of Ara-2 relative to Ara+2, based on 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show the corresponding, non-parametric 95% confidence. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness.

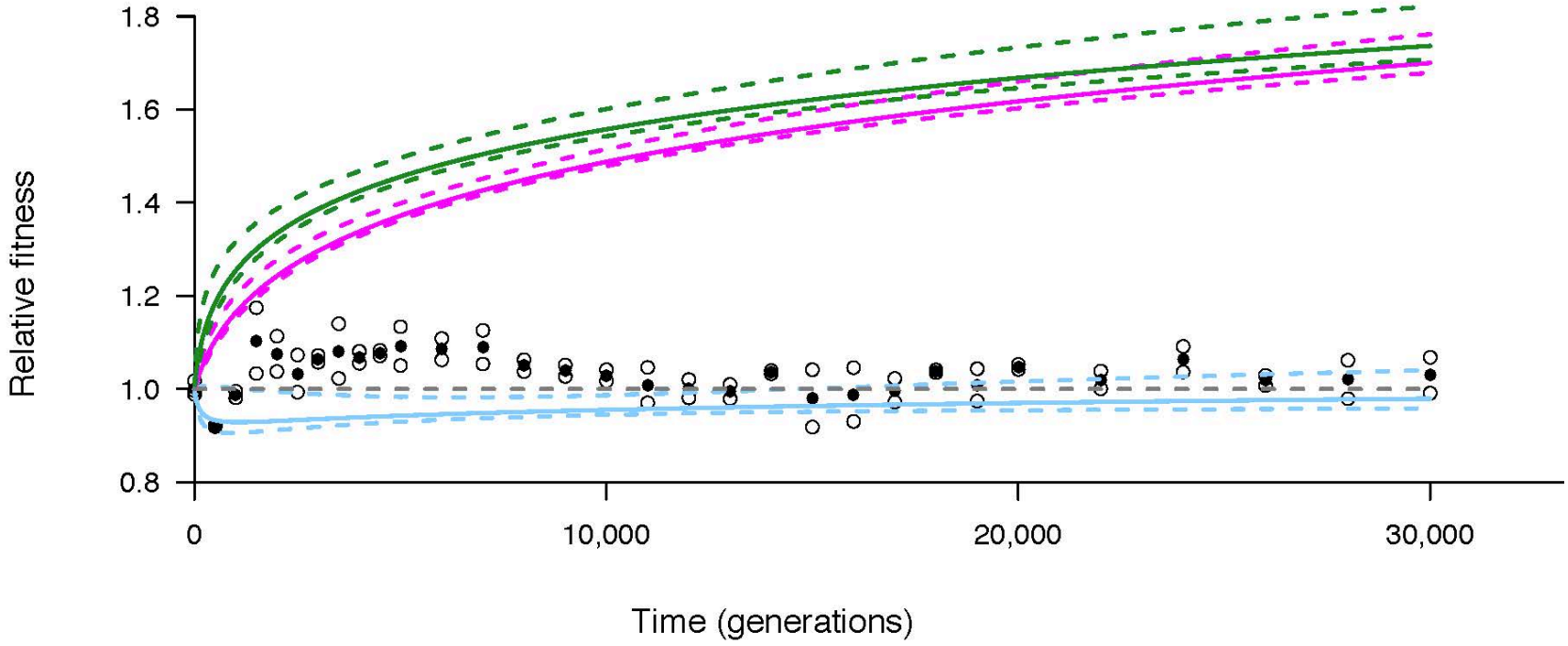

Figure 3.19: **Ara-2 v Ara+2**. The solid magenta curve shows the mean Power Law fitness trajectory for population Ara-2. The solid green curve shows the mean Power Law fitness trajectory for population Ara+2. The solid light blue curve shows the mean fitness of Ara-2 relative to Ara+2. Each of these lines is the mean across 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show corresponding non-parametric 95% confidence intervals around the solid curves. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point.

Figure 3.20 shows the pairing of populations Ara-3 and Ara+3. Between generations 32,000 and 34,000, population Ara-3 went through a massive population expansion, as it developed the ability to metabolize citrate in the presence of oxygen. Citrate is present in our growth medium DM 25 at a high enough concentration that this population has a roughly 7-fold larger population size than other populations in the LTEE (*16*). Therefore, we have restricted our analysis to just those time points before the citrate-utilizing population expansion. In Figure 3.20, we can see that population Ara+3 has a steadily widening fitness advantage over population Ara-3 for the first 25,000 generations, at which point it maintains a roughly 10% fitness advantage over population Ara-3 through 32,000 generations.

Figure 3.21 shows how well these measurements match our predictions. From the trajectories of populations Ara-3 and Ara+3 competed against the ancestor, we would expect population Ara+3 to have a substantial and continual fitness advantage over population Ara-3. Our data reflect this, with many of the individual measurements falling within the 95% confidence interval of our expectation. The deviations from expectation are small relative to the two population's individual trajectories, as is shown in Figure 3.22.

Looking across the set of these population pairs, we see that fitness differences between a chosen pair of populations do not always accumulate monotonically – a population may increase its fitness relative to another for a while, only to lose that advantage, and then gain it back again. Nor do the changes in relative fitness always perfectly track those we would expect based

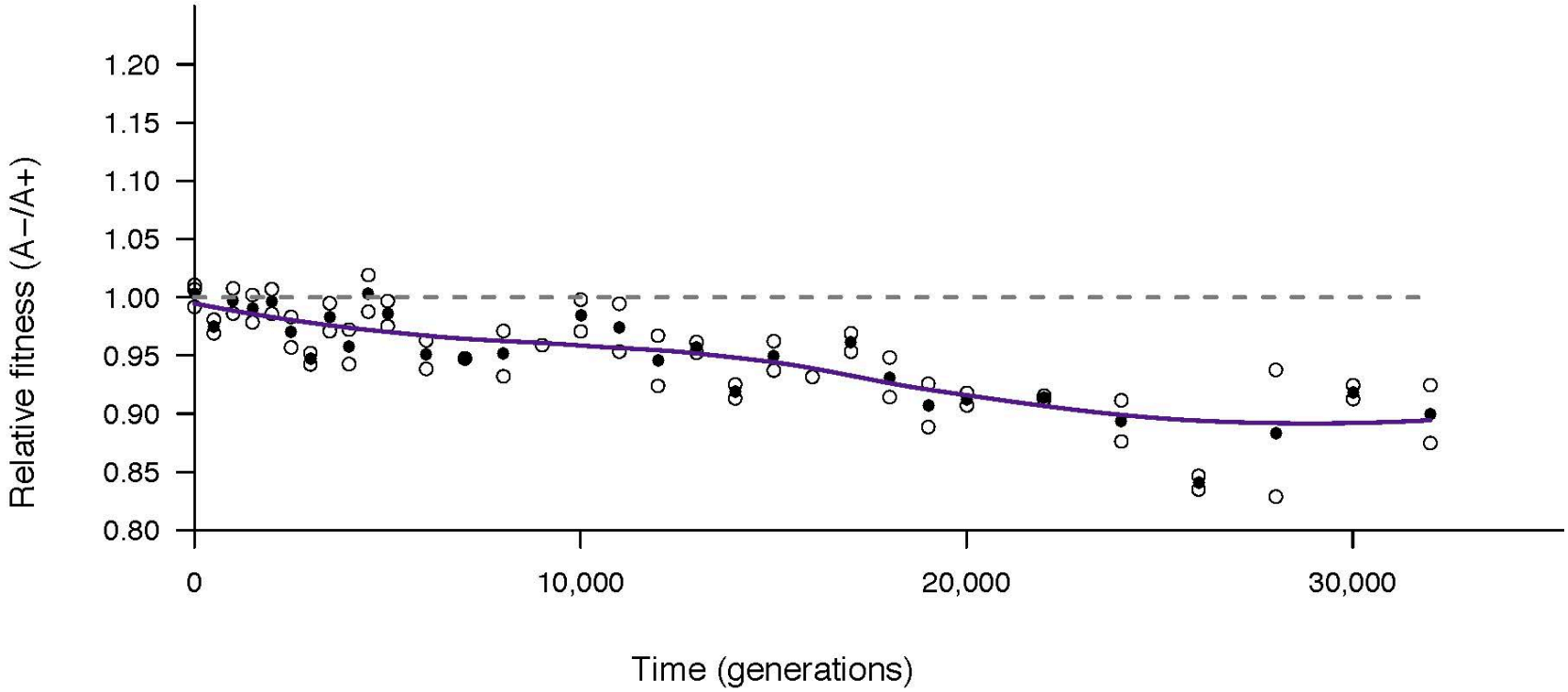

Figure 3.20: **Ara-3 v Ara+3**. Open symbols show each measured value of relative fitness from head-to-head competitions. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness. The solid purple curve is a local smoothing function showing the general trend of the data.

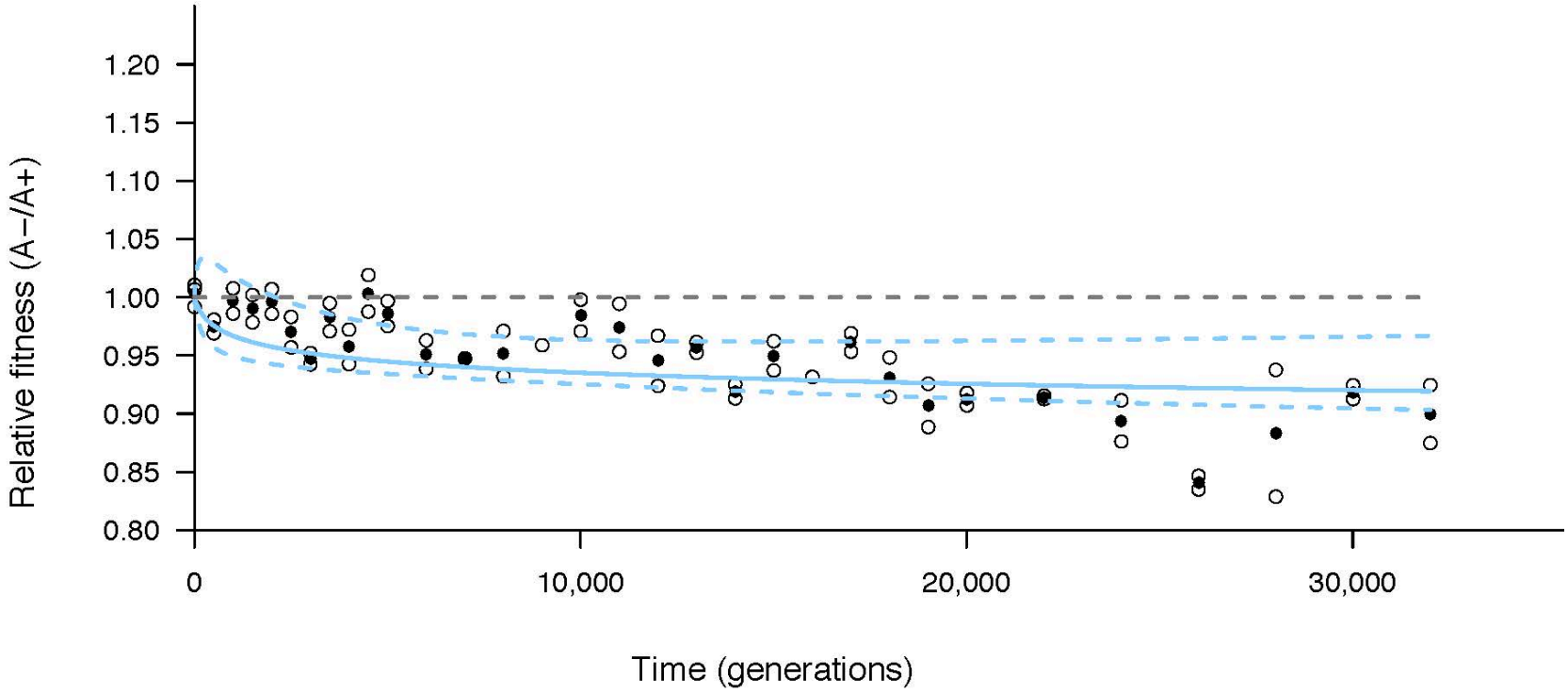

Figure 3.21: **Ara-3 v Ara+3**. The solid light blue line shows the mean fitness of Ara-3 relative to Ara+3, based on 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show the corresponding, non-parametric 95% confidence. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point. The dashed gray line at 1.00 is the level at which the competitors have equal fitness.

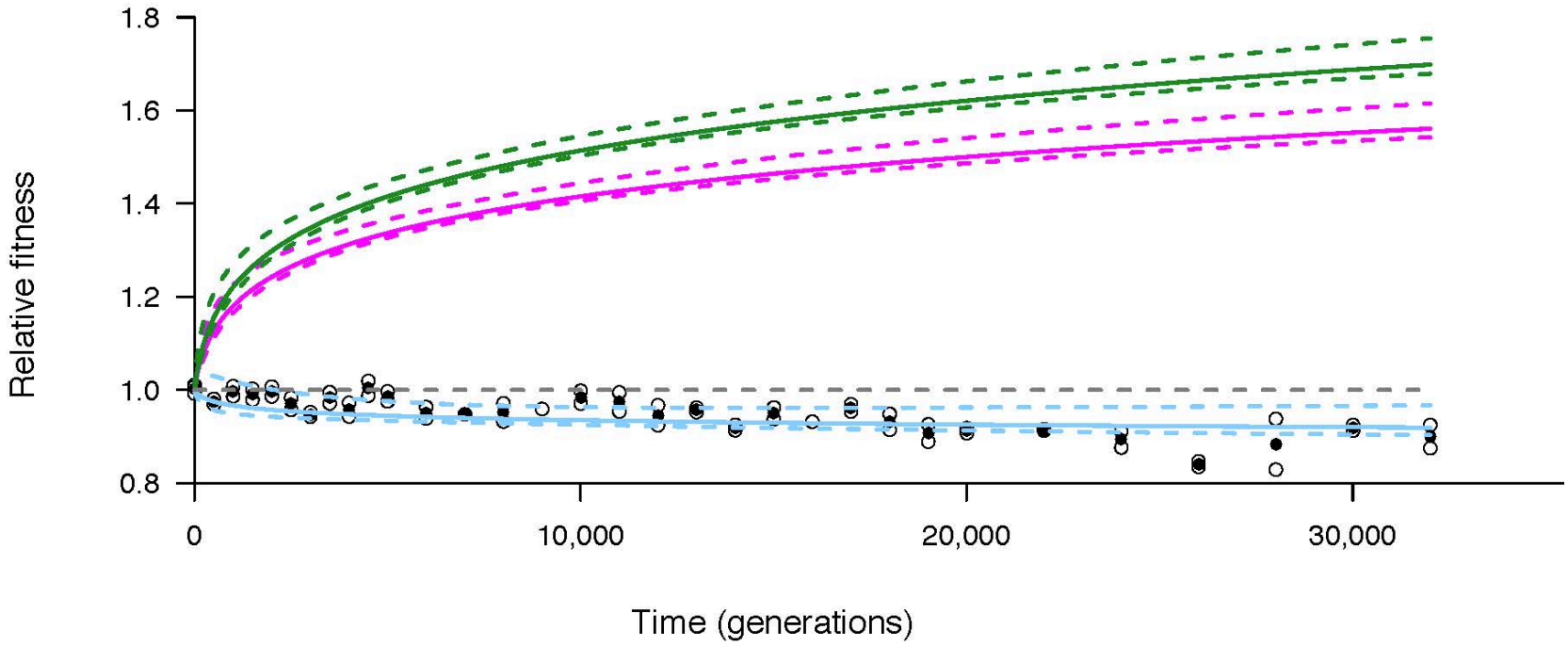

Figure 3.22: **Ara-3 v Ara+3**: The solid magenta curve shows the mean Power Law fitness trajectory for population Ara-3. The solid green curve shows the mean Power Law fitness trajectory for population Ara+3. The solid light blue curve shows the mean fitness of Ara-3 relative to Ara+3. Each of these lines is the mean across 10,000 bootstrap re-samplings of the data from Wiser et al (2013). Dashed lines show corresponding non-parametric 95% confidence intervals around the solid curves. Open symbols show each relative fitness measured in a head-to-head competition. The solid symbols are the mean at each time point.

on how each population diverged from a common competitor. Instead, the population pairs show more dynamic changes in relative fitness than our simple expectations predict. This dynamism suggests that it is unlikely that each of our populations is climbing a parallel slope in the adaptive landscape, whether on the same peak, or on different peaks. Combined with our data from chapter 2 that the populations in the LTEE are so far from reaching fitness peaks that we cannot detect evidence of an eventual asymptote to fitness, we are left with two possibilities of how the populations are traversing the fitness landscape. One, they may be climbing different peaks, with trajectories that are not parallel in fitness. In this scenario, they will not necessarily converge to similar fitness values at any point in the future, as the different peaks may be of different heights. Two, they may be climbing the same peak, but along very different paths. In this scenario, if the populations eventually get close enough to the top of this peak, we would expect population fitness to converge, and among-population variance in fitness to decline. We do not yet see evidence that this second scenario is occurring, but recognize that we may be so far from the top of whatever peak(s) the populations are climbing that we cannot rule out this possibility either.

Summary:

We examined several sources of data to look for patterns in among-population variance in fitness within a long-term evolution experiment and, from those patterns, learn about the adaptive landscape for populations in this experiment. We first analyzed data from Wiser et al (2013), calculating among-population variance estimate at each time point considered. We also gathered new data from a smaller number of

time points, but with greater replication. These data show an among-population standard deviation in fitness consistent with what had been observed through 10,000 generations in Lenski and Travisano (1994). Our data lack sufficient statistical power to state whether the among-population standard deviation in fitness continues to rise or reaches a plateau after this point.

In order to estimate the among-population variance in fitness, we perform an ANOVA at each time point measured. Though most of these ANOVAs are not individually significant, there is a significant overrepresentation of small p values within the set, demonstrating that there is still significant among-population variance in fitness. We next competed pairs of evolved populations against each other, and compared those results to the expectations derived from each population's individual fitness trajectory compared to the generation 0 ancestor. We find that the population pair estimates largely follow the predicted trends, but that these greater precision measurements are often subtly different from the expectations, allowing us to observe finer-scale patterns of relative fitness change between populations.

From previous work, we already knew that populations within this experiment had not reached a fitness peak by 50,000 generations. Whether peaks even exist in real adaptive landscapes is itself a contested point; real populations face selection on far more than two dimensions at once, and our intuition about geometry in 3-dimensional space may not apply to much higher-dimensional spaces (*18*, *19*). Nevertheless, it is possible to use information about the variance among populations within an evolution experiment to infer the likely topology of the adaptive landscape experienced by the populations.

Conclusions:

Among-population variance in fitness remains at appreciable levels even after 50,000 generations of evolution. This is further evidence that populations are not converging at the top of a single fitness peak in their adaptive landscape. Although we lack sufficient statistical power to define a function of among-population variance over evolutionary time, we can state firmly that it is not being driven down to insignificant levels over the course of 50,000 generations. However, even in the absence of clear patterns of how variance is changing, looking at the cumulative distribution of significance values in the ANOVAs used to calculate among-population variance demonstrates that there is significant signal of persistent variance despite the noise.

Broad-scale patterns of fitness differences in populations are consistent with expectations. Fitness is largely transitive in this system: if A > B, and B > C, then A > C in the majority of cases. Populations that show generally larger gains in fitness compared to their ancestor over long periods of time also show higher relative fitness when competed directly against evolved populations with smaller gains relative to the ancestor. This is in spite of known cases of frequency-dependent fitness within individual populations (*20*–*22*), which could easily disrupt transitivity of fitness.

Divergence in fitness between different populations is most often quite low compared to their divergence in fitness from the ancestor. This allows for measurements comparing evolved population pairs to extend over additional generations, and consequently reach higher levels of precision. It also demonstrates a substantial degree of parallelism in fitness – which is, essentially, an integrated

measurement of competitive ability within a given environment – despite populations being isolated for an extended period of time and having many differences in specific mutations.

Competitions between evolved populations reveal greater detail than can be observed from populations competing against their ancestor. The empirically-measured relative fitness of a pair of evolved populations is most often near what the expected value is based on their fitness trajectories relative to the ancestor. However, this measured relative fitness is still often outside the confidence interval of expectations. Further, these theoretical expectations are constrained to have relatively simple dynamics. Actual population pair measurements often show more complex dynamics, such as having more inflection points, or abrupt changes in slope than the expectations. Our power law models explain large-scale changes in fitness both across and within populations, but individual populations have time frames in which their actual fitness either accelerates or decelerates relative to the power law, and competitions between evolved populations can reveal these deviations from expectation.

Future Work:

The material in this chapter is almost exclusively empirical and statistical. A collaborator, Noah Ribeck, is working on simulating evolving populations using the population genetics framework published in Wiser et al (2013). We plan to compare our empirical measurements of variance over time to models in which the parameter that describes diminishing-returns epistasis parameter is either constant or changes over time as a population encounters different regions of the genetic space that underlies the

fitness landscape. These models will provide a description of how we should expect the among-population variance in fitness to change over time, to which we can then compare our empirical measurements.

Acknowledgements:

We thank Caroline Turner, Alita Burmeister, Noah Ribeck, Rohan Maddamsetti, Anya Vostinar, and Brian Goldman for discussion and feedback in the drafting of this chapter. We also thank Neerja Hajela for technical assistance. This work was supported, in part, by an NSF grant (DEB-1451740) and by the BEACON Center for the Study of Evolution in Action (NSF Cooperative Agreement DBI-0939454).

REFERENCES

REFERENCES

1. C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, C. Adami, Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*. **412**, 331–333 (2001).

2. P. C. Phillips, Epistasis -- the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*. **9**, 855–867 (2008).

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

4. S. Kinoshita, S. Kageyama, K. Iba, Y. Yamada, H. Okada, Utilization of a Cyclic Dimer and Linear Oligomers of ε-Aminocaproic Acid by Achromobacter guttatus KI 72. *Agric. Biol. Chem.* **39**, 1219–1223 (1975).

5. B. Koskella, J. N. Thompson, Gail M. Preston, A. Buckling, Local Biotic Environment Shapes the Spatial Scale of Bacteriophage Adaptation to Bacteria. *Am. Nat.* **177**, 440–451 (2011).

6. M. C. Urban, P. L. Zarnetske, D. K. Skelly, Moving forward: dispersal and species interactions determine biotic responses to climate change. *Ann. N. Y. Acad. Sci.* **1297**, 44–60 (2013).

7. A. H. Melnyk, R. Kassen, Adaptive Landscapes in Evolving Populations of Psuedomonas fluorescens. *Evolution*. **65**, 3048–3059 (2011).

8. M. J. Wiser, N. Ribeck, R. E. Lenski, Long-term dynamics of adaptation in asexual populations. *Science*. **342**, 1364–1367 (2013).

9. R. E. Lenski, M. Travisano, Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci.* **91**, 6808–6814 (1994).

10. Relative Fitness Data Through Generation 10,000, (available at http://myxo.css.msu.edu/ecoli/relfit.html).

11. R. Sokal, F. J. Rohlf, *Biometry* (W. H. Freeman and Company, New York, ed. 3rd, 1995).

12. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria, 2013; http://www.R-project.org/).

13. M. J. Wiser, N. Ribeck, R. E. Lenski, Data from: Long-term dymanics of adaptation in asexual populations. (2013), (available at http://dx.doi.org/10.5061/dryad.0hc2m).

14. P. D. Sniegowski, P. J. Gerrish, R. E. Lenski, Evolution of high mutation rates in experimental populations of E. coli. *Nature*. **387**, 703–705 (1997).

15. J. E. Barrick *et al.*, Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*. **461**, 1243–1247 (2009).

16. Z. D. Blount, J. E. Barrick, C. J. Davidson, R. E. Lenski, Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*. **489**, 513–518 (2012).

17. M. J. Wiser, R. E. Lenski, A Comparison of Methods to Measure Fitness in Escherichia coli. *PLoS ONE*. **10**, e0126210 (2015).

18. S. Gavrilets, Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12**, 307–312 (1997).

19. D. M. McCandlish, Visualizing Fitness Landscapes. *Evolution*. **65**, 1544–1558 (2011).

20. R. Maddamsetti, R. E. Lenski, J. E. Barrick, Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Experiment with Escherichia coli. *Genetics*. **200**, 619–631 (2015).

21. D. E. Rozen, R. E. Lenski, Long-Term Experimental Evolution in Escherichia coli. VIII. Dynamics of a Balanced Polymorphism. *Am. Nat.* **155**, 24–35 (2000).

22. N. Ribeck, R. E. Lenski, Modeling and quantifying frequency-dependent fitness in microbial populations with cross-feeding interactions. *Evolution*. **69**, 1313–1320 (2015).

CHAPTER 4: LONG-TERM DYNAMICS OF ADAPTATION IN ASEXUAL DIGITAL POPULATIONS.

Authors: Michael J. Wiser, David M. Bryson, Charles Ofria, and Richard E. Lenski

Abstract:

Previous work has shown that experimental evolution populations of bacteria exhibit power law dynamics, implying that improvements will continue indefinitely. Computational systems offer us the chance to study evolving populations for more generations than are ever feasible in microbial experimental evolution studies. Here we evolve populations of digital organisms in Avida for either 200,000 or 1,000,000 generations, across three different environments. We find that in both the most complex and the simplest of these environments, fitness obeys power law dynamics. In the intermediate case, fitness is better described by a hyperbolic model, but fitness still increases over long time scales. Our work suggests that power law fitness dynamics may be a general feature of evolving systems.

Introduction:

Wiser et al (2013) previously showed that in populations of *Escherichia coli* evolved for 50,000 generation, fitness over time exhibited power law dynamics (*1*). The long-term evolution experiment (LTEE) from which those data

are derived is the biological experimental evolution study that has the run for the largest number of generations (*2*). It would therefore appear that we cannot investigate whether similar patterns arise in other systems. However, explicitly biological experiments are not the only ones that can provide insight into evolutionary dynamics.

Artificial systems allow us to study whether certain properties are shared across evolving systems, independent of the details of cellular machinery. This has been happening for decades. John Maynard Smith (1992) stated “So far, we have been able to study only one evolving system and cannot wait for interstellar flight to provide us with a second. If we want to discover generalizations about evolving systems, we will have to look at artificial ones.” (*3*). We therefore turn to a computational system of evolving populations, and ask whether this system exhibits similar patterns of fitness over evolutionary time as the LTEE.

Study System:

We conducted computational evolution experiments with the digital evolution software platform Avida. This platform has been detailed extensively elsewhere (*4*), but a brief summary follows. Organisms are self-replicating asexual computer programs, composed of sequences of instructions. Users define a mutation rate, controlling the per-site probability that a new organism will be different from its parent. As with biological organisms, these mutations may be beneficial, neutral, deleterious, or lethal. Organisms within Avida compete for space in their virtual world with other organisms. Additionally, in most

environments, organisms compete with each other for resources by completing tasks that are rewarded with additional CPU cycles. These extra CPU cycles allow organisms to copy themselves faster than less fit competitors.

Evolution by natural selection is substrate neutral; any system of organisms that exhibits variation, inheritance, selection, and time across generations will undergo evolution by natural selection (*5*). This applies to artificial life, as well as natural organisms. Populations in Avida gain variation through mutation, organisms inherit parental variations, and the environment imposes a selective pressure. Because experiments in Avida extend across many generations, Avida thus meets all of the criteria for evolution by natural selection. Avida is not merely a simulation of evolution, but an instance of it.

Fitness in Avida is calculated as Execution Rate[1] divided by Generation Length[2], the amount of time it takes an organism to copy itself. Organisms thus can increase their fitness either by executing instructions more rapidly (increasing their Execution Rate) or by requiring fewer executed instructions to replicate (reducing their Generation Length). Fitness therefore measures the number of offspring produced in a given amount of time.

Experimental conditions:

We performed experiments in three different environmental reward regimes: No Task, Logic-9, and Logic-77. In the No Task environment, there were no tasks that organisms could perform to gain additional CPU cycles. In

[1] Listed as Merit in the Avida data files
[2] Listed as Gestation Time in the Avida data files

this environment, competition between organisms was solely for space. In the Logic-9 environment, nine different one- or two-input Boolean logic tasks are rewarded. More complex tasks are given greater rewards; the simplest tasks reward the organism by doubling the Execution Rate, while the most complex task rewards the organism by multiplying Execution Rate by 32. In the Logic-77 environment, 77 different one-, two-, or three-input Boolean logic tasks are rewarded. Each task performed doubles the organism's Execution Rate, regardless of complexity of the task.

In each of the three environments, we conducted experiments for a defined number of generations. We ran populations for 200,000 $\pm$ 1 generations in the Logic-77 and Logic-9 environments, and 1,000,000 $\pm$ 1 generations in the No Task environment.

Statistical methods:

We performed all statistical analyses in R version 3.0.2 (*6*). We calculated relative fitness by dividing population fitness by the fitness of the ancestor. When necessary (in the Logic-77 and Logic-9 environments), we transformed fitness as log 2 fitness; means across replicates were calculated after transformation. We fit linear models with the lm() command, and we fit non-linear models with the nls() command. We calculated posterior odds ratios from difference in BIC value, according to Raftery (1995) (*7*).

Results and Discussion:

Logic-77 Environment:

Of the environments that we tested, the Logic-77 environment is the most complex. This complexity makes the Logic-77 environment the most like biological environments, where complexity is a common. Even in extremely simple laboratory environments, different organisms in a population can specialize on different resources (*8*), and there are many internal cellular processes that can be optimized.

We first tested whether evolution in this environment reaches an optimum. Previous research in Avida has generally run for durations of $\leq$ 150,000 Updates (*9–12*) – an internal measure of time within Avida – which corresponds to less than 15,000 generations. Because the rate of adaptation slows with time, other researchers have concluded that populations are approaching a fitness peak (*10*, *13*).

By looking at the same data over a range of time scales, we can examine whether the appearance of an early plateau actually signals a halt in the adaptive process. Figure 4.1 show the mean fitness across 20 replicate runs in the Logic-77 environment. Different panels in the figure show the same data examined over different numbers of generations. The dashed vertical lines show the end points of previous panels. As we can see, the curve has the same basic shape in each of the panels. What appears to be a plateau in one panel is revealed to be part of the upward trajectory in a later panel. In fact, the appearance of a plateau is an artifact of sampling. Were the run to extend over considerably

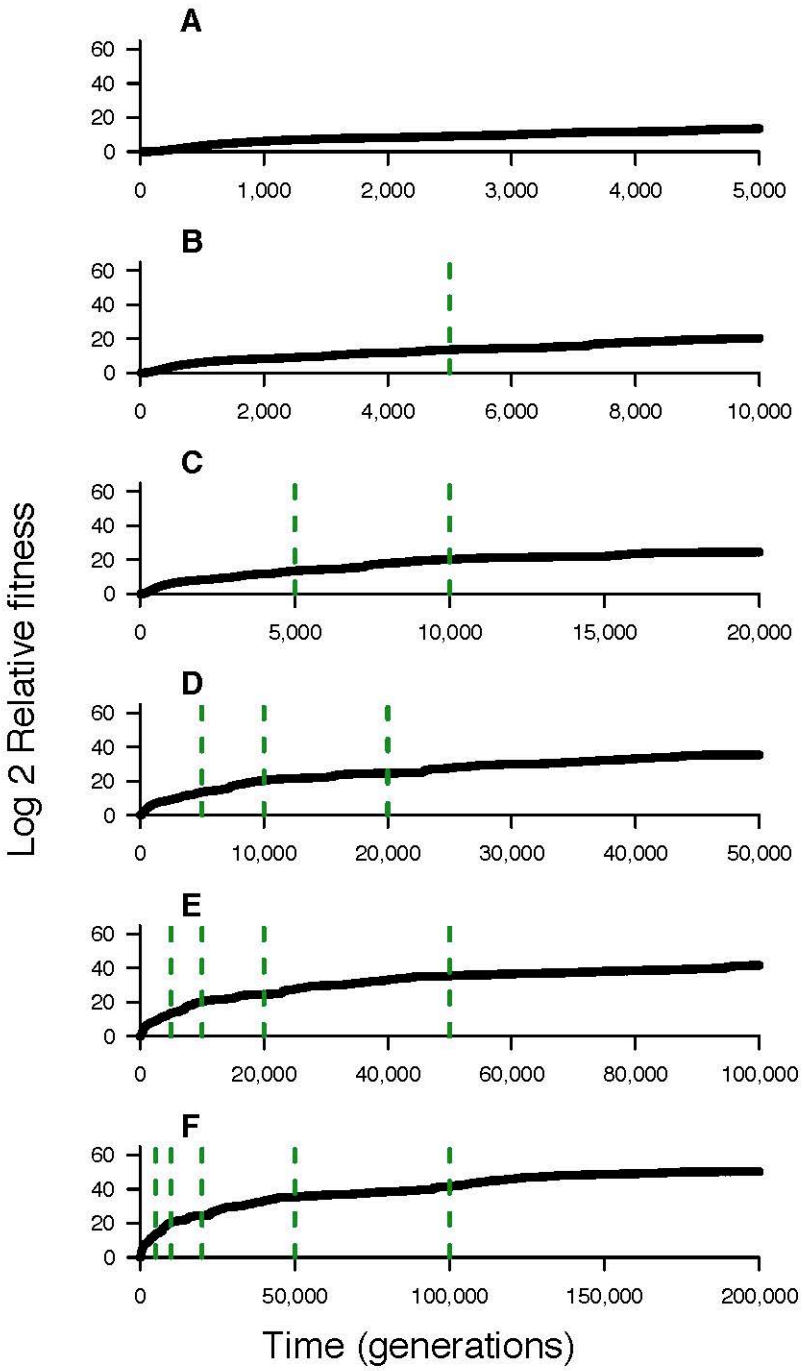

Figure 4.1: **Fitness over time in the Logic-77 environment**. The solid (black) curve is the mean log 2 relative fitness across 20 replicates. Different panels show different numbers of generations. Dashed (green) vertical lines show ends of previous panels.

more generations, we would expect the new apparent plateau to be of a higher value, and reached later in the run. This is evidence that the populations have not reached an evolutionary optimum, but are still adapting to their environment.

We can also look at changes over time within individual replicates. Figure 4.2 shows a scatterplot of fitness for the Logic-77 data, with each individual point being one replicate run. From these data, we notice two striking facts. First, points predominantly fall above the y = x line, which shows that they are reaching higher fitness at the end of the evolutionary run than they are 2/3 of the way through a run. Indeed, log 2 relative fitness is higher at 200,000 generations than at 133,333 generations (one-tailed t test, t = 2.7414, df = 19, p = 0.00649). We also find a significant, positive slope to a linear regression in this late time period for log 2 relative fitness over time (slope estimate = $4.081 * 10^{-5}$, t = 11.95, $p < 2 * 10^{-16}$). Note that we are not arguing that this late slope is linear, but merely that a significant, positive linear slope indicates that fitness is increasing in some fashion (see Figure S4.1). Second, different replicates reach very different levels of fitness. Because each task performed in the Logic-77 environment doubles the organism's fitness, the log 2 relative fitness provides an approximation of how many tasks the organism performs. Many of the replicates still have fitness values indicating more than a dozen additional tasks could be performed by the average members of their populations. This means that at least 19 of the 20 replicates could reach a higher fitness, indicating clearly that they have not reached a global optimum.

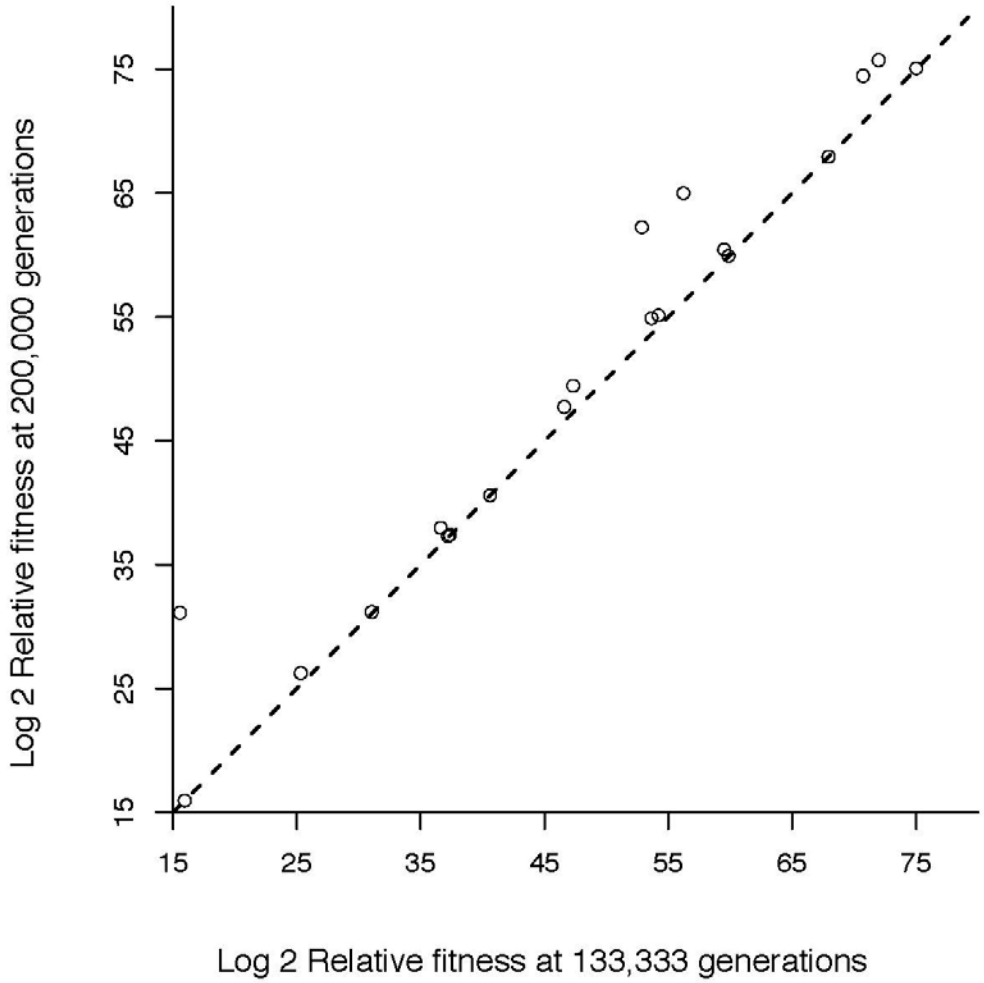

Figure 4.2: **Late fitness v final fitness in the Logic-77 environment**. Each point is one replicate. The dashed line is at y = x; points on this line have the same fitness at the end of the run as at 2/3 of the run.

The analysis above examines just specific points in the fitness trajectories; we gain additional information by looking at the entire trajectories (Figure 4.3). In this figure, each of the individual replicates are shown as gray points, with the mean across runs as the black curve. The mean fitness trajectory appears as a smooth curve, while many individual trajectories appear to be made of step-like combinations of rapid increases and long-term stability. From these trajectories, we can see not only that different populations reach different final fitness values – as we saw in Figure 4.2 – but that even populations which achieve similar final fitness do not necessarily do so in similar time frames. For example, of the three populations that achieved the highest final fitness, one of them had gotten to roughly this final fitness prior to 50,000 generations, while the other two did not until after 150,000 generations.

We next examine the functional shape of how fitness changes over evolutionary time. We have previously shown that in a long-term evolution study in bacteria, fitness over time is better fit by a power law than a hyperbola, indicating that fitness is expected to increase indefinitely (*1*). In Figure 4.4, we compare the best fit power law to the best fit hyperbola in the Logic-77 environment in Avida. When considering all of the data, the power law model substantially outperforms the hyperbolic model (difference in BIC = 10341.01, posterior odds ratio < $e^{-5170} << 10^{-10}$). Like with the LTEE, we also fit models to the first 40% of the generations, and project what those models predict for the rest of the data. Here, the power law model somewhat overestimates future fitness, while the hyperbolic model somewhat underestimates future fitness

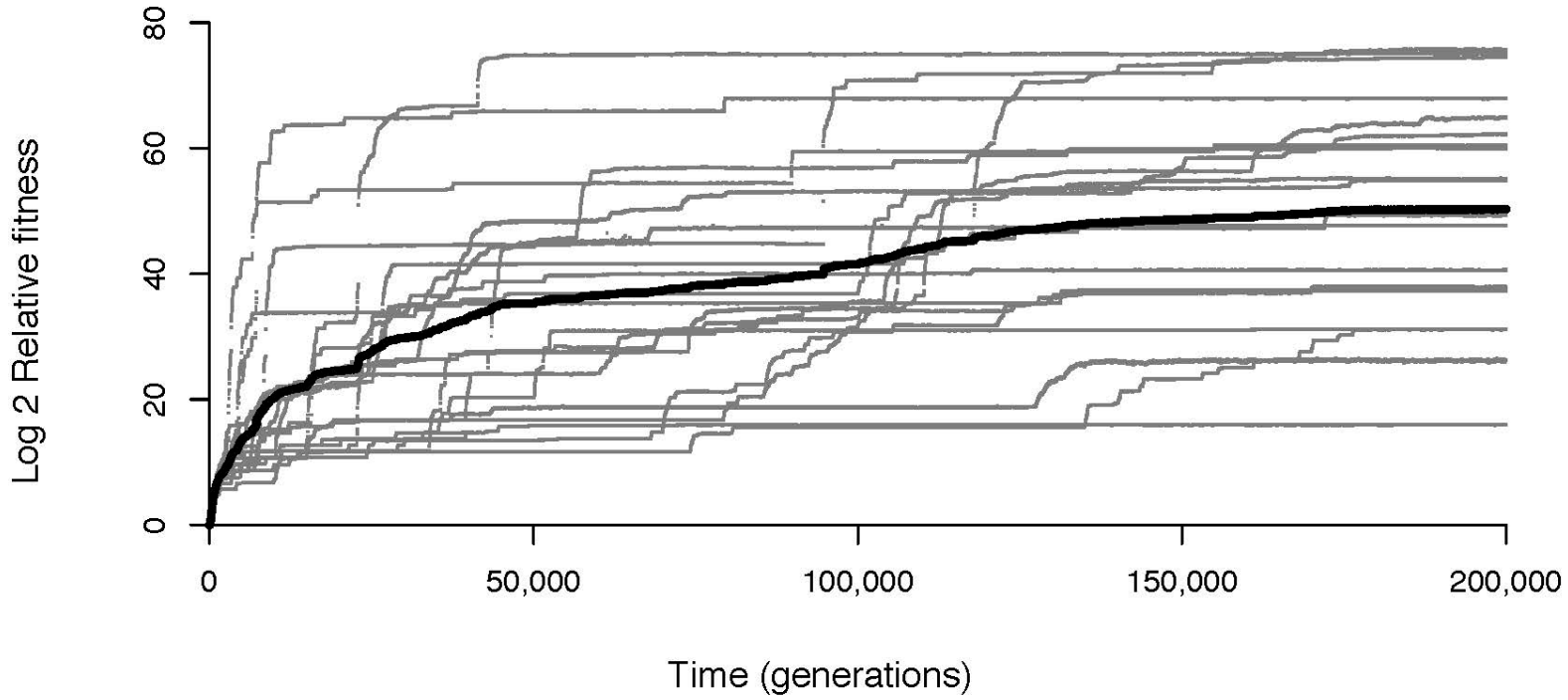

Figure 4.3: **Fitness over time in the Logic-77 environment**. The gray points show each of the 20 replicates. The black curve shows the mean log 2 relative fitness over time.

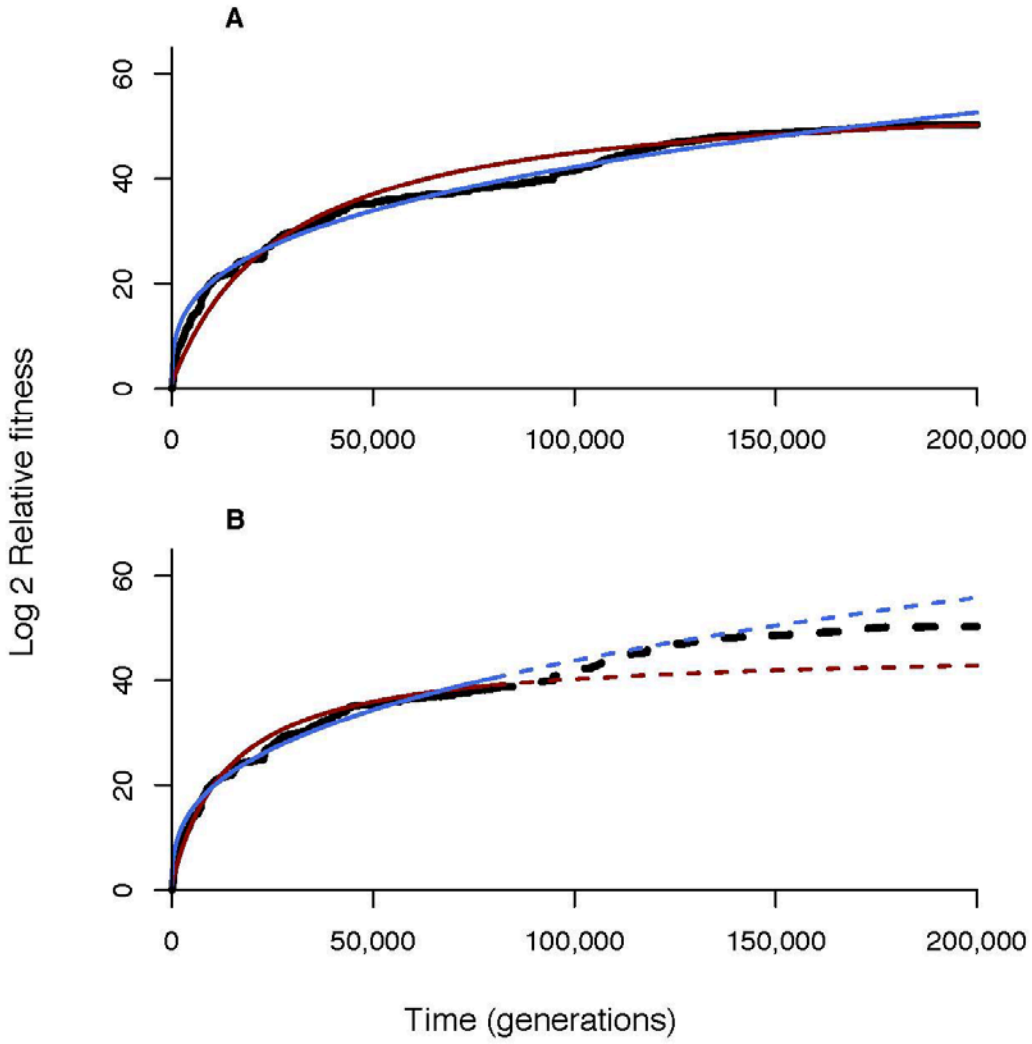

Figure 4.4: **Comparison of model fits in the Logic-77 environment**. (**A**) Hyperbolic (red) and power-law (blue) models fit to the set of mean log 2 fitness values (black symbols) from all 20 replicates. (**B**) Fit of hyperbolic (solid red) and power-law (solid blue) models to data from first 80,000 generations only (solid black), with model predictions (dashed red and blue curves) and later data (dashed black curve).

The weight of the data in the Logic-77 environment strongly indicates that fitness has not reached a final plateau. At least 19 of the 20 replicates have not reached a global optimum, as they have lower fitness than the most fit population. Fitness values at 200,000 generations are significantly higher than at 133,333 generations. The time frame from 133,333 generations to 200,000 generations displays a significant, positive slope in fitness over time. Log 2 relative fitness exhibits power law dynamics in this environment. The power law finding is particularly striking, as relative fitness also exhibits power law dynamics in the LTEE. The fact that we get similar dynamics – albeit, in a log 2 transformation of fitness – in a completely different system lends support to the idea that power law fitness dynamics may not be an idiosyncrasy of the LTEE. Instead, these dynamics may be a more general feature of evolving systems.

No Task Environment:

The No Task environment is at other extreme of complexity from the Logic-77 environment. Here, the only way for organisms to improve their fitness is to lower their Generation Length, and thus replicate faster. For a self-replicating organism this is one of the simplest environments conceivable. We also extended the evolutionary runs much further, out to 1,000,000 generations. If any of our environments would lead to evolutionary stagnation, we would expect it to be this one: a simpler environment, fewer ways to improve, and small selection coefficients for those mutations which are beneficial than in the logic environments.

Despite this extreme simplicity, evolution does not reach a maximum and stop in this environment. From Figure 4.5, we see that apparent plateaus in relative fitness are – just as in the Logic-77 environment – artifacts of sampling. If we allow the runs more evolutionary time, what previously appeared to be a plateau in relative fitness now becomes a steep portion of the curve. Note that in this case, we are analyzing relative fitness, not a log 2 transformation of it, because fitness gains are much smaller when there are not tasks to be evolved.

We likewise get similar results when we look at changes over time within replicates. Figure 4.6 shows a scatterplot of fitness in the No Task environment. Again, the points fall predominantly above the y = x line of equal fitness at the end of the evolutionary run as 2/3 of the way through the run. Fitness is higher at 1,000,000 generations than at 666,667 generations (one-tailed t test, t = 2.2666, df = 19, p = 0.0176). A linear regression of fitness over time between generations 666,667 and 1,000,000 yields a significant, positive slope to fitness (slope = 23.81, t = 5.267, p = 1.39 * 10^{-7}), indicating a rise in fitness over the final third of this experiment. In this case, we certainly don't interpret this positive linear slope as indicating a linear increase in fitness – the diagnostic plots for the model reveal that the data do not meet the assumptions for a linear model (see Figure S4.2) – but merely note that a significant, positive slope for a linear regression of fitness indicates an increase in fitness in this time frame.

In this case, though, some of the points are actually below the y = x line, indicating replicates where the final population fitness is lower than the population fitness two thirds of the way through the evolutionary run. What can

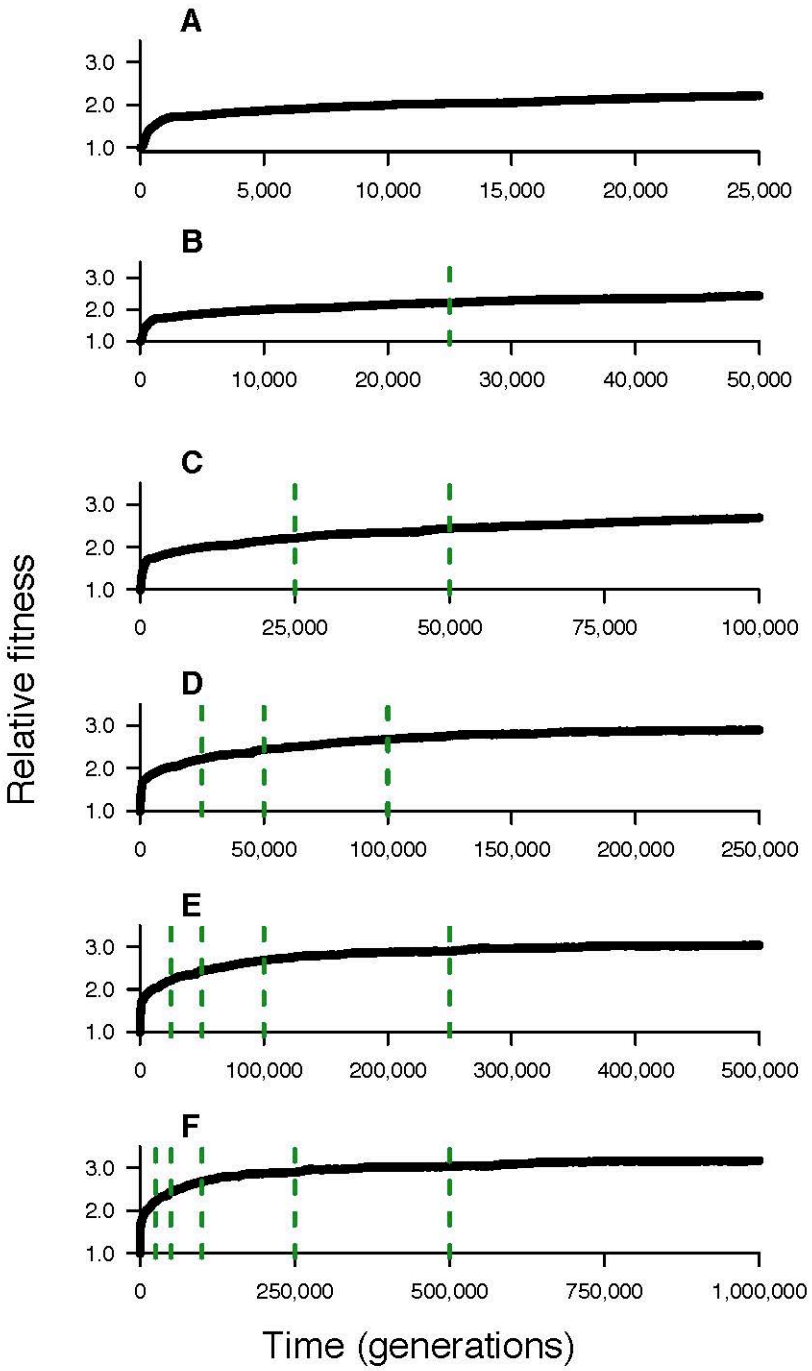

Figure 4.5: **Fitness over time in the No Task environment**. The solid (black) curve is the mean log 2 Relative fitness across 20 replicates. Different panels show different numbers of generations. Dashed (green) vertical lines show ends of previous panels.

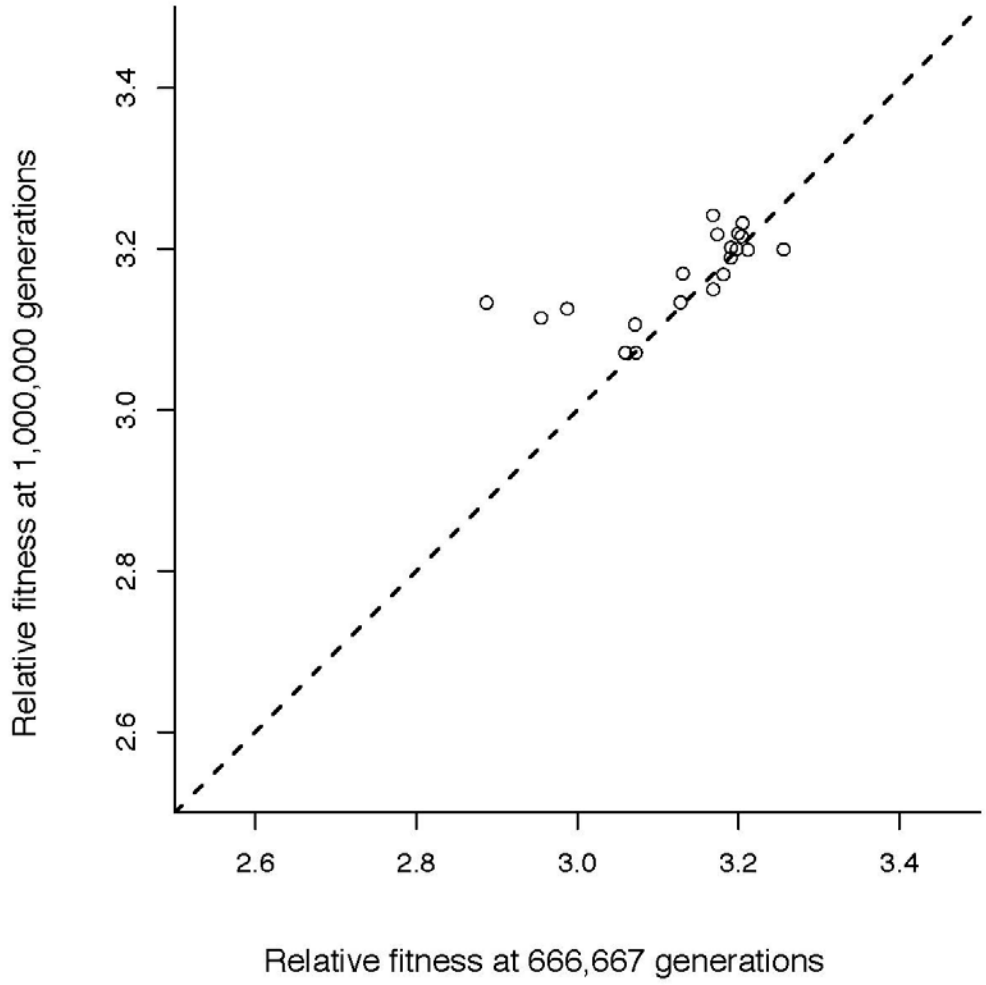

Figure 4.6: **Late fitness v final fitness in the No Task environment**. Each point is one replicate. The dashed line is at y = x; points on this line have the same fitness at the end of the run as at 2/3 of the run.

account for this unexpected result? One possibility is the appearance of additional, detrimental mutations. The population is not homogenous; at any given time, there will be at least some genetic variation because some organisms will be one or two mutational steps away from their parent. When a beneficial variant begins to spread through a population, the spreading clade will initially be close to clonal. However, as the population size of this beneficial clade increases, so does the probability that it will contain individuals both with the beneficial mutation and other, not beneficial, mutations. If the rate of detrimental or lethal mutations is high enough, and the rate of new beneficial mutations is low enough, population fitness may rise as the beneficial mutation spreads, but then decline until reaching a mutation-selection equilibrium later. Because the fitness gains of individual beneficial mutations are so small in this environment, and the declines we observe in fitness are so small, this seems like a likely explanation.

These changes in population fitness are shown in detail in Figure 4.7. Populations exhibit short periods of rapid rise in fitness, followed by long periods of little change in fitness. However, during these periods of relative stability, fitness still fluctuates. Unlike with physical organisms, these are not explained by measurement error – for any given population, at any given time point, we can measure the exact fitness within Avida. Instead, they show actual small changes in population fitness, due to changes in population composition, either from existing genotypes changing in frequency, or new genotypes arising through mutation. These small scale fluctuations exist in all populations with non-zero

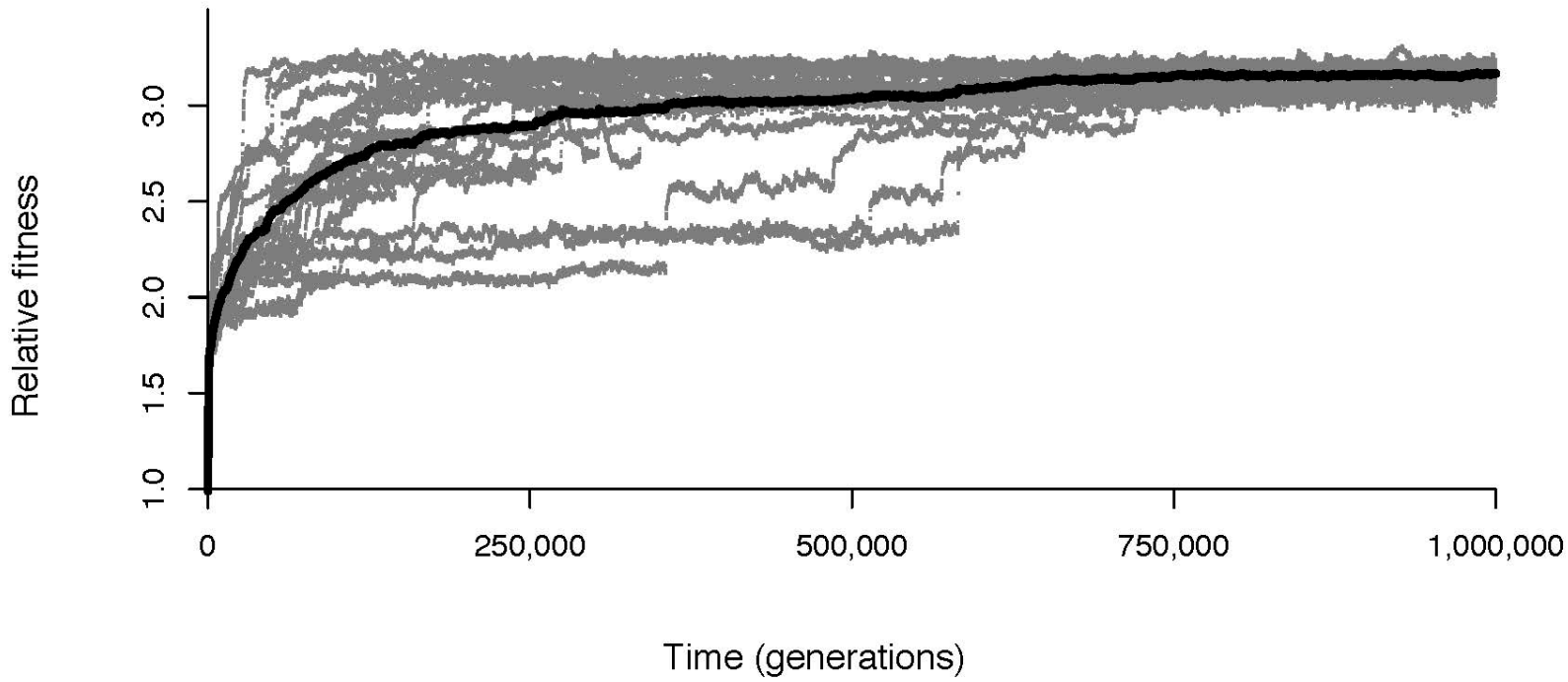

Figure 4.7: **Fitness over time in the No Task environment**. The gray points show each of the 20 replicates. The black curve shows the mean relative fitness over time.

mutation rates, but are more visible in this case because there are few mutations of large enough effect to obscure the dynamics.

As before, we then fit two different models to fitness over time in this system. From Figure 4.8, we observe that the two models do a similar job in predicting future changes in fitness, with the power law model overestimating future fitness to a similar extent that the hyperbolic model underestimates it. However, we can again see that even over these long time frames in a simple environment, fitness is better fit by a power law than by a hyperbola (difference in BIC = 13712.11, posterior odds ratio < e^{-6856} << 10^{-10}).

It is particularly striking that fitness continues to increase over long time scales in this experiment. For one, the time scales here are five-fold greater than in the logic environments, which themselves are four-fold greater than the number of generations we examined in the LTEE in Chapter 2 and Chapter 3. Combined with the simplicity of the environment, it would be easy to assume that all the populations in this environment would reach a global optimum in fitness, and stop improving. Yet this is not the case. Instead, populations stuck in regions of relatively low fitness for extended periods of time eventually find their way to regions of higher fitness. Populations in regions of relatively high fitness themselves experience fluctuations in fitness, sometimes including a rise to an even higher fitness region. That our measurements in even this very simple environment support an unbounded fitness function lends credence the possibility that such unbounded increases are a general feature of evolving populations.

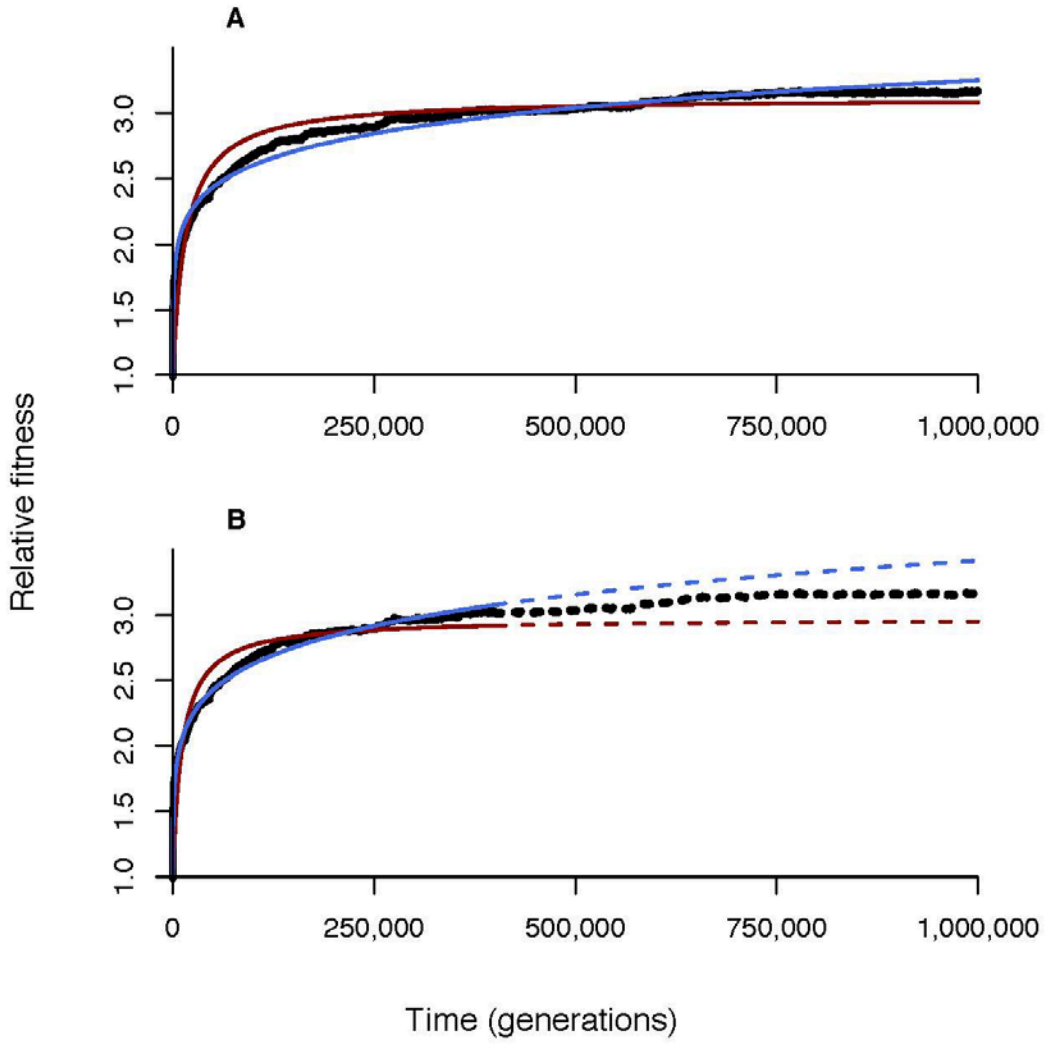

Figure 4.8: **Comparison of model fits in the No Task environment**. (**A**) Hyperbolic (red) and power-law (blue) models fit to the set of mean fitness values (black symbols) from all 20 replicates. (**B**) Fit of hyperbolic (solid red) and power-law (solid blue) models to data from first 400,000 generations only (solid black), with model predictions (dashed red and blue curves) and later data (dashed black curve).

Logic-9 Environment:

In the Logic-9 environment in Avida there are a small number of rewarded behaviors that organisms can evolve, putting it between the extremes of the other environments. Yet the rewards for these behaviors are very large; all else being equal, the lowest-reward behaviors double the organism's fitness, while the most-rewarded behavior multiplies it by a factor of 32. We therefore expect the fitness gains from evolving tasks to mask small changes from improved replication efficiency. Despite this, the Logic-9 environment is the most extensively used in previous work in Avida (*9*, *14*, *15*), which is why we chose to examine fitness dynamics in this environment.

Figure 4.9 shows log 2 relative fitness over time in the Logic-9 environments. Unlike in the Logic-77 environment (Figure 4.1) or the No Task environment (Figure 4.5), here we see that the appearance of a plateau in fitness does not disappear simply by looking at longer time frames. While there is a slight upward trajectory from 20,000 to 200,000 generations, the increase is small enough that it isn't immediately obvious. The value of this plateau is also telling. Holding everything else constant, an organism in the Logic-9 environment gets a 2^{25}- fold improvement in fitness by performing all nine logic tasks. The plateau in fitness is very close to 2^{25}, as the y-axis is the log 2 of relative fitness. Therefore, what improvements remain in fitness will be predominantly those from decreasing Generation Length, and each mutation that

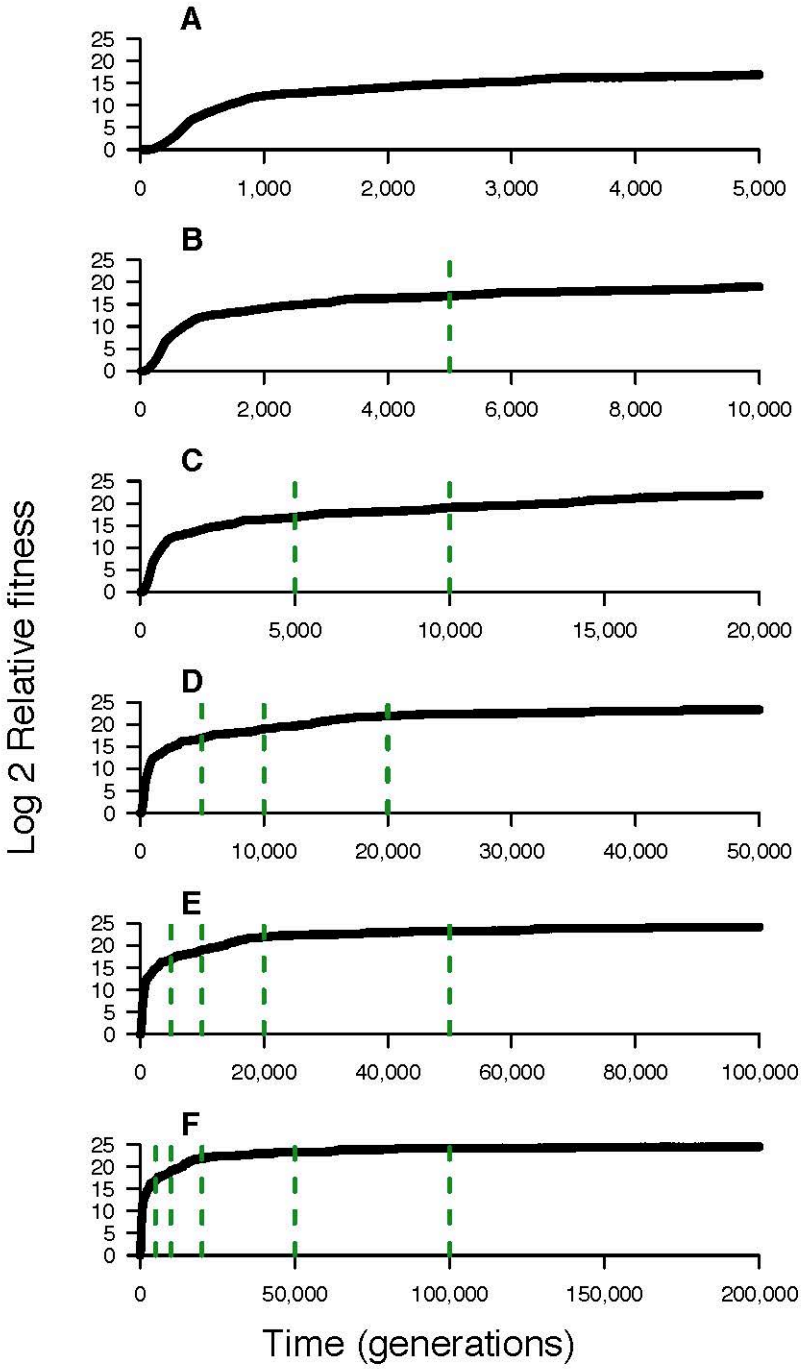

Figure 4.9: **Fitness over time in the Logic-9 environment**. The solid (black) curve is the mean log 2 Relative fitness across 20 replicates. Different panels show different numbers of generations. Dashed (green) vertical lines show ends of previous panels.

does so will have a smaller individual effect than any of the mutations which provided solutions to new tasks.

From Figure 4.10, we can see that 16 of the 20 replicates have fitness values consistent with most individuals in the population performing all nine logic tasks by 200,000 generations. The improvement in fitness in this environment from generation 133,333 to 200,000 is only marginal (one-tailed t test, t = 1.3402, df = 19, p = 0.0980). A linear regression of log 2 relative fitness over time from 133,333 generations to 200,000 generations yields a significant, positive slope (slope = 2.59 * 10^{-6}, t = 4.65, p = 3.33 * 10^{-6}). As in the No Task environment, though, diagnostic plots for this model reveal that a linear model is not a good fit for these data (see Figure S4.3).

The fact that most fitness gains in this environment are driven by task acquisition is further underscored by the individual replicate fitness trajectories shown in Figure 4.11. Most populations achieve a log 2 relative fitness of slightly more than 25, and then stop visibly improving in fitness from that point onward. This saturation is further corroborated by comparing the two model fits in Figure 4.12. In this environment, log 2 relative fitness is better explained by a hyperbola than a power law (difference in BIC = 14010.79, posterior odds ratio < e^{-7005} << 10^{-10}). In Figure 4.12B, we can see that the hyperbolic model does a strikingly better job of predicting future fitness than the power law model does.

What accounts for this major difference between the Logic-9 environment in Avida and the other ones we have examined? One possibility is that is the nature of the rewards for task completion. In the Logic-77 environment, each

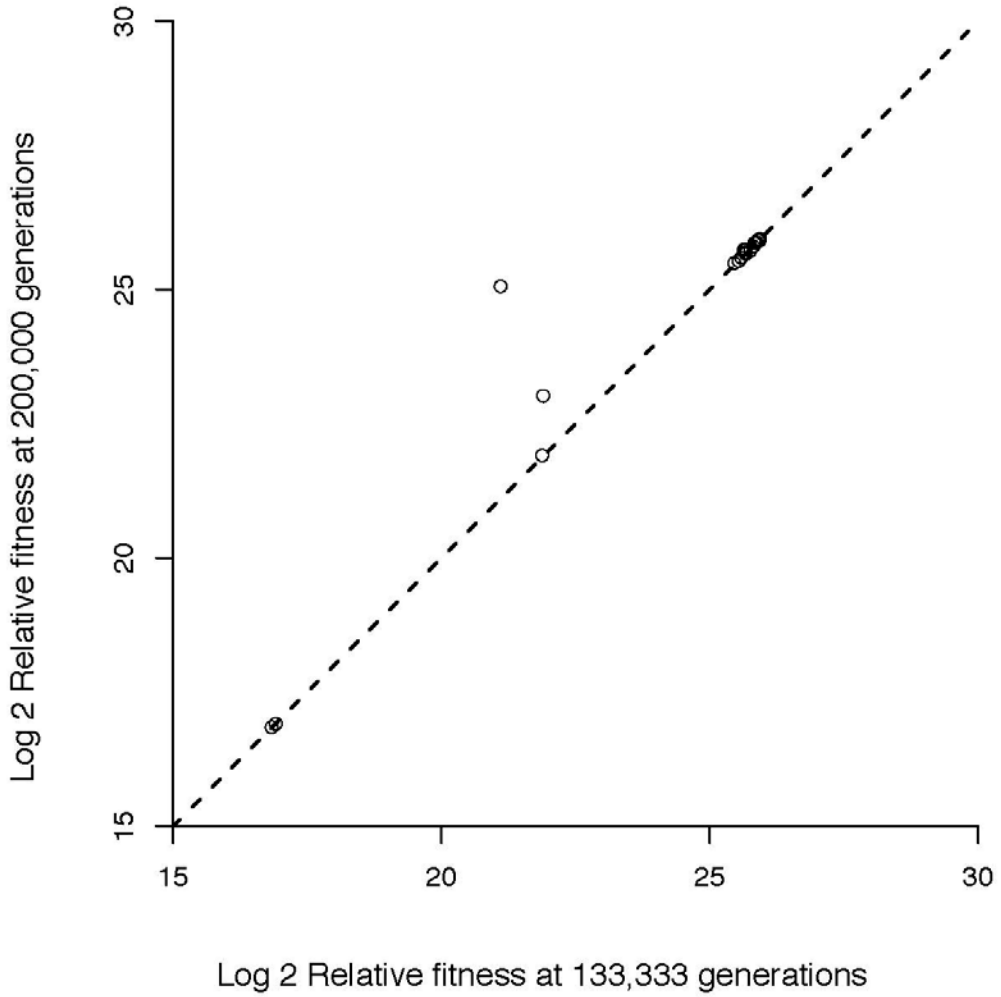

Figure 4.10: **Late fitness v final fitness in the Logic-9 environment**. Each point is one replicate. The dashed line is at y = x; points on this line have the same fitness at the end of the run as at 2/3 of the run.

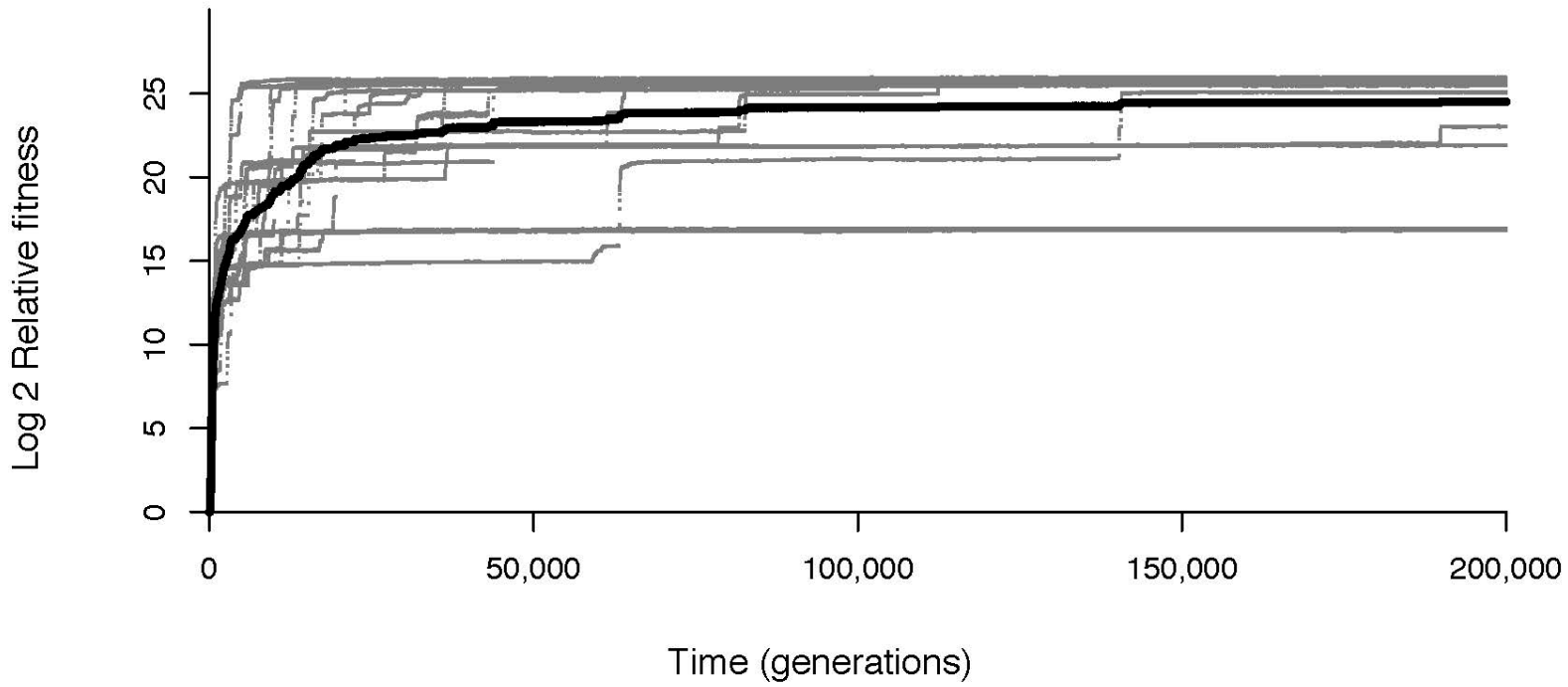

Figure 4.11: **Fitness over time in the Logic-9 environment**. The gray points show each of the 20 replicates. The black curve shows the mean log 2 relative fitness over time.

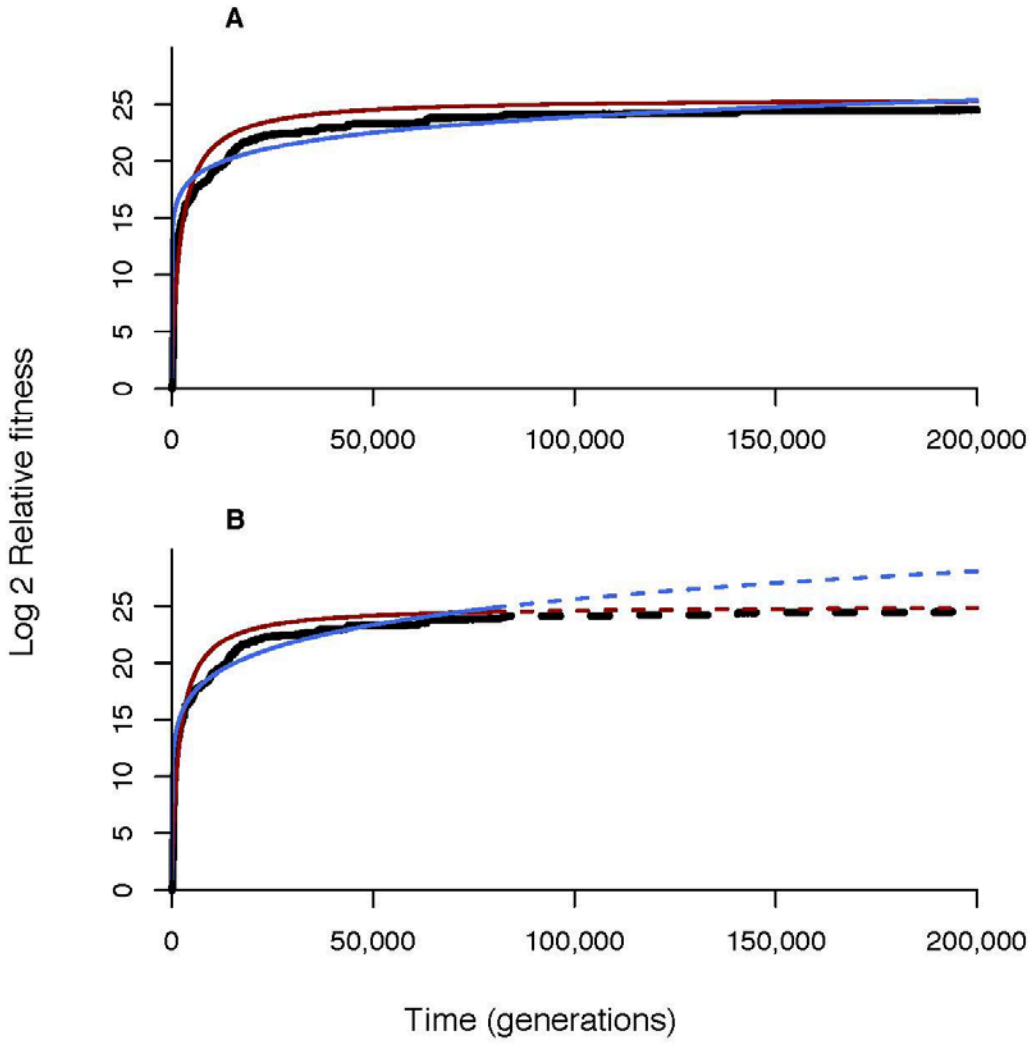

Figure 4.12: **Comparison of model fits in the Logic-9 environment**. (**A**) Hyperbolic (red) and power-law (blue) models fit to the set of mean log 2 fitness values (black symbols) from all 20 replicates. (**B**) Fit of hyperbolic (solid red) and power-law (solid blue) models to data from first 80,000 generations only (solid black), with model predictions (dashed red and blue curves) and later data (dashed black curve).

new task doubles fitness of an organism. In the No Task environment, fitness gains are substantially smaller. In the LTEE, the largest known beneficial mutations were on the order of a 13% fitness boost (*16*). In the Logic-9 environment, though, individual mutations can increase fitness up to multiplying it by 32. These large effect mutations will, by necessity, drive the pattern of fitness change over time, and minimize the impact of small mutational steps such as those that drive the pattern in the No Task environment. In fact, given that the No Task environment exhibits power law dynamics in fitness over time, we would expect the Logic-9 environment to do the same starting from the point where all nine logic tasks are being performed. A related explanation lies in the fact that in both of the logic environments in Avida, the ancestor is drastically unfit compared to its eventual descendants. In the LTEE, fitness gains were on the order of 60-80% over 50,000 generations (*1*). In the Logic-77 environment, fitness gains are on the order of ~ 2^{50} by 200,000 generations; in the Logic-9 environment, they are on the order of ~ 2^{25} by 200,000 generations, and in the No Task environment they're on the order of ~ 320% by 1,000,000 generations. With a small number of large effect mutations, and a large number of drastically-smaller effect mutations available, population fitness will tend to rise rapidly when the large effect mutations are spreading, and move only small amount otherwise. This will cause the trajectory to look more like a hyperbola. In future work, we will address these explanations by 1) starting runs with evolved ancestors, and 2) changing the task rewards so that individual mutations do not have as large of an impact as in the Logic-9 environment.

Conclusions:

In the Logic-77 and No Task environments of Avida, fitness obeys power law dynamics, much as it does in the LTEE. In the Logic-9 environment of Avida, fitness is better explained by a hyperbolic model. Even over hundreds of thousands of generations, fitness continues to increase in this system across these environments. This suggests that unbounded increases in fitness over evolutionary time scales may be general to evolving systems as a whole, and not due to the specifics of the LTEE.

Future Work:

We will extract the numerically-dominant organism from the end of each of ten runs in each of the three environments tested. We will use these organisms as the ancestor for additional (replicated) evolutionary runs, both within the environment in which they had evolved, and within simpler environments. We will test if these new bouts of evolution, starting from a more adapted ancestor, exhibit unbounded fitness increases over time.

Acknowledgements:

We thank Noah Ribeck, Alita Burmeister, Rohan Maddamsetti, Anya Vostinar, and Emily Dolson for discussion and feedback in the drafting of this chapter. We also thank Neerja Hajela for technical assistance. This work was supported, in

part, by the BEACON Center for the Study of Evolution in Action (NSF Cooperative Agreement DBI-0939454).

APPENDIX

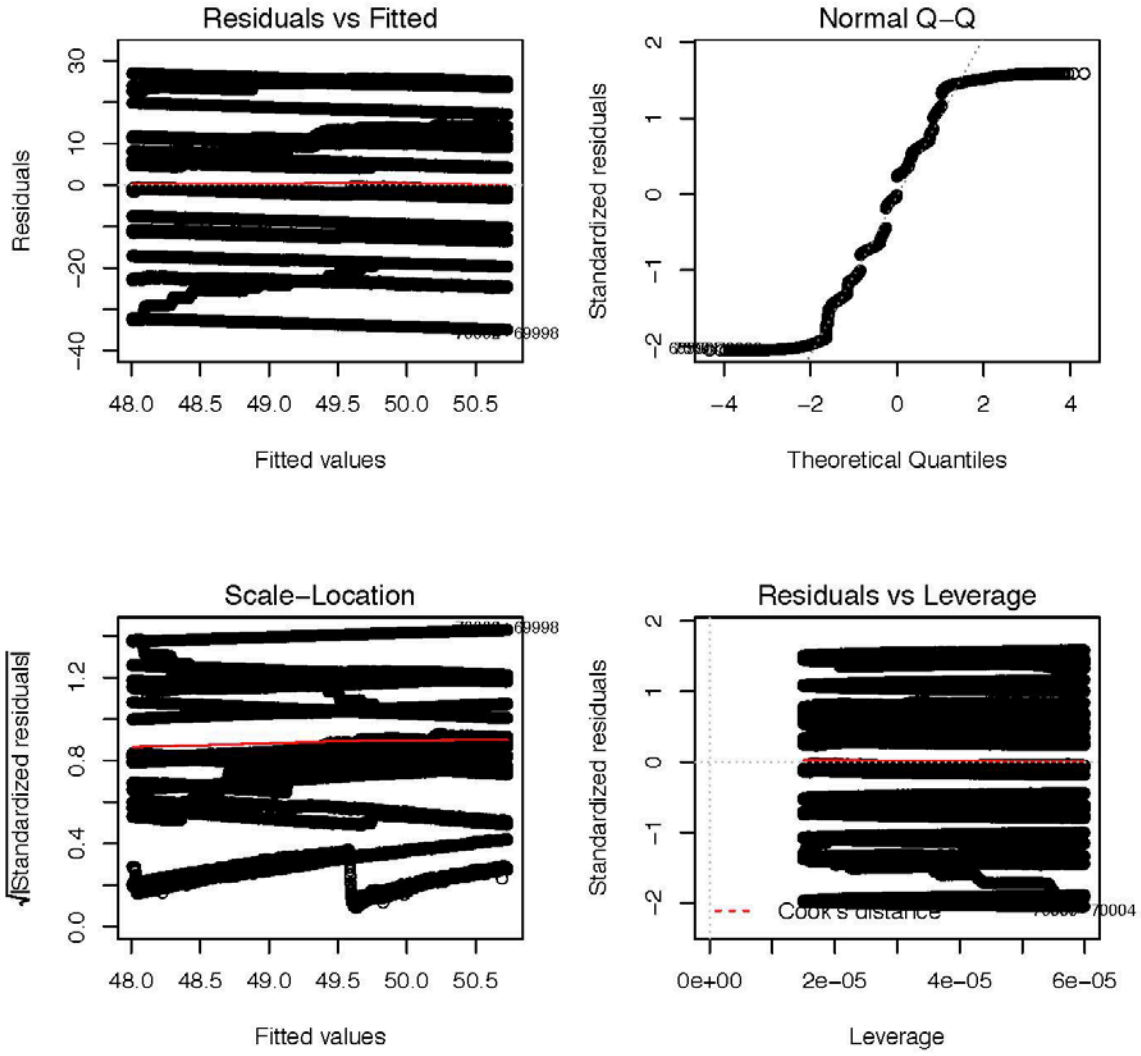

Figure S4.1: **Diagnostic plots for Logic-77 environment, late fitness as linear model**.

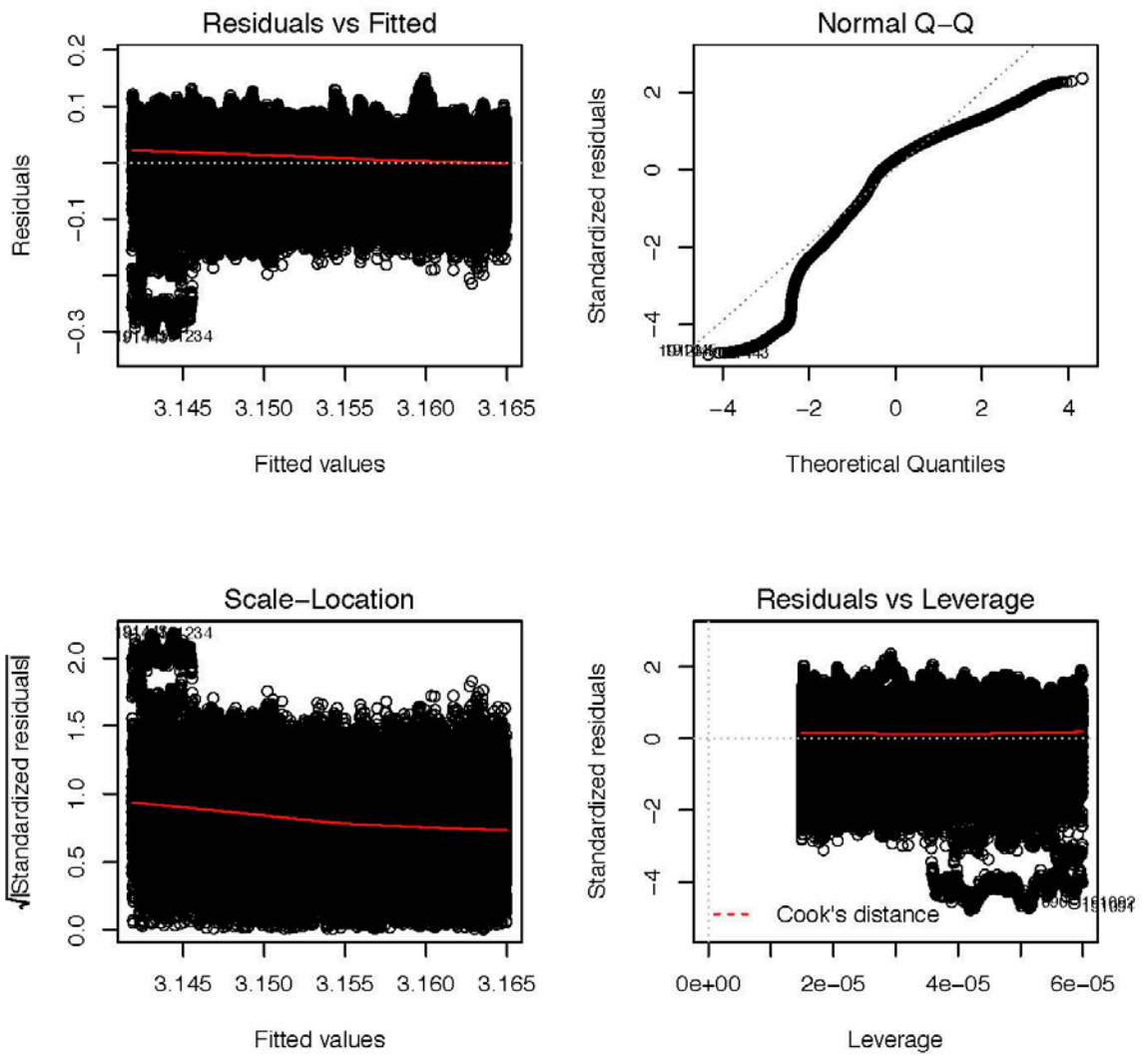

Figure S4.2: **Diagnostic plots for No Task environment, late fitness as linear model**.

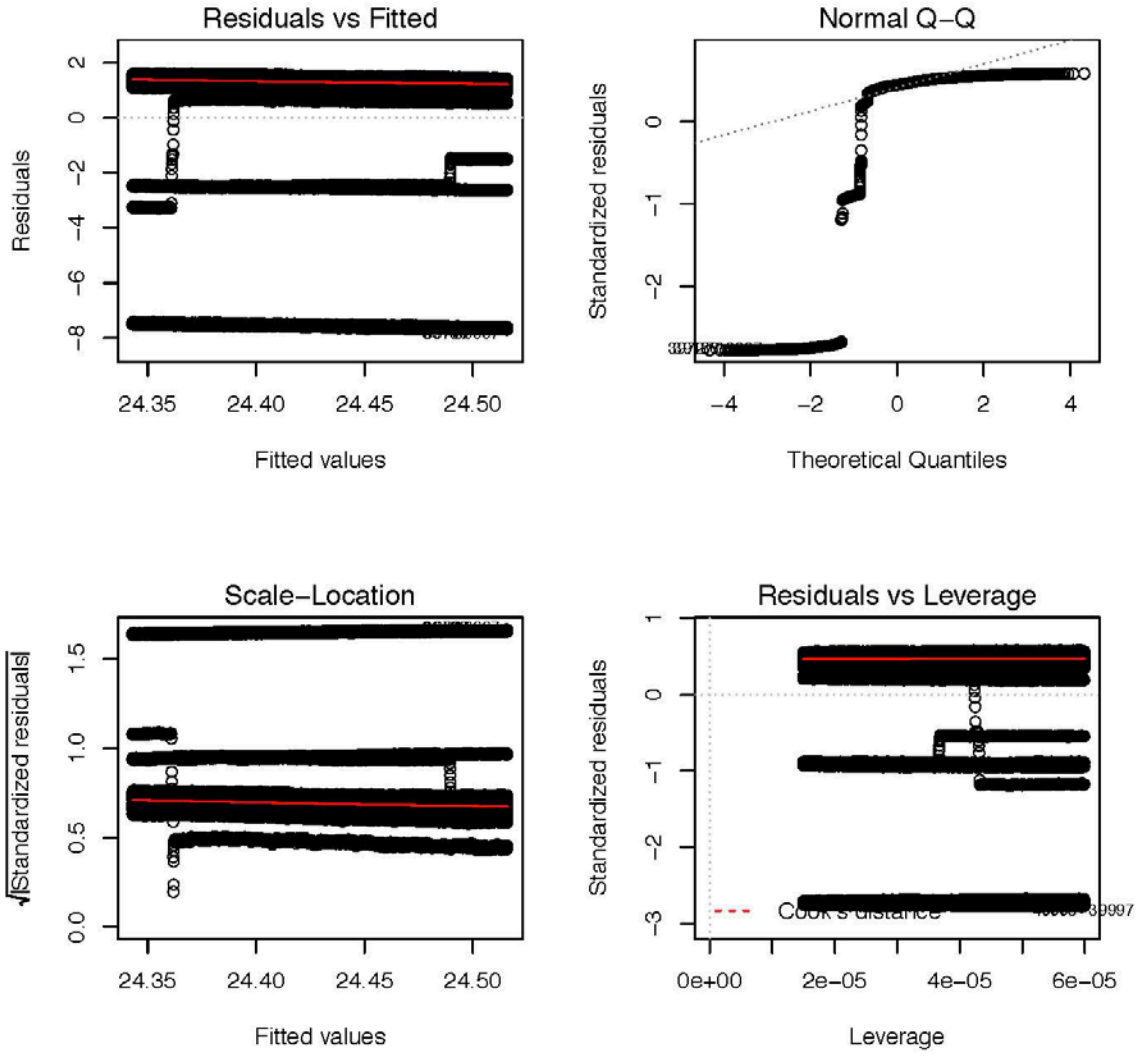

Figure S4.3: **Diagnostic plots for Logic-9 environment, late fitness as linear model**.

REFERENCES

REFERENCES

1. M. J. Wiser, N. Ribeck, R. E. Lenski, Long-term dynamics of adaptation in asexual populations. *Science*. **342**, 1364–1367 (2013).

2. M. J. Wiser, R. E. Lenski, A Comparison of Methods to Measure Fitness in Escherichia coli. *PLoS ONE*. **10**, e0126210 (2015).

3. J. Maynard Smith, Byte-sized evolution. *Nature*. **355**, 772–773 (1992).

4. C. Ofria, D. M. Bryson, C. O. Wilke, in *Artificial Life Models in Software* (Springer London, 2009), pp. 3–35.

5. D. C. Dennett, Darwin's dangerous idea. *The Sciences*. **35**, 34–40 (1995).

6. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria, 2013; http://www.R-project.org/).

7. A. E. Raftery, Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–164 (1995).

8. C. B. Turner, Z. D. Blount, D. H. Mitchell, R. E. Lenski, Evolution and coexistence in response to a key innovation in a long-term evolution experiment with Escherichia coli. *bioRxiv* (2015), doi:10.1101/020958.

9. R. E. Lenski, C. Ofria, R. T. Pennock, C. Adami, The evolutionary origin of complex features. *Nature*. **423**, 139–144 (2003).

10. H. Zhang, M. Travisano, in *Artificial Life, 2007. ALIFE '07. IEEE Symposium on* (2007), pp. 39–46.

11. C. Adami, C. Ofria, T. C. Collier, Evolution of biological complexity. *Proc. Natl. Acad. Sci.* **97**, 4463–4468 (2000).

12. J. Clune *et al.*, Natural Selection Fails to Optimize Mutation Rates for Long-Term Adaptation on Rugged Fitness Landscapes. *PLoS Comput Biol*. **4**, e1000187 (2008).

13. R. K. Standish, Open-Ended Artifical Evolution. *Int. J. Comput. Intell. Appl.* **03**, 167–175 (2003).

14. D. Misevic, C. Ofria, R. E. Lenski, Sexual reproduction reshapes the genetic architecture of digital organisms. *Proc. R. Soc. Lond. B Biol. Sci.* **273**, 457–464 (2006).

15. S. S. Chow, C. O. Wilke, C. Ofria, R. E. Lenski, C. Adami, Adaptive Radiation from Resource Competition in Digital Organisms. *Science*. **305**, 84–86 (2004).

16. E. Crozat, N. Philippe, R. E. Lenski, J. Geiselmann, D. Schneider, Long-term experimental evolution in Escherichia coli. XII. DNA topology as a key target of selection. *Genetics*. **169** (2005), doi:10.1534/genetics.104.035717.