EVOLUTION OF LABORATORY AND NATURAL POPULATIONS OF *ESCHERICHIA COLI*

By

Rohan Maddamsetti

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

2016

ABSTRACT

EVOLUTION OF LABORATORY AND NATURAL POPULATIONS OF *ESCHERICHIA COLI*

By

Rohan Maddamsetti

My dissertation spans two dichotomies: evolution in the laboratory versus evolution in nature, and asexual versus sexual evolutionary dynamics. In Chapter 1 I describe asexual evolutionary dynamics in one population of Lenski's long-term evolution experiment with *Escherichia coli*. I describe cohorts of mutations that sweep to fixation together as characteristic of clonal interference dynamics. I also describe an ecological interaction that evolved and then went extinct after thousands of generations, and discuss how such interactions affect cohorts of mutations. In Chapter 2 I report that conserved core genes tend to be targets of selection in the long-term experiment. In Chapter 3, I investigate the surprising observation that synonymous genetic diversity is not uniform across the genomes of natural *E. coli* isolates. This observation is surprising because in clonal organisms with a constant point mutation rate, synonymous diversity should be constant across the genome. I use patterns of synonymous mutations in the long-term experiment to argue that genome-wide variation in the mutation rate does not adequately explain patterns of synonymous genetic diversity. In Chapter 4, I propose that recombination and gene flow could account for genome-wide variation in synonymous genetic diversity. In Chapter 5, I analyze *E. coli* genomes isolated from an evolution experiment with recombination in which *E. coli* K-12 with known growth defects could donate genetic material to recipient populations founded by long-term experiment clones. The degree of recombination varied dramatically across sequenced clones. The strongest predictor of successful transfer was proximity to the *oriT* origin of transfer in the K-12 donors. Donor alleles close to *oriT* replaced their recipient

counterparts at a high rate, and in many of those cases, known beneficial mutations in the recipients were replaced by donor alleles.

ACKNOWLEDGEMENTS

Many people have helped me complete my Ph.D. and write my dissertation. First, I thank my mother and father for raising me and coming to America. My privilege stems from their struggles on three continents. My parents made enormous personal sacrifices to make sure that my brother and I have more freedom of choice than they had in India. I owe them for surrounding me with books, and for encouraging my curiosity and creativity from a very young age. I thank my brother, Pavan, whom I love the most. I have many fond memories of our membership in the secretive and mysterious organization known only as the Brotherhood.

At times, graduate school was a struggle. I have since discovered that struggle is a great teacher, second only to failure. But much of my struggle would have been fruitless without my partner Jihea's unconditional love, care, and support. I see now that I made it surprisingly far in life without a good understanding of some obvious, basic truths about being a human being. I thank her for continually warning me of the dangers of self-absorption, and for encouraging me to be more generous as a friend, professional, and human being. Jihea taught me that many of my accomplishments are grounded in privilege that I have largely taken for granted. She also taught me to be aware of how privilege, class, and race invisibly shape the world. Thanks to her, I want to build better, more inclusive scientific communities and institutions, which of course starts with personal and professional relationships built on trust and respect. In short, I am a better person for her, and I love her for that.

I must thank my adviser, Richard Lenski. Rich is a brilliant scientist and a generous person, and I aspire to be the kind of person that he is. However, the best part of being a graduate student in the Lenski lab is access to the many wonderful minds that pass through. I thank Neerja for her role as lab mother, and for making sure that everything works smoothly. I am especially

grateful to Alita Burmeister, Mike Wiser, and Caroline Turner for many comments on earlier versions of my dissertation in writing group. I also thank Jeff Morris, Zack Blount, Noah Ribeck, Luis Zaman and Justin Meyer for many helpful comments and advice over the years. I look up to Justin in particular as a great role model for what an early career scientist should be, in terms of asking important questions, getting things done, and reaching out to others for help as well as helping out many other scientists (including myself). I also owe a special thank you to Jeff Barrick for getting me started in the lab, and mentoring me a great deal on the project that he conceived, started, and funded that became Chapter 1 of my dissertation (and an award-winning paper!).

My past scientific mentors have all given me a tremendous amount of intellectual freedom. I thank Michael Lynch for letting me do bioinformatics research in his lab: even though my hypothesis turned out to be wrong, it was a valuable learning experience. I thank Dan Weinreich for first showing me the world of computational and experimental evolution and for letting my design my own project as an undergraduate, and Sorin Istrail for urging me to study computational biology—even though it was a struggle and I hated it at first, it has paid dividends years later. I thank Vaughn Cooper for reading and supporting my work and putting me in touch with Phil Hatcher, who turned out to be a great collaborator. And finally I have to thank Steven Valenziano, without whose Adobe Illustrator skills the Muller plot that is the centerpiece of Chapter 1 would have been impossible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: ADAPTATION, CLONAL INTERFERENCE, AND FREQUENCY-DEPENDENT INTERACTIONS IN A LONG-TERM EVOLUTION EXPERIMENT WITH *ESCHERICHIA COLI*

Authors: Rohan Maddamsetti, Richard E. Lenski, and Jeffrey E. Barrick

Originally published in the journal *Genetics*, 200: 619–631.

Abstract

Twelve replicate populations of *Escherichia coli* have been evolving in the laboratory for more than 25 years and 60,000 generations. We analyzed bacteria from whole-population samples frozen every 500 generations through 20,000 generations for one well-studied population, called Ara–1. By tracking 42 known mutations in these samples, we reconstructed the history of this population's genotypic evolution over this period. The evolutionary dynamics of Ara–1 show strong evidence of selective sweeps as well as clonal interference between competing lineages bearing different beneficial mutations. In some cases, sets of several mutations approached fixation simultaneously, often conveying no information about their order of origination; we present several possible explanations for the existence of these mutational cohorts. Against a backdrop of rapid selective sweeps both earlier and later, two genetically diverged clades coexisted for over 6000 generations before one went extinct. In that time, many additional mutations arose in the clade that eventually prevailed. We show that the clades evolved a frequency-dependent interaction, which prevented the immediate competitive exclusion of either clade, but which collapsed as beneficial mutations accumulated in the clade that prevailed. Clonal interference and frequency dependence can occur even in the simplest microbial populations. Furthermore, frequency dependence may generate dynamics that extend the period of coexistence that would otherwise be sustained by clonal interference alone.

Introduction

The long-term evolution experiment (LTEE) spans more than 25 years and 60,000 generations of bacterial evolution. In this experiment, 12 replicate populations of *Escherichia coli* have been propagated in a simple environment, and samples of each population frozen at 500-generation intervals. This experiment originally focused on whether and to what extent the populations would diverge in their mean fitness and other phenotypic properties as they adapted to identical environments (Lenski *et al.* 1991; Lenski and Travisano 1994). Over time, this experiment has become a model for exploring many other aspects of evolution, including the emergence of new functions (Blount *et al.* 2008), the evolution of mutation rates (Sniegowski *et al.* 1997), the maintenance of genetic diversity (Elena and Lenski 1997; Rozen and Lenski 2000; Le Gac *et al.* 2012), and the structure of the fitness landscape (Woods *et al.* 2011; Khan *et al.* 2012; Wiser *et al.* 2013). The ability to examine these and other issues has grown tremendously as data that were difficult or impossible to obtain when the LTEE began have yielded to new technologies, particularly genome sequencing (Barrick and Lenski 2009; Barrick *et al.* 2009; Blount *et al.* 2012; Barrick and Lenski 2013; Wielgoss *et al.* 2013).

The LTEE has also inspired theoretical work, especially on the dynamics of adaptation in large asexual populations (Gerrish and Lenski 1998; Hegreness *et al.* 2006; Desai and Fisher 2007; Schiffels *et al.* 2011; Park and Krug 2013; Wiser *et al.* 2013). The LTEE populations are subject to clonal interference, a phenomenon that limits the rate of adaptation by natural selection in large asexual populations. In the absence of recombination, two or more beneficial mutations that appear in different lineages in the same population cannot recombine into a single background; instead, the lineages that possess alternative beneficial mutations compete with one another. As a consequence, each beneficial mutation will interfere with the progress of other

contending beneficial mutations toward fixation, though the mean fitness of the population will nonetheless rise as the beneficial alleles collectively displace their progenitors. Although some early theory on clonal interference was developed with the LTEE in mind (Gerrish and Lenski 1998), other evolution experiments using bacteria and yeast have provided compelling demonstrations of this phenomenon by combining dense temporal sampling with intensive discrimination using genetic markers (de Visser and Rozen 2006; Hegreness *et al.* 2006; Woods *et al.* 2011; Barroso-Batista *et al.* 2014; Levy et al. 2015), in-depth analysis of genes under positive selection (Lee and Marx 2013), or whole-genome sequencing (Lang *et al.* 2011; Lang *et al.* 2013).

Without clonal interference, the classic model of periodic selection in asexual populations involves selective sweeps of beneficial mutations that arise singly and fix sequentially. Neutral and nearly neutral mutations that would otherwise accumulate in an evolving population are swept out (or occasionally to fixation) as each successive beneficial mutation goes to fixation (Atwood *et al.* 1951). As a consequence, within-population genetic diversity rises and falls in conjunction with the successive sweeps, and no specific polymorphism is maintained indefinitely. Clonal interference can increase genetic diversity in three ways. First, the multiple beneficial mutations, all rising to moderate frequencies, increase diversity relative to a single allele rising to high frequency. Second, the beneficial alleles remain at intermediate frequencies longer than would a single beneficial mutation of comparable effect-size during a classic sweep. Third, while the beneficial mutations remain at intermediate frequencies, their associated lineages can accumulate neutral and nearly neutral polymorphisms that also persist longer than they would in the face of selective sweeps that progress to fixation. Nonetheless, the diversity-promoting effects of clonal interference are only transient because eventually one lineage or another will

prevail, either because the beneficial mutation it carries is superior to the others (even though all lineages are more fit than their predecessors) or because one lineage acquires additional beneficial mutations that eventually break the logjam.

Although clonal interference can prolong polymorphic states only transiently, ecological interactions between genetically diverged individuals in a population that result in negative frequency-dependent effects on fitness can sustain polymorphisms indefinitely, at least in principle. The LTEE environment was designed to minimize the potential for frequency-dependent interactions to arise, in order to simplify measuring fitness and assessing the repeatability of evolution. In particular, there are no predators, parasites, or other competing species in the LTEE environment; there is no sexual recombination or horizontal gene transfer; the physical environment is well mixed and lacks spatial structure; and the density-limiting resource, glucose, is provided at a low concentration, which limits the concentration of metabolic byproducts that could support cross-feeding specialists. Nonetheless, frequency-dependent interactions have emerged in at least two LTEE populations (Elena and Lenski 1997; Rozen and Lenski 2000; Blount *et al.* 2008; Rozen *et al.* 2009; Le Gac *et al.* 2012; Plucain *et al.* 2014; Ribeck and Lenski 2015), and it is possible that such interactions have evolved in some or all of the other populations as well.

In previous work, we sequenced individual clones and population samples taken at generations 2000, 5000, 10,000, 15,000, 20,000, 30,000, and 40,000 from one LTEE population, designated Ara–1, that has served as the focal population for many in-depth analyses (Barrick and Lenski 2009; Barrick *et al.* 2009). Using these genomic data, we designed genotyping assays for many derived alleles present in one or more clones or population samples. In this study, we use population samples taken at 500-generation intervals to examine at high resolution the

dynamics of 42 mutant alleles over the first 20,000 generations of this population. These data show compelling evidence for both adaptive fixations and clonal interference. The data also suggest, and competition assays confirm, that a negative frequency-dependent interaction evolved that delayed the fixation of some mutations for several thousand generations—in effect extending the duration of a particular bout of clonal interference. Nevertheless, one clade's slight advantage when rare was eventually overcome by the consolidation of further beneficial mutations in the other clade, and this bubble of transient diversity collapsed. Our results are broadly consistent with other evolution experiments that have examined the dynamics of *E. coli* populations as they adapt to simple (chemostat) and complex (mouse gut) environments (Maharjan *et al.* 2015; Barroso-Batista *et al*. 2014).

Materials and Methods

Ara–1 Population Samples

The LTEE is described in detail elsewhere (Lenski *et al.* 1991; Lenski and Travisano 1994; Lenski 2004). In brief, 12 replicate populations have been propagated in a glucose-limited medium called DM25. Daily 100-fold dilutions and re-growth allow ~6.6 generations per day. Every 500 generations (75 days), after the transfers into fresh medium, glycerol was added as a cryoprotectant to the remaining cultures, which were then stored for later research at –80°C. In this study, we analyzed the 40 samples of population Ara–1 collected from 500 to 20,000 generations.

Sample Preparation for Genotypic Analyses

We briefly thawed the top portion of each frozen sample and removed 0.1 ml. We washed the cells in a saline solution and centrifuged them to remove residual glycerol; we then inoculated the cells into flasks containing 10 ml of DM100 medium (the same medium used in the LTEE, except with a higher glucose concentration to yield more cells,) which were incubated for 24 h at 37°C. We also mixed fully-grown cultures of the Ara^+ ancestral strain, REL607, and a 20,000-generation clone from population Ara–1, REL8593A, at several cell ratios for use as standards for calibrating the genotypic analyses. We then isolated genomic DNA from all of these cultures using an Invitrogen PureLink *Pro* 96 Genomic DNA Purification kit.

In a separate experiment, we revived the frozen samples from generations 7500 and 10,000 using the same protocol as above. The next day, we diluted and plated samples from the cultures, and the following day we picked 90 clones (single-colony isolates) from each sample. We inoculated each of these clones into a flask containing 10 ml of DM1000 medium and isolated genomic DNA from these cultures after growth, as described for the population samples.

We quantified genomic DNA concentrations for all samples using an Invitrogen Quant-iT PicoGreen dsDNA Assay kit, and we then submitted the samples for genotyping to the MSU Genomics Core facility. We prepared and submitted three different biological replicates for the population-level analyses, each of which was separately revived from the frozen sample and re-cultured before DNA isolation.

Allele Frequency Measurements

We performed an Illumina GoldenGate Genotyping Assay with Veracode technology to measure allele frequencies in the Ara–1 population. We designed allele-specific oligonucleotides for

single nucleotide polymorphisms (SNPs), insertions, deletions, and rearrangements that were previously discovered by sequencing either clones or whole-population samples (Barrick and Lenski 2009; Barrick et al. 2009). These oligonucleotides contain universal PCR primer sequences and barcode sequences targeting specific Veracode microbead types. Allele-specific, fluorescent-labeled PCR products hybridize to the microbeads, and the fluorescent signal provides an indicator of allele frequency. Of the assays we designed, 42 SNPs and indels yielded useful data about the history of the Ara–1 population.

Illumina's GenomeStudio software gave initial estimates of allele frequencies. We used the allele frequency estimates generated from known mixtures of REL8593A and REL607 to correct the frequencies of 26 of the 42 mutations that were present in REL8593A (all those that had fixed in the population by 20,000 generations). These DNA samples consisted of 21 known mixtures of the two strains designed to contain every 5% increment, based on culture volumes, in the percentage of REL8593A from 0% to 100%. We estimated the true ratio of cell numbers in each mixed sample by plating a dilution on tetrazolium arabinose (TA) indicator agar and counting the red and white colonies made by REL8593A and REL607, respectively.

Average cell size increased over time in the LTEE (Lenski and Travisano 1994), and the evolved bacteria also contain more DNA per cell than the ancestral strain (Lenski *et al.* 1998). This change in DNA content meant that the ratio of REL8593A to REL607 cells in the control mixtures had to be corrected to reflect the actual number of copies of the evolved and ancestral alleles in these samples. To account for this difference, we stained stationary-phase cultures of the ancestral Ara$^-$ clone, REL606, and the 20,000-generation clone, REL8593A, with a PicoGreen fluorescent DNA stain (Ferullo *et al.* 2009) and used flow cytometry to quantify the average genomic DNA content per cell in DM100 media. We used six replicate measurements to

calculate a correction factor for the allele frequencies in the REL8593A/REL607 mixtures. On average, the evolved REL8593A cells had 1.67 ± 0.23 (± 95% confidence interval) times as much DNA as the ancestral REL606 cells under these conditions.

With this information, we further corrected the raw frequencies estimated for each allele (θ_0) from the GenomeStudio genotyping software to the known frequencies of that allele (θ) in the control DNA samples. From the triplicate assays of the REL8593A/REL607 mixtures, we fit a calibration curve for each allele to an empirical function that corrected for a symmetric convex bias that captured the deviation from linearity. Specifically, we fit five coefficients (c_n) in a linear model of the form $\theta = \theta_0 + c_1(\theta_0 - \theta_0^2) + c_2(\theta_0 - \theta_0^4) + c_3(\theta_0 - \theta_0^4) + c_5(\theta_0 - \theta_0^8) + c_5(\theta_0 - \theta_0^{10})$ to each calibration curve. Then, we used these curves to correct the GenomeStudio estimates of allele frequencies in all of the Ara–1 population samples that we analyzed. These calculations, as well as the code used to plot the temporal dynamics of the allele frequencies, are available as R analysis scripts at the Dryad Digital Depository.

For the 17 transient alleles we analyzed that were not present in the REL8593A genome, we could not correct their estimated allele frequencies in this way. These allele estimates are thus expected to be slightly less accurate. In cases where these unsuccessful mutations were competing with a contemporary allele that fixed by 20,000 generations, we constrained their frequencies to being no more than 100% minus the prevalence of the most abundant successful mutation. The Muller plot in Figure 1.1A is based on a simple interpretation of how the estimated frequencies of each genotype sum to 100%.

Pyrosequencing

We confirmed the existence of large temporal fluctuations in the two coexisting clades in the Ara–1 population by using pyrosequencing to estimate the *rpsA*, *yghJ*, and *gltB* allele frequencies in samples from 5,000 to 14,000 generations. Standard assays and analysis of peak intensities using a Qiagen Pyromark 24 instrument were used to estimate the ratios of alternative alleles involving single-base substitutions in the cases of *rpsA* and *yghJ* and a 16-bp deletion in the case of *gltB*. Again, control mixtures of REL607 and REL8953A cultures were used to verify the accuracy of inferred allele frequencies, including the correction for the relative DNA content in each strain as described above.

Competitions to Test for Frequency-Dependent Interactions

The ancestral strain REL606 and the cells from population Ara–1 cannot grow on the sugar L-arabinose owing to a point mutation in the *araA* gene (Studier *et al.* 2009). However, one can obtain Ara^+ mutants by inoculating large numbers of cells onto minimal arabinose (MA) plates; most of the resulting mutants are selectively neutral under the conditions of the LTEE, and the Ara phenotype serves as a useful marker in competition experiments used to measure relative fitness (Lenski *et al.* 1991). Thus, we isolated Ara^+ mutants of certain evolved clones of interest by inoculating those clones onto MA plates. We then confirmed the neutrality of those mutations with competition assays involving a 1:1 volumetric ratio of the Ara^+ mutants and their Ara^- parents. These Ara^+ mutants and their Ara^- parents were then used to test for frequency-dependent interactions per the following procedure; this procedure is identical to that used to test for neutral Ara^+ mutants, with the exception of using three volumetric ratios instead of one.

To begin a competition assay, two clones (one Ara^- and the other Ara^+) were inoculated separately from freezer stocks into flasks containing LB medium and grown for 24 h at 37C. These cultures were then diluted 100-fold into saline solution, and 0.1 ml was transferred into 9.9 ml of DM25 (i.e., the same medium as used in the LTEE), where the cultures were again incubated for 24 h at 37C. This step ensured that the competitors were physiologically acclimated to the same conditions where they would compete.

The two clones were then mixed at three initial volumetric ratios (1:9, 1:1, and 9:1) and a combined 0.1 ml was added to 9.9 ml of DM25. An initial sample was taken immediately from this mixture and plated on TA agar (where Ara^- and Ara^+ cells produce red and white colonies, respectively) to estimate the initial number of each type (N_i). The mixture was then propagated for six days with daily 1:100 dilutions in DM25 medium. At the end of the experiment, a second sample was plated onto TA agar to estimate the final number of each type of cell (N_f). The Malthusian parameter (m) was then calculated as the realized growth rate of a competitor over the competition, as follows: $m = \ln(100^t \times N_f/N_i)/t$, where t is the number of days of dilution and re-growth, and where cell numbers reflect equivalent end-of-cycle values based on the dilutions used for plating. The fitness of one clone relative to the other was then calculated simply as the ratio of their Malthusian parameters.

For each pair of competitors, we performed 12 replicate assays at each initial ratio. The plate counts, fitness calculations, and R scripts used to analyze and plot the data are available at the Dryad Digital Depository.

Results and Discussion

Muller Plot of Allele Dynamics

Figure 1.1A shows the dynamics across 20,000 generations of 42 spontaneous mutations that arose and reached detectable frequencies in the Ara–1 population of the LTEE. Each mutation is identified by the name of the gene in which it occurred (e.g., *topA*) or, for mutations in intergenic regions, by the adjacent genes (e.g., *yedW-yedX*). In a few cases, a numeral follows the name of a gene in order to distinguish mutations that affected the same gene (e.g., several *pykF* alleles). Twenty-five of these 42 mutations – all of them with labels preceded by dots (e.g., *topA* at the left and *rpsD* on the right) – eventually reached fixation (or nearly so) in the population. In other words, these mutations were on the line of descent leading to fully sequenced clones from generations 30,000 and 40,000. The other 17 mutations were transient (e.g., *fabR* on the left). That is, these alleles reached detectable frequencies – and in a few cases even achieved majority status (e.g., *acs-nrfA* between generations 8,500 and 12,500) – but they later went extinct.

Not All Mutations Were Tracked

Figure 1.1A contains a great deal of data, but when interpreting these dynamics it is important to realize that some information is missing because only a subset of all mutations were targeted in our genotyping assays. Many alleles were undoubtedly rare and transient; if an allele was not moderately abundant (at least several percent) in one of the generational samples that were deeply sequenced (Barrick and Lenski 2009), and if it was not on the line of descent leading to the clones sequenced in later generations (Barrick *et al.* 2009), then we could not design a targeted genotyping assay to detect it. Moreover, even some alleles that were known to fix in the population could not be accurately quantified using our genotyping method, (e.g., new

transposon insertions and other structural mutations) and so information for those mutations is also missing from Figure 1.1A. We attempted to assay all 37 mutations in the 20,000-generation clone, REL8593A, that were on the line of descent leading to clones sequenced at 30,000 and 40,000 generations (Barrick *et al.* 2009). However, twelve assays failed: six of the failures were structural mutations mediated by insertion-sequence elements, one was a 1-bp insertion, and five were SNPs. Including mutations off the line of descent, REL8593A has 45 mutations in total, comprising 29 SNPs and 16 insertions, deletions, and other rearrangements (Barrick *et al.* 2009). The fact that a single clone had almost twice as many mutations as the number successfully tracked in that same clone implies that many other variants were overlooked. In the future, improved estimates of the cumulative number of polymorphisms might be derived by more frequent whole-population sequencing, especially if this data is analyzed using methods that can ably detect both SNPs and non-SNP mutations (Deatherage *et al.*, 2015).

Pure Drift Cannot Explain Evolutionary Rate

Random genetic drift alone cannot explain the large number of fixations observed or, for that matter, even one fixation. Each LTEE population began from a single haploid cell, and therefore any variant arose as a new mutation with an initial frequency of $1/N$, where N is the population size. For a neutral mutation that eventually fixes in a population, the expected time to fix by random drift alone (i.e., without hitchhiking) is on the order of N generations (Kimura 1983). In the LTEE, N fluctuates between ~5×10^6 and ~5×10^8 as a consequence of the daily dilutions and regrowth (Lenski *et al.* 1991). These numbers are far too large to allow mutations to fix by pure drift in 20,000 generations, much less within the many fewer generations observed for the earliest fixations.

Beneficial Drivers

We also know that mean fitness increased over the course of the LTEE. It increased in an almost step-like manner over the first few thousand generations (Lenski *et al.* 1991; Lenski and Travisano 1994), implying a series of fixations of beneficial driver mutations under positive selection, and it has continued to increase more gradually during the subsequent tens of thousands of generations (de Visser and Lenski 2002; Wiser *et al.* 2013). We also know or infer that particular mutations were beneficial because isogenic strains were constructed to measure the fitness effects of single evolved mutations, because similar mutations affecting the same gene reached high frequencies or fixed in many replicate LTEE populations, or both. These known or suspected beneficial drivers include 15 of the mutations in Figure 1.1A including *topA* (Crozat *et al.* 2005, 2010; Woods *et al.* 2011), *fabR* (Woods *et al.* 2011; Deatherage *et al.* 2015), *spoT* (Cooper *et al.* 2003; Woods *et al.* 2011), three in *pykF* (Woods *et al.* 2006, 2011; Barrick *et al.* 2009), *mrdB* (Woods *et al.* 2006), *mrdA* (Woods *et al.* 2006), *malT* (Pelosi *et al.* 2006), *infB* (Barrick *et al.* 2009), *fis* (Crozat *et al.* 2005), *nadR* (Woods *et al.* 2006; Barrick *et al.* 2009), *pcnB* (Barrick *et al.* 2009), *iclR* (Barrick and Lenski 2009), and *rpsD* (Barrick *et al.* 2009). It is likely that additional mutations in Figure 1.1A were also beneficial drivers, while some others may have been neutral or weakly deleterious hitchhikers that achieved transiently high frequency or fixation by virtue of beneficial mutations present in the same genomes.

Clonal Interference

Clonal interference refers to the effect of competition between beneficial mutations that occur in different lineages in the same asexual population; its effects are clearly seen in Figure 1.1A as wedges that open, expand, and then close. Owing to the absence of genetic exchange, two or

more beneficial mutations that arise contemporaneously, but in different lineages, cannot be combined in the same genome. Instead, one will eventually prevail and the others must go extinct. The winning lineage might prevail because its initial driver mutation is more beneficial than the initial drivers in the contending lineages. However, the winner might also depend on the effects of later beneficial mutations that arise in one or more of the contending lineages. Such an outcome is especially likely if the initial drivers in the different lineages have similar fitness effects, because that similarity means it would take longer for one lineage to exclude the other, thus providing more time for later beneficial mutations to occur and affect the outcome.

One interference wedge is seen near the very start as a mutation in *fabR* rises in frequency over the first 1000 generations, but that lineage declines precipitously during the next 500 generations before disappearing entirely (Figure 1.1A). Five more wedges rise and fall between generations 1000 and 2500. Strikingly, three of them involve different mutations in one gene, *pykF*. Yet another mutation in *pykF* later went to fixation in this population (Schneider *et al.* 2000, Barrick *et al.* 2009), but it is not shown here because the type of mutation – an insertion of an IS*150* element – was not amenable to the method for detecting genetic polymorphisms used in this study. Moreover, different *pykF* mutations fixed in all 12 LTEE populations (Woods *et al.* 2006). Thus, there were numerous parallel increases of *pykF* mutations, both within and across populations, and this parallel evolution provides strong evidence that these were beneficial mutations.

Some additional interference wedges are also present later in the Muller plot including, most notably, an extremely large one that begins with a mutation in *rpsA* around generation 5000. This lineage became numerically dominant by generation 8000; it remained the majority for thousands of generations before a precipitous decline and sharp recovery between ~11,000 and

13,000 generations; and it finally petered out to extinction by generation 15,000. We will return to this episode in a later section on "Evidence for Frequency Dependence".

Rapid Increases in Mutation Frequencies

Except for the period from 5000 to 15,000 generations, most mutations that fixed in the Ara–1 population did so very quickly, usually within a few thousand generations. Although these dynamics are visually striking, they are not unexpectedly fast. Three issues come into play when we consider these dynamics. First, the alleles tracked in the Muller plot arose by mutation long before they are shown as being present. Only alleles with frequencies well above 1% could be reliably detected using our methods, but when any new mutation occurred its frequency was $1/N$, a mere 0.0000002% to 0.00002% depending on when the mutational event happened in the transfer cycle. In any case, each successful allele had to reach at least 0.00002% after surviving the first transfer event, which took place within the first day after it arose. Thus, mutations were hidden from view until they had increased in frequency by 50,000-fold or more.

Second, the fitness effects of beneficial mutations that evolved in the LTEE, and which have been measured by constructing and competing isogenic strains, are typically between ~1% and ~10% depending on the mutation (Barrick *et al.* 2009). Assuming a constant fitness difference, the logarithm of the ratio of the beneficial mutant to its progenitor should increase linearly (Dykhuizen and Hartl 1983). For a mutation that confers a 10% advantage, it takes only ~33 generations (doublings, as used in the LTEE) to rise from 1:100 to 1:10, another 33 generations to increase from 1:10 to 1:1 (i.e., 50%), another to go from 1:1 to 10:1, and one more to achieve near-fixation at 100:1. For this mutation, the entire 'visible' process would occur in ~130 generations; however, the time steps in Figure 1.1A are 500 generations each, and so the

fixation would likely be manifest in a single interval. For a mutation that gives a 1% benefit, each order-of-magnitude change in the ratio would take 10-fold longer, or ~330 generations. In that case, the visible fixation process would require ~1300 generations, which still falls within a mere three time steps in the Muller plot.

Third, in many of the fixation events, the initial rate of increase in a lineage was much steeper than the final rate. For example, the lineage with the *ybaL* allele went from undetectable in the 2000-generation sample to being the vast majority by generation 2500, but it did not fix until after generation 5000. Stated the other way, many lineages are unexpectedly persistent, such as the lineage that carried the *spoT* allele but lacked the *yegI* and *ybaL* mutations, which hung on at low frequency from 2500 to 5000 generations. We can posit two explanations for this effect. First, as a consequence of clonal interference, a lineage that has a new beneficial mutation is initially competing primarily with its progenitor, but over time that lineage will increasingly compete against other lineages with different beneficial alleles, thereby slowing its ascent (Lang *et al.* 2011).

A second possibility is frequency-dependent selection, such that the unexpectedly persistent lineage has acquired some mutation that gives it a fitness advantage when it is rare; for example, it may more efficiently use a metabolic byproduct of the other lineage. This advantage when rare allows the lineage to survive longer than it otherwise could; in the absence of on-going evolution, its advantage when rare would allow it to persist forever. However, in this scenario the new majority lineage evolves, sooner or later, additional beneficial mutations that overcome the minority type's advantage when rare, thus completing the extinction of the once-dominant lineage by the successor lineage. Incidentally, a somewhat different scenario occurred in another of the LTEE populations. In that population, two lineages stably coexisted by frequency-

dependent selection. Both continued to fix beneficial mutations, and their relative abundance fluctuated dramatically over time. However, neither gained a sufficient advantage to drive the other extinct, and they have coexisted now for tens of thousands of generations (Rozen and Lenski 2000; Le Gac *et al.* 2012).

Nested Fixations and Cohorts that Fix Together

In the Muller plot (Figure 1.1A), we see many examples of nested fixations, in which one mutation begins its sweep to fixation in the background of a prior mutation that has not yet completed its own sweep to fixation. For example, before the *topA* allele has fixed, the *spoT* mutation has begun its rise; similarly, the *ybaL* mutation begins to spread before the *yegI* mutation has fixed. These nested fixations are not surprising; like clonal interference, they imply that mutations with fairly large beneficial effects are sufficiently common, given the population size, that more than one highly beneficial mutation is often present in contending subpopulations at the same time.

In other cases, however, we see what appear to be simultaneous fixations of two or more mutant alleles – what have been called 'cohorts' (Lang *et al.* 2013). One conspicuous example involved four mutations in the *mrdA*, *malT*, *nagC*, and *infB* genes that rose together starting at ~3,000 generations in the background carrying the *ybaL* mutant allele. Another case started around 12,500 generations with three mutations in the *pcnB*, *arcB*, and *ebgR* genes that spread and fixed together.

These quasi-simultaneous fixations may, at first glance, seem surprising. In fact, however, there are several plausible explanations for their occurrence, as explained below. An important consideration is that most of the time these mutant alleles spent rising in frequency occurred

before they reached a frequency at which they could be detected; thus, the relevant dynamics for determining the order in which the mutations happened were hidden from our view. Also, the hypotheses below are not mutually exclusive; instead, two or more of the processes might be involved in any given cohort fixation.

Simultaneous mutations. The mutations could have occurred simultaneously; that is, the two or more changes in the DNA sequence might have taken place in the same replicating cell. However, given what we know about typical mutation rates in bacteria, in general, and the LTEE, in particular (Wielgoss *et al.* 2011, 2013), this explanation seems very unlikely. It is relevant to note that the Ara–1 population did not evolve a hypermutator phenotype during the 20,000 generations studied here, although it did so later (Barrick *et al.* 2009). It is also relevant that most mutations we were able to track using our methods were simple point mutations, rather than mutations involving mobile elements or other genetic mechanisms that may occur at much higher rates (Moxon *et al.* 1994; Cooper *et al.* 2001).

Classic hitchhiking. Neutral or even slightly deleterious mutations can fix in a population if they are physically linked to a beneficial mutation that goes to fixation. There is no horizontal gene exchange in the LTEE (Lenski *et al.* 1991; Lenski 2004), and so the LTEE populations have perfect linkage. If the beneficial mutation that is eventually fixed occurs in a background that already has the neutral or deleterious mutation, then the hitchhiker and beneficial driver will fix at the same time. If the order is reversed, such that the neutral or slightly deleterious mutation occurs in the background of the beneficial mutation, but one or more generations later, then the situation becomes more complex. If the hitchhiker is deleterious, then it may rise in frequency for a while but will not fix, because the lineage that carries the beneficial mutation alone will out-compete the lineage with both mutations. If the hitchhiker is neutral, then the double mutant

will rise in parallel with the lineage that carries only the beneficial driver, but the hitchhiker should not fix. For example, if a neutral hitchhiker occurs when the lineage with the beneficial mutation has expanded to four cells, then we would expect the double mutant to reach a frequency of ~25% when the beneficial mutation fixes in the population. In both cases, these scenarios are simplifications because they ignore the effects of clonal interference and random drift. The latter is important especially while the beneficial allele is still very rare. In particular, although a hitchhiker might occur after the beneficial driver, if the genotype that contains only the beneficial mutation was lost by drift (e.g., during a transfer soon after it appeared), then the hitchhiker and driver mutations would fix simultaneously.

Beneficial co-drivers. The term hitchhiker is usually applied only to neutral or deleterious mutations, as discussed in the preceding paragraph. However, two or more beneficial mutations that occur in the same lineage can help drive one another to fixation, including cases where one or both mutations might not be able to fix by themselves owing to clonal interference from single beneficial mutations of larger effect (Cooper *et al.* 2001; Schiffels *et al.* 2011). In essence, such mutations are co-drivers that receive a boost from their partner, similar to the boost that a hitchhiker would get. Although the beneficial mutations may occur sequentially and, in fact, many generations apart, they may nonetheless appear to fix simultaneously. The reason for the apparent simultaneity of fixation hinges on the dynamics of selection coupled with the limits of detection of rare genotypes.

Consider a lineage bearing a mutation that confers a 5% fitness advantage, and which survived extinction by random drift. Using the same framework as in the section on "Rapid Increases in Mutation Frequencies," the ratio of that lineage to its progenitor should increase by an order of magnitude in ~66 generations. After 132 generations, ~100 cells will have the

beneficial mutation. Now imagine that, at this point, a second beneficial mutation occurs in the background with the first mutation, and that the double mutant has a 5% advantage relative to the single mutant and a 10% advantage over the numerically dominant type without either mutation. Again, we will assume that the double mutant escapes extinction by drift, because we are only interested in those cases that leave a record of fixations. The double mutant should increase by an order of magnitude relative to the single mutant in ~66 generations, and relative to the non-mutant progenitor in 33 generations. Let us now assume that the mutant alleles can be detected only when they become 1% of the total population. For purposes of these calculations, we will use 10^7 as the population size; this value is conservative because the effective population size, taking into account the daily transfer cycle in the LTEE, is somewhat larger (Lenski *et al.* 1991). Absent the second beneficial mutation, the single-mutant lineage would require ~330 generations to increase five orders of magnitude and reach the 1% frequency where it could be detected. Although the double mutant appeared 132 generations later, it requires only ~165 generations to reach the 1% threshold in the population, which remains numerically dominated by the genotype with neither beneficial mutation. Then, after 99 generations more—or 396 generations since the first mutation and 264 since the second—the double mutant would be ~100-fold more common than the single mutant. With the 1% limit of resolution, and with samples taken at 500-generation intervals, it would thus appear as though the two mutations had simultaneously fixed, even though they arose many generations apart and were, strictly speaking, nested.

Positive epistasis. This explanation is, in essence, an extension of the previous one. In the LTEE, there is an overall tendency toward diminishing-returns (i.e., negative) epistasis between beneficial mutations with respect to their fitness effects (Khan *et al.* 2013; Wiser *et al.* 2013), although there are also exceptions where mutations exhibit positive epistasis (Blount *et al.* 2012).

Even if negative epistasis is more common, cases of positive epistasis may play a disproportionate role in cohort fixations. This effect can be easily understood by imagining a scenario like the one discussed in the previous paragraph, except now imagine that the two mutations together give a 15% benefit, whereas either alone would provide only a 5% benefit. In this case, the double mutant would overtake the single mutant that much sooner, obscuring the sequential order faster and more completely; it would do so even if the two mutations were more separated in time, if their individual beneficial effects were smaller, or both.

Frequency-dependent selection. The scenarios presented above might allow cohorts of two or even several mutations to fix simultaneously. However, these scenarios become increasingly unlikely as the size of the fixation cohorts becomes very large. For example, in a study of experimentally evolving yeast populations, Lang *et al.* (2013) observed cohorts of 7-10 mutations that fixed simultaneously. In such cases, another explanation is needed, and frequency-dependent selection—coupled with its subsequent breakdown—provides a mechanism whereby mutational cohorts of any size can be fixed. Two lineages can stably coexist if, as a result of negative frequency dependent interactions, each type has a fitness advantage when it is rare. As long as the stable equilibrium persists, each lineage can accumulate an arbitrarily large number of distinguishing mutations by natural selection and hitchhiking (Rozen and Lenski 2000; Herron and Doebeli 2013). Now imagine that, at some later time, a beneficial mutation or combination of mutations arises in one lineage, the fitness effect of which is large enough to disrupt the stable equilibrium and drive the other lineage extinct. At that time, this hyper-beneficial mutation will not only fix but also drive to fixation in the total population all the other alleles that previously had fixed only in its lineage. In the next section, we will present evidence that such a scenario played out in population Ara–1.

Evidence for Frequency Dependence

We now turn to the striking dynamics that occurred in this population between 5000 and 15,000 generations (Figure 1.1A). To better resolve the dynamics, including especially the linkage relationships among various alleles, we genotyped 90 clonal isolates sampled from generations 7500 (Figure 1.1B) and 10,000 (Figure 1.1C) for the presence or absence of the various mutations. The earliest discernible events involved the simultaneous rise of two distinct clades from within the blue-colored background that bears a cohort of four mutations (including *mrdA*) that eventually fixed (Figure 1.1A). One of the two clades began with a mutation in the *rpsA* gene; it is aqua-colored at first, then various shades of green as more mutations accumulated with time (Figure 1.1A). The nested pattern of the first two mutations in *rpsA* and *acs-nrfA* was evident in the clones from generation 7500 (Figure 1.1B), but by generation 10,000 these two mutations were fully concordant (Figure 1.1C). Three more mutations – in *atoS*, *nuoG*, and *nuoM* – emerged as distinct sub-clades in the later sample, but they did not persist (Figure 1.1C). The other main clade appeared first as a cohort of four mutations (including *gltB*); it is shown in chartreuse initially, with shades of yellow, red, pink, and purple used as more mutations arose (Figure 1.1A). In the clones sampled at 7500 generations, it was impossible to discern the order of those first four mutations, although the next three mutations that arose in this lineage were already present as nested subsets at that point (Figure 1.1B). By generation 10,000, six mutations in this clade had become fully concordant; two mutations that would eventually fix (*pflC* and *nadR*-2) were nested, as were four others that did not persist (Figure 1.1C). Following their appearances after 5000 generations, both major clades increased in number, with the one that carried the *rpsA* mutation becoming the clear majority by 8000 generations and remaining so for most of the next few thousand generations. It then experienced a precipitous decline, a sharp

recovery, and another precipitous decline, followed by its extinction. By the end of this episode, at least 14 mutations (from *gltB* to *iclR*, and probably others not detected by our methods) had piled up on the line of descent before the lineage that ultimately prevailed finally drove the other clade extinct by 15,000 generations.

Clonal interference alone is insufficient. One might be tempted, at first glance, to suggest that this episode was simply a protracted case of clonal interference. We know that the rate of fitness increase decelerated over time (Wiser *et al.* 2013) and that the beneficial fitness effects of mutations that fixed later tend to be smaller than those that fixed earlier owing to diminishing-returns epistasis (Khan *et al.* 2011). Together, these facts imply that later fixations usually took longer than earlier ones, thereby providing the opportunity for more drawn-out bouts of clonal interference. However, this explanation does not account for two important features of the data: (i) there were rapid fixations after, as well as before, this episode; and (ii) sharp reversals in the relative abundance of the two lineages occurred within this episode. Moreover, as explained earlier, these dynamics are entirely consistent with the smaller selection coefficients that prevailed in the later generations. Therefore, we think that some other process than simple clonal interference must have contributed to this extremely drawn-out set of fixations.

Hypothetical scenario with frequency-dependent fitness. We hypothesized that a negative frequency-dependent interaction stabilized the relative abundance of the two clades, while beneficial mutations within each clade buffeted their abundances until, eventually, one drove the other extinct (Figure 1.2). Under the null hypothesis, the relative fitness of members of the two clades, taken at the same point in time, does not depend on their relative abundance in the population (Figure 1.2A). The alternative is that each type has an advantage when rare, such that there is a stable equilibrium (i.e., a relative fitness equal to unity) at some intermediate frequency

(Figure 1.2B). Now imagine that beneficial mutations occur in both lineages that alter the shape of the fitness function (Figure 1.2C). We show these changes as affecting the intercept, but not the slope, of the fitness function, as though the mutations provide a general benefit that does not affect the frequency-dependent interaction; however, one can imagine similar scenarios where the slope changes. In the scenario illustrated, the initial equilibrium frequency of clade C2 is ~0.2. A beneficial mutation in clade C1 drives the equilibrium frequency of C2 down to ~0.05 (green line), and then a later beneficial mutation in C2 drives its equilibrium frequency up to ~0.5 (blue line); finally, though, another mutation in C1 pushes the equilibrium frequency of C2 into negative territory (red line) and C2 then goes extinct, assuming it gets no further beneficial mutations, because it is less fit than C1 at all frequencies (Figure 1.2C).

Experimental tests of frequency dependence. We ran competition experiments to test whether representatives of the two lineages exhibited frequency-dependent interactions at two time points. We chose clones from the 7500-generation sample that represented the most derived genotypes within clades C1 and C2; they differed by at least nine mutations (Figure 1.1B). They competed head-to-head starting at three different ratios, with 12-fold replication for each ratio. We calculated their relative fitness as the ratio of their realized growth rates during the competition assay. We saw no evidence of frequency dependence (Figure 1.3A); of course, we cannot exclude the possibility of a very small effect, on the order of 1% or less. It is also noteworthy that there was no detectable difference in fitness between the two competitors, even though they differed by so many mutations and both were increasing in abundance relative to the overall population in which they arose (Figure 1.1A).

We repeated this experiment except using the most derived genotypes from clades C1 and C2 at 10,000 generations. In particular, the C1 clone had a mutation in *nuoM*, and the C2

clone had a mutation in *nadR* that was the latest mutation that subsequently fixed in the Ara–1 population. In this case, we observed strong frequency dependence; the C2 clone had a fitness advantage of 5.4% when its starting frequency was 10%, an advantage of 2.9% when its initial frequency was 50%, and an advantage of 0.9% when it started at 90% (Figure 1.3B). Although the C2 clone was fitter than the C1 clone at all three initial ratios tested, the trend suggests the possibility of stable coexistence with C1 at an equilibrium frequency of <5%. Note, however, that this hypothetical equilibrium frequency does not match the frequency of clade C1 observed in the 10,000-generation sample, when it was roughly half the total population (Figure 1.1A). This discrepancy is not surprising, though, because the competitions involved single representatives from two clades that were continuing to evolve and adapt over time. The important point to take away from the frequency-dependent interaction is this: as clade C2 spread, its advantage over C1 diminished. The frequency dependence thus helps to explain, at least in part, why this selective sweep took so much longer, and involved many more mutations, than other sweeps that both preceded and followed it.

Failed speciation. The clades C1 and C2 diverged ecologically and accumulated different beneficial mutations over several thousand generations, and one can imagine that they might have coexisted indefinitely. Although we do not know the ecological mechanism promoting coexistence of the two clades, some form of ecological divergence is necessary for negative frequency-dependent selection in haploid organisms. One possible mechanism is a demographic tradeoff on glucose, such that one ecotype is a superior competitor when glucose is abundant and the other has an advantage when glucose is scarce; an alternative mechanism is a crossfeeding interaction, whereby one ecotype is a superior competitor for glucose and the other is better at competing for metabolic byproducts (Turner *et al.* 1996; Ribeck and Lenski 2015). It is tempting

to speculate that such ecological divergence might be an example of failed speciation. Defining speciation is often problematic, and especially so for asexual organisms like those in the LTEE. Nonetheless, one interesting way to think about speciation in bacteria posits that it occurs when lineages have diverged to occupy ecological niches that are sufficiently different that selective sweeps can happen in one lineage without driving the other extinct (Cohan 2006; Cohan and Perry 2007). This model fits with the general idea that species are cohesive groups that have diverged irreversibly (Templeton 1989). The importance of reproductive isolation between sexual species is that it locks in the requisite divergence, but asexual organisms can also diverge irreversibly by adapting to different ecological opportunities (Cohan 2002). Of course, any definition that depends on a multistep temporal process is likely to involve intermediate states – grey areas – of ambiguous status. The divergence and eventual fates of the clades C1 and C2 illustrate that ambiguity. In any case, despite the ecological divergence that led to their negative frequency-dependent interaction, the two clades continued to compete and their fates remained entangled. In time, the further adaptation of one clade caused the community of two nascent species to collapse to a mono-specific population. Without the "fossil record" of frozen samples, studies of later generations would reveal nothing of this incipient, but ultimately failed, process of speciation. In contrast to this failed speciation, more persistent ecotypic divergence has evolved in two other LTEE populations, Ara–2 and Ara–3 (Rozen and Lenski 2000; Blount *et al.* 2012).

Conclusions

We tracked the dynamics of several dozen mutations in an *E. coli* population as it evolved in and adapted to a simple laboratory environment for 20,000 generations. The population started from a single haploid cell, and the bacteria in the experiment lack any mechanism for horizontal gene transfer. Glucose was the limiting resource, and its concentration in the culture medium was set low to reduce the cell density and thus, it was thought, simplify the evolutionary dynamics in two respects (Lenski 2004). First, a low population density should reduce the concentration of metabolic byproducts and thereby limit the opportunity for frequency-dependent interactions. Second, a low population size should reduce the rate at which beneficial mutations arise and thereby reduce the impact of clonal interference. With the benefit of hindsight, it is now clear that frequency-dependent interactions and clonal interference were important forces even at this low glucose concentration.

Thus, despite the apparently simple conditions of the LTEE, the dynamics are rich and complex. As expected, we observed many selective sweeps in which derived alleles replace their ancestral counterparts. The speed with which the sweeps occurred is inconsistent with pure genetic drift, but consistent with what is known about the fitness effects of individual mutations in this experiment, given also the limits of detection for the mutations. We also observed, as expected, many cases of clonal interference. We undoubtedly documented far fewer instances of interference than actually occurred, because we tracked only mutations that were known to have either fixed in the population or been present at substantial frequency in a few generational samples that were previously screened for polymorphisms (Barrick and Lenski 2009).

Two other features of the genome dynamics were equally conspicuous but more surprising, especially given the presumed simplicity of the experimental conditions. One such

feature was the drawn out and temporally complex interaction of two clades, which required several thousand generations before one clade eventually excluded the other. This episode reflected a combination of frequency-dependent selection and the rise of new beneficial mutations in both clades that buffeted their relative numbers before one clade finally gained an insurmountable advantage. Previous work has revealed an even longer-lasting coexistence of two clades in another of the LTEE populations (Rozen and Lenski 2000; Le Gac *et al.* 2013). In that population, beneficial mutations also buffeted the abundances of the clades, but the frequency dependence was sufficiently strong that they have continued to coexist for several tens of thousands of generations (Le Gac *et al.* 2013). Our study, along with the work on that other population, shows that frequency-dependent interactions emerged in the LTEE, despite the low resource concentration intended to limit their importance. Our study also shows that such interactions can be transient and might appear, at least superficially, to be cases of very drawn out clonal interference.

The second unexpected feature of the genome dynamics that we observed is the prevalence of cases in which two or more mutations appear to have fixed more or less simultaneously in the population. These cohorts seem to be at odds with the expectation that fixations should be sequential or nested (i.e., one mutation arising in the background that contains the other before it reaches fixation). In a recent study of yeast populations, Lang *et al.* (2013) found similar cohort fixations by deeply sequencing genomes at many time points. However, the explanation for these mutational cohorts remains unclear. We proposed several possibilities, ranging from the conceptually simple (but unlikely) idea that the mutations occurred simultaneously to more complicated scenarios that invoke epistasis and frequency dependence. We think the most parsimonious explanation – the null hypothesis from a

population-genetic perspective – is that these cohorts are an illusion caused by the limited resolution of the actual genome dynamics, at least in the case of small cohorts of two, and perhaps even several, mutations. That is, mutations are only detected in these experiments after they have become fairly common – at least one percent – and by that time they have already increased in frequency by many orders of magnitude. As a consequence, given two beneficial mutations that occurred sequentially – but in sufficiently close temporal proximity – the double mutant may reach a detectable frequency before the first mutation alone or even when the first mutation alone never reaches a detectable frequency (especially given the 500-generation interval between successive samples in our study).

It is not obvious to us whether the different scenarios, including the null hypothesis, can always be distinguished empirically given the limited resolution of population-genomic methods, the temporal gaps between population samples, and the very small fitness effects (including frequency-dependent and epistatic interactions) that are likely to be relevant to any particular case. Instead, we suggest that numerical simulations could provide an important next-step toward resolving the causes of the cohorts. By performing simulations, one can explore what rates of mutations and distributions of fitness effects would give rise to the appearance of cohorts for various allele detection limits and sampling schemes. The inferred rates and distributions could then be compared to corresponding rates and distributions estimated in other ways, such as from population mean-fitness trajectories (Wiser *et al.* 2013), to determine if they are compatible. Whatever the findings from these theoretical studies might be, the results of our study and the rapidly growing body of data from the LTEE and other evolution experiments show the exciting challenges that remain for describing and understanding the dynamics of genomic and phenotypic evolution in microbial populations, even under seemingly simple conditions.

Acknowledgments

We thank Zachary Blount and Noah Ribeck for discussions; Steven Valenziano for help with producing figures; Jeff Landgraf and Cecil Harkey for assistance with the genotyping and pyrosequencing assays; Jeff Morris, Magdalena Felczak, and Louis King for protocols and help with the flow cytometry; Neerja Hajela for assistance in the laboratory; and the editor and reviewers for helpful comments. This work was supported by the National Science Foundation (DEB-1019989 to R.E.L.), the National Institutes of Health (R00-GM087550 to J.E.B.), the NSF BEACON Center for the Study of Evolution in Action (DBI-0939454), and a National Defense Science and Engineering Graduate Fellowship to R.M.

APPENDIX

Figure 1.1: Dynamics of mutant alleles during a long-term evolution experiment with *E. coli*.

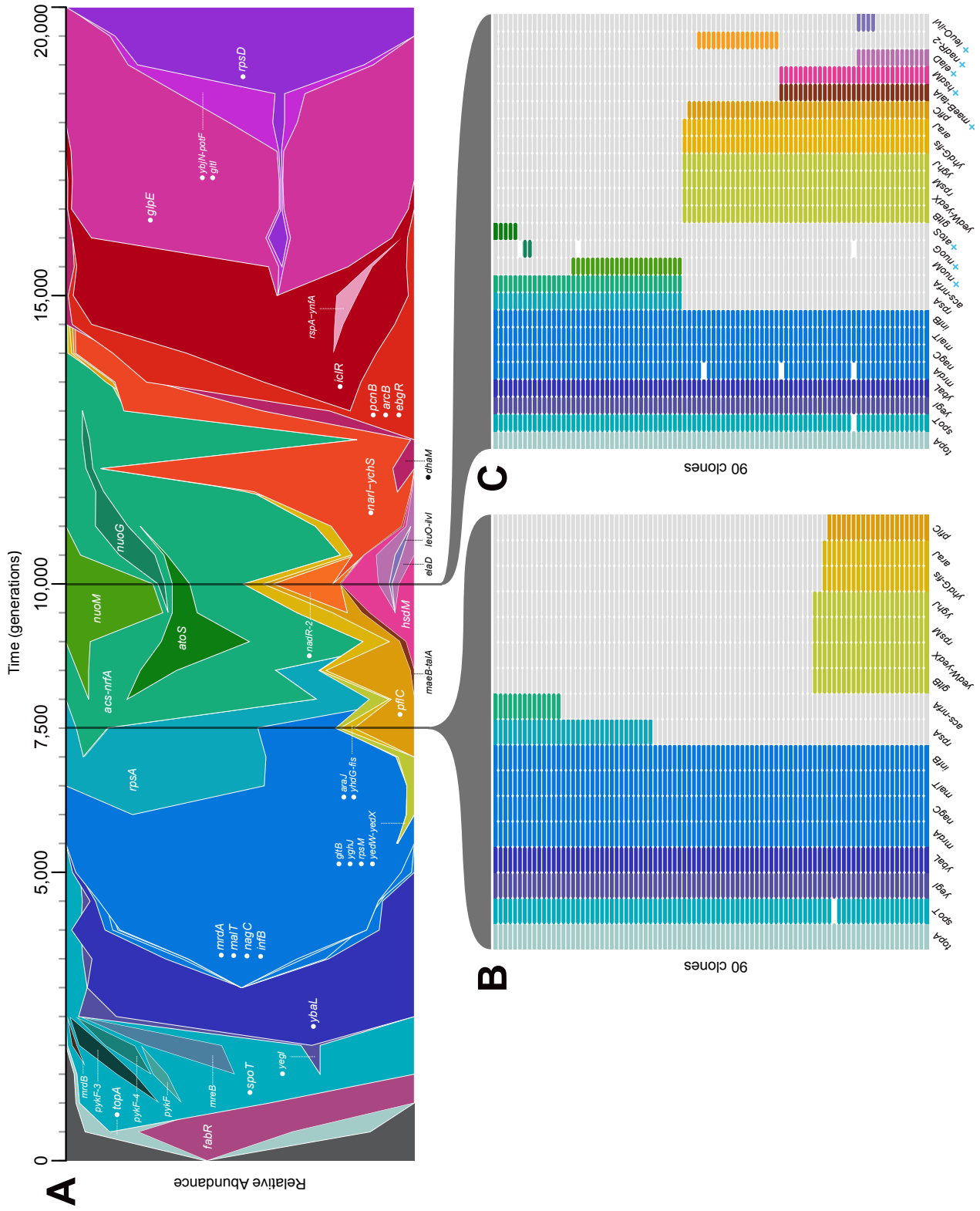


Figure 1.1 (cont'd): (A) The Muller plot shows estimated frequencies of 42 mutations in population Ara–1 over 20,000 generations. Mutations are identified by the gene in which they occurred or, if intergenic, by the adjacent genes; trailing numbers indicate that the same gene had multiple alleles. Labels preceded by a dot indicate the mutation fixed in the population, as indicated by its presence on the line of descent leading to sequenced clones from 30,000 and 40,000 generations. Genotypes of 90 clones sampled at (B) 7500 and (C) 10,000 generations, showing previously fixed and variable alleles only. Each row in each panel represents a clone. Mutations are colored as in (A); light grey fill shows the mutation is not present (i.e., the clone has the ancestral allele), and white fill indicates missing data. Mutations labeled with the blue + symbol were detected in the clones sampled at 10,000 generations but not at 7500 generations.

Figure 1.2: Hypothetical scenario showing the interplay between negative frequency dependence and ongoing beneficial mutations. (A) Null case, in which the fitness of clade C2 relative to clade C1 is independent of their relative frequencies. (B) Static negative frequency dependence, where the fitness of C2 relative to C1 declines as the frequency of C2 in the population increases. (C) Dynamic frequency dependence, where ongoing mutations in each clade provide generic fitness benefits. Although the mutations do not affect the frequency-dependent interaction per se, they affect the location and existence of the stable equilibrium. In the example shown here, the first beneficial mutation (green line) occurs in C1 and lowers the equilibrium frequency of C2; the second beneficial mutation (blue line) occurs in C2 and raises its equilibrium frequency; and a third beneficial mutation (red line) occurs in C1 that eliminates the equilibrium and drives C2 extinct.

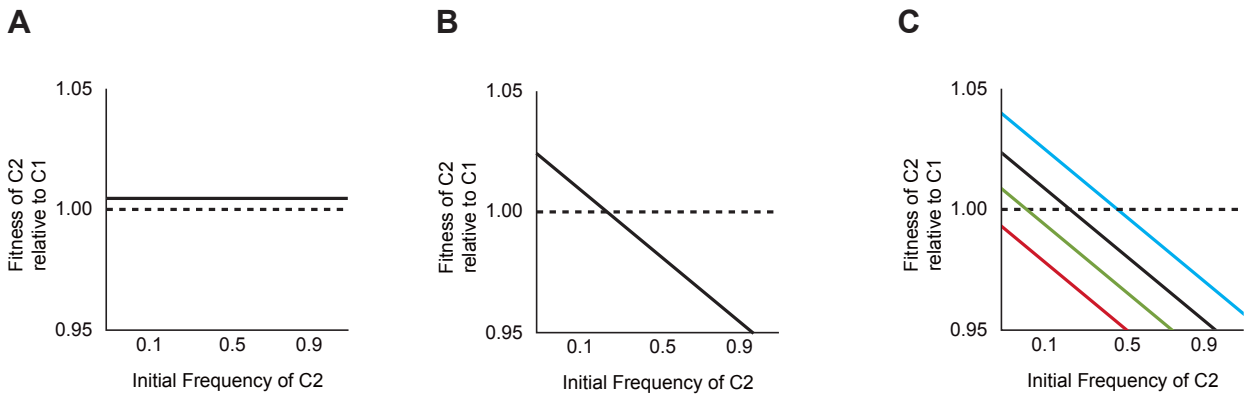
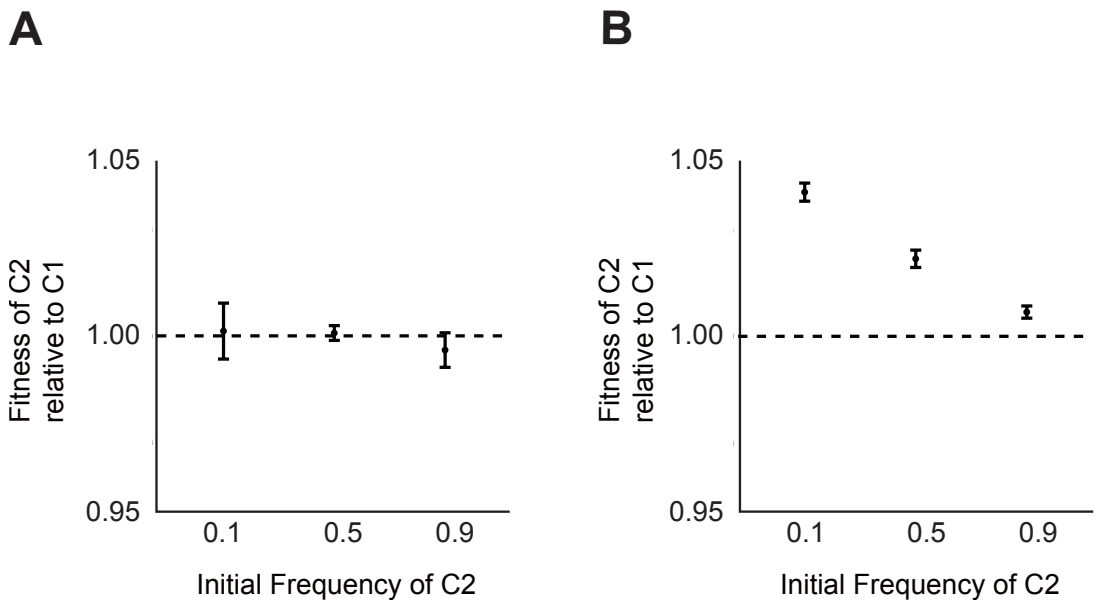


Figure 1.3: Changing nature of interaction between clones from two clades over time. (A) Clonal isolates from clades C1 and C2 sampled from population Ara–1 at 7500 generations show no frequency-dependent interaction. Although they differ by at least 9 mutations, there is no significant difference in their fitness at any of the three initial ratios tested. (B) Clonal isolates from the same clades at 10,000 generations show a strong negative frequency-dependent interaction, with the fitness of the C2 clone declining as its frequency increases. The C2 clone has a significant advantage at all three initial ratios tested, although there may be a stable equilibrium with C2 at a frequency above those tested. Although C2 eventually drove C1 extinct, ongoing beneficial mutations evidently affected the dynamics (Figure 1.1A). Error bars show 95% confidence intervals.



LITERATURE CITED

LITERATURE CITED

Atwood KC, Schneider LK, Ryan FJ. 1951. Periodic selection in *Escherichia coli*. *Proc Natl Acad Sci USA* 37: 146–155.

Barrick JE, Lenski RE. 2009. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Spring Harb Symp Quant Biol* 74: 119–129.

Barrick JE, Lenski RE. 2013. Genome dynamics during experimental evolution. *Nat Rev Genet* 14: 827–839.

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK *et al.* 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461: 1243–1247.

Barroso-Batista J, Sousa A, Lourenço M, Bergman M-L, Demengeot J *et al.* 2014. The first steps of adaptation of *Escherichia coli* to the gut are dominated by soft sweeps. *PLoS Genet*. 10: e1004182.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489: 513–518.

Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli***.** *Proc Natl Acad Sci USA* 105: 7899–7906

Cohan FM. 2002. What are bacterial species? Ann Rev Microbiol 56: 457–487.

Cohan FM. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Phil Trans R Soc B* 361: 1985–1996.

Cohan FM, Perry EB. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* 17: R373–R386.

Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc Natl Acad Sci USA* 100: 1072–1077.

Cooper VS, Schneider D, Blot M, Lenski RE. 2001. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *E. coli* B. *J Bacteriol* 183: 2834–2841.

Crozat E, Philippe N, Lenski RE, Geiselmann J, Schneider D. 2005. Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics* 169: 523–532.

Crozat E, Winkworth C, Gaffé J, Hallin PF, Riley MA *et al.* 2010. Parallel genetic and phenotypic evolution of DNA superhelicity in experimental populations of *Escherichia coli*. *Mol Biol Evol* 27: 2113–2128.

Deatherage DE, Traverse CC, Wolf LN, Barrick JE. 2015. Detecting rare structural variation in evolving microbial populations from new sequence junctions using *breseq*. *Front Genet* 5:468.

Desai MM, Fisher DS. 2007. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* 176: 1759–1798.

de Visser JAGM, Lenski RE. 2002. Long-term experimental evolution in *Escherichia coli*. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evol Biol* 2:19.

de Visser JAGM, Rozen DE. 2006. Clonal interference and the periodic selection of new beneficial mutations in *Escherichia coli*. *Genetics* 172: 2093–2100.

Dykhuizen DE, Hartl DL. 1983. Selection in chemostats. *Microbiol Rev* 47: 150–168.

Elena SF, Lenski RE. 1997. Long-Term Experimental Evolution in *Escherichia coli*. VII. Mechanisms Maintaining Genetic Variability Within Populations. *Evolution* 51: 1058–1067.

Ferullo DJ, Cooper DL, Moore HR, Lovett ST. 2009. Cell cycle synchronization of *Escherichia coli* using the stringent response, with fluorescence labeling assays for DNA content and replication. *Methods* 48: 8–13.

Gerrish PJ, Lenski RE. 1998. The fate of competing beneficial mutations in an asexual population. *Genetica* 102/103: 127–144.

Hegreness M, Shoresh N, Hartl D, Kishony R. 2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* 311: 1615–1617.

Herron MD, Doebeli M. 2013. Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biol* 11: e1001490.

Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* 332: 1193–1196.

Lang GI, Botstein D, Desai MM. 2011. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* 188: 647–661.

Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM *et al.* 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500: 571–574.

Le Gac M, Plucain J, Hindré T, Lenski RE, Schneider D. 2012. Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with *Escherichia coli*. *Proc Natl Acad Sci USA* 109: 9487–9492.

Lee MC, Marx CJ. 2013. Synchronous waves of failed soft sweeps in the laboratory: remarkably rampant clonal interference of alleles at a single locus. *Genetics* 193: 943–952.

Lenski RE. 2004. Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium *Escherichia coli*. *Plant Breeding Rev* 24: 225–265.

Lenski RE, Mongold JA, Sniegowski PD, Travisano M, Vasi F *et al.* 1998. Evolution of competitive fitness in experimental populations of *E. coli*: What makes one genotype a better competitor than another? *Antonie van Leeuwenhoek* 73: 35–47.

Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am Nat* 138: 1315–1341.

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

Levy SF, Blundell JR, Venkataram SV, Petrov DA, Fisher DS, *et al.* 2015. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* 519: 181–186.

Maharjan RP, Liu B, Feng L, Ferenci T, Wang L. 2015. Simple phenotypic sweeps hide complex genetic changes in populations. *Genome Biol Evol* 13: 531–544.

Moxon ER, Rainey PB, Nowak MA, Lenski RE. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol* 4: 24–33.

Park S-C, Krug J. 2013. Rate of adaptation in sexuals and asexuals: a solvable model of the Fisher–Muller effect. *Genetics* 195: 941–955.

Pelosi L, Kühn L, Guetta D, Garin J, Geiselmann J, *et al.* 2006. Parallel changes in global protein profiles during long-term experimental evolution in *Escherichia coli*. *Genetics* 173: 1851–1869.

Plucain J, Hindré T, Le Gac M, Tenaillon O, Cruveiller S *et al.* 2014. Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science* 343: 1366–1369.

Ribeck N, Lenski RE. 2015. Modeling and quantifying frequency-dependent fitness in microbial populations with cross-feeding interactions. *Evolution* doi: 10.1111/evo.12645

Rozen DE, Lenski RE. 2000. Long-term experimental evolution in *Escherichia coli*. VIII. Dynamics of a balanced polymorphism. *Am Nat* 155: 24–35.

Rozen DE, Philippe N, de Visser JA, Lenski RE, Schneider D. 2009. Death and cannibalism in a seasonal environment facilitate bacterial coexistence. *Ecol Lett* 12: 34–44.

Schiffels S, Szöllösi G, Mustonen V, Lässig M. 2011. Emergent neutrality in adaptive asexual evolution. *Genetics* 189: 1361–1375.

Schneider D, Duperchy E, Coursange E, Lenski RE, Blot M. 2000. Long-term experimental evolution in *Escherichia coli*. IX. Characterization of IS-mediated mutations and rearrangements. *Genetics* 156: 477–488.

Sniegowski PD, Gerrish PJ, Lenski RE. 1997. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387: 703–705.

Studier FW, Daegelen P, Lenski RE, Maslov S, Kim JF. 2009. 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.

Templeton AR. 1989. The meaning of species and speciation: a genetic perspective, pp. 3-27 in *Speciation and Its Consequences*, edited by D. Otte and J. A. Endler. Sinauer Associates, Sunderland, MA.

Turner PE, Souza V, Lenski RE. 1996. Tests of ecological mechanisms promoting the stable coexistence of two bacterial genotypes. *Ecology* 77: 2119–2129.

Wielgoss S, Barrick JE, Tenaillon O, Cruvellier S, Chane-Woon-Ming B *et al.* 2011. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3* 1: 183–186.

Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ *et al.* 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci USA* 110: 222–227.

Wiser MJ, Ribeck N, Lenski RE. 2013. Long-term dynamics of adaptation in asexual populations. *Science* 342: 1364–1367.

Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR *et al*. 2011. Second-order selection for evolvability in a large *Escherichia coli* population. *Science* 331: 1433–1436.

Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. 2006. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc Natl Acad Sci USA* 103: 9107–9112.

CHAPTER 2: CORE GENES EVOLVE RAPIDLY IN THE LONG-TERM EVOLUTION EXPERIMENT WITH *ESCHERICHIA COLI*

Authors: Rohan Maddamsetti, Philip J. Hatcher, Barry L. Williams, and Richard E. Lenski

Abstract

Conserved genes evolve slowly in nature, by definition, but we find that some conserved genes are among the fastest-evolving genes in the long-term evolution experiment with *Escherichia coli* (LTEE). We identified the set of almost 2000 core genes shared among sixty clinical, environmental, and laboratory strains of *E. coli*. During the LTEE, these core genes accumulated significantly more nonsynonymous mutations than did flexible (i.e., noncore) genes after accounting for the mutational target size. Furthermore, the core genes under strongest positive selection in the LTEE are more conserved in nature than the average core gene based both on sequence diversity among *E. coli* strains and divergence between *E. coli* and *Salmonella enterica*. We conclude that the conditions of the LTEE are novel for *E. coli*, at least in relation to the long sweep of its evolution in nature. We suggest that what is most novel about the LTEE for the bacteria is the constancy of the environment, its biophysical simplicity, and the absence of microbial competitors, predators, and parasites.

Introduction

By combining experimental evolution and genomic technologies, researchers can study in great detail the genetic underpinnings of adaptation in the laboratory (Barrick and Lenski 2013). However, questions remain about how the genetic basis of adaptation might differ between experimental and natural populations (Bailey and Bataillon 2016).

To explore that issue, we examined whether the genes that evolve most rapidly in the long-term evolution experiment with *Escherichia coli* (LTEE) also evolve and diversify faster than typical genes in nature. If so, the genes involved in adaptation in the LTEE might also be involved in local adaptation to diverse environments in nature. On the other hand, it might be the case that the genes involved in adaptation during the LTEE diversify more slowly in nature than typical genes. Perhaps these genes are highly constrained in nature by purifying selection. For example, they may play important roles in balancing competing metabolic demands or fluctuating selective pressures in the complex and variable natural world, but they can be optimized to fit the simplified and stable conditions of the LTEE.

To test these alternative hypotheses, we compare the signal of positive selection across genes in the LTEE to the sequence diversity in a set of 60 clinical, environmental, and laboratory strains of *E. coli*—henceforth, the "*E. coli* collection"—and to the divergence between *E. coli* and *Salmonella enterica* genomes, respectively. We find that the genes that have evolved the fastest in the LTEE tend to be conserved core genes in the *E. coli* collection. We can exclude recurrent selective sweeps at these loci in nature as an explanation for their limited diversity because the genes and the particular amino-acid residues under positive selection in the LTEE have diverged slowly since the *Escherichia–Salmonella* split.

Materials and Methods

Panortholog Identification in the E. coli *Collection*

We downloaded the nucleotide and amino-acid sequences from GenBank for 60 fully sequenced *E. coli* genome accessions (accessions available upon request). We refer to this diverse set of clinical, environmental, and laboratory strains as the *E. coli* collection. We identified 1968 single-copy orthologous genes, or panorthologs, that are shared by all 60 strains in the *E. coli* collection using the pipeline described in Cooper *et al.* (2010). To guard against recent gene duplication or horizontal transfer events, we confirmed that none of these panorthologs had better local BLAST hits in any given genome. We refer to these panorthologs as core genes, and other genes that are present in only some of the *E. coli* collection as flexible genes.

The NCBI Refseq accession for the ancestor for the LTEE, *E. coli* B strain REL606, is NC_012967. The accession for the *S. enterica* strain used as an outgroup is NC_003197. We downloaded *E. coli* and *S. enterica* orthology information from the OMA orthology database (Altenhoff *et al.* 2015), only examining the one-to-one matches. For internal consistency, we also used the panortholog pipeline to generate one-to-one panorthologs between *E. coli* B strain REL606 and *S. enterica*. We analyzed the 2853 panortholog pairs that the pipeline and the OMA database called identically.

Analysis of the Keio Collection

We downloaded data on essentiality and growth yield in rich and minimal media for the Keio collection of single-gene knockouts in *E. coli* K-12 from the supplementary tables in the original paper describing the collection (Baba *et al.* 2006). We classified the knocked-out genes as

panorthologs (i.e., core) or not (i.e., flexible), and we compared differences in essentiality and growth yield between the two sets of genes.

Nonsynonymous and Synonymous Substitutions in the LTEE at 40,000 Generations

We identified all substitutions in protein-coding genes in the genome sequences of single clones isolated from each of the 12 independently evolving populations of the LTEE at 40,000 generations. These data were reported in two previous studies (Maddamsetti *et al.* 2015; Tenaillon *et al.* 2016). Six of the 12 populations descend from REL606, and six descend from REL607 (Lenski *et al.* 1991). These ancestral strains differ by point mutations in the *araA* and *recD* genes (Tenaillon *et al.* 2016), and the two mutations were thus excluded from our analysis.

G *Scores and Positive Selection on Genes in the LTEE*

We use the *G*-score statistics reported in Supplementary Table 2 of Tenaillon *et al.* (2016) as a measure of positive selection at the gene level in the LTEE. The *G*-score for each gene reflects, in a likelihood framework, the number of independent nonsynonymous mutations in nonmutator lineages relative to the number expected given the length of that gene's coding sequence (relative to all coding sequences) and the total number of such mutations. In this analysis, the nonmutator lineages included the six LTEE populations that never evolved point-mutation hypermutability as well as lineages in the other populations before they became mutators. This analysis included whole-genome sequences from 264 clones isolated at 11 time points through 50,000 generations of the LTEE.

Sequence Diversity and Divergence

We adapted Nei's nucleotide diversity metric (Nei and Li 1979) for use with amino-acid sequences to reflect nonsynonymous differences. Specifically, we calculated the mean number of differences per site between all 1770 (i.e., 60 x 59 / 2) pairs of sequences in the protein alignments from the 60 genomes in the *E. coli* collection. In the site-specific analysis, we calculated this diversity metric separately for the sites that evolved in the LTEE and those that did not, and we compared the values to see if the former also tended to vary in nature. For the sequence divergence between *E. coli* and *S. enterica*, we used the ancestral strain of the LTEE, REL606, as the representative *E. coli* genome in order to maximize the number of orthologous genes available in our analysis. The divergence for each gene was calculated as the proportion of amino-acid residues that differ between the two aligned proteins, where an amino-acid difference implies at least one nonsynonymous change in the corresponding codon since the most recent common ancestor of the two alleles.

Statistical Analyses

All data tables and analysis scripts will be deposited in the Dryad Digital Repository upon acceptance.

Results

Core Genes Are Functionally Important

To make consistent comparisons, we analyzed single-copy genes with homologs in all 60 fully sequenced genomes in the *E. coli* collection. For the purpose of our study, we define this set of

panorthologous genes as the *E. coli* core genome and the set of all other genes as the flexible genome (Materials and Methods). We used published data from the Keio collection of single-gene knockouts in *E. coli* K-12 (Baba *et al.* 2006) to test whether the core genes tend to be functionally more important than the flexible genes based on essentiality and growth yield. Core genes are indeed more essential than flexible genes (Welch's $t = 6.60$, d.f. = 3387.8, one-tailed $p < 10^{-10}$), and knockouts of core genes cause larger growth-yield defects than flexible genes in both rich (Welch's $t = 3.79$, d.f. = 3379, one-tailed $p < 0.0001$) and minimal media (Welch's $t = 4.95$, d.f. =3457.3, one-tailed $p < 10^{-6}$).

Core Genes Evolve Faster than Flexible Genes in the LTEE

We first examined the substitutions in single genomes sampled from each of the 12 LTEE populations after 40,000 generations. These genomes are part of a large dataset comprising 264 fully sequenced genomes from the first 50,000 generations of the LTEE (Tenaillon *et al.* 2016). Six of these populations had evolved greatly elevated point-mutation rates by 40,000 generations (Sniegowski *et al.* 1997; Wielgoss *et al.* 2013; Tenaillon *et al.* 2016). As a consequence of their much higher mutation rates, a much larger fraction of the mutations seen in hypermutable populations are expected to be neutral or even deleterious passengers (hitchhikers), as opposed to beneficial drivers, in comparison to those populations that retained the low ancestral point-mutation rate (Tenaillon *et al.* 2016). In genomes from the nonmutator populations, we observe an excess of nonsynonymous substitutions in the core genes. The core genes constitute ~48.5% of the total coding sequence in the genome of the LTEE ancestral strain, but 69% (105/152) of the nonsynonymous substitutions are in the core genes (Table 2.1, row 1, $p < 10^{-6}$). By contrast, the frequency of synonymous mutations does not differ significantly between the core and

flexible genes (Table 2.1, row 2). Also, the frequencies of both nonsynonymous and synonymous substitutions in core versus flexible genes are close to the null expectations in the populations that evolved hypermutability (Table 2.1, rows 3 and 4).

These results indicate that core genes are evolving faster, on average, than the flexible noncore genome in the LTEE populations that retained the ancestral point-mutation rate. This faster evolution is consistent with some subset of the core genes being under positive selection to change from their ancestral state during the LTEE. To examine this issue more closely, we compared the rates of evolution of core genes observed in the LTEE with the rates of evolution of the same genes in the *E. coli* collection. As a measure of the rate of evolution of each core gene in the LTEE, we used a *G*-score, as calculated by Tenaillon *et al.* (2016), that expresses the excess number of independent nonsynonymous mutations in nonhypermutable lineages relative to the number expected given the length of that gene's coding sequence (relative to all coding sequences) and the total number of such mutations.

We used two different measures for the rate of evolution of each core gene in nature. The first one is based on the level of nonsynonymous sequence diversity in the gene across the 60 sequenced genomes in the *E. coli* collection. There is a negative correlation between a core gene's *G*-score in the LTEE and its diversity in in the *E. coli* collection (Spearman-rank correlation $r = –0.0701$, two-tailed $p = 0.0019$; Figure 2.1A). That is, core genes that evolved faster in the LTEE (i.e., higher *G*-scores) are significantly less diverse in the *E. coli* collection than those that evolved more slowly. Only 163 genes in the core genome had positive *G*-scores (i.e., one or more nonsynonymous mutations in nonhypermutable lineages) in the LTEE, and we do not find a significant correlation between the *G*-score and sequence diversity using only those genes (Spearman-rank correlation $r = –0.0470$, two-tailed $p = 0.5515$; Figure 2.1B). However,

the 163 core genes with positive *G*-scores have significantly lower diversity in the *E. coli* collection than do the 1805 with zero *G*-scores (Mann-Whitney U = 125,660, two-tailed p = 0.0020; Figure 2.1C). Hence, the difference between core genes with and without nonsynonymous substitutions in the nonmutator LTEE lineages largely drives the overall negative correlation.

By using segregating polymorphisms in the *E. coli* collection, our first measure of the rate of evolution of core genes in nature might be dominated by transient variation or local adaptation. By contrast, divergence between core genes found in different species has occurred over a longer timescale and should be less affected by these issues. Therefore, our second measure for the rate of evolution of core genes in nature uses the sequence divergence between *E. coli* and *Salmonella enterica*. We repeated the above analyses using the set of 2853 panorthologs—single-copy genes that map one-to-one across species (Lerat *et al.* 2003; Cooper *et al.* 2010)—for *E. coli* and *S. enterica*. We found a negative correlation across genes between their *G*-scores in the LTEE and interspecific divergence (Spearman rank-correlation $r = -0.0911$, two-tailed $p < 10^{-5}$; Figure 2.2A). This negative correlation remains significant even if we consider only those 210 panorthologs with positive *G*-scores in the LTEE (Spearman rank-correlation $r = -0.2567$, two-tailed $p = 0.0002$; Figure 2.2B). In addition, the panorthologs with positive *G*-scores are less diverged between *E. coli* and *S. enterica* than the 2643 panorthologs with zero *G*-scores (Mann-Whitney U = 223,330, two-tailed $p < 10^{-5}$; Figure 2.2C).

Taken together, these analyses contradict the hypothesis that those genes that have evolved fastest in the LTEE are ones that also evolve and diversify faster than typical genes in nature. Instead, they support the hypothesis that the genes involved in adaptation during the LTEE tend to be more conserved than typical genes in nature, presumably because they are

constrained in nature by purifying selection. When the bacteria evolve under the simple and stable ecological conditions of the LTEE, these previously conserved genes undergo adaptive evolution that fits them to their new environment.

Protein Residues that Changed in the LTEE are also Conserved in Nature

It is possible that the substitutions in the LTEE occurred at highly variable sites in otherwise conserved proteins. To examine this issue, we asked whether nonsynonymous changes found in the nonmutator LTEE lineages at 40,000 generations tended to occur in fast-evolving codons. For the 66 proteins with such substitutions in the LTEE, we calculated the diversity at the mutated sites and in the rest of the protein for the 60 genomes in the *E. coli* collection. The sites that had changed in the LTEE were significantly less variable than the rest of the protein in that collection (Wilcoxon signed-rank test, $p < 10^{-5}$). In fact, only 7 of these 66 proteins had any variability at those sites in the *E. coli* collection, and they account for only 9 of the 105 amino-acid substitutions in those proteins in the 40,000-generation LTEE clones. We obtained similar results for the divergence between *E. coli* and *Salmonella.* In the 40,000-generation LTEE clones, 128 nonsynonymous substitutions occurred in 86 panorthologs, and only 5 of the LTEE substitutions were at diverged sites. These results demonstrate that particular residues under positive selection in the LTEE are ones that tend to be conserved in nature.

Discussion

It has been long known that, in nature, some genes evolve faster than others. In most cases, the more slowly evolving genes are core genes—ones possessed by most or all members of some

species or higher taxon—and their relative sequence conservation reflects functional constraints that limit the potential for the encoded proteins to change while retaining their functionality. As a consequence, the ratio of nonsynonymous to synonymous substitutions also tends to be low in these core genes. By contrast, we found that most nonsynonymous substitutions in nonmutator lineages of the LTEE occurred in core genes that are shared by all *E. coli* (Table 2.1). Moreover, even among the core genes, those that experienced positive selection to change in the LTEE are both less diverse over the *E. coli* species (Figure 2.1) and less diverged between *E. coli* and *S. enterica* (Figure 2.2) than core genes without substitutions in any of the nonmutator LTEE populations. Also, the particular sites where substitutions occurred during the LTEE are usually more conserved than the rest of the corresponding protein, excluding the possibility that substitutions occurred at a subset of fast-evolving positions in otherwise slow-evolving genes.

It is clear, then, that the specific conditions of the LTEE have favored new alleles in core genes that are usually highly conserved in nature. From one perspective, this result is surprising—the 37°C temperature of the LTEE is typical for the human and many other mammalian bodies in which *E. coli* lives; the limiting resource is glucose, which is *E. coli*'s preferred energy source, such that it will repress the expression of genes used to catabolize other resources when glucose is available; and the LTEE does not impose other stressors such as pH, antibiotics, or the like. However, the very simplicity and constancy of the LTEE conditions are presumably novel, or at least atypical, in the long sweep of *E. coli* evolution. In other words, the uniformity and simplicity of the laboratory conditions—including the absence of microbial competitors and parasites as well as host-dependent factors—stand in stark contrast to the variable and complex communities that are *E. coli*'s natural habitat (Blount 2015).

Given the importance and even essentiality of many core genes, it seems unlikely that most of the nonsynonymous mutations in the LTEE cause complete losses of function. Instead, we suspect that the mutations are beneficial because they fine-tune the regulation and expression of functions that contribute to the bacteria's competitiveness and growth in the simple and predictable environment of the LTEE. By contrast, some other genes that were repeatedly mutated in the LTEE—not by point mutations, but instead by deletions and transposable-element insertions—typically encode noncore, nonessential functions including prophage remnants, plasmid-derived toxin-antitoxin modules, and production of extracellular structure that are probably important for host colonization (Tenaillon et al. 2016). Both types of change have been shown to be adaptive in the LTEE environment—the former by affecting a gene's function and the expression of interacting genes (Cooper et al. 2003, Philippe et al. 2007), and the latter by eliminating unused and potentially costly functions (Cooper et al. 2001).

Of course, other evolution experiments would generate different types of genomic changes, including in some cases probably a preponderance of point mutations in noncore genes. For example, if the experimental environment involves lethal agents such as phages or antibiotics, then perhaps only a few noncore genes might be the targets of selection, and the resulting mutations might be different from and even at odds with adaptation to other aspects of the environment (Scanlan et al. 2015). Similarly, adaptation to exploit novel resources—such as the ability to use the citrate that has been present throughout the LTEE, but which only one population has discovered how to use (Blount et al. 2008, Blount et al. 2012)—may produce a different genetic signature of adaptation. Yet other signatures might emerge if horizontal gene transfer from other strains or species provided another source of variation (Souza et al. 1997). Imagine a scenario in which gene flow allowed *E. coli* to obtain DNA from a diverse natural

community; in that case, a transporter acquired from another bacterial species might well provide an easier pathway to use the citrate in the LTEE environment.

We can turn the question around from asking why core genes evolve so quickly in the LTEE, to asking why they usually evolve slowly in nature. Core genes encode functions that, by definition, are widely shared, and so their sequences have had substantial time to diverge across taxa (Biller *et al.* 2015) and become fine-tuned to different niches. As a consequence, there are fewer opportunities for new alleles of core genes to provide an advantage. Moreover, given the diversity of species (including transients) in most natural communities, extant species may usually fill any vacant niches that appear faster than *de novo* evolution. Nonetheless, mutations in conserved core genes might sometimes provide the best available paths for adaptation to new conditions, such as when formerly free-living or commensal bacteria become pathogens (Lieberman et al. 2011). In such cases, finding parallel or convergent changes offers a way to identify adaptive mutations when they occur in core genes. For example, *E. coli* and *S. enterica* have been found to undergo convergent changes at the amino-acid level in core genes when strains evolve pathogenic lifestyles (Chattopadhyay *et al.* 2009; Chattopadhyay *et al.* 2012).

In summary, the genetic signatures of adaptation vary depending on circumstances including the novelty of the environment from the perspective of the evolving population, the complexity of the biological community in which the population exists, the intensity of selection, and the number and types of genes that can produce useful phenotypes. In the LTEE, nonsynonymous mutations in core genes that encode conserved and even essential functions for *E. coli* have provided a major source of the large fitness gains in the evolving populations over many thousands of generations (Wiser et al. 2013, Lenski et al. 2015).

Acknowledgments

We thank Alita Burmeister, Michael Wiser, and Kyle Card for discussions and comments on earlier versions of our manuscript. This work was supported, in part, by a National Defense Science and Engineering Graduate Fellowship to R.M.; a grant from the National Science Foundation (DEB-1451740) to R.E.L.; and the BEACON Center for the Study of Evolution in Action (National Science Foundation Cooperative Agreement DBI-0939454).

APPENDIX

Table 2.1: Nonsynonymous mutations are overrepresented in the core genome of nonmutator LTEE populations.

Category and Population	Core	Flexible	Odds Ratio	Significance
Nonsynonymous mutations in nonmutator populations	105	47	2.37	$p < 10^{-6}$
Synonymous mutations in nonmutator populations	10	15	0.71	$p = 0.4297$
Nonsynonymous mutations in mutator populations	2038	2247	0.96	$p = 0.2273$
Synonymous mutations in mutator populations	845	880	1.02	$p = 0.6822$

NOTE—The length of the core and flexible (i.e., noncore) portions of the coding sequences in the genome of the LTEE ancestor (*E. coli* strain REL606) are 1,944,921 and 2,066,263 bp, respectively. Data show the numbers of mutations found in the core and flexible portions in genomes sampled and sequenced at 40,000 generations from six nonmutator populations that retained the ancestral point-mutation rate and six mutator populations that evolved hypermutability. The odds ratio expresses the extent to which the category of mutation is overrepresented (>1) or underrepresented (<1) in the core genome relative to the flexible genome in the indicated populations. The *p*-value is based on a two-tailed binomial test comparing the observed numbers of mutations to the expectations based on the relative lengths of the core and flexible genomes.

Figure 2.1: Relationship between positive selection in the LTEE and nonsynonymous sequence diversity of core genes in the *E. coli* collection of 60 clinical, environmental, and laboratory strains. The *G*-score provides a measure of positive selection based on the excess of nonsynonymous substitutions in the LTEE lineages that retained the ancestral point-mutation rate. The $\log_{10}$ and square-root transformations of the *G*-score and sequence diversity, respectively, improve visual dispersion of the data for individual genes, but they do not affect the nonparametric tests performed, which depend only on rank order. (A) *G*-scores and sequence diversity are negatively correlated across all 1968 core genes (Spearman-rank correlation, p = 0.0019). (B) The correlation becomes not significant using only the 163 genes with positive *G*-scores (Spearman-rank correlation, p = 0.5515). (C) The 163 core genes with positive *G*-scores in the LTEE have significantly lower nonsynonymous sequence diversity in natural isolates than the 1805 genes with zero *G*-scores (Mann-Whitney test, p = 0.0020). Error bars show 95% confidence intervals around the median.

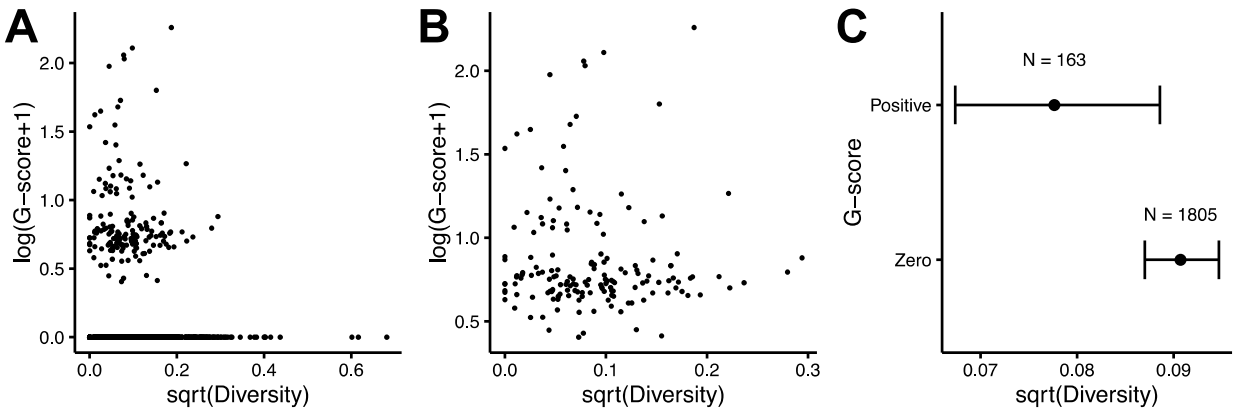
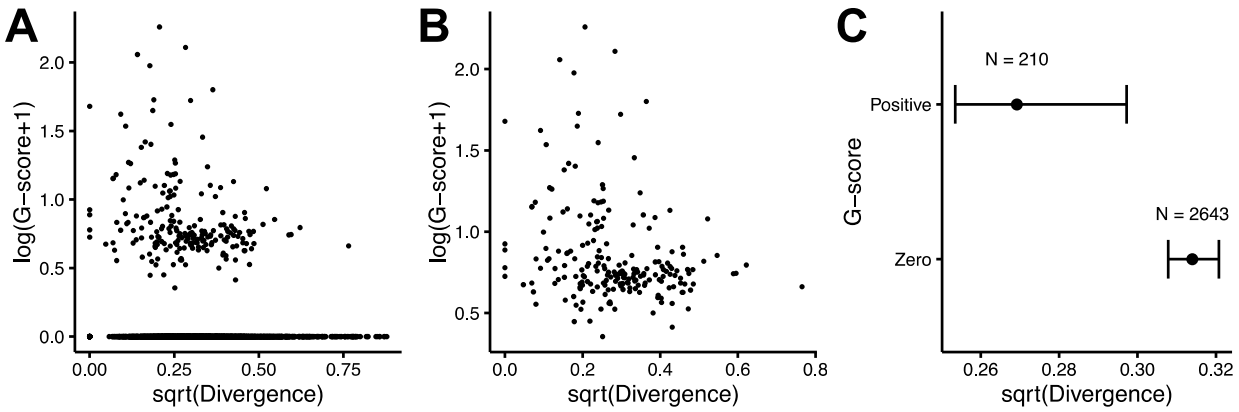


Figure 2.2: Relationship between positive selection in the LTEE and nonsynonymous sequence divergence of panorthologs between *E. coli* (strain REL606) and *S. enterica*. REL606 is the common ancestor of the LTEE populations. See Figure 2.1 for additional details. (A) *G*-scores and divergence are negatively correlated across all 2853 panorthologs (Spearman-rank correlation, $p < 10^{-5}$). (B) The correlation remains significant even using only the 210 panorthologs with positive *G*-scores (Spearman-rank correlation, $p = 0.0002$). (C) The 210 panorthologs with positive *G*-scores in the LTEE are significantly less diverged between *E. coli* and *S. enterica* in natural isolates than the 2643 panorthologs with zero *G*-scores (Mann-Whitney test, $p = < 10^{-5}$). Error bars show 95% confidence intervals around the median.



LITERATURE CITED

LITERATURE CITED

Altenhoff AM, Škunca N, Glover N, Train CM, Sueki A, Piližota I, Gori K, Tomiczek B, Müller S, Redestig H, *et al*. 2015. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* 43:D240–249.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol.* 2:2006.0008.

Bailey SF, Bataillon T. 2016. Can the experimental evolution program help us elucidate the genetic basis of adaptation in nature? *Mol Ecol.* 25:203–218.

Barrick JE, Lenski RE. 2013. Genome dynamics during experimental evolution. *Nat Rev Genet.* 14:827–839.

Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. *Prochlorococcus:* the structure and function of collective diversity. *Nat Rev Microbiol.* 13:13–27.

Blount ZD. 2015. The unexhausted potential of *E. coli eLife* 4:e05826.

Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV. 2009. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc Natl Acad Sci U S A.* 106:12412–12417.

Chattopadhyay S, Paul S, Kisiela DI, Linardopoulou EV, Sokurenko EV. 2012. Convergent molecular evolution of genomic cores in *Salmonella enterica* and *Escherichia coli*. *J Bacteriol.* 194:5002–5011.

Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 100:1072–1077.

Cooper VS, Schneider D, Blot M, Lenski RE. 2001. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *E. coli* B. *J Bacteriol.* 183:2834–2841.

Cooper VS, Vohr SH, Wrocklage SC, Hatcher PJ. 2010. Why genes evolve faster on secondary chromosomes in bacteria. *PLoS Comput Biol.* 6:e1000732.

Lenski RE, Rose MR, Simpson SC, Tadler SC. 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am Nat.* 138:1315–1341.

Lenski RE, Wiser MJ, Ribeck N, Blount ZD, Nahum JR, Morris JJ, Zaman L, Turner CB, Wade BD, Maddamsetti R, Burmeister AR, Baird EJ, Bundy J, Grant NA, Card KJ, Rowles M, Weatherspoon K, Papoulis SE, Sullivan R, Clark C, Mulka JS, Hajela N. 2015. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. *Proc R Soc Lond B.* 282:20152292.

Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the γ-Proteobacteria. *PLoS Biol.* 1: e19.

Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Skurnik D, Leiby N, LiPuma JJ, Goldberg JB, McAdam AJ, Priebe GP, Kishony R. 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Gen.* 43:1275–1280.

Maddamsetti R, Hatcher PJ, Cruveiller S, Médigue C, Barrick JE, Lenski RE. 2015. Synonymous genetic variation in natural isolates of *Escherichia coli* does not predict where synonymous mutations occur in a long-term evolution experiment with *Escherichia coli*. *Mol Biol Evol.* doi:10.1093/molbev/msv161.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 76:5269–5273.

Philippe N, Crozat E, Lenski RE, Schneider D. 2007. Evolution of global regulatory networks during a long-term experiment with *Escherichia coli*. *BioEssays* 29:846–860.

Scanlan PD, Hall AR, Blackshields G, Friman VP, Davis MR Jr, Goldberg JB, Buckling A. 2015. Coevolution with bacteriophages drives genome-wide host evolution and constrains the acquisition of abiotic-beneficial mutations. *Mol Biol Evol.* 32:1425–1435.

Sniegowski PD, Gerrish PJ, Lenski RE. 1997. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387:703–705.

Souza V, Turner PE, Lenski RE. 1997. Long-term experimental evolution in *Escherichia coli*. V. Effects of recombination with immigrant genotypes on the rate of bacterial evolution. *J Evol Biol.* 10:743–769.

Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Médigue C, *et al.* 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. bioRXiv preprint http://dx.doi.org/10.1101/036806

Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D. 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A.* 110:222–227.

Wiser MJ, Ribeck N, Lenski RE. 2013. Long-term dynamics of adaptation in asexual populations. *Science* 342:1364–1367.

CHAPTER 3: SYNONYMOUS GENETIC VARIATION IN NATURAL ISOLATES OF *ESCHERICHIA COLI* DOES NOT PREDICT WHERE SYNONYMOUS SUBSTITUTIONS OCCUR IN A LONG-TERM EXPERIMENT

Authors: Rohan Maddamsetti, Philip J. Hatcher, Stéphane Cruveiller, Claudine Médigue, Jeffrey E. Barrick, and Richard E. Lenski

Originally published in the journal *Molecular Biology and Evolution*, 32: 2897–2904.

Abstract

Synonymous genetic differences vary by more than 20-fold among genes in natural isolates of *Escherichia coli*. One hypothesis to explain this heterogeneity is that genes with high levels of synonymous variation mutate at higher rates than genes with low synonymous variation. If so, then one would expect to observe similar mutational patterns in evolution experiments. In fact, however, the pattern of synonymous substitutions in a long-term evolution experiment with *E. coli* does not support this hypothesis. In particular, the extent of synonymous variation across genes in that experiment does not reflect the variation observed in natural isolates of *E. coli*. Instead, gene length alone predicts with high accuracy the prevalence of synonymous changes in the experimental populations. We hypothesize that patterns of synonymous variation in natural *E. coli* populations are instead caused by differences across genomic regions in their effective population size that, in turn, reflect different histories of recombination, horizontal gene transfer, selection, and population structure.

Introduction

According to the neutral theory of molecular evolution, mutation and random genetic drift are largely responsible for shaping the patterns of genetic variation in nature (Kimura 1968). The generality of the empirical predictions of this theory remains contentious (Hahn 2008), but it does provide a useful quantitative framework for generating falsifiable hypotheses. One of the central predictions of neutral theory is that synonymous variation in protein-coding sequences should reflect the underlying mutation rate and the time passed as populations diverge.

Based on patterns of synonymous variation across the genomes of diverse *Escherichia coli* isolates, Martincorena *et al*. (2012) hypothesized that natural selection has optimized local mutation rates such that physiologically important, highly expressed genes that experience strong purifying selection mutate at lower rates than less important, lowly expressed genes. They used synonymous nucleotide diversity, θ_s, to estimate the mutation rate for each gene. The expected number of neutral mutations in a given lineage is $t\mu$, where t is time in generations and μ is the mutation rate over the relevant genomic sites. The expected divergence time (that is, the time to coalescence) between two lineages is N_e generations. Therefore, the expected number of neutral mutations separating two sampled genomes is $2N_e\mu$ (Figure 3.1). In their analysis, Martincorena *et al*. (2012) implicitly assumed that all of the genes in the core genome of *E. coli* have experienced the same coalescence time and effective population size, so that any significant variation among those genes in the quantity θ_s was attributed to differences in gene-specific mutation rates. They excluded non-core genes from their analysis owing to the recognition that such genes probably have different coalescence times as the result of horizontal transfer between species.

The hypothesis of local optimization of mutation rates comes from Martincorena *et al.* (2012), but their empirical findings of non-random patterns in synonymous substitutions find precedence in earlier studies. Comparing *E. coli* and *Salmonella typhimurium* gene sequences and controlling for gene expression, Sharp *et al.* (1989) found a significant relationship between synonymous divergence and distance from *oriC*, the chromosomal origin of replication. They proposed that genes farther from *oriC* tend to experience higher mutation rates than those closer to *oriC* because genes closer to *oriC* have higher copy numbers in growing cells and therefore more opportunity for recombination-based repair. Eyre-Walker (1994) reported that synonymous substitutions tend to be clustered in genomes, and he proposed several possible explanations: template-based mutational events that can introduce multiple base-pair changes, interspecific recombination, and selection acting on the secondary structure of nucleic acids.

As in many bacterial species, gene content varies substantially among *E. coli* strains. In a sample of 20 *E. coli* genomes, ~18,000 different genes were found in at least one strain, whereas only ~2000 were found in all 20 strains (Touchon *et al*. 2009). The latter set of genes forms the core genome of *E. coli*, and the synonymous variation in that core genome is the subject of the study by Martincorena *et al*. (2012) as well as our own.

If the point-mutation rate varies across the *E. coli* genome, and in particular if mutations are selectively neutral, then we should see neutral mutations accumulate at different rates across different genes in evolution experiments. On the timescale of experiments with large asexual populations that begin without any standing variation, increases in the frequency of synonymous mutations should occur almost entirely when they hitchhike with beneficial driver mutations. Because beneficial mutations will also drive neutral mutations in other backgrounds extinct, the net effect is a wash – that is, comparing two asexual genomes separated by *t* generations, there

will have been *t* opportunities for any given neutral mutation to occur, regardless of whether other mutations were under selection. Therefore, the expected rate of accumulation of neutral mutations should reflect their underlying mutation rate. Not all synonymous changes are perfectly neutral, but any fitness effects they have—even if beneficial—are generally far too small for these mutations to increase in frequency on their own over the course of even the longest experiment (Wielgoss *et al.* 2011). In fact, even strongly beneficial mutations rarely fix alone during experiments with large asexual populations owing to a phenomenon called clonal interference (Lang *et al.* 2013b; Maddamsetti *et al.* 2015). Clonal interference occurs because, in the absence of recombination, beneficial mutations that arise in different lineages compete with one another, thus slowing the progress of each one toward fixation (Gerrish and Lenski 1998; Barrick and Lenski 2013). As a consequence, only the most highly beneficial mutations can drive selective sweeps in the clonal interference regime (Levy *et al.* 2015), and secondary beneficial mutations that arise after the contending lineages reach high frequencies often determine which lineage ultimately prevails (Woods *et al.* 2011; Maddamsetti *et al.* 2015). Most synonymous mutations, even if they are not perfectly neutral, should have very small selection coefficients; as a consequence, they will have a negligible effect on the fixation probability of lineages that also have beneficial mutations with much larger fitness effects. Therefore, the rate of accumulation of synonymous substitutions—whether they are strictly neutral or not—provides a reasonable proxy for the point-mutation rate in evolution experiments.

From mutation-accumulation experiments and fluctuation tests, it is clear that both the rate and spectrum of spontaneous mutations vary across the tree of life (Luria and Delbrück 1943; Keightley *et al.* 2009; Ossowski *et al.* 2010; Lee *et al*. 2012; Sung *et al.* 2012; Ford *et al.* 2013). Multiple studies have reported variation in the mutation rate depending on chromosomal

location, local sequence context, and chromatin structure (Sharp *et al.* 1989; Lang and Murray 2008; Lang and Murray 2011; Warnecke *et al.* 2012; Foster *et al.* 2013; Long *et al.* 2015). Also, the process of transcription has been reported to be mutagenic (Kim and Jinks-Robertson 2009; Park *et al.* 2012; Paul *et al.* 2013).

Other patterns where substitution rates vary with chromosomal location have been seen elsewhere. In *Burkholderia* and *Vibrio* species that have primary and secondary chromosomes, genes on the secondary chromosome have higher rates of both nonsynonymous and synonymous substitutions (Cooper *et al.* 2010). This variation appears to indicate that fast-evolving genes have disproportionately migrated to the secondary chromosome. This finding also raises the possibility that selection has operated on the rate and spectrum of mutations in such a way that more important genes mutate less often.

Natural selection can also shape synonymous variation directly. For example, purifying selection on synonymous sites has been seen in *Drosophila melanogaster* (Lawrie *et al.* 2013), and synonymous substitutions that are beneficial because they increase gene expression have been reported in an evolution experiment with *Pseudomonas fluorescens* (Bailey *et al.* 2014). Natural selection also affects codon usage, and recoding a functionally important gene via synonymous changes can affect fitness (Agashe *et al.* 2013). Other more indirect evidence also implicates selection as an important force shaping synonymous variation in bacteria. Although mutation is universally biased towards increased AT-content in bacteria, genomic GC-content varies among species from less than 20% to more than 70%. GC-content at synonymous sites strongly correlates with genomic GC-content; the fact that genome composition is inconsistent with the mutational bias suggests that selection has acted in opposition to the mutational bias

even at synonymous sites (Hershberg and Petrov 2010; Hildebrand *et al.* 2010; Rocha and Feil 2010).

In this paper, we confirm the finding by Martincorena *et al*. (2012) that synonymous nucleotide diversity varies by more than an order of magnitude across the core genome of *E. coli*. In other words, some gene trees have much longer branches, on average, than other gene trees, even in the core genome. This result means that different genes give different estimates of when *E. coli* isolates diverged from each other, assuming that mutation rates do not vary across the genome. However, Martincorena *et al.* argued instead that this heterogeneity is caused by local genomic differences in the mutation rate. If their hypothesis were correct, then we would expect such mutation-rate heterogeneity to occur and be evident in the long-term evolution experiment (LTEE) with *E. coli* that has been running for more than 25 years (Lenski and Travisano 1994, Wiser *et al.* 2013). To test that prediction, one must focus on the effects of mutation rate rather than natural selection. To that end, we count the number of synonymous substitutions that have accumulated in almost 3,000 genes after 40,000 generations in clones (i.e., individuals) from 12 replicate populations, while also controlling for gene length. Most of the synonymous mutations occurred in populations that evolved hypermutator phenotypes owing to defects in DNA repair (Sniegowski *et al.* 1997; Wielgoss *et al.* 2013). However, we will show that the base substitution signatures of different types of hypermutability do not affect our results.

In brief, we find no evidence from these experimental populations that those core genes with low synonymous nucleotide diversity in nature have lower mutation rates than those with high synonymous nucleotide diversity. Instead, we find a close correspondence between the number of synonymous substitutions in different genes and the length of those genes, consistent with the null hypothesis of a point-mutation rate that is homogeneous across the genome. We

also find a weak, positive relationship between a gene's level of expression and its rate of synonymous substitution, but this relationship is not significant when controlling for gene length; that is, longer genes tend to have slightly higher gene expression levels but also more sites at risk for mutation.

Results

We identified a total of 1,069 synonymous substitutions in the core genome (described in the Materials and Methods) of clones sampled from 12 independently evolved populations after 40,000 generations of the LTEE (Lenski and Travisano 1994, Wiser *et al.* 2013). To control for variation in gene length, we compared the observed cumulative distribution of synonymous substitutions across this gene set with the distribution expected under the null hypothesis of a uniform point-mutation rate (Figure 3.2). Despite a large number of events, there is no significant difference between the observed distribution and the null hypothesis (Kolmogorov-Smirnov test, $D = 0.0281$, $P = 0.21$). In broad terms, therefore, the accumulation of synonymous mutations in the LTEE is consistent with a uniform rate of point mutation across the *E. coli* genome.

Under the alternative hypothesis, the variation among genes in the quantity θ_s reflects differences in their underlying mutation rates. In that case, we would expect μ—and thus the distribution of synonymous mutations in the evolution experiment—to be directly proportional to θ_s. However, the difference between that expectation and the distribution of synonymous mutations observed in the evolution experiment is extremely significant (Figure 3.2; Kolmogorov-Smirnov test, $D = 0.244$, $P < 10^{-15}$). Importantly, this difference holds when clones

from the four mismatch-repair (*mutS* or *mutL*) and two base-excision repair (*mutT*) hypermutator lineages are analyzed separately ($P < 10^{-15}$ and $P < 10^{-8}$, respectively). Hence, rejection of this hypothesis does not depend on the particular mutational signature of one or the other class of hypermutator (Figure 3.3). None of these results change when we use the θ_s estimates from Martincorena *et al*. (2012) instead of our own estimates. Therefore, the data from the LTEE do not support the hypothesis of Martincorena *et al.* (2012) that the mutation rate has been locally optimized. Instead, the point-mutation rate is remarkably uniform across the core genome (Figure 3.2). Of course, we cannot prove that such uniformity would persist if we had equally large samples of synonymous changes from non-mutator populations. It is possible that the hail of mutations caused by hypermutability obscures subtle differences among genes in their point-mutation rate; for example, defects in mismatch repair in yeast mask variation in the mutation rate associated with replication timing (Lang and Murray 2011; Lang *et al.* 2013a). Nevertheless, the concordance of the results across two functionally distinct classes of hypermutators indicates that the uniformity we observe in the location of synonymous changes is not a peculiar feature caused by one or the other affected mutational process. Furthermore, hypermutators occur in natural *E. coli* populations, and there is evidence of recurrent losses and reacquisitions of functional DNA repair genes (including *mutS*) during *E. coli* evolution (Denamur *et al*. 2000). Thus, hypermutators likely also contribute to the natural sequence variation analyzed by Martincorena *et al.* (2012), although the extent of this contribution is unclear.

After seeing an earlier version of our analysis above, Martincorena and Luscombe (2012) pointed out that the presence of synonymous substitutions in the LTEE seems to be correlated with gene expression. To examine this issue and its relevance to the issue at hand, we grouped all of the genes from the LTEE ancestral genome into two categories: those with a synonymous

substitution in at least one of the 12 evolved genomes, and those without any synonymous substitutions. As indicated in the Materials and Methods, we used gene expression data obtained under the same conditions as used in the LTEE (Cooper *et al.* 2003), because those expression levels would be the ones relevant to any effect on mutation rate in our study. Indeed, there is a small (3.7%) but significant difference in mean gene expression between these two sets of genes (Welch's *t*-test, $P = 0.0137$). However, gene expression itself is also weakly correlated with gene length ($r = 0.09$), so the difference in expression between genes with and without synonymous changes might be driven by gene length. To examine that possibility, we calculated Kendall's partial coefficient of rank-correlation between synonymous substitutions and gene expression controlling for gene length (Kendall 1942), and we tested its significance assuming normality (Kim and Yi 2006). In fact, the relationship between gene expression and synonymous mutations is not significant ($P = 0.73$) when gene length is taken into account. On average, genes with synonymous substitutions are 1296 bp long, whereas genes without synonymous substitutions are only 850 bp long (Figure 3.4), and this difference is highly significant (Wilcoxon rank-sum test, $P < 10^{-15}$). Taken together, these analyses clearly show that gene length is the main factor determining where synonymous substitutions have accumulated in the genomes of the LTEE populations.

Discussion

Martincorena *et al*. (2012) proposed that natural selection has optimized point-mutation rates at the level of genes within genomes, such that more important genes mutate at lower rates than do less important genes. They used the gene-specific level of synonymous nucleotide variation in

diverse *E. coli* isolates to estimate the mutation rate for each gene. On theoretical grounds, Chen and Zhang (2013) argued that the locally optimized mutation rate hypothesis is untenable, owing to the extremely small size of the relevant selection coefficients, and they presented empirical evidence that also cast doubt on the hypothesis. Moreover, we now show that the accumulation of synonymous changes—a proxy for the underlying mutation rate—in a long-term evolution experiment with *E. coli* is extremely well correlated with gene length, but not with the extent of synonymous diversity in natural isolates (Figure 3.2). Taken together, these theoretical and empirical considerations indicate the need for some alternative explanation to explain why synonymous diversity varies so much across the core genome of *E. coli*.

In general, synonymous nucleotide diversity in natural populations depends not only on the mutation rate but also on the effective population size, which in turn depends on the rate and history of recombination and horizontal gene transfer (HGT). Intragenomic variation in effective population size has been found in many eukaryotic species (Gossmann *et al.* 2011). Furthermore, population structure can cause variation in effective population size (Nordborg 1997), including different histories of HGT at different loci. Research with *Drosophila melanogaster* has shown that nucleotide diversity in natural isolates positively correlates with local recombination rates (Begun and Aquadro 1992), and this relationship has been seen in many other species including human, mouse, *Caenorhabditis elegans*, mosquito, *Arabidopsis thaliana*, and tomato (Hahn 2008). In a related vein, a population genomic analysis of *D. simulans* compared to sister species *D. yakuba* and *D. melanogaster* found that nucleotide diversity and divergence fluctuate on large scales across the genome; these fluctuations are probably related to natural selection, and not the mutation rate (Begun *et al*. 2007). If these fluctuations were caused by recombination being mutagenic, then nucleotide divergence between species should be positively correlated with

recombination, which is not the case in the *D. simulans* dataset. Instead, genomic regions with more recombination may allow polymorphic loci to escape the effects of selection at other sites (Hahn 2008). A recent study mapped recombination rates at fine scales over a significant portion of the *D. pseudoobscura* and *D. miranda* genomes, supporting the hypothesis that such patterns of nucleotide diversity are caused by recombination preserving variation that would otherwise be eliminated by selection operating at linked sites (McGaugh *et al.* 2012).

Empirical work has also shown the importance of recombination and HGT for microbial genome evolution, even over short timescales. In natural populations of *E. coli*, recombination between related strains generates substantially more nucleotide substitutions than does mutation (Guttman and Dykhuizen 1994; Dixit *et al.* 2015). A population genomic study of *Vibrio cyclitrophicus* found that recombination plays a fundamental role in ecological differentiation as positively selected genes, rather than entire genomes, sweep through evolving populations (Shapiro *et al*. 2012). Direct measurements of substitution rates in nature reveal the success of hybrid genotypes containing alleles from distinct *Leptospirillum* groups over mere decades (Denef and Banfield 2012). In Archaea, species are determined largely by ecological differentiation, rather than by physical or genetic barriers to gene flow (Cadillo-Quiroz *et al*. 2012). These empirical studies demonstrate that recombination and HGT play important roles in microbial evolution over short timescales. Moreover, simulations of evolving populations show that the topology of a bacterial phylogeny can be recovered in the presence of recombination, but the branch lengths can be badly distorted (Hedge and Wilson 2014).

In the study by Martincorena *et al*. (2012) and in our work, each *E. coli* genome sampled from nature contains information not only about the mutation rate that its ancestors experienced, but also its particular history of recombination, HGT, and natural selection. This information is

more or less distinct, depending on its genealogical history, from that contained in the other *E. coli* genomes, even if we consider only their shared core. Owing to ecological and genetic differences between strains and related species, some *E. coli* genes may be more readily transferred between diverged lineages than other genes, even among those genes that constitute the core genome. Indeed, experiments have shown that some genes—including those that encode ribosomal proteins often used as phylogenetic markers—are more resistant to HGT between species than others (Sorek *et al.* 2007). Also, computational work has shown that highly expressed genes tend to be more resistant to HGT (Park and Zhang 2012).

Recombination and HGT can also affect the evolution of mutation rates in interesting and important ways. First, recombination can directly impact mutation rates. Functional mismatch repair genes in natural isolates of *E. coli* show high sequence mosaicism relative to housekeeping genes, indicating that repair genes have undergone frequent HGT (Denamur *et al*. 2000). Second, recombination affects how selection operates on mutation rates, with even rare recombination reducing selection for hypermutable phenotypes (Tenaillon *et al.* 2000). Hypermutators often evolve in experiments with bacteria, presumably because they reduce the waiting time for new beneficial mutations, although at the cost of an increased load of harmful mutations (Sniegowski *et al.* 1997; Wielgoss *et al.* 2013); unlike in nature, however, the bacteria in these experiments lack the potential for HGT that could restore a functional gene from another strain.

These issues are important because they support the possibility that variation in θ_s among the core genes of *E. coli* reflects differences in their histories of recombination and HGT, rather than gene-specific differences in their mutation rates. Martincorena *et al.* (2012) showed that genes with low θ_s tend to have functional characteristics typical of housekeeping genes subject to strong purifying selection, and they used that as evidence to argue that mutation rates have been

locally optimized. Their observations are also consistent with recombination and HGT, however, because highly conserved genes should also resist the influx of foreign alleles more effectively than genes that face weak or variable selection. In summary, our analyses offer no support for the hypothesis that point-mutation rates vary among genes and have been optimized, as postulated by Martincorena *et al.* Instead, we think a more plausible explanation is that the variation among genes in their synonymous diversity reflects different histories of recombination and HGT.

Materials and Methods

Calculating synonymous diversity for the core genome of E. coli

Using procedures described elsewhere (Cooper *et al.* 2010), we identified a total of 2,837 single-copy orthologous genes that were shared by all of the *E. coli* strains listed in Table 3.1. We realize that three of the strains (REL606, BL21-DE3, and K-12-MG1655) have been in laboratories for many years, but the vast majority of their mutations accumulated in nature. We also recognize that two of them (REL606 and BL21-DE3) derive from the same natural isolate (Daegelen *et al.* 2009; Jeong *et al*. 2009), but that redundancy does not affect our substantive conclusions because we obtained essentially the same results when we replicated our analyses using the θ_s estimates from Martincorena *et al*. (2012). We used the SATé package (Liu *et al.* 2009) to align the gene sequences. We then performed the θ_s estimation procedure of Martincorena *et al.* (2012) for these alignments using OmegaMap (Wilson and McVean 2006). Using OmegaMap, we could estimate θ_s for 2,835 of the 2,837 single-copy orthologous genes; another gene did not pass a filter for pseudogenes and proteins containing selenocysteine. We consider the resulting set of 2,834 protein-coding genes to be the core genome for our study.

Synonymous substitutions in the LTEE

We identified all synonymous substitutions in the genome sequences of single clones isolated from each of the 12 independently evolved populations after 40,000 generations of the LTEE (Lenski and Travisano 1994; Wiser *et al.* 2013). Six of these clones derived from lineages that had evolved mutations in *mutS*, *mutL*, or *mutT* (Sniegowski *et al.* 1997; Barrick *et al.* 2009; Wielgoss *et al.* 2013). The genomic reads for all 12 populations have been deposited at the NCBI Sequence Read Archive where the accession numbers are SRP001369 (Barrick *et al.* 2009), SRP004752 (Blount *et al*. 2012), SRP045228 (Raeside *et al*. 2014), SRP060289 (Wielgoss *et al.* 2011), and SRP060314 (this study). Across the entire genome, we identified a total of 1,518 synonymous substitutions, which are summarized by population in Figure 3.3. However, in our other analyses we used only the 1,069 synonymous substitutions present in the core genome, including 1,055 in the hypermutator lineages.

Gene expression analyses

We compared the levels of gene expression in the ancestor between those genes that either had or lacked synonymous substitutions in any of the 12 experimentally evolved genomes. We used previously reported gene expression data that was measured under the same conditions as used in the LTEE (Cooper *et al.* 2003).

Statistical analyses, computer code and figures

The data and analysis scripts have been deposited in the Dryad Digital Repository (doi:10.5061/dryad.266g4).

Acknowledgments

We thank Iñigo Martincorena, Vaughn Cooper, Luis Zaman, Justin Meyer, Mike Wiser, and Caroline Turner for discussions and comments. This work was supported, in part, by a National Defense Science and Engineering Graduate Fellowship to RM; a grant from the National Science Foundation (DEB-1019989) to REL; and the BEACON Center for the Study of Evolution in Action (National Science Foundation Cooperative Agreement DBI-0939454).

APPENDIX

Figure 3.1: The expected time to coalescence for individuals from an evolving haploid population is N_e generations. Tick marks show neutral mutation events along two lineages, which occur at some rate μ per generation. The expected number of mutations separating Individuals 1 and 2 is $2N_e\mu$. If all genes in the genome have experienced the same N_e, then significant variation among genes in the per-site rate of accumulation of neutral mutations would imply gene-specific heterogeneity in the underlying mutation rate.

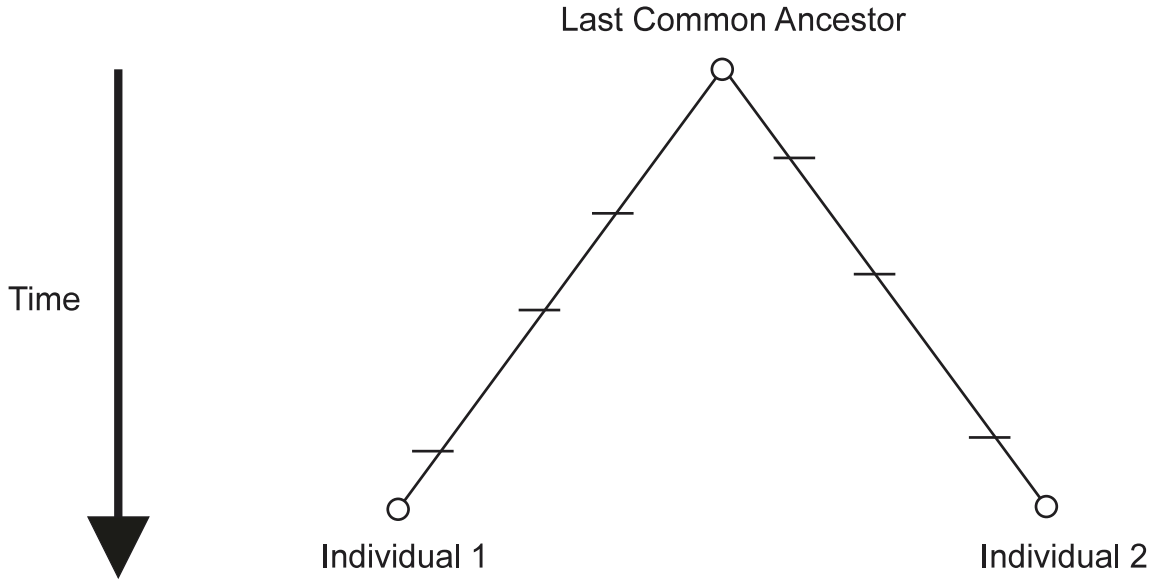


Figure 3.2: Synonymous substitutions observed in experimental populations of *E. coli* closely match the null hypothesis of a uniform point-mutation rate across genes, but not the distribution expected if the variability in θ_s across genes in natural isolates is explained by gene-specific differences in the point-mutation rate. Each observed or hypothetical series shows the cumulative proportion of 1,069 synonymous substitutions in 2,834 genes that have been sorted and ranked by their θ_s values (i.e., the synonymous nucleotide diversity seen in natural isolates for each gene). The red line shows the observed distribution of synonymous mutations in 12 independently evolved genomes after 40,000 generations. The dashed curve shows the null hypothesis of a uniform point-mutation rate, where gene length alone predicts the occurrence of synonymous changes. The dotted curve shows the alternative hypothesis where each gene's point-mutation rate is proportional to θ_s.

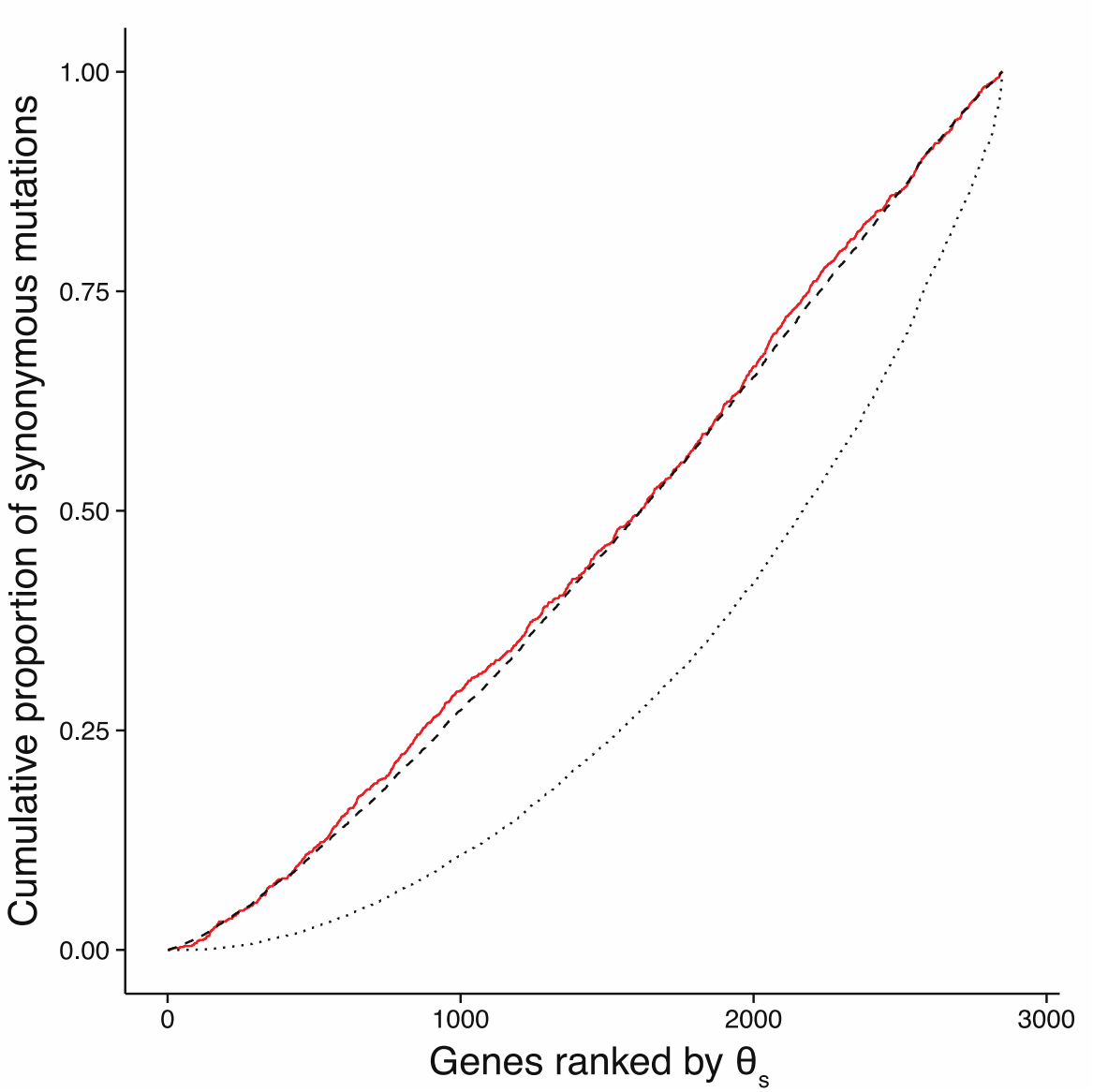


Figure 3.3: Hypermutator clones have distinctive spectra of synonymous mutations in addition to elevated mutation rates. Clones with defective *mutS* or *mutL* genes (Ara–2, Ara–3, Ara–4, Ara+3) have large numbers of C:G to T:A and A:T to G:C transitions, whereas clones with defects in *mutT* (Ara–1, Ara+6) have large numbers of A:T to C:G transversions.

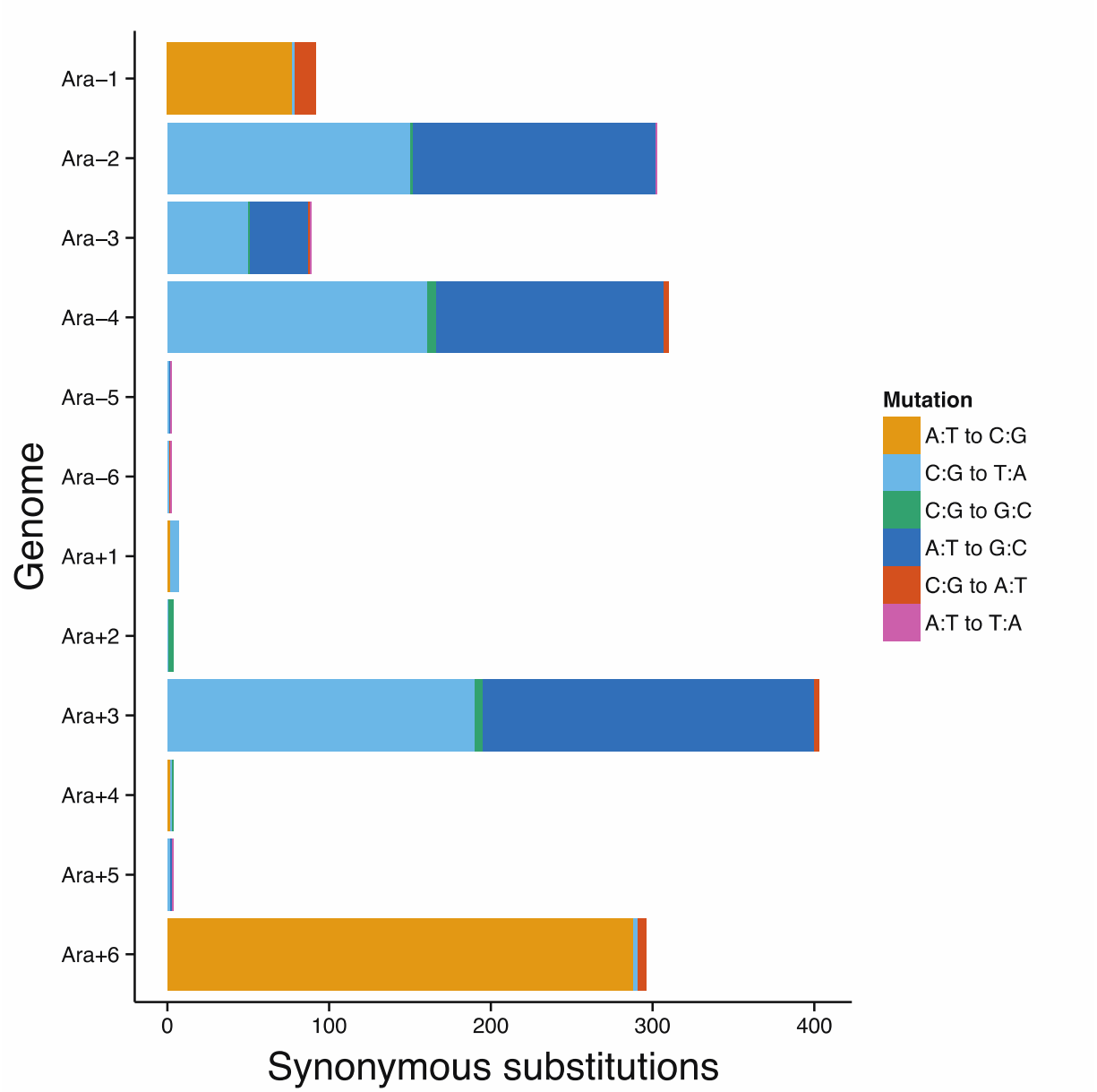


Figure 3.4: Synonymous substitutions tend to be found in longer genes. Genes with at least one synonymous substitution after 40,000 generations (green) are on average 1296 bp long, whereas those without any synonymous substitutions (purple) are on average only 850 bp long. The bin width is 50 bp.

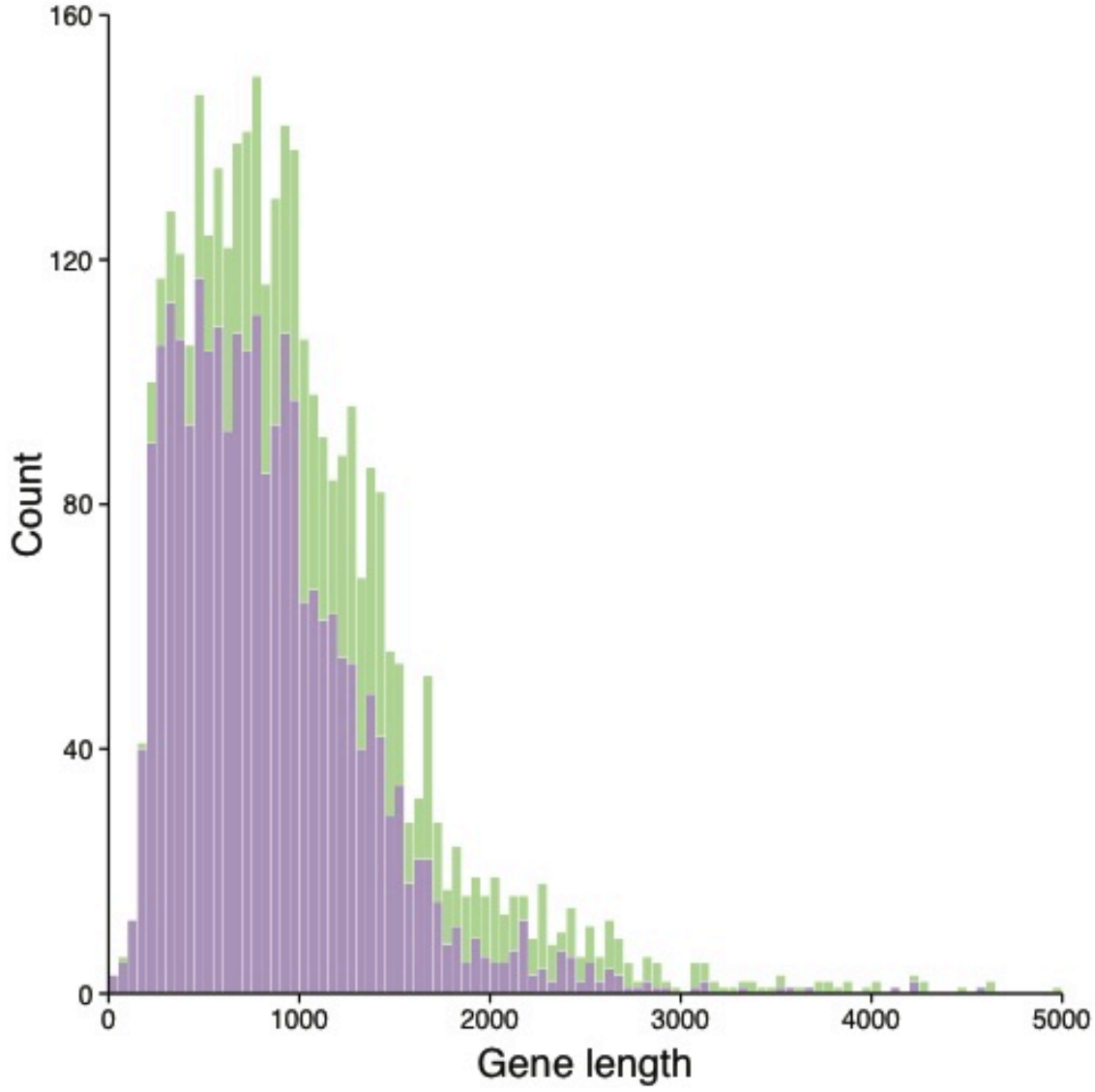


Table 3.1: *E. coli* genomes used in this study.

Strain	NCBI Accession	JGI Taxon ID	Size (bp)	Coding Sequences
E. coli B str. REL606	NC_012967	644736359	4,629,812	4,404
E. coli BL21(DE3)	NC_012892	646862324	4,558,947	4,360
E. coli K-12, MG1655	NC_000913	646311926	4,641,652	4,140
E. coli O157:H7 Sakai (EHEC)	NC_002695	637000108	5,498,450	5,204
E. coli O26:H11 str. 11368	NC_013361	648028025	5,697,240	5,528
E. coli UMN026	NC_011751	644736365	5,202,090	4,819
E. coli SMS-3-5	NC_010498	641522624	5,068,389	4,773
E. coli HS	CP000802	640753025	4,643,538	5,228
E. coli 536	NC_008253	637000104	4,938,920	4,553
E. coli O111:H- str. 11128	NC_013364	646311924	5,371,077	5,167

LITERATURE CITED

LITERATURE CITED

Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. 2013. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol Biol Evol.* 30:549–560.

Bailey SF, Hinz A, Kassen R. 2014. Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. *Nat Commun.* 5:4076.

Barrick JE, Lenski RE. 2013. Genome dynamics during experimental evolution. *Nat Rev Genet.* 14:827–839.

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461:1243–1247.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519–520.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*. 5:e310.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489:513–518.

Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ. 2012. Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol*. 10:e1001265.

Chen X, Zhang J. 2013. No gene-specific optimization of mutation rate in *Escherichia coli*. *Mol Biol Evol.* 30:1559–1562.

Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 100:1072–1077.

Cooper VS, Vohr SH, Wrocklage SC, Hatcher PJ. 2010. Why genes evolve faster on secondary chromosomes in bacteria. *PLoS Comput Biol.* 6:e1000732.

Daegelen P, Studier FW, Lenski RE, Cure S, Kim JF. 2009. Tracing ancestors and relatives of *Escherichia coli* B, and the derivation of B strains REL606 and BL21(DE3). *J Mol Biol.* 394:634–643.

Denamur E, Lecointre G, Darlu P, Tenaillon O, Acquaviva C, Sayada C, Sunjevaric I, Rothstein R, Elion J, Taddei F, et al. 2000. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* 103:711–721.

Denef VJ, Banfield JF. 2012. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* 336:462–466.

Dixit P, Pang TY, Studier FW, Maslov S. 2015. Recombinant transfer in the basic genome of *E. coli*. *Proc Natl Acad Sci U S A.* forthcoming.

Eyre-Walker A. 1994. Synonymous substitutions are clustered in enterobacterial genes. *J Mol Evol.* 39:448–451.

Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM. 2013. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet*. 45:784–790.

Foster PL, Hanson AJ, Lee H, Popodi EM, Tang H. 2013. On the mutational topology of the bacterial genome. *G3 (Bethesda)*. 3:399–407.

Gerrish PJ, Lenski RE. 1998. The fate of competing beneficial mutations in an asexual population. *Genetica* 102/103:127–144.

Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective population size within a genome. *Genetics* 189:1389–1402.

Guttman DS, Dykhuizen DE. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266:1380–1383.

Hahn MW. 2008 Toward a selection theory of molecular evolution. *Evolution* 62:255–265.

Hedge J, Wilson DJ. 2014. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *mBio* 5:e02158.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6:e1001115.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6:e1001107.

Jeong H, Barbe V, Lee CH, Vallenet D, Yu DS, Choi SH Couloux A, Lee SW, Yoon SH, Cattolico L, et al. 2009. Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J Mol Biol*. 394:644–652.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res*. 19:1195–1201.

Kendall MG. 1942. Partial rank correlation. *Biometrika* 32:277–284.

Kim N, Jinks-Robertson S. 2009. dUTP incorporation into genomic DNA is linked to transcription in yeast. *Nature* 459:1150–1153.

Kim SH, Yi SV. 2006. Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*. *Mol Biol Evol.* 23:1068–1075.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.

Lang GI, Murray AW. 2008. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics* 178:67–82.

Lang GI, Murray AW. 2011. Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome Biol Evol*. 3:799–811.

Lang GI, Parsons L, Gammie AE. 2013a. Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. *G3 (Bethesda)*. 3:1453–1465

Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. 2013b. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500:571–574.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* 5:e1003527.

Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc Natl Acad Sci U S A*. 109:E2774–E2783.

Lenski RE, Travisano M. 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc Natl Acad Sci U S A*. 91:6808–6814.

Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. 2015. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* 519:181–186.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324:1561–1564.

Long H, Sung W, Miller SF, Ackerman MS, Doak TG, Lynch M. 2015. Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient *Pseudomonas fluorescens* ATCC948. *Genome Biol Evol*. 7:262–271.

Luria SE, Delbrück M. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511.

Maddamsetti R, Lenski RE, Barrick JE. 2015. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics* 200:619–631.

Martincorena I, Seshasayee ASN, Luscombe NM. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485:95–98.

Martincorena I, Luscombe NM. 2012. Response to Horizontal gene transfer may explain variation in theta_s. arXiv:1211.0928

McGaugh SE, Heil CSS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, Noor MA. 2012. Recombination modulates how selection affects linked sites in *Drosophila*. *PLoS Biol*. 10:e1001422.

Nordborg M. 1997. Structured coalescent processes on different time scales. *Genetics* 146:1501–1514.

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92–94.

Park C, Zhang J. 2012. High expression hampers horizontal gene transfer. *Genome Biol Evol*. 4:523–532.

Park C, Qian W, Zhang J. 2012. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep*. 13:1123–1129.

Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, Merrikh H. 2013. Accelerated gene evolution through replication–transcription conflicts. *Nature* 495:512–515.

Raeside C, Gaffé J, Deatherage DE, Tenaillon O, Briska AM, Ptashkin RN, Cruveiller S, Médigue C, Lenski RE, Barrick JE, et al. 2014. Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *mBio* 5:e01377-14.

Rocha EP, Feil EJ. 2010. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *PLoS Genet.* 6:e1001104.

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science* 336:48–51.

Sharp PM, Shields DC, Wolfe KH, Li WH. 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* 246:808–810.

Sniegowski PD, Gerrish PJ, Lenski RE. 1997. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387:703–705.

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318:1449–1452.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A*. 109:18488–18492.

Tenaillon O, Le Nagard H, Godelle B, Taddei F. 2000. Mutators and sex in bacteria: conflict between adaptive strategies. *Proc Natl Acad Sci U S A*. 97:10465-10470.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet*. 5:e1000344.

Warnecke T, Supek F, Lehner B. 2012. Nucleoid-associated proteins affect mutation dynamics in *E. coli* in a growth phase-specific manner. *PLoS Comput Biol*. 8:e1002846.

Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D. 2011. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)*. 1:183–186.

Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D. 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A*. 110:222–227.

Wilson DJ, McVean G. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* 172:1411–1425.

Wiser MJ, Ribeck N, Lenski RE. 2013. Long-term dynamics of adaptation in asexual populations. *Science* 342:1364–1367.

Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE. 2011. Second-order selection for evolvability in a large Escherichia coli population. *Science* 331:1433–1436.

CHAPTER 4: GENE FLOW IN MICROBIAL COMMUNITIES COULD EXPLAIN UNEXPECTED PATTERNS OF SYNONYMOUS VARIATION IN THE *ESCHERICHIA COLI* CORE GENOME

Author: Rohan Maddamsetti

Originally published in the journal *Mobile Genetic Elements*, 6: e1137380.

Abstract

Researchers contest the importance of gene flow in bacterial core genomes, as traditionalists view microbes as predominantly clonal, asexually reproducing organisms. Contrary to the traditional perspective, *Escherichia coli* core genes vary greatly in their levels of synonymous genetic diversity. This observation indicates that the relative importance of evolutionary forces such as mutation, selection, and recombination varies from gene to gene. In this paper, I highlight why the synonymous diversity observation is broadly relevant to researchers interested in the evolutionary dynamics of microbial populations and communities. I explain how a model of evolution called the coalescent relates neutral diversity (i.e. mutations with negligible fitness effects) to mutation rates, evolutionary time, and a parameter called effective population size. I then describe the possible ways in which mutation, selection, and recombination can explain observed patterns of synonymous diversity in *E. coli.* Finally, I describe a model for *E. coli* genome evolution in which different loci are subject to varying levels of gene flow among co-occurring microbes and viruses in the environment. Researchers can falsify the gene flow hypothesis by sequencing genes and strains isolated from stable microbiomes or by carrying out evolution experiments that trace gene genealogies in real-time.

Evolutionary dynamics of the *Escherichia coli* genome

As for many microbes, gene content across *Escherichia coli* strains is quite variable. The *E. coli* genome comprises a core set of genes shared by all *E. coli* isolates, and a set of flexible genes found in some but not all *E. coli* isolates. A commonplace assumption is that core genes share a common history of vertical descent. Over time, *E. coli* lineages accumulate mutations that have negligible effects on fitness. The rate at which these neutral mutations accrue is roughly proportional to the mutation rate. Synonymous mutations are a reasonable proxy for truly neutral mutations, because their fitness effects are usually (but not always) negligible compared to nonsynonymous mutations that change amino acid sequence (Maddamsetti *et al.* 2015b). From this line of reasoning it follows that levels of synonymous genetic diversity in core genes should be roughly proportional to the mutation rate at those core genes.

However, levels of synonymous genetic diversity vary by more than an order of magnitude over core *E. coli* genes (Martincorena *et al.* 2012, Maddamsetti *et al.* 2015b). Such variation in levels of synonymous diversity causes the branch lengths of some gene trees to be uniformly longer than the branches of other gene trees without affecting tree topology. Trees for highly expressed, important housekeeping genes tend to have shorter branch lengths (less synonymous diversity) than less important core genes. The implication is that either the mutation rate unexpectedly varies over orders of magnitude over core *E. coli* genes, or that there is a serious flaw in the preceding argument linking synonymous diversity to mutation rates. The rest of this paper delves into the evolutionary theory behind synonymous diversity, and goes into the evolutionary forces that could cause synonymous diversity to vary over *E. coli* core genes. I argue that it is a mistake to assume that core genes in the same *E. coli* genome share the same history of vertical descent, when in fact recombination and gene transfer can cause the history of

core genes (or pieces of core genes) present in the same genome to differ substantially without affecting the topologies of bacterial phylogenies (Hedge and Wilson 2014).

The Wright-Fisher model and the coalescent: neutral models of molecular evolution

In this section, I explain how neutral models of evolution help in understanding patterns of synonymous diversity. The neutral theory of molecular evolution makes clear predictions for how genetic drift, in the absence of all other evolutionary forces, shapes genetic diversity (Kimura 1983). Neutral theory has become an essential tool for studying genome evolution because it is the null hypothesis that must be rejected before considering more complicated explanations for patterns of molecular variation (Lynch 2007).

The Wright-Fisher model of neutral evolution describes an idealized population of N organisms (Figure 4.1A). In the absence of natural selection, all organisms are equally fit. We measure time in discrete generations, and the population size is fixed at N. Every generation, we randomly pick organisms from the current generation to leave offspring in the next generation. As in all neutral models, evolution reduces to random sampling of a finite population.

Due to random sampling, eventually the whole population descends from a single organism. If we trace the ancestry of a population backwards in time, eventually we come to this individual: the most recent common ancestor (MRCA) of the population. The basic premise of the coalescent is that we run a model of neutral evolution backwards in time to the MRCA (Figure 4.1B). The history of two given individuals coalesces in the generation in which they share a common ancestor. At any point in time, the probability that a second organism has the same ancestor as a first organism is $\frac{1}{N}$. Therefore, the probability that two specific individuals

coalesce in one generation is $\frac{1}{N}$, and the probability that they do not coalesce is $1 - \frac{1}{N}$.

Eventually, the histories of all individuals in the population coalesce to that of the MRCA.

The probability that two specific individuals in the current generation coalesce *t* generations in the past is the probability that they do not coalesce for $t - 1$ generations backwards in time and then coalesce in the t^{th} generation: $P(X = t) = \left(1 - \frac{1}{N}\right)^{t-1}\left(\frac{1}{N}\right)$. The coalescence of a pair of organisms is thus described by a geometric random variable *X* with a mean of *N* generations. The mathematics is identical to flipping a coin until reaching a flip of heads (Figure 4.1C). Intuitively, it takes two coin flips on average to flip heads once. Flipping a long stretch of tails before flipping heads is unlikely with a fair coin, because the probability of flipping a long stretch of tails before flipping heads decreases geometrically $\left(\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \ldots\right)$. Coalescence for two specific individuals is like flipping a biased coin where the probability of heads (coalescence) is $\frac{1}{N}$, and the probability of tails (no coalescence) is $1 - \frac{1}{N}$.

Effective population size, coalescence times, and neutral diversity

It is important to remember that *N* is not the population size for organisms evolving in the real world, but the population size of organisms in an idealized model of neutral evolution. For this reason, researchers add a subscript to make it clear that N_e is the population size of the idealized model of neutral evolution that best fits molecular data. Much of the power of coalescent theory derives from the fact that more complicated models of evolution involving recombination, natural selection, and population structure make predictions for patterns of molecular variation

that are identical to a neutral model with an appropriately scaled effective population size N_e (Charlesworth 2009). In general, effective population sizes are usually orders of magnitude smaller than actual census population sizes in nature (Charlesworth 2009). For example, a population that has experienced a recent selective sweep or population bottleneck coalesces to the MRCA after a short period of time, causing a dramatically lower effective population size with regard to levels of neutral genetic diversity. Researchers interested in bacterial speciation have used computer simulations to demonstrate that recombination, mutation, and population structure (i.e. dividing a population into many subpopulations) can cause populations to cluster or diverge genetically in the absence of natural selection. In these models, effective population size is simply the number of organisms in the simulation, and levels of neutral genetic diversity depend on the relative importance of recombination, mutation, and population structure in the model. In neutral models, clusters of diverged genotypes ("species") do not easily form in recombining populations, implying a strong role for either natural selection or strong population subdivision (or both) in bacterial speciation (Fraser *et al.* 2007; Fraser *et al.* 2009).

In clonal populations, neutral genetic diversity should accumulate uniformly across the genome because all genes in a genome are completely linked, and thus equally affected by evolutionary forces such as mutation or natural selection. Variation in synonymous genetic diversity among core genes allows us to reject the null hypothesis that core *E. coli* genes experience the same evolutionary forces. Neutral theory applies equally well to genes as to individuals, so on average, the MRCA for two neutrally evolving sequences existed N_e generations in the past. If the mutation rate μ is constant over the genome, then the number of neutral genetic differences between two sequences in the present day is $\theta = 2\mu N_e$. If we use synonymous variation as a proxy for neutral genetic changes, then synonymous diversity θ_s =

$2\mu N_e$ is a natural statistical estimator for both the effective population size as well as the coalescence time for pairs of sequences. In the next section, I discuss possible explanations why synonymous genetic diversity varies so much across the core genome of *E. coli*.

Explanations for variation in synonymous diversity in *E. coli* core genes

Many evolutionary forces, including mutation, selection, and recombination, have similar as well as correlated effects on both μ and N_e. Disentangling the contributions of these forces to patterns of natural variation remains challenging. I discuss the effects of these evolutionary processes on μ and N_e in turn (Figure 4.2).

Mutation. One explanation for why some genes are more variable than others is mutation rate variation. While there is good evidence for local differences in the point mutation rate in bacterial genomes, explanations that solely rely on local mutation rate variation are implausible because no studies to date have found a correlation between mutation rates and patterns of synonymous variation in *E. coli* (Maddamsetti *et al.* 2015b). In short, variation in the mutation rate does not appear to be strong enough to explain orders of magnitude differences in synonymous genetic diversity across *E. coli* core genes.

Natural selection. Selection plays an important role in determining genetic variability across loci. When a highly beneficial mutation sweeps through a population (positive selection), it also reduces genetic variability at linked sites and decreases the time to coalescence to the MRCA. Because a selective sweep reduces variation at all linked sites, this explanation cannot account for patterns in synonymous genetic diversity in *E. coli* without sufficient recombination, because a selective sweep uniformly reduces standing genetic diversity in completely clonal populations.

Background selection is a more satisfying explanation for patterns of synonymous diversity in *E. coli*. Housekeeping core genes are more conserved on the amino acid level than other core genes, because mutations in these most essential core genes can have large effects on organismal fitness. This form of selection is known as purifying selection because it promotes sequence conservation. Purifying selection on deleterious mutations also decreases variability at nearby sites in the genome, and selection on neutral mutations due to purifying selection on nearby sites is called background selection. Background selection is the most parsimonious explanation for variation in synonymous diversity, although Martincorena *et al.* (2012) rejected it as a sufficient explanation.

Negative frequency-dependent selection (balancing selection) on a locus preserves genetic diversity. Such beneficial mutations do not complete selective sweeps because the fitness advantage conferred by the mutation decreases as it increases in frequency in the population. Mutations conferring frequency-dependent advantages are common in evolution experiments (Maddamsetti *et al.* 2015a), and are probably even more common in complex and heterogeneous environments such as the animal gut. However, this explanation again requires recombination, otherwise frequency-dependent selection would maintain synonymous variation at similar levels across the genome.

Recombination. Many studies have estimated the relative contributions of recombination and mutation to *E. coli* diversity (Dixit *et al.* 2015; Bobay *et al.* 2015). An important open question outside the scope of this paper is how and why diverse bacterial species and populations vary in their propensity toward freely-recombining and clonal lifestyles. Some natural populations of *Synechococcus* have enough homologous recombination to generate quasisexual

evolutionary dynamics (Rosen *et al.* 2015), while some *Pseudomonas* populations appear to be largely clonal (Sarkar *et al.* 2004).

Recombination can affect synonymous diversity because a recombination event between diverged sequences causes multiple changes to appear simultaneously, while recombination between closely related or even identical sequences may not be detectable at all. If some genes have had a history of more successful recombination events with diverged homologs compared to other genes in the genome, then those genes will be more diverse than genes with a history of fewer successful recombination events. However, recombination with diverged homologs cannot explain observed patterns of synonymous diversity in *E. coli.* Any recombination event with an outgroup will either change the topology of the gene tree or cause anomalously long branches (Figure 4.2B), while observed patterns of synonymous diversity in *E. coli* core genes are inconsistent with these predictions (Martincorena *et al*. 2012). Nonetheless, a combination of recombination and positive selection or negative frequency-dependent selection could account for some of the observed variation in synonymous diversity.

Mutagenic effects of recombination. Recent sequencing studies have found that new mutations correlate with the location of recent crossover events in human sperm as well as in plants and honeybees (Arbeithuber *et al.* 2015; Yang *et al.* 2015). It is unclear whether the molecular mechanisms responsible for elevated mutated rates in these studies also occur in *E. coli*. Nonetheless, error-prone repair of double-strand breaks associated with recombination events could contribute to higher levels of synonymous diversity at loci with a history of many successful but undetected recombination events in *E. coli*.

Population structure. Population structure measures the degree to which populations are not well-mixed. A simple case is a metapopulation, or a population subdivided into a large

number of subpopulations. Populations can be structured at multiple spatial scales (i.e. subpopulations of subpopulations), and population structure generally maintains genetic diversity by restricting the scope of selective sweeps. Population structure can also reduce effective population sizes and coalescence times due to local extinctions, colonization events and local population bottlenecks (Fraser *et al.* 2009).

Gene flow could explain patterns of synonymous genetic variation in *E. coli*

In this section, I present a model that combines aspects of recombination, selection, and population structure to explain patterns of synonymous genetic variation in *Escherichia coli*. Although this model is not parsimonious, it is testable and consistent with existing molecular and ecological observations in the literature (Polz *et al.* 2013).

While it is well-known that flexible *E. coli* genes differ in their histories of recombination and selection across diverged microbial species in gut communities, the same may hold true for many *E. coli* core genes. Imagine a "wind" of diverse alleles blowing into a population of *E. coli,* this "wind" being the migration of alleles into the population from other *E. coli* populations, viral populations, or other microbes in the community. Resident genes under purifying selection can resist this "wind" more strongly, and they will have a shorter coalescence time than genes that cannot effectively resist replacement by diverse alleles. In terms of the Wright-Fisher model, gene flow between species within a community increases the effective population size of that gene compared to species-specific genes (Figure 4.3). This argument is general in that it holds for subpopulations of a single bacterial species, or for populations of co-evolving phage and bacteria. For instance, imagine two subpopulations of *E. coli*, each adapted to different parts of an animal's gut. Genes under stronger purifying selection in one subpopulation would better

resist gene flow from the other subpopulation. The key point in this model is that gene flow within microbial communities can change effective population sizes and coalescence times at core genes without changing the topology of gene trees constructed with single isolates from diverse ecological sources. In the most extreme cases, between-species divergence and within-species polymorphism may be indistinguishable. One likely mechanism for gene flow in microbial communities are phage-bacteria infection networks in which generalized transducing phage infect multiple microbial species and act as viral vectors (Modi *et al.* 2013; Dixit *et al.* 2015). The gene flow model makes a strong prediction: genes with high synonymous diversity should tend to cluster according to microbial community, while genes with low synonymous diversity should tend to cluster by species (Figure 4.3). Evolution experiments or appropriate sampling of microbiomes could test this prediction to falsify the gene flow model.

The gene flow model has some support in the literature. Retchless *et al.* (2007) proposed the fragmented speciation model in which different segments of bacterial chromosomes become genetically isolated at different times. Species-specific alleles become isolated first; alleles can sweep across species boundaries, and gene flow stops earlier at earlier diverging loci. This study came to the conclusion that in some cases, it may not be possible to make a clear distinction between intraspecific and interspecific variability in microbes. Sheppard *et al*. (2008) found evidence of increasing gene flow between previously distinct *Campylobacter* species. Retchless *et al.* (2010) argued that phylogenetic incongruence in gene trees made with genes found in *Escherichia*, *Salmonella*, and *Citrobacter* provides further evidence for the fragmented speciation model. Luo *et al.* (2011) described the genomes of environmental isolates of *E. coli* and found little evidence of gene exchange with gut commensal *E. coli* due to plausible ecological barriers. Although they found within-clade transfer of core genes, this paper rejected

the fragmented speciation model because fragmented speciation posits gene flow across *E. coli* clades except at niche-specific adaptive mutations or genetic incompatibilities restricting gene flow. Karberg *et al.* (2011) found that recently acquired genes in *Salmonella* and *Escherichia* genomes have similar codon usage frequencies, while core genes in *Salmonella* and *Escherichia* have noticeably diverged in codon usage. Therefore, it appears that *Salmonella* and *Escherichia* strains acquire genes from a common pangenome shared among enterobacterial species. Smillie *et al.* (2011) built a database of horizontally transferred sequences among 2,235 full bacterial genomes to explore the effects of phylogeny, geography, and ecology on horizontal gene transfer. This study found that shared ecology is far more important than phylogenetic relatedness in structuring networks of gene flow across bacterial species.

Conclusion

Synonymous genetic diversity depends on both the mutation rate and effective population size. In neutral models of evolution, effective population size has a second interpretation as the average time for two lineages to coalesce. Many evolutionary forces, including mutation, selection, and recombination can affect genome-wide variation in synonymous genetic diversity. While researchers recognize the importance of gene flow in structuring the flexible genome of microbes, gene flow may also affect the core genome of microbes. If so, gene flow could explain why highly important *E. coli* core genes have less synonymous genetic diversity than other core genes. While the importance of gene flow in microbial genome evolution depends strongly on ecological context, many important microbiomes, such as the animal gut, might be effectively described as metapopulations of genes that interact within and across genomes over multiple spatial and temporal scales.

Acknowledgments

I thank Alita Burmeister and Michael Wiser for comments on the manuscript, and I thank John Wakeley, Sergey Kryazhimskiy, and Justin Meyer for critical comments and discussions. This work was supported by the BEACON Center for the Study of Evolution in Action (National Science Foundation Cooperative Agreement DBI-0939454).

APPENDIX

Figure 4.1: The population size in a neutral model of evolution also describes the average time for two lineages to coalesce in that model. A) One run of the Wright-Fisher model over four generations for a population of four individuals. B) The coalescent for the run of the Wright-Fisher model in part A). C) The probability that it takes *t* generations for two lineages to coalesce is identical to the probability of flipping *t – 1* tails before flipping heads using a biased coin that has a probability of flipping heads (i.e. coalescence) of 1/*N*. A geometric distribution with mean *N* describes both processes.

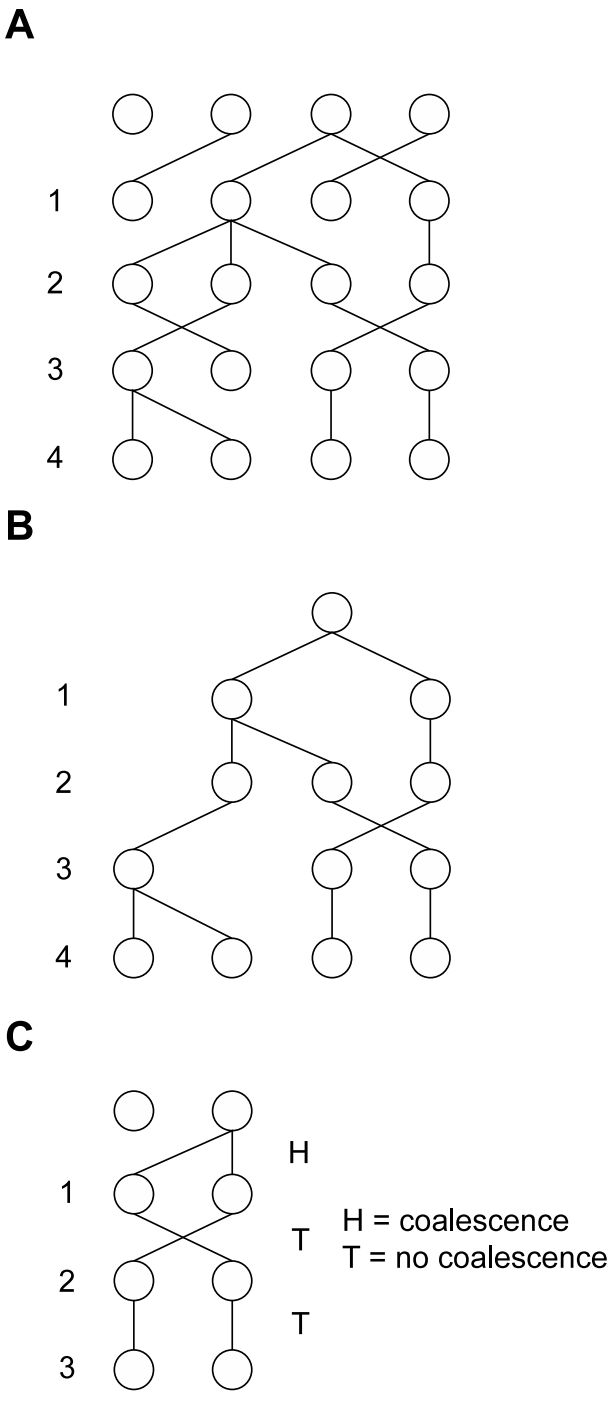


Figure 4.2: Mutation, selection, and recombination affect the branch lengths and topology of phylogenetic trees. A) Differing selection pressures or mutation rates can lengthen or shorten branch lengths. B) Recombination with an ingroup will not change the tree, while recombination with an outgroup always changes either the topology of the tree or disproportionately changes the length of some branches.

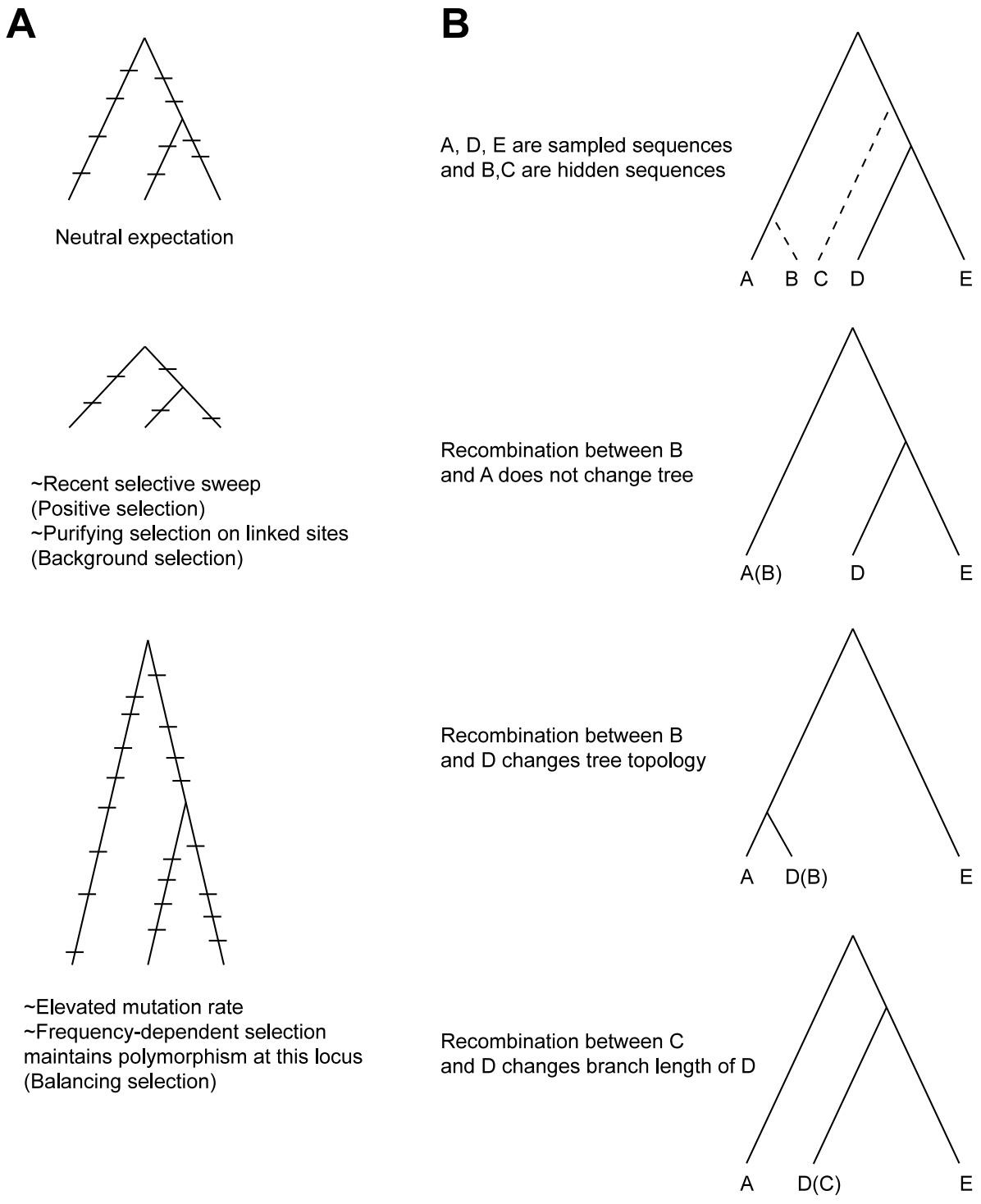
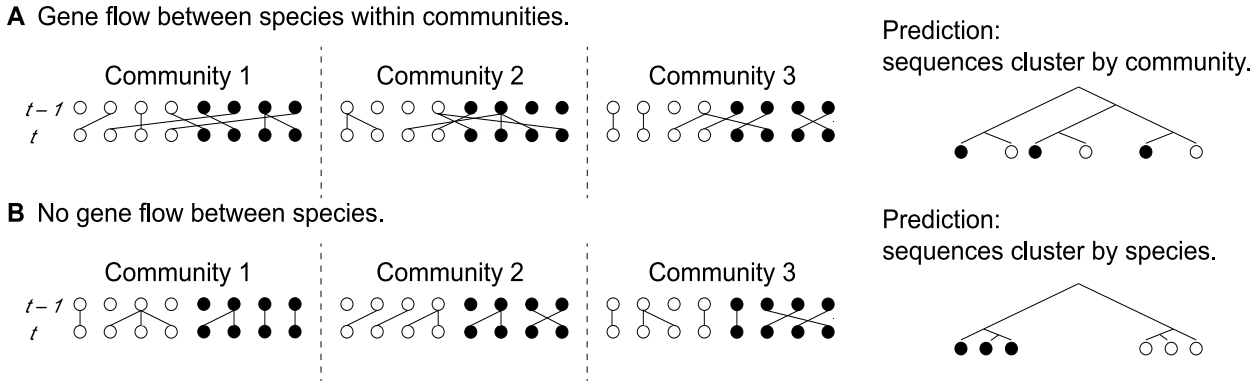


Figure 4.3: Different rates of gene flow at different loci causes effective population size to vary at these loci, in turn affecting gene tree coalescence times without changing tree topology for genes co-occurring in the same genome. A) Gene flow at this locus occurs between species within communities, increasing the effective population size of this locus. In this case, communities cluster in the gene tree. B) Gene flow does not occur between species at this second locus. The effective population size at this locus is the population size of the species in which it is found. In this case, species cluster in the gene tree.



LITERATURE CITED

LITERATURE CITED

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Acad Natl Sci U S A* 112: 2109–2114.

Bobay LM, Traverse CC, Ochman H. 2015. Impermanence of bacterial clones. *Proc Natl Acad Sci U S A* 112: 8893–8900.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10: 195–205.

Dixit PD, Pang TY, Studier FW, Maslov S. 2015. Recombinant transfer in the basic genome of *Escherichia coli*. *Proc Natl Acad Sci U S A* 112: 9070–9075.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. *Science* 315: 476–480.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323: 741–746.

Hedge J, Wilson DJ. 2014. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio* 5: e02158.

Karberg KA, Olsen GJ, Davis JJ. 2011. Similarity of genes horizontally acquired by *Escherichia coli* and *Salmonella enterica* is evidence of a supraspecies pangenome. *Proc Natl Acad Sci U S A* 108: 20154–20159.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press. 384 p.

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci U S A* 108: 7200–7205.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates. 494 p.

Maddamsetti R, Lenski RE, Barrick JE. 2015. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics* 200: 619–631.

Maddamsetti R, Hatcher PJ, Cruveiller S, Médigue C, Barrick JE, Lenski RE. 2015. Synonymous genetic variation in natural isolates of *Escherichia coli* does not predict where synonymous mutations occur in a long-term experiment. *Mol Biol Evol* 32: 2897–2904.

Martincorena I, Seshasayee AS, Luscombe NM. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485: 95–98.

Modi SR, Lee HH, Spina CS, Collins JJ. 2013. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499: 219–222.

Polz MF, Alm EJ, Hanage WP. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet* 29: 170–175.

Retchless AC, Lawrence JG. 2007. Temporal fragmentation of speciation in bacteria. *Science* 317:1093–1096.

Retchless AC, Lawrence JG. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proc Natl Acad Sci U S A* 107: 11453–11458.

Rosen MJ, Davison M, Bhaya D, Fisher DS. 2015. Fine-scale diversity and extensive recombination in a quasi-sexual bacterial population occupying a broad niche. *Science* 348: 1019–1023.

Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Appl Environ Microbiol* 70: 1999–2012.

Sheppard SK, McCarthy ND, Falush D, Maiden MC. 2008. Convergence of Campylobacter species: implications for bacterial evolution. *Science* 320: 237–239.

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480: 241–244.

Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen JQ, Hurst LD, Tian D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* 523: 463–467.

CHAPTER 5: GENOMIC ANALYSIS OF *ESCHERICHIA COLI* FROM AN EVOLUTION EXPERIMENT WITH INTERGENOMIC RECOMBINATION

Authors: Rohan Maddamsetti and Richard E. Lenski

Abstract

We analyze the genomes of clones isolated from an experiment in which *E. coli* K-12 Hfr donor strains were periodically introduced into 12 populations of *E. coli* B derived from the long-term evolution experiment. Contrary to expectations, intergenomic recombination did not speed up adaptation, despite demonstrably increasing genetic variation over the course of the experiment. In this study, we sequenced 24 genomes isolated after 1000 generations from the recombination experiment. The effects of recombination were highly variable across the 12 recipient populations: one lineage was mostly derived from the K-12 donors, while another had almost no K-12 ancestry. We found some evidence of parallel evolution in the genomic architecture of the remaining recombinant lineages, with some regions showing repeated introgression and others almost none. These broad-scale changes were largely consistent with the molecular basis of the donors' conjugative functions. That is, regions with high or low introgression of donor DNA tended to be near or far, respectively, from the donors *oriT* origin of transfer sites. Moreover, the donors carried mutations that made them auxotrophic, which were effectively lethal in the experimental environment, and these mutations appear to have further shaped the patterns of introgression. We also asked whether beneficial mutations that previously arose in the LTEE-derived recipient populations made nearby regions of the genome impervious to recombination. The results were inconclusive in this respect, with beneficial mutations replaced by donor alleles

in some recombinant lineages but not others. We found evidence of new beneficial mutations in some recombinant populations. We also found clusters of identical mutations across replicate lineages that suggest gene conversion or some other process of non-homologous recombination. We are currently examining whether merodiploidy—in which both donor and recipient alleles are found in the same (normally haploid) genome—might account for some genomic features observed in recombinant lineages. On balance, though, our analyses are consistent with a scenario in which recombination pressure was sufficiently strong to allow donor alleles to invade the recipient populations without those alleles providing much, if any, selective advantage.

Introduction

An open question in microbial evolution is why some populations of bacteria seem to have extensive intergenomic recombination (Rosen *et al.* 2015) while others seem to have very little (Sarkar and Guttman 2004). Some symbiotic bacteria appear to be entirely clonal owing to their tight associations with hosts that preclude contact with other lineages (Moran *et al.* 2009), but otherwise the reasons for differences in recombination rates across bacterial taxa are unclear. Intergenomic recombination unlinks beneficial and deleterious mutations from the rest of the genome. Under conditions of high recombination and strong selection, individual genes rather than entire genomes can go to fixation. When recombination is infrequent or absent but selection is strong, highly beneficial mutations can drive large genomic regions or even whole genomes to fixation. Recent work shows that both gene-specific and genome-wide selective sweeps occur in microbial communities (Bendall *et al.* 2016).

Horizontally transmitted viruses and conjugative plasmids mediate recombination in many species of bacteria (Levin and Lenski 1983). In this context, recombination is a special kind of evolutionary process. Like mutation rates, intergenomic recombination rates can evolve because many genes involved in DNA replication and repair affect the rates of mutation and recombination. But unlike mutation rates, recombination rates in bacteria are also subject to the direct effects of coevolution owing to the association with plasmids and viruses. Intergenomic recombination qualitatively changes evolutionary dynamics and can speed up adaptive evolution under some circumstances (Keightley and Otto 2006; Cooper 2007; McDonald *et al.* 2016). However, recombination—especially as it occurs in bacteria—may have originated as a byproduct of the spread of selfish elements rather than for the purpose of increasing the efficiency of natural selection in the genome as a whole (Hickey and Rose 1986).

In this paper, we revisit an evolution experiment in which recombination did not appear to increase the efficiency of natural selection and did not speed up adaptation. Souza, Turner, and Lenski (Souza *et al.* 1997) conducted an evolution experiment with *E. coli* in which they periodically introduced strains of *E. coli* K-12 that could donate genetic material but not themselves grow (owing to mutations that caused nutritional deficiencies) into populations of *E. coli* B that had previously evolved in and adapted to their minimal-glucose environment for 7,000 generations. We refer to their recombination treatment as the Souza-Turner-Lenski experiment (STLE) and to the asexual experiment that generated the recipient populations as the long-term evolution experiment (LTEE). The stated goal of the STLE was to test whether recombination would increase the rate of adaptation by increasing the genetic variation available to natural selection (Souza *et al.* 1997). In the absence of complex selection dynamics (such as frequency-dependent selection), the expected rate of adaptation of a population is proportional to

the genetic variance in fitness in a population (Fisher 1958). The outcome of the STLE was puzzling in that recombination demonstrably introduced substantial genetic variation (as determined by tracking ~10 genetic markers available to the authors at that time), but it had no significant effect on the rate of adaptation compared to control populations that evolved without recombination. Two possible explanations were proposed that might explain those results. According to one hypothesis, the recombination treatment was so effective that recombination acted like a strong mutational force, replacing many neutral alleles (or even overwhelming natural selection, if the donor alleles were actually maladaptive) through the sheer flux of donor alleles into the recipient populations. Alternatively, the interactions between donor and recipient genes, and the ecological context in which those interactions occurred, might have somehow generated strong frequency-dependent selection such that the assays used to measure fitness gains were undermined. In fact, the Ara-3 recombinant population of the STLE actually appeared to decline substantially in fitness, an effect that was shown to reflect an evolved frequency-dependent interaction (Turner *et al.* 1996). In this study, we explore the genomic basis of the puzzling results of the STLE. We find that recombination acted more like an increased mutational pressure, and less like a sieve separating beneficial from deleterious mutations.

Materials and Methods

Overview of STLE

The experiments performed by Souza *et al.* (1997) are described fully in that paper. In brief, 12 recombinant populations and 12 control populations were started from clones isolated after 7000 generations of the LTEE (Table 5.1). These populations were propagated daily for 1000

generations (150 days) following the same transfer regime and using the same DM25 medium and other conditions as the LTEE (Lenski *et al.* 1991). However, on day 0 and every fifth day following (~33 generations), 0.01 ml of an equal mix of four K-12 Hfr donor strains (REL288, REL291, REL296, and REL298) were added to the recombinant treatment populations and allowed to mate (conjugate) for 1 hour. All four donors were auxotrophic for one or more essential amino acids (arginine, leucine, or isoleucine-valine), so that they could transfer their genetic material but not grow and persist in the population.

Isolation of clones for genomic analysis

Samples of all 12 recombination-treatment populations from generation 1000 of the STLE (Table 5.1) were revived according to the standard LTEE protocol: 15 μL from each frozen stock were pipetted into 10 mL of LB medium, grown for 24 h, diluted and grown in DM25 for another 24 h, and then spread on LB agar plates. The two colonies that grew closest to randomly placed marks were re-streaked on LB agar plates and then grown in LB medium. These 24 STLE-derived recombinant clones were stored at -80°C (Table 5.1). We sometimes refer to these clones as "odd" and "even" (based on the freezer numbers for each member of a pair) when we present data for just one set of 12 clones.

Genome sequencing and analysis

The 4 Hfr donor strains (REL288, REL291, REL296, and REL298), 12 LTEE-derived recipient clones used as the ancestors in the STLE, and 24 STLE-derived recombinant clones (two per population) were thawed and grown in LB medium, and samples of genomic DNA were isolated from each one. The genomic DNA was then sequenced on an Illumina MiSeq by the MSU

RTSF Genomics Core Facility. We used the *breseq* pipeline (Deatherage *et al.* 2015) to analyze the genomes in this project, using the ancestral LTEE strain REL606 as the reference genome unless otherwise specified. We used the K-12 MG1655 reference genome to identify mutations specific to the donor strains. We used the *gdtools* utility in *breseq* to compute a table listing the union of all the mutations found in the K-12 donor genomes in comparison to REL606.

A custom python program called *label_mutations.py* was written that performs four tasks. First, this program labels all of the mutations found in the STLE-derived recombinant genomes relative to the REL606 genome by looking at the corresponding site in the donor, recipient, and recombinant genomes. A total of five distinct labels are used. (i) Mutations that are present in both a recombinant clone and its parent recipient clone (but not in REL606 or the union of donor strains) are labeled as LTEE mutations. (ii) Mutations that are found in both a recombinant clone and the union of K-12 donors are labeled as K-12 mutations. (iii) Any mutations that are present in a recombinant clone, but not found in either the donors or recipient clone, are labeled as new mutations. (iv) Mutations present in the union of K-12 donors that are not found in the recipient clone are labeled as REL606 mutations: these sites are genetic markers that distinguish the *E. coli* B-derived REL606 strain used to start the LTEE from K-12. (v). Mutations found in the recipient but not in the recombinant are labeled as deleted mutations; they were removed by recombination with the donors or otherwise lost during the STLE. The program ignores all sites that are identical between the recombinant, recipient, and donors because they provide no useful information. As a second task, the *label_mutations.py* program produces a table of the markers that distinguish K-12 from REL606 (see label iv above). Third, this program generates a table of the LTEE-specific mutations found in the recipient genomes (see labels i and v above). Fourth, the *label_mutations.py* program identifies cases of parallel evolution among the new mutations

found in the recombinant genomes by running the *parallyze* analysis program (in development by Elizabeth Baird and R.M.).

We also performed a preliminary exploratory analysis to examine merodiploidy in the STLE recombinant genomes. First, we ran the *breseq* pipeline in polymorphism mode to estimate the allele frequencies for all mutations. Most mutations have a frequency of 100%, which should exclude merodiploidy as a complicating factor at those sites. However, some mutations had intermediate frequencies, and they therefore are candidate merodiploids because true merodiploids should contain both K-12 and either ancestral B or new LTEE alleles in the sequencing data. We used the *label_mutations.py* program to make a table of labeled mutations as before, but here with an additional field for allele frequency.

An R script called *dissertation_analysis*.R makes figures and does statistical tests using the tables of labeled mutations that *label_mutations.py* produces. All code is freely available upon request to R.M.

Manual annotation of specific donor genome features

Each donor strain also has two special features: the transposon-generated mutations that made it an amino-acid auxotroph, and the Hfr transfer origin and orientation. Surprisingly, *breseq* version 0.26 did not find the auxotrophy mutations present in the donor strains REL288, REL291, REL296, and REL298. However, these strains were shown to be auxotrophs in Souza *et al.* (1997), and so we used the information in their Table 5.1 and in a traditional linkage map of *Escherichia coli* K-12 (Berlyn 1998) to find the location of the auxotrophy mutations in these strains; these mutations had been generated and annotated in Wanner *et al.* (1986). *E. coli* K-12 and *E. coli* B strain REL606 are largely syntenic (Studier *et al.* 2009); therefore, we used the

linkage map for K-12 to place the auxotrophy mutations and Hfr *oriT* transfer origin sites in the donor K-12 strains with respect to their homologous genes in REL606; this mapping of K-12 elements to the REL606 genome is only approximate but appears to perform reasonably well. Further work will be needed to find the auxotrophy mutations and Hfr transfer origins in the short-read sequences derived from the donor genomes and using K-12 as the reference genome.

Calculation of lengths of donor and recipient segments in recombinant genomes

Recombination breakpoints occur somewhere in the interval between donor-specific and recipient-specific markers. A minimal estimate of the length of a donor segment would place recombination breakpoints at the donor markers on each end. A maximal estimate would place the breakpoints at the flanking recipient markers. In fact, the true breakpoints cannot be known exactly. Our approach uses the minimal estimate on the left, but the maximal estimate on the right, and so it will produce intermediate values that should tend to an overall average length similar to what would be obtained by averaging the minimal and maximal segment lengths. In particular, our algorithm alternates between K-12 and B segments along the genome, switching states when reaching the alternate marker type. This approach thus takes the maximal estimate on the right, but a maximal estimate on the right then yields the minimal estimate on the left for the following segment.

We used the following algorithm. Each genome is a list of labeled mutations. First, we initialize a list of '0' with the length of the genome. We keep track of two state variables: the index of the last breakpoint (transition-state) and a Boolean state variable called 'in.K12.chunk' that is initialized to FALSE under the assumption that the first segment of the genome comes from the *E. coli* B recipient. For every labeled mutation in the genome, we check whether the

current mutation has a label that changes the state of 'in.K12.chunk'. If 'in.K12.chunk' is FALSE and its state changes, the current mutation is labeled '1-2'. If 'in.K12.chunk' is TRUE and its state changes, the current mutation is labeled '2-1.' At the end of the loop, we check our initial assumption that 'in.K12.chunk' was FALSE. Because we stored the position of the last transition-state, we check if the last transition-state in the genome is '1-2', in which case the first '1-2' transition should be set to '0' since the *E. coli* genome is circular. All sites marked '0' are removed from the genome. We then calculate the differences between the N-1 pairs of transition-state mutations; '1-2' on the left and '2-1' right gives the length of a K-12 segment, whereas '2-1' on the left and '1-2' on the right gives the length of a B segment. The final pair of transition-state mutations is the last and first element of the list. In this way, we calculate the lengths of segments in a recombinant genome that were derived from the donor and recipient. The code as currently written does not adjust for any deletions or insertions that may have occurred in those segments.

Results

Architecture of recombinant genomes

Figure 5.1 summarizes the rich and complex information on genomic changes that occurred before and after the 1000-generation STLE. Two clones were sampled at random from each of the 12 populations in the recombination treatment, and their genomes were sequenced and analyzed. The genomic sites marked in red show mutations that distinguish the 12 clones that were used as recipients in the STLE from the ancestor of the LTEE, and which were also present in the genome of the clone that was sequenced after the STLE. These mutations thus arose

during the 7000 generations of the LTEE that preceded the start of the STLE, and they persisted for the 1000 generations of the STLE. Three pairs of recombinant clones (those from populations Ara+3, Ara+6, and Ara-2) have far more such mutations than the clones from the other nine populations because the recipient clones used to start these populations came from populations that had evolved hypermutable phenotypes during the early generations of the LTEE (Sniegowski *et al.* 1997; Tenaillon *et al.* 2016).

The genomic sites marked in yellow show mutations that are shared by the pool of K-12 donor strains and the sequenced recombinant clone, but are not found in the recipient used to start a population. These mutations were introduced by intergenomic recombination during the STLE. These sites reveal several features. First, the majority of sites in most genomes are derived from the recipients, not from the Hfr donor strains. In fact, single clones from populations Ara+3 (Figure 5.1A) and Ara-6 (Figure 5.1D) of the STLE appear to lack any DNA regions that derive from the donors, and both clones from population Ara+2 only have one very short donor segment (~1 kbp in length, barely visible at ~2.5 Mbp in Figure 5.1A). Second, there is one striking exception to the above pattern: the genomes of both clones sampled from population Ara-3 are largely comprised of DNA derived from the Hfr donor strains (Figure 5.1C). We also sequenced two other clones (REL4397 and REL4398) from this population that were used in a previous study of frequency-dependent selection (Turner *et al.* 1996). These clones, too, are predominantly K-12, but with many small regions that descend from the LTEE recipient clone. Third, in most STLE populations, the pattern of introgression of donor DNA is very similar in the two recombinant clones that we sequenced. However, there are differences in several cases including Ara+3 and Ara-6 where, as noted above, one of each pair appears to lack any donor DNA, and Ara+4 (Figure 5.1B) and Ara-5 (Figure 5.1D), where the pairs of

recombinant clones share some regions of donor DNA but not others. More generally, we sampled and sequenced only two clones from each STLE population, and so we have a limited view of within-population diversity.

Fourth, there is an almost complete absence of donor DNA in all of the recombinant populations (except Ara-3, which has mostly donor DNA) between the positions ranging from ~1 to ~3 Mbp on the circular genomic map. Figure 5.2 shows this point clearly as the sum of the number of introgressions of donor DNA into the odd-numbered clone from each STLE population, excluding the aberrant population Ara-3. Figure 5.2 also shows that donor DNA appears to be concentrated in two distinct regions of the recombinant genomes. One region is centered at ~3.5 Mbp on the map, falling off more or less symmetrically on either side. The other region peaks at ~0.25 Mbp to ~0.5 Mbp and appears to extend farther to the left, eventually wrapping around the circular chromosome.

In addition to the yellow and red marks that indicate donor-derived and LTEE-derived mutations, respectively, Figure 5.1 also has some black and green marks. Black marks indicate mutations that do not exist in either the Hfr donor pool or the recipient clone that was used to start a given population. These marks therefore indicate new mutations that arose during the 1000-generation STLE. Not surprisingly, there are many more new mutations in the three populations founded by the hypermutable recipient clones. Green marks indicate mutations present in the LTEE-derived recipient clone but absent from the STL-derived recombinant clone. These marks imply that the mutations at those sites were deleted during the STLE. Such deletions are most common in populations Ara+1 (Figure 5.1A) and Ara-4 (Figure 5.1D).

Probable beneficial mutations

Figure 5.1 also has labels showing the names of certain genes that harbor mutations in one or both clones from a particular recombinant population. The mutations are marked by filled symbols that are colored in a similar manner to the lines: red symbols indicate mutations that were present in the LTEE-derived recipient and retained by the recombinant clone; black symbols are new mutations that were not present in the recipient or the donor but were found in the recombinant; and green symbols are mutations that were present in the recipient but absent from the recombinant, either because they were replaced by donor DNA (in which case the green symbol is surrounded by yellow) or otherwise because that site was deleted during the STLE. A total of 31 different genes are labeled in one or more of the recombinant populations. Table 5.2 provides some additional information on each of the genes.

These 31 genes are all probable targets of positive selection under the conditions of the LTEE. They were previously identified (along with some other genes that do not have any mutations in the clones in our study) by sequencing a total of 264 genomes from the 12 LTEE populations at various time points through 50,000 generations, and finding that they accumulated an unexpectedly large number of independent nonsynonymous mutations in lineages that had not become hypermutable (Tenaillon *et al.* 2016). The *G* scores shown in Table 5.2 indicate the strength of the evidence for excessive parallelism in a gene, relative to the length of its coding sequence. For some of the genes, genetic manipulations and competition assays have directly confirmed that mutations indeed provide a fitness benefit under the conditions of the LTEE (Barrick *et al.* 2009; Khan *et al.* 2011).

The STLE's 1000-generation duration is short relative to the 50,000 generations of the LTEE, and so we might not expect to see many new beneficial mutations rising to high

frequency in these genes. However, we see some examples including four in population Ara+1 in the *fabR*, *trkH, hslU,* and *iclR* genes (Figure 5.1A) and three in population Ara-4 in the *topA, pykF*, and *hslU* genes (Figure 5.1D). Curiously, these two are also the two non-mutator populations that had the most gene deletions (green hash marks), although we do not know whether this relationship is coincidental or meaningful. In the next section, we consider the fate of those presumptively beneficial mutations that were present in the LTEE-derived recipient at the start of the STLE.

Possible sources of variation in introgression across genomic regions, and the fate of previously evolved beneficial mutations

What is the source of the variation in the extent of introgression of the Hfr donors' DNA into the recombinant populations? There are several distinct hypotheses that rely either on differences in the propensity for genomic regions to be transferred by the donors or on the fitness effects of integrating different regions into the recipient's chromosome. These hypotheses are not mutually exclusive, and so two or more of them may contribute to the observed patterns of introgression (Figures 5.1 and 5.2). Hypothesis 1: Some regions of donor DNA were transferred more often than other regions, leading to overrepresentation of the former regions in the recombinant genomes. Hypothesis 2: Some regions of donor DNA contained alleles that were beneficial to the recipient, leading to overrepresentation of those regions in the recombinant genomes. Hypothesis 3: Some regions of donor DNA contained alleles that were deleterious to the recipient, leading to underrepresentation of those regions in the recombinant genomes. This hypothesis can be subdivided into two variant hypotheses. According to Hypothesis 3A, the donor alleles were maladaptive regardless of the beneficial mutations that arose during the LTEE. According to Hypothesis 3B, the donor alleles were maladaptive specifically because the

recipient genomes had acquired beneficial mutations in those regions during the 7000 generations of the LTEE that preceded the STLE.

By way of background, it is should be noted that *E. coli* K-12—the strain that gave rise to the Hfr donors—and *E. coli* B—the strain from which the recipients derive—are fairly closely related, at least as far as *E. coli* strains go. These two source strains were independently isolated from nature many years ago (Daegelen *et al.* 2009). However, about half their shared genes encode proteins that have identical amino-acid sequences (Studier *et al.* 2009). On the other hand, several hundred genes are present in only one or the other strain, including so-called "genomic islands" that are thought to have been acquired by horizontal gene transfer in the phylogenetic networks leading to one or the other strain (Studier *et al.* 2009). In addition to these more or less ancient differences, the four K-12 donors were deliberately modified by transposon mutagenesis to make them auxotrophic (for different amino acids in the four donors) and by introducing the F-plasmid (at different locations in the four donors) into their chromosomes to make them Hfr (high-frequency recombination) strains; and, as described in the section above, the B-derived recipients accumulated beneficial mutations during the LTEE.

Hypothesis 2 was, in essence, the original motivation for the STLE, with Souza *et al.* (1997) suggesting that intergenomic recombination with the K-12 donors might increase the rate of adaptation (relative to control populations that evolved asexually) by providing an additional source of genetic variation to the LTEE-derived populations. We lack *a priori* information about what sites in the K-12 donor genomes could provide beneficial alleles to the recombinant populations, but *a posteriori* we expect them—if they exist—to be in those regions where the introgression scores are high (Figure 5.2A). On the other hand, Hypothesis 2 seems unlikely, because Souza *et al.* (1997) found that fitness gains were not greater in the recombinant

populations than the control populations, which implies that intergenomic recombination did not increase the supply of beneficial alleles.

Figure 5.2A shows the inferred location and direction of the Hfr origins of transfer of the four K-12 donor strains as well as the inferred location of their auxotrophy mutations, which bear on Hypotheses 1 and 3A, respectively. With respect to Hypothesis 1, Hfr strains transfer their DNA in a unidirectional manner, and the probability that donor genes are transferred to recipients is expected to decline at greater distances from the origin. In the STLE, the cultures in which the donors and recipients were mixed were not shaken for one hour (Souza *et al.* 1997), which would, in principle, allow the transfer of ~60% of the entire chromosome if the conjugative mating began immediately after the strains were mixed. However, not all matings would begin immediately, shaking interrupts DNA transfer, and the efficiency of DNA transfer generally declines with distances from the transfer origin even without shaking. The peak in the introgression scores between ~3 and ~4 Mbp fits very well with the locations and directions of the *oriT* transfer origin sites for two of the Hfr donors: REL288 and REL298 have *oriT* sites near the edges of this peak that point inward from opposite directions. Most of the second, less defined peak in introgression scores (which wraps around the "end" of the chromosome back to the beginning) seems to fit moderately well with the other two Hfr donors, REL296 and REL291, whose *oriT* sites are at ~0.5 and ~0.1 Mbp, respectively, with the former transferring in the direction of the peak introgression scores and the latter transferring in the same direction toward the broad shoulder between ~4.1 and ~0.1 Mbp. However, the other shoulder of the second, less defined peak—from ~0.5 to ~0.9 Mbp—is not explained by the logic of Hfr donor transfer. On the other hand, the near absence of introgression between ~1.0 and ~3.0 Mbp—representing over

40% of the genome—fits well with the Hfr donor *oriT* sites and directionality. On balance, then, patterns of introgression provide strong, albeit imperfect, support for Hypothesis 1.

We also found compelling evidence of strong purifying selection at the sites of the auxotrophy mutations in the K-12 Hfr donors. Recall that these mutations mean that the cells cannot produce essential amino acids, and therefore the cells cannot grow and persist in the minimal medium used for the STLE. The two donors whose transfer properties well account for the introgression peak between ~3 and ~4 Mbp have auxotrophic mutations located at positions that would sharpen the peak by limiting introgression at each edge. In particular, REL288 has an auxotrophy mutation in the *ilv* gene that lies just beyond the *oriT* site for REL298; and REL298 has an auxotrophy mutation in the *argA* gene that lies a short distance past the *oriT* site for REL288. The other two Hfr donors, REL291 and REL296, have auxotrophy mutations in the *argE* and *leuB* genes, respectively, that would contribute to the observed decline in introgression scores on the broader shoulder of the less defined peak from ~4.0 to ~0.2 Mbp on the circular map. (REL298 also has a second auxotrophy mutation in *leuB*, but this gene is very far from its *oriT* site and thus probably not relevant to the observed patterns of introgression.) On balance, we also find support for Hypothesis 3A, whereby selection against the effectively lethal auxotrophy mutations in the donor strains reinforces and sharpens the patterns of introgression generated by the mechanics of gene transfer according to Hypothesis 1.

Hypothesis 3B offers another plausible explanation for the observed patterns of introgression. It rests on the idea that selection should also act against donor alleles in those genes where beneficial mutations arose in the LTEE and were present in a given recipient at the start of the STLE. If this hypothesis were correct, then we would expect to see few, if any cases, where these presumably beneficial mutations were removed and replaced by donor DNA. The

evidence in support of this hypothesis is ambiguous, at best, because of the considerable variation among the recombinant clones, in terms of both the proportion of their DNA that comes from the donor strains and the extent to which the LTEE-derived beneficial mutations have been retained or replaced. For example, in population Ara-1 (Figure 5.1C), only 1 of the 9 presumed LTEE-derived beneficial mutations present in the recipient was replaced by donor DNA (in both recombinant clones), but ~22% of the recombinant genomes was donor DNA. This pattern is consistent with Hypothesis 3B. By contrast, consider population Ara+1 (Figure 5.1A), in which 4 of the 5 presumed beneficial mutations in the recipient were replaced, when only ~32% of the overall recombinant genomes came from the donors. Across the 12 STLE populations, in fact, we see a slight tendency for these presumed beneficial mutations to have been replaced by donor alleles more often than the average genomic site, contrary to our hypothesis. On balance, the evidence does not support Hypothesis 3B.

The interpretation could get more complex, however, if we consider an explanation for the patterns of introgression that combines Hypothesis 3B with Hypothesis 2. That is, imagine that the hypothetical beneficial donor alleles presumed to promote introgression by accelerating adaptation under Hypothesis 2 involve some of the same genes in which beneficial mutations arose in the LTEE. In that case, the replacements of the LTEE-derived beneficial by the donor alleles might be effectively neutral or perhaps even provide a small net benefit. To address this possibility, we examined the recombinant clone sequences in detail for these replacements to determine whether they simply reverted the gene back to its pre-LTEE ancestral state (i.e., the corresponding sequence in REL606) or, alternatively, introduced a different allele.

Gene conversion and new mutations

The following analyses focus, for simplicity, on the odd-numbered recombinant clones from the STLE populations that were not hypermutable; however, the even-numbered clones carry similar information. We noticed that these recombinant clones often had more new mutations than typical LTEE clones that had evolved for 1000 generations (Tenaillon *et al.* 2016). When we looked for evidence of parallel evolution among these new mutations, we found strong but spurious signals in two genes, *ECB_03438* and *nohB*; in particular, we saw that multiple identical mutations had appeared in those genes in multiple lineages. The most likely explanation for these events is gene conversion, in which recombination occurs between *non-homologous* genes in the K-12 donors and B recipients. To investigate the possibility of gene conversion further, we scored all genes that had three or more new mutations in the same recombinant genome as putative gene conversion events (Table 5.3). Because so many apparently multi-mutation events occurred, and usually in multiple lineages, we think they are best explained by single non-homologous recombination events, not by multiple mutations.

We also used data on the strength of selection on genes from the LTEE (Tenaillon *et al.* 2016) to ask whether new mutations tended to occur in genes where beneficial alleles are known to often arise. The genes affected by the multi-site events that we infer were caused by gene conversions had a mean *G*-score of 0.42, while the genes affected by all other new mutations had a mean *G*-score of 13. This difference, though only marginally significant (two-sided Welch's *t*-test, p-value = 0.045), suggests that typical single-site new mutations that arose during the STLE were under stronger positive selection than the multi-site mutations that occurred by non-homologous recombination.

One of the most puzzling findings is that many LTEE-derived mutations, including mutations that were almost certainly beneficial, had been lost in the Ara+1 and Ara-4 STLE populations (Figure 5.1). One possibility is that these genes exist in a merodiploid state (i.e., with both K-12 and B alleles in the genome), and so they were missed by our initial analysis. We re-examined all 24 recombinant genomes using the polymorphism mode in *breseq*, and we do see some polymorphic regions consistent with merodiploidy. However, there is no evidence of merodiploidy at the sites of the deleted mutations in Ara+1 and Ara–4. A second possibility is that recombination with the K-12 donors removed these mutations from the Ara+1 and Ara-4 populations. When we looked for evidence of parallel evolution among new mutations, we found new alleles of *fabF*, *trkH*, *hslU,* and *iclR* in Ara+1, and new alleles of *topA*, *pykF*, and *hslU* in Ara–4. These genes are among those under strong positive selection in the LTEE (Tenaillon *et al.* 2016). Another piece of evidence suggests that something interesting and unexpected is involved with the putative deletions, in which LTEE-derived mutations that were present in the recipient genomes are absent in the recombinant genomes. That is, the three STLE lineages with the most "deleted" mutations (Ara+1 with 21, Ara–3 with 25, and Ara–4 with 16) also had the most new mutations after excluding the multisite gene conversion events and the hypermutable populations (Ara+1 with 21, Ara–3 with 10, and Ara–4 with 19 new mutations).

No characteristic distribution of lengths of recombinant segments

We examined the distributions of recombinant segment lengths to see whether conjugation left a consistent signature in this respect. Figure 5.3 shows the distribution of lengths of DNA segments derived from the K-12 donor, except for population Ara-3, which is mostly donor DNA and where we show the lengths of the remaining B-derived segments. To explore the possibility

of characteristic segment lengths, we excluded not only Ara-3 but also Ara+2 (which had almost no K-12 ancestry) and the three mutator lineages (Ara-2, Ara+3, Ara+6), because differences in DNA repair processes also affect recombination. Even focusing on just the remaining seven lineages, the recombinants show significant heterogeneity in the distributions of the lengths of their donor-derived segments (Kruskal-Wallis rank sum test, chi-squared = 22.03, df = 6, p = 0.001). At this time, the code that calculates the length of recombinant segments does not handle deletions properly, but this seems unlikely to account for such heterogeneity.

Discussion

In this chapter, we have presented an analysis of 24 recombinant genomes from the STLE. Much work remains to be done to understand certain results, especially (i) the explanation for the apparent losses of some presumably beneficial mutations and (ii) the extent and effects, if any, of merodiploidy on our conclusions thus far. Nonetheless, two main results are clear. First, we found parallel evolution in the structure of recombinant *E. coli* genomes. This pattern appears to be largely explained by the biases caused by the molecular biology of conjugation, coupled with selection against the introduction of the effectively lethal auxotrophy mutations carried by the donor strains.

What remains to be done? In particular, the analyses in this chapter have involved comparing mutations in recombinant genomes to REL606 as a reference. Our analyses miss evolutionary change occurring in parts of the recombinant genomes with no homology to REL606. We plan to repeat the analyses using K-12 as a reference genome in order to make sure that the main results and conclusions do not depend on the specific reference genome we used.

We know that Souza *et al.* (1997) saw that some genes in the recombinant clones appeared to have both K-12 and REL606 alleles (based on banding patterns in allozyme electrophoretic gels) indicative of partial diploidy, and our analyses would miss the effects of these gene duplications. We can (and have begun to) run *breseq* in polymorphism mode to find genes that have both LTEE and K-12 alleles consistent with merodiploidy.

We also would like to better understand the origin of the Ara-3 recombinant clones, which completely lack the LTEE markers present in their ancestral recipient, but which still have small segments that derive from *E. coli* B. Three possible explanations stand out, and they are not mutually exclusive, so that some combination of factors might be involved. First, a B recipient might have been converted into an Hfr donor and delivered small segments to a donor strain that then survived. Second, recombination between the K-12 donor and B recipient genomes might have activated an otherwise latent or even "dead" prophage, leading to virus-mediated transduction in the opposite direction to conjugation. Third, a K-12 donor strain might have reverted its auxotrophy mutation, allowing it to grow and persist in the minimal medium of the STLE. It is known that the Tn10-transposon mutagenesis used to construct the donor strains (Wanner 1986) yields unstable genotypes, in which the transposons can move to other locations in the genome. Fourth, a related hypothesis is that one K-12 donor might have recombined with a second K-12 genotype (which had perhaps lost its F-plasmid and thus become a recipient) in such a way as to repair the nutritional defect. Fifth, some mutation in the Ara-3 recipient genome may have allowed for vastly more efficient conjugation and DNA incorporation. However, we found that the two recipient strains with defects in their mismatch repair (Ara+3 and Ara-2) did not have more K-12 ancestry than the other strains, even though previous research has shown

that *E. coli* strains with defective mismatch repair have relaxed DNA homology requirements for molecular recombination (Rayssiguier *et al.* 1989).

Evolution experiments with both bacteria and yeast have shown that intergenomic recombination can speed up adaptation (Cooper 2007, McDonald *et al.* 2015). In contrast, the STLE shows how such recombination in bacteria can act in a manner analogous to a mutational pressure, leading to the introduction of neutral, and perhaps even deleterious, changes. The most striking evidence of recombination as deleterious force is the number of beneficial mutations in the recipient clones in populations Ara+1, Ara-3, and Ara-4 that were evidently "erased" during the course of the STLE. If many donor alleles were neutral or maladaptive in the environment of the STLE, it is not surprising that those alleles did not speed up adaptation (Souza *et al.* 1997). What is surprising, though, is the extent to which those alleles could evidently invade and replace better-adapted recipient alleles. The most likely explanation, in our view, is that donor genes physically linked to the *oriT* transfer sites could replace the homologous genes in the recipients (even sometimes more fit ones) by virtue of the transmission advantage caused by the Hfr conjugative force, i.e., the high frequency of gene transmission. This explanation may also lend credence to the hypothesis that the evolutionary origin of recombination might lie in a transmission advantage for recombining genes (Hickey and Rose 1986), and not in an organismal or population-level advantage related to the efficiency of natural selection.

Acknowledgments

We thank Neerja Hajela for technical assistance; Valeria Souza and Paul Turner for their prior work on the STLE; Jeff Barrick for advice on *breseq* and the design of this genome sequencing project; and Elizabeth Baird for work on the *parallyzer* program. We also thank Elizabeth Baird, Chris Marx, Joshua Mell, Jeff Morris, Rosemary Redfield, Valeria Souza, Jim Stapleton, and Paul Turner, along with other past and present members of the Lenski lab, for valuable discussions. This work was supported by a grant from the National Science Foundation (DEB-1451740 to R.E.L.), the BEACON Center for the Study of Evolution in Action (NSF Cooperative Agreement DBI-0939454), a National Defense Science and Engineering Graduate Fellowship (to R.M.), and a dissertation completion fellowship from the MSU Graduate School (to R.M.).

APPENDIX

Table 5.1: The freezer identifying numbers and relationships of the 12 sequenced recipient clones used to start the 12 STLE populations and the 24 recombinant clones isolated at the end of the STLE.

Population name	Recipient clone	Final population	Final clones
Ara+1	REL2537	REL4361	REL11734, REL11735
Ara+2	REL2538	REL4362	REL11736, REL11737
Ara+3	REL2539	REL4363	REL11738, REL11739
Ara+4	REL2540	REL4364	REL11740, REL11741
Ara+5	REL2541	REL4365	REL11742, REL11743
Ara+6	REL2542	REL4366	REL11744, REL11745
Ara-1	REL2543	REL4367	REL11746, REL11747
Ara-2	REL2544	REL4368	REL11748, REL11749
Ara-3	REL2545	REL4369	REL11750, REL11751
Ara-4	REL2546	REL4370	REL11752, REL11753
Ara-5	REL2547	REL4371	REL11754, REL11755
Ara-6	REL2548	REL4372	REL11756, REL11757

Figure 5.1: Genomes of recombinant clones isolated after 1000 generations of the STLE.

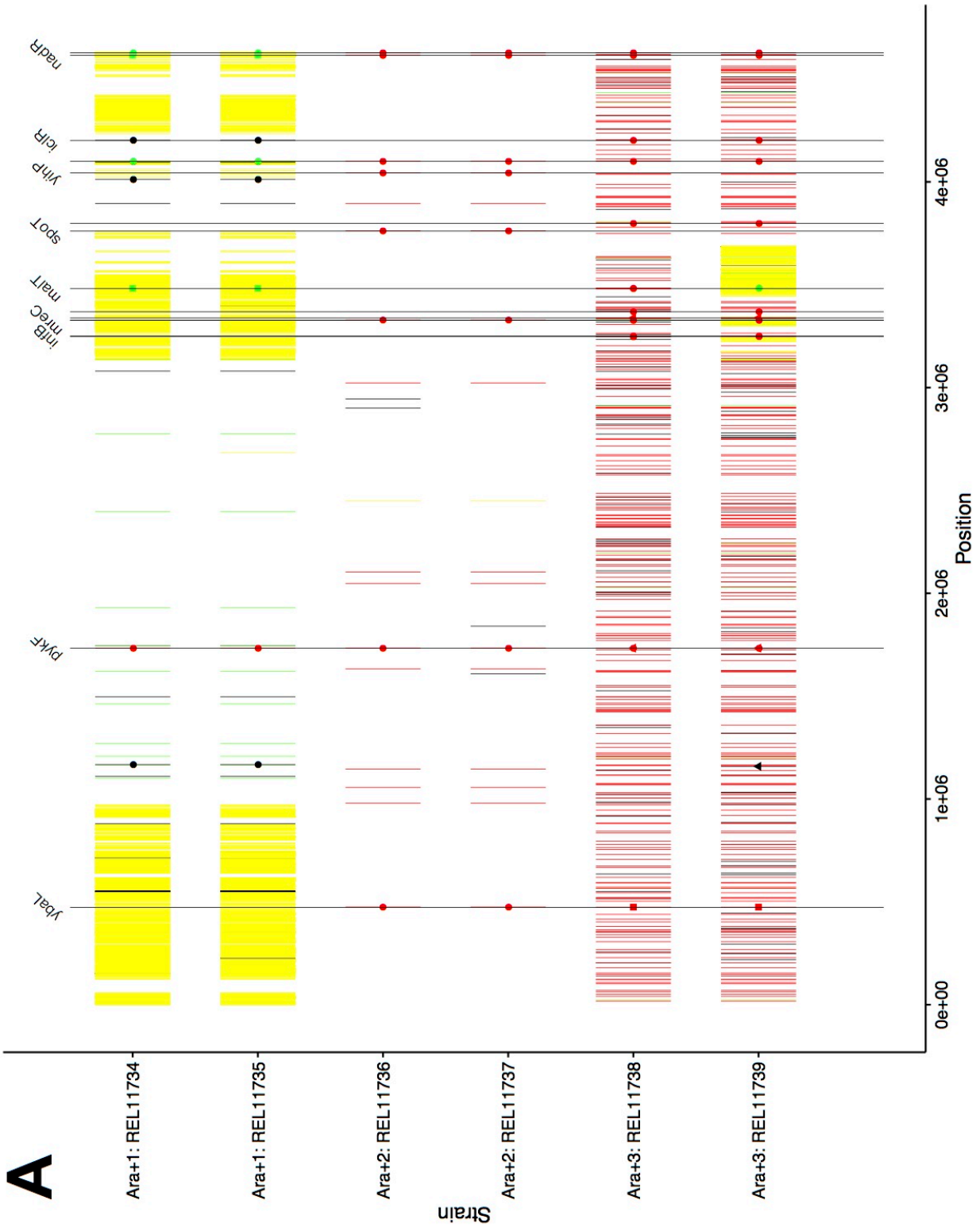


Figure 5.1 (cont'd):

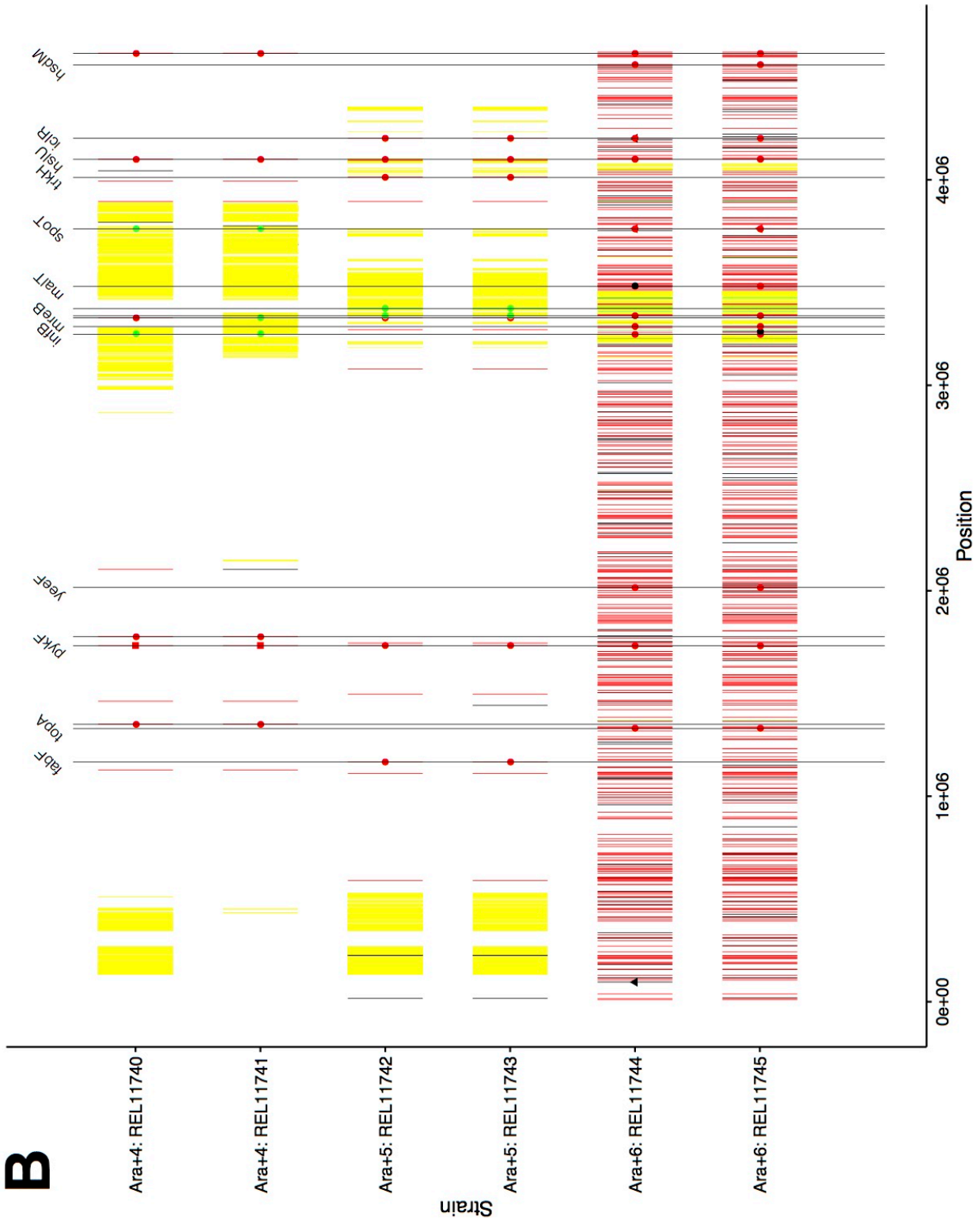


Figure 5.1 (cont'd):

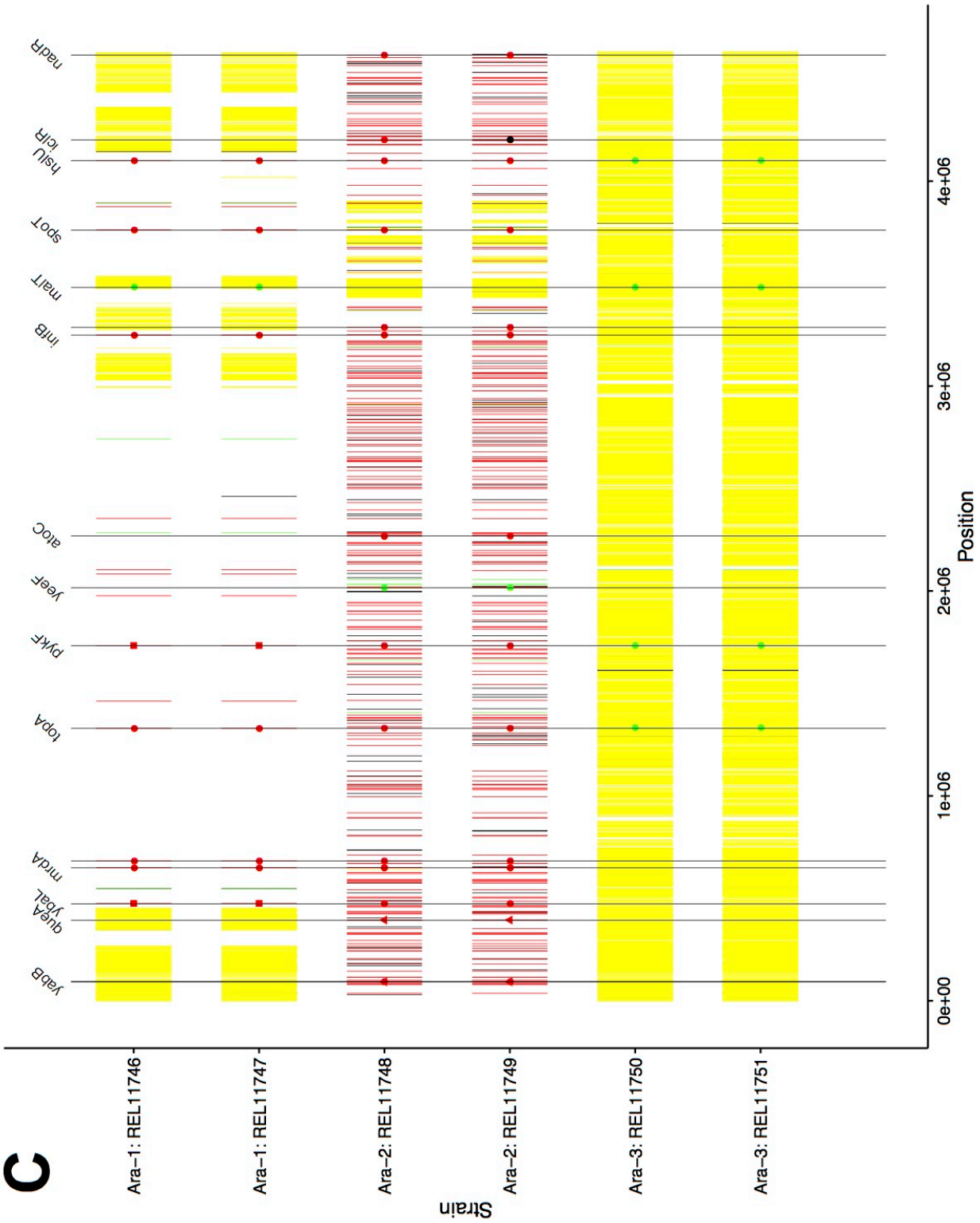


Figure 5.1 (cont'd):

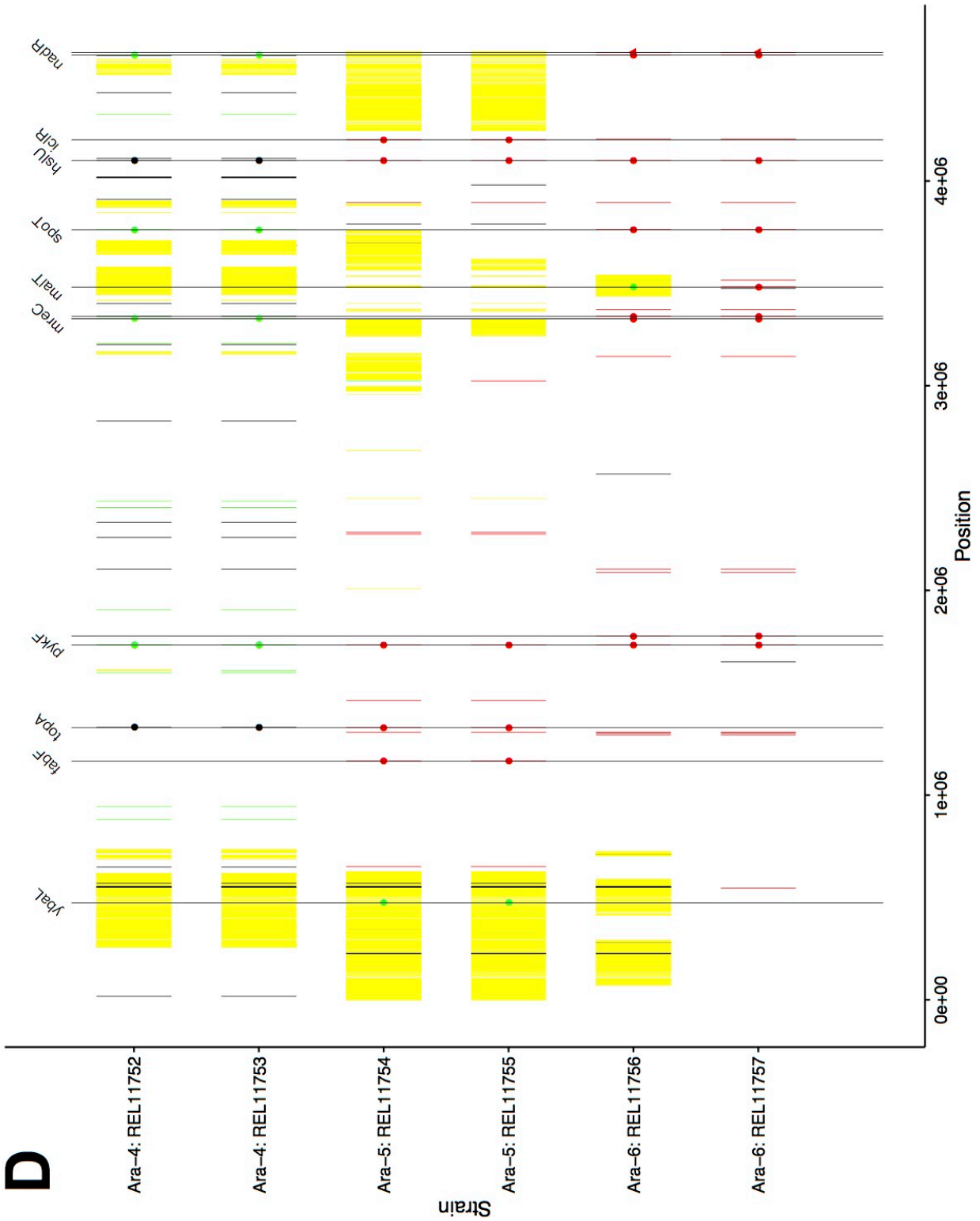


Figure 5.1 (cont'd): REL606 genomic coordinates are plotted on the x-axis and two clones from each of the 12 replicate populations are shown on the y-axis. K-12 genetic markers are shown in yellow, recipient mutations that arose during the LTEE are red, new mutations that arose during the STLE are black, and recipient mutations that were replaced by donor DNA or deleted are green. Mutations in 31 genes with strong evidence of positive selection in the LTEE are marked by symbols of the same color; circles indicate nonsynonymous point mutations, triangles synonymous mutations, and squares other types of mutation. These mutations are also labeled by the gene name (Table 5.2). (A) Genomes from STLE populations Ara+1, Ara+2, and Ara+3. (B) Genomes from STLE populations Ara+4, Ara+5, Ara+6. (C) Genomes from STLE populations Ara–1, Ara–2, and Ara–3. (D) Genomes from STLE populations Ara–4, Ara–5, and Ara–6.

Table 5.2: 31 genes under strong positive selection in the LTEE that were present in the recipient clones used to start the STLE.

Index	Gene	Start position	Coding	G score
1	*yabB*	92438	459	14.48
2	*ftsI*	94217	1767	16.07
3	*queA*	393434	1071	11.09
4	*ybaL*	473629	1677	9.30
5	*mrdA*	648900	1902	48.33
6	*nagC*	682469	1221	10.57
7	*fabF*	1166508	1242	10.50
8	*topA*	1329420	2598	52.38
9	*sapF*	1350197	807	12.23
10	*pykF*	1732965	1413	180.42
11	*infC*	1777363	435	14.70
12	*yeeF*	2016147	1359	17.65
13	*atoC*	2268566	1382	17.55
14	*infB*	3248576	2673	43.57
15	*nusA*	3251273	1488	9.78
16	*arcB*	3285924	2337	14.39
17	*mreC*	3326766	1103	27.50
18	*mreB*	3327935	1044	46.77
19	*yhdG*	3338171	966	11.51
20	*rpsD*	3368960	621	13.28
21	*rplF*	3373512	534	33.30
22	*malT*	3481685	2706	51.73
23	*spoT*	3760757	2109	113.18
24	*yicL*	3798180	924	11.69
25	*trkH*	4011518	1452	25.30
26	*yihP*	4044063	1407	10.00
27	*hslU*	4099899	1332	93.73
28	*iclR*	4201735	825	127.57
29	*hsdM*	4559434	1586	9.52
30	*nadR*	4615529	1233	106.19
31	*arcA*	4627750	717	40.91

Figure 5.2: The number of parallel introgressions of K-12 markers summed over the odd-numbered STLE clones. We omit the Ara-3 clone, which is almost completely derived from K-12 donor DNA. The locations of auxotroph mutations in the donor genomes are shown as dashed vertical lines, and the location and orientation of the Hfr *oriT* transfer origin sites are labeled below the x-axis.

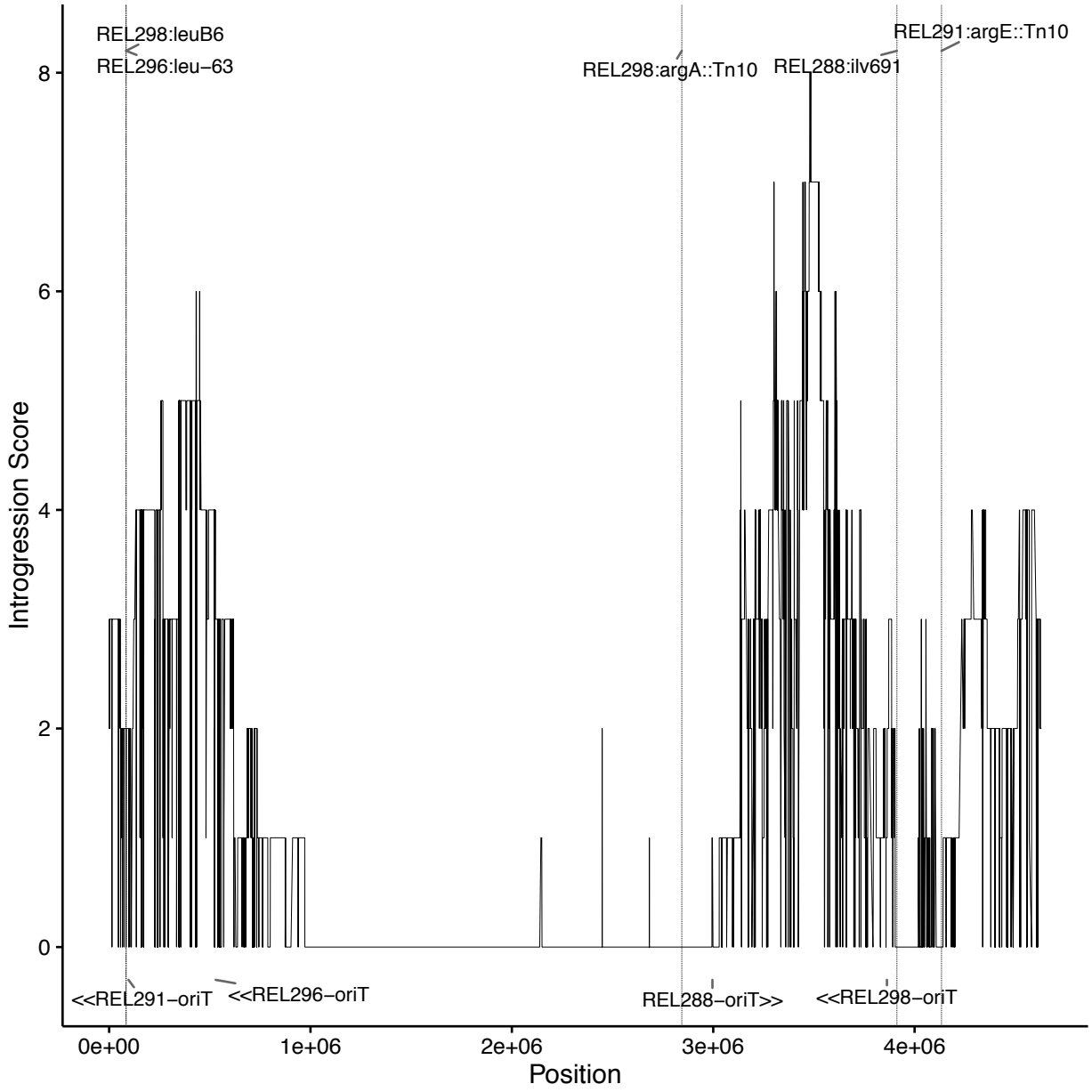
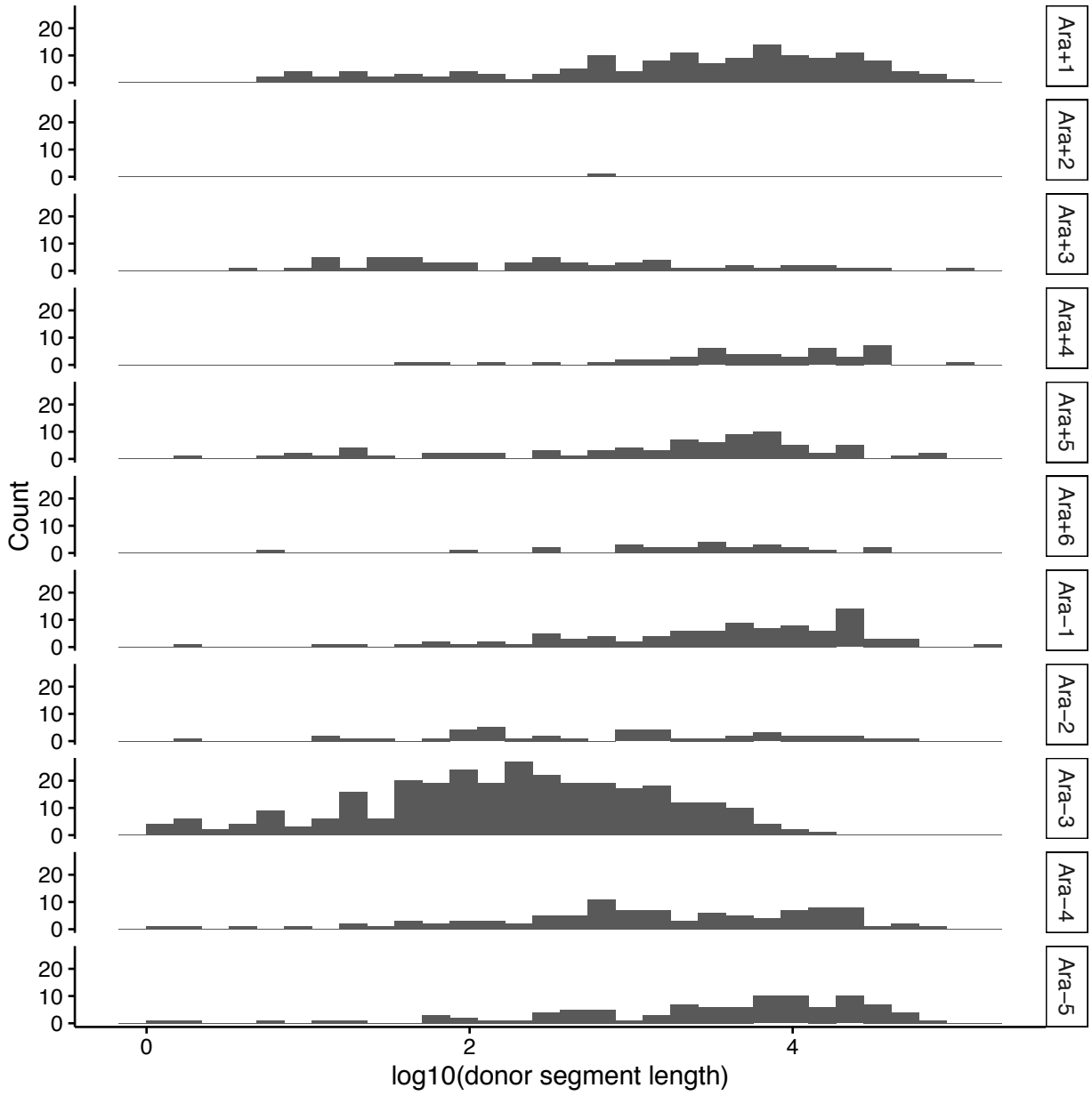


Table 5.3: Putative gene conversion events in recombinant genomes.

Gene annotation	Lineages mutated	Mutation count	Positions mutated
ECB_00510/nohB	3	12	4
ECB_03438	3	51	17
ECB_03538	1	14	14
[ECB_03438]	3	3	1
caiF/caiE	4	12	3
essQ	1	6	6
gmhB/rrsH	3	8	3
nohB	3	15	5
rrlA	1	4	4
thrW/yagP	3	3	1
waaQ	3	16	6
ydfU	1	7	7

Figure 5.3: Length distributions of segments of donor-derived DNA in the odd-numbered recombinant genomes. For population Ara-3, we show the length distribution of segments derived from the recipient genome because it is largely comprised of donor DNA (as seen in Figure 5.1C). The Ara-6 recombinant genome had no donor DNA and is not shown.



LITERATURE CITED

LITERATURE CITED

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK *et al.* 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461: 1243–1247.

Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P. *et al*. 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J* doi: 10.1038/isme.2015.241

Beryln MK. 1998. Linkage map of *Escherichia coli* K-12, edition 10: the traditional map. *Microbiol Mol Biol Rev* 62: 814–984.

Cooper TF. 2007. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. *PLoS Biol* 5: e225.

Deatherage DE, Traverse CC, Wolf LN, Barrick JE. 2015. Detecting rare structural variation in evolving microbial populations from new sequence junctions using *breseq*. *Front Genet* 5:468.

Daegelen P, Studier FW, Lenski RE, Cure S, Kim JF. 2009. Tracing ancestors and relatives of *Escherichia coli* B and the derivation of B strains REL606 and BL21(DE3). *J Mol Biol* 394: 634-643.

Fisher RA. 1958. The genetical theory of natural selection, Ed. 2. New York: Dover. 291 p.

Hickey DA, Rose MR. 1988. The role of gene transfer in the evolution of eukaryotic sex. Pp. 161-175 in *The Evolution of Sex: An examination of Current Ideas* (B.R. Levin & R.E. Michod, Eds.) Sunderland, (MA): Sinauer Associates.

Keightley PD, Otto SP. 2006. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* 443:89–92.

Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* 332:1193–1196.

Levin, BR, Lenski RE. 1985. Bacteria and phage: a model system for the study of the ecology and coevolution of hosts and parasites. Pp. 227-242 in *Ecology and Genetics of Host-Parasite Interactions* (D Rollinson and RM Anderson, Eds.) London: Academic Press.

Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323: 379–382.

McDonald MJ, Rice DP, Desai MM. 2016. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* 531: 233–236.

Rosen MJ, Davison M, Bhaya D, Fisher DS. 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science* 348: 1019–1023.

Rayssiguier C, Thaler DS, Radman M. 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* 342: 396–401.

Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal endemic plant pathogen. *Appl Environ Microbiol* 70: 1999–2012.

Sniegowski PD, Gerrish PJ, Lenski RE. 1997. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387: 703–705.

Souza V, Turner PE, Lenski RE. 1997. Long-term experimental evolution in *Escherichia coli*. V. Effects of recombination with immigrant genotypes on the rate of bacterial evolution. *J Evol Biol.* 10:743–769.

Studier FW, Daegelen P, Lenski RE, Maslov S, Kim JF. 2009. 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.

Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Médigue C, *et al.* 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. bioRXiv preprint http://dx.doi.org/10.1101/036806

Turner PE, Souza V, Lenski RE. 1996. Tests of ecological mechanisms promoting the stable coexistence of two bacterial genotypes. *Ecology* 77: 2119–2129.

Wanner BL. 1986. Novel regulatory mutants of the phosphate regulon in *Escherichia coli* K-12. *J Mol Biol* 191: 39–58.