ARTICLES

Genome evolution and adaptation in a long-term experiment with *Escherichia coli*

Jeffrey E. Barrick[1]*, Dong Su Yu[2,3]*, Sung Ho Yoon[2], Haeyoung Jeong[2], Tae Kwang Oh[2,4], Dominique Schneider[5], Richard E. Lenski[1] & Jihyun F. Kim[2,6]

The relationship between rates of genomic evolution and organismal adaptation remains uncertain, despite considerable interest. The feasibility of obtaining genome sequences from experimentally evolving populations offers the opportunity to investigate this relationship with new precision. Here we sequence genomes sampled through 40,000 generations from a laboratory population of *Escherichia coli*. Although adaptation decelerated sharply, genomic evolution was nearly constant for 20,000 generations. Such clock-like regularity is usually viewed as the signature of neutral evolution, but several lines of evidence indicate that almost all of these mutations were beneficial. This same population later evolved an elevated mutation rate and accumulated hundreds of additional mutations dominated by a neutral signature. Thus, the coupling between genomic and adaptive evolution is complex and can be counterintuitive even in a constant environment. In particular, beneficial substitutions were surprisingly uniform over time, whereas neutral substitutions were highly variable.

Adaptation has often been viewed as a gradual process. Darwin[1] wrote that ''We see nothing of these slow changes in progress, until the hand of time has marked the long lapse of ages…''. Theoretical work in quantitative genetics supported this view by showing that gradual adaptation would result from constant selection on many mutations of small effect[2]. However, an alternative model of evolution on rugged fitness landscapes challenged this perspective[3] and, later, empirical evidence was found for alternating periods of rapid phenotypic evolution and stasis in some lineages[4,5]. The causes of variation in the rate of adaptation remain controversial and are probably diverse. They may include changes in the environment, in circumstances promoting or impeding gene flow, and in opportunities for refinement following the origin of key innovations or the invasion of new habitats, among other factors[6–11].

Genomic changes underlie evolutionary adaptation, but mutations—even those substituted (fixed) in evolving populations—are not necessarily beneficial. Variation in the rate of genomic evolution is also subject to many influences and complications. On the one hand, theory predicts that neutral mutations should accumulate by drift at a uniform rate, albeit stochastically, provided the mutation rate is constant[12]. On the other hand, rates of substitution of beneficial and deleterious mutations depend on selection, and hence the environment, as well as on population size and structure[13,14]. Moreover, the relative proportions of substitutions that are neutral, deleterious and beneficial are usually difficult to infer given imperfect knowledge of any organism's genetics and ecology, in the past as well as in the present.

Experiments with tractable model organisms evolving in controlled laboratory environments minimize many of these complications and uncertainties[15,16]. Moreover, new methods have made it feasible to sequence complete genomes from evolution experiments with bacteria[17–20]. To date, such analyses have focused on finding the mutations responsible for particular adaptations. However, the application of comparative genome sequencing to experimental evolution studies also offers the opportunity to address major conceptual issues, including whether the dynamics of genomic and adaptive evolution are coupled very tightly or only loosely[10,12,13,21,22].

Genome dynamics and adaptation

To examine the tempo and mode of genomic evolution, we sequenced the genomes of *E. coli* clones sampled at generations 2,000, 5,000, 10,000, 15,000, 20,000 and 40,000 from an asexual population that evolved with glucose as a limiting nutrient for almost 20 years as part of a long-term experiment. The complete sequence of the ancestral strain served as a reference for identifying mutations in the evolved clones, which we refer to by their generation abbreviations 2K, 5K, 10K, 15K, 20K and 40K.

Figure 1 shows all mutations identified in the evolved clones through 20,000 generations. The 45 mutations in the 20K clone include 29 single-nucleotide polymorphisms (SNPs) and 16 deletions, insertions and other polymorphisms (DIPs). Figure 2 shows that the number of mutational differences between the ancestral and evolved genomes accumulated in a near-linear fashion over this period. Any deviation from linearity was not statistically significant based on randomization tests.

The near-linearity of the trajectory for genomic evolution is rather surprising, given that such constancy is widely taken as a signature of neutral evolution[12], whereas the fitness trajectory for this population[23] shows profound adaptation that is strongly nonlinear. In particular, the rate of fitness improvement decelerates over time (Fig. 2), which indicates that the rate of appearance of new beneficial mutations is declining, their average benefit is becoming smaller, or both. These effects, in turn, should cause the rate of genomic evolution to decelerate.

To understand this point, consider a simple model of the substitution of beneficial mutations in a clonal population of haploid organisms. A beneficial mutation has an initial frequency of $1/N$,

[1]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA. [2]Industrial Biotechnology and Bioenergy Research Center, Korea Research Institute of Bioscience and Biotechnology, Yuseong, Daejeon 305-806, Korea. [3]Department of Computer Science and Engineering, Chungnam National University, Yuseong, Daejeon 305-764, Korea. [4]21C Frontier Microbial Genomics and Applications Center, Yuseong, Daejeon 305-806, Korea. [5]Institut Jean Roget, Laboratoire Adaptation et Pathogénie des Microorganismes, CNRS UMR 5163, Université Joseph Fourier, Grenoble 1, BP 170, F-38042 Grenoble cedex 9, France. [6]Functional Genomics Program, School of Science, University of Science and Technology, Yuseong, Daejeon 305-333, Korea.
*These authors contributed equally to this work.

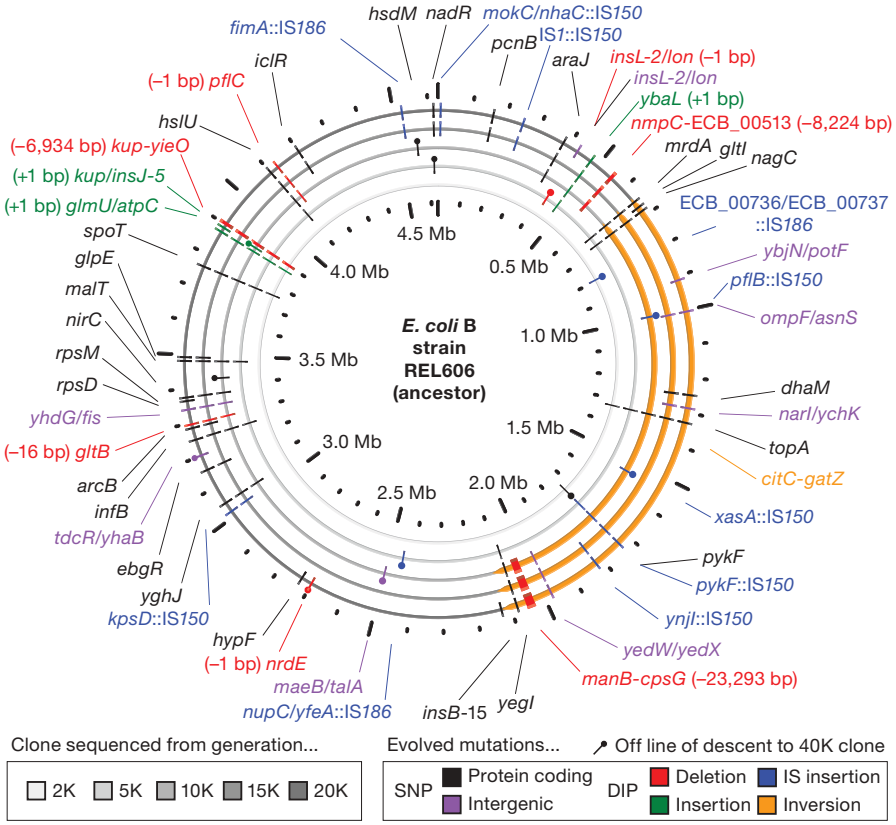

Figure 1 | Mutations found by sequencing genomes sampled between 2,000 and 20,000 generations from an evolution experiment with *E. coli*. The outermost ring represents the genome sampled at 20,000 generations, and labels all genes with SNP mutations in coding (black) and intergenic (purple) regions, and those with DIP mutations including deletions (red), insertions (green), insertion sequence (IS) element insertions (blue), and an inversion between *citC* and *gatZ* (orange). Insertion sequences are transposable elements present in bacterial genomes. The next four rings, from outer to inner, show mutations present in genomes sampled at 15,000, 10,000, 5,000, and 2,000 generations. The innermost circle shows the genome position and scale in megabase pairs (Mb). Mutations that are off the line of descent to a genome sampled at 40,000 generations are capped with a circle. Only one mutation (*kup*/*insJ*-5), a 1-base-pair (bp) insertion near an IS*150* element, shows an aberrant homoplastic distribution, being present in clones 10K and 20K but not 15K. Precise molecular details for all mutations are shown in Supplementary Tables 1 and 2.

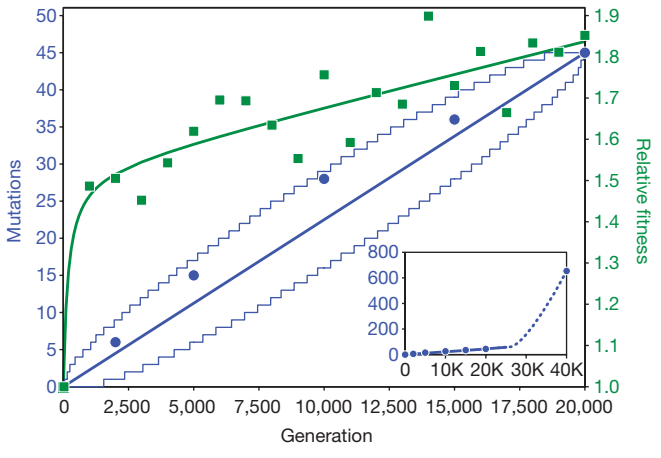

Figure 2 | Rates of genomic evolution and fitness improvement. Blue circles show the total number of genomic changes relative to the ancestor in each sampled clone. The blue line represents a model where mutations accumulate uniformly over time. The light blue curves define the 95% confidence interval for this linear model. Green squares show the improvement of this population's mean fitness relative to the ancestor over time, and the green curve is a hyperbolic plus linear fit of this trajectory. Each fitness estimate is the mean of three assays; most of the spread of points around the fitness trajectory reflects statistical uncertainty inherent to the assays. The inset shows the number of mutations in the 40,000-generation clone; the dashed curve approximates the change in the time course of genomic evolution after a mutator phenotype appeared by about generation 26,500.

where N is population size, and it confers a selective advantage S over its progenitor. Nevertheless, there is some probability that the mutant is lost by drift while it is rare. Given large N, small S and a Poisson distribution of offspring[24], a beneficial mutant has a probability of escaping extinction of $\sim 2S$. If the mutant survives, it takes $\tau \approx \log_2(0.5N)/S$ cell generations to increase to 50% frequency. These dynamics thus have two phases. In the first, a population waits for the appearance of a beneficial mutation that avoids extinction by drift with an expected waiting time of $\omega \approx 1/(2SNv)$, where v is the beneficial mutation rate. In the second phase, the mutant spreads by selection, becoming the new majority type after τ generations.

We can explore the relationship between rates of adaptation and genomic evolution under three scenarios. In the first, the substitution of any beneficial mutation has no effect on either the selection coefficient, S, or the beneficial mutation rate, v. The rates of fitness improvement and genome evolution should therefore be constant over the long term. Under the second scenario, the number of possible sites for beneficial mutations is finite, so that v declines with increasing prior substitutions. The expected wait for a beneficial mutation becomes progressively longer, and the trajectories for adaptation and genomic evolution should decelerate in parallel. In the third scenario, the advantage of new beneficial mutations declines as fitness increases. The waiting and sweep times are both inversely proportional to S, so the total expected time between substitutions is also inversely proportional to S. The rate of fitness gain will decelerate with the reduced rate of beneficial substitutions as well as their declining effects, although the trajectories may not be parallel. Under all three scenarios, this model thus predicts declining rates

of both adaptive and genomic evolution or, alternatively, no deceleration in either trajectory.

Predominance of beneficial substitutions

The simplest hypothesis that could explain the discrepancy between the nearly constant rate of genomic change and the sharply decelerating fitness trajectory posits that only a small fraction of all substitutions are beneficial, whereas most are neutral or nearly so[12,14]. Accordingly, the beneficial substitutions would be concentrated in the early phase of rapid adaptation to the conditions of the experiment, but over time that initial burst would be swamped by the constant accumulation of neutral mutations by drift. However, four lines of evidence allow us to reject this explanation.

First, under this drift hypothesis, one expects disproportionately more synonymous than non-synonymous mutations, because the former have no effect on protein sequence and thus are more likely to be neutral. In fact, all 26 point mutations we found in coding regions (22 in clone 20K, and 4 off the line of descent) are non-synonymous. The probability of observing no synonymous substitutions is only 0.07% if the same base changes were distributed randomly in the coding regions of the ancestral genome.

Second, if mutations had spread by random drift, we would not expect to see mutations in the same genes in the other independently evolved populations of the long-term experiment, because only ~1% of the >4,000 genes in *E. coli* harbour mutations in the population studied here. By contrast, selection should target the same genes in the replicate lines because they started from the same ancestor and evolved in identical environments. Fourteen genes in which mutations were found in our study population have been sequenced in all the other populations after 20,000 generations. There is substantial parallelism, with three cases where all eleven other populations have substituted mutations in the same gene, nine additional genes with mutations in other lines, and only two cases where no other line has a mutation in the same gene (Table 1). In almost all cases, the evolved alleles differ between the populations, so accidental cross-contamination cannot explain these parallel changes.

Third, under the drift hypothesis, we would expect many mutations in individual clones that did not become fixed in the population as a whole. However, almost all mutations in the earlier clones were present in clones from all subsequent generations. For example, four of the six mutations in clone 2K are present in all later clones, and all thirty-four mutations in clone 15K occur in clones 20K and 40K. Moreover, two of the thirteen mutations through 20K that are off the line of descent to the 40K clone occur in genes (*pykF* and *nadR*) where different mutations arose and were substituted later. Both of these genes also have substitutions in all of the other populations, so even these early unsuccessful alleles were probably beneficial, but were nonetheless eliminated because competing sub-lineages had even more beneficial mutations[25,26].

Fourth, strains with these mutations should have no fitness advantage under the neutral drift hypothesis. To date, isogenic strains with ancestral and derived alleles have been constructed at nine loci. In all but one case, the derived allele confers a significant advantage in competition (Table 2). The exception (*ompF*) might also be beneficial in combination with other mutations present in the genetic background in which it evolved, especially because parallel mutations arose in other populations (Table 1). By contrast, another study found that none of 26 random insertion mutations conferred a significant advantage in the same environment[27].

Other explanations for rate discordance

Taken together, these four lines of evidence demonstrate that discordance in rates of genomic and adaptive evolution in this experiment cannot be explained by assuming a preponderance of neutral substitutions. Another plausible explanation for the disparity is an ecological one. Fitness levels were measured, at all generations, in competition with the ancestor. In an evolution experiment with yeast, non-transitive ecological interactions gave rise to complex dynamics, such that the cumulative adaptation measured across successive episodes of selection was greater than that measured directly from start to finish[28]. However, there is no significant discrepancy between the fitness gains summed over shorter intervals and the overall improvement measured from start to finish for the population in our study[23], allowing us to reject this hypothesis.

Clonal interference occurs in asexual organisms when sub-lineages with beneficial mutations are driven extinct by competition with other sub-lineages bearing mutations that are even more beneficial[25,26], and this process might contribute to the relatively constant rate of genomic change. In particular, the most beneficial mutations should dominate the early phase of evolution for large populations in a new environment[26], but there are more potential mutations that confer small advantages than large ones[2,13,29,30]. Thus, the supply of contending beneficial mutations may increase enough to sustain a uniform rate of overall genomic change. It may also be relevant that some early substitutions, which contributed the most to fitness improvement, involve global regulatory functions including the stringent response and DNA supercoiling[31,32]. These mutations have pleiotropic effects on the expression of many genes, and although these changes are beneficial on balance, some of their side effects are probably deleterious. These maladaptive side effects may introduce new opportunities for compensatory changes that restore appropriate expression of other genes and thereby further increase the supply of mutations conferring small advantages.

Emergence of a hypermutable phenotype

Several of the long-term populations evolved mutator phenotypes by 20,000 generations, but the population in our study retained the low ancestral mutation rate to at least that time point[33,34]. In later generations, however, this population exhibited a greatly elevated rate of genomic evolution (Fig. 2 inset). The 40K genome contains 627 SNP and 26 DIP mutations (Supplementary Tables 3 and 4). As a consequence of the DIP mutations (including six new insertions of

Table 1 | Frequency of parallel mutations in 11 other independently evolved lines

Gene or region	Function	Parallel mutations (%)	Source
nadR	Transcriptional regulator	100	Ref. 42
pykF	Pyruvate kinase	100	Ref. 42
rbs operon	Ribose catabolism	100	Ref. 43
malT	Transcriptional regulator	64	Ref. 44
spoT	Stringent response regulator	64	Ref. 31
mrdA	Cell-wall biosynthesis	45	Ref. 42
infB	Translation initiation factor 2	45*	This study
fis	Nucleoid-associated protein	27	E. Crozat, D.S., unpublished
topA	DNA topoisomerase I	27	E. Crozat, D.S., unpublished
pcnB	Poly(A) polymerase	27	This study
ompF	Outer-membrane porin	18*	This study
rpsD	30S ribosomal protein	18*	This study
rpsM	30S ribosomal protein	0	This study
glmU promoter	Cell-wall biosynthesis	0	M. Stanek, R.E.L., unpublished

* In addition to populations with substitutions, one or more others were polymorphic.

Table 2 | Tests of fitness effect in competition between isogenic constructs

Gene or region	Fitness effect (%)	Significance	Source
topA	13.3	***	Ref. 32
*pykF**	11.1	***	D.S., R.E.L., unpublished
spoT	9.4	***	Ref. 31
nadR†	8.1	***	D.S., R.E.L., unpublished
glmU promoter	4.9	***	M. Stanek, T. Cooper, R.E.L., unpublished
fis	2.9	***	Ref. 32
rbs operon†	2.1	***	Ref. 43
malT	0.4	**	Ref. 44
ompF‡	−9.7	**	D.S., R.E.L., unpublished

* For this mutation, isogenic constructs were made by replacing the evolved allele with the ancestral allele in the evolved genetic background. For all other mutations, isogenic constructs were made in the ancestral background.
† In these two cases, artificial deletions of the genes were constructed in the ancestral background and the fitness effects of those deletions are reported.
‡ The deleterious effect of this mutation could indicate that it hitchhiked to high frequency. Alternatively, its fitness effect was tested only in the ancestral background, and it might be beneficial in association with one or more other evolved alleles.
** $P < 0.01$; *** $P < 0.001$. All significance levels are based on multiple independent competition assays.

IS*150*, three of IS*186* and one of IS*1*) the genome size of the 40K clone is 4.57×10^6 bp, representing a reduction of 1.2% from the ancestor. Of particular interest is a 1-bp insertion causing a frameshift mutation in the *mutT* gene. Defects in *mutT* specifically cause A·T→C·G transversions[35], and 92.3% (553 of 599) of the new point mutations at 40K have this signature, significantly higher than the 23.5% (8 of 34) frequency among earlier mutations (one-tailed Fisher's exact test, $P = 3 \times 10^{-20}$).

We sequenced the site of the *mutT* frameshift in clones from other generations to determine the time-course of its fixation in the population. The mutation occurs in 0 of 3 clones tested at generations 20,000, 25,000, 25,500 and 26,000; 1 of 3 clones at generation 26,500; 2 of 3 clones at generations 27,000, 27,500, 28,000 and 28,500; and 3 of 3 clones at generations 29,000, 29,500, 30,000, 35,000 and 40,000. Thus, the *mutT* mutant appeared by generation 26,500 and soon dominated the population. Luria–Delbrück fluctuation tests[33,34] indicate that the mutation rate to nalidixic acid resistance increased about 50- to 100-fold in later generations.

The large number of new SNP mutations at 40K is presumably the result of drift coupled with the elevated mutation rate. Unlike the SNP mutations occurring before 20,000 generations, only a small fraction of these new mutations are likely to be beneficial. To test this prediction, we examine below the proportion of synonymous mutations after the mutator phenotype evolved to determine if it is consistent with a random distribution across sites.

Synonymous changes and mutation rates

The fact that no synonymous mutations fixed in the first 20,000 generations is consistent with the low point-mutation rate in *E. coli* and population-genetic theory if those mutations are selectively neutral. According to theory, the expected rate of neutral substitutions equals the rate of neutral mutations[12]. We calculate an upper bound for the mutation rate from the Poisson distribution, which specifies a 5% chance that no synonymous substitutions would occur even if three were expected. That upper bound corresponds to a mutation rate of 1.6×10^{-10} per bp per generation given 20,000 generations, a genome length of 4.63×10^6 bp, and the fact that 20.4% of all possible point mutations are synonymous. This inferred rate lies between earlier estimates from mutation studies[36] and comparative analyses of *E. coli* and *Salmonella enterica*[37].

In the 40K genome, by contrast, 13.9% (83 of 599) of the new base substitutions (those not in the 20K genome) are synonymous. This fraction is lower than would be expected if 20.4% of random substitutions were synonymous. However, mutations in the 40K genome are highly skewed towards A·T→C·G transversions, which have a lower probability of causing synonymous changes than other point mutations. To reflect this mutational bias, we grouped point mutations into two categories: *mutT*, either A→C or T→G; and non-*mutT*, all other base substitutions. These categories have probabilities of synonymous mutations of 11.3% and 22.1%, respectively, and they are represented by 553 and 46 new mutations, respectively, in the 40K genome. The observation of 83 synonymous substitutions is slightly higher than the random expectation of 71.6 based on the sum of these two binomial distributions, although the excess of synonymous changes is small and only marginally significant at best (one-tailed $P = 0.105$). The small excess of synonymous substitutions implies that a high proportion of late-arising non-synonymous changes are also neutral or nearly so under the conditions of the evolution experiment.

We can also use the number of synonymous substitutions to estimate the point mutation rate after the mutator phenotype evolved. The precise time of origin for the *mutT* subpopulation is unknown, but it was present in the 26,500-generation sample; we assume it arose at generation 25,000 for this estimation. As previously noted, neutral mutations should accumulate at a rate equal to their underlying mutation rate. The lineage leading to the 40K clone accumulated all or almost all of its 83 synonymous substitutions after it became a mutator. Given roughly 15,000 generations, a final genome length of 4.57×10^6 bp, and the fact that only 11.3% of *mutT* point mutations should produce synonymous changes, the 83 such instances imply a point-mutation rate of 1.1×10^{-8} per bp per generation. This rate is about 70-fold higher than the upper bound estimated before the mutator phenotype evolved.

Perspective and outlook

Genome re-sequencing in the context of experimental evolution provides new opportunities for quantifying evolutionary dynamics. We observed discordance between the rates of genomic change and fitness improvement during a 20-year experiment with *E. coli* in two respects. First, mutations accumulated at a near-constant rate even as fitness gains decelerated over the first 20,000 generations. Second, the rate of genomic evolution accelerated markedly when a mutator lineage became established later. The fluid and complex coupling observed between the rates of genomic evolution and adaptation even in this simple system cautions against categorical interpretations about rates of genomic evolution in nature without specific knowledge of molecular and population-genetic processes. Our results also call attention to new opportunities for population-genetic models to explore the long-term dynamic coupling between genome evolution and adaptation, including the effects of clonal interference, compensatory adaptation, and changing mutation rates.

METHODS SUMMARY

Evolution experiment. Twelve *E. coli* populations were propagated at 37 °C for 6,000 days in minimal medium supplemented with limiting glucose at 25 mg l^{-1} by transferring 0.1 ml of culture into 9.9 ml of fresh medium each day[38,39]. The population designated Ara-1 is the focus of this study. Samples were stored periodically at –80 °C and later revived for sequencing and phenotypic analyses.
Genome re-sequencing. On the basis of the sequence[40] of the ancestral strain REL606 (GenBank accession number NC_012967.1), NimbleGen microarray-based comparative genome sequencing[41] was first used to screen for mutations in the 2K and 20K clones. All six evolved clones were then sequenced to >50× coverage using the Illumina 1G platform. Mutations were identified using BRESEQ, a custom computational pipeline. Targeted sequencing was used to

confirm almost all mutations in the 20K and earlier clones, and to find parallel mutations in the other long-term populations.

Mutation trajectory. We performed two randomization tests for deviations from linearity in the rate of genome evolution through 20,000 generations, one based on the cumulative distribution of mutations and the other on a time-weighted test statistic. Each test was performed using three data sets: the total number of mutations, only mutations on the line of descent, and only SNP mutations. None of the one-tailed tests was significant at $P < 0.05$.

Fitness trajectory. Fitness levels were previously measured in competition assays between the Ara-1 population samples and a genetically marked variant of the ancestor[23]. We fit these data to a hyperbolic plus linear model: $w = at/(b + t) + ct$, where w is mean fitness, t is time, and a, b and c are free parameters. The inclusion of the linear term c significantly improves the fit relative to a hyperbolic-only (a, b) model ($F_{1,17} = 7.715$, $P = 0.0129$).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 July; accepted 28 August 2009.
Published online 18 October; corrected 29 October 2009 (see full-text HTML version for details).

1. Darwin, C. *On the Origin of Species by Means of Natural Selection* (Murray, 1859).
2. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, 1930).
3. Wright, S. in *Proc. 6th Int. Cong. Genet.* **1,** 356–366 (1932).
4. Gould, S. J. & Eldredge, N. Punctuated equilibrium: the tempo and mode of evolution reconsidered. *Paleobiol.* **3,** 115–151 (1977).
5. Eldredge, N. *et al.* The dynamics of evolutionary stasis. *Paleobiol.* **31,** 133–145 (2005).
6. Simpson, G. G. *The Major Features of Evolution* (Columbia Univ. Press, 1953).
7. Charlesworth, B., Lande, R. & Slatkin, M. A neo-Darwinian commentary on macroevolution. *Evolution* **36,** 474–498 (1982).
8. Reznick, D. N., Shaw, F. H., Rodd, F. H. & Shaw, R. G. Evaluation of the rate of evolution in natural populations of guppies (*Poecilia reticulata*). *Science* **275,** 1934–1937 (1997).
9. Schluter, D. *The Ecology of Adaptive Radiations* (Oxford Univ. Press, 2000).
10. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314,** 119–121 (2006).
11. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **105,** 7899–7906 (2008).
12. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
13. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, 1991).
14. Ohta, T. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23,** 263–286 (1992).
15. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Rev. Genet.* **4,** 457–469 (2003).
16. Buckling, A., Maclean, C. R., Brockhurst, M. A. & Colegrave, N. The *Beagle* in a bottle. *Nature* **457,** 824–829 (2009).
17. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).
18. Fiegna, F., Yu, Y. T., Kadam, S. V. & Velicer, G. J. Evolution of an obligate social cheater to a superior cooperator. *Nature* **441,** 310–314 (2006).
19. Herring, C. D. *et al.* Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nature Genet.* **38,** 1406–1412 (2006).
20. Hegreness, M. & Kishony, R. Analysis of genetic systems using experimental evolution and whole-genome sequencing. *Genome Biol.* **8,** 201 (2007).
21. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
22. Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392,** 917–920 (1998).
23. de Visser, J. A. G. M. & Lenski, R. E. 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 (2002).
24. Haldane, J. B. S. A mathematical theory of natural and artificial selection. V. Selection and mutation. *Proc. Camb. Philos. Soc.* **23,** 838–844 (1927).
25. Muller, H. J. Some genetic aspects of sex. *Am. Nat.* **66,** 118–138 (1932).
26. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102/103,** 127–144 (1998).
27. Remold, S. K. & Lenski, R. E. Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **98,** 11388–11393 (2001).
28. Paquin, C. E. & Adams, J. Relative fitness can decrease in evolving asexual populations of *S. cerevisiae*. *Nature* **306,** 368–371 (1983).
29. Orr, H. A. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* **56,** 1317–1330 (2002).
30. Perfeito, L., Fernandes, L., Mota, C. & Gordo, I. Adaptive mutations in bacteria: high rate and small effects. *Science* **317,** 813–815 (2007).
31. Cooper, T. F., Rozen, D. E. & Lenski, R. E. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **100,** 1072–1077 (2003).
32. Crozat, E., Philippe, N., Lenski, R. E., Geiselmann, J. & Schneider, D. Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics* **169,** 523–532 (2005).
33. Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* **387,** 703–705 (1997).
34. Cooper, V. S. & Lenski, R. E. The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature* **407,** 736–739 (2000).
35. Friedberg, E. C., Walker, G. C. & Siede, W. *DNA Repair and Mutagenesis* (ASM Press, 1995).
36. Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl Acad. Sci. USA* **88,** 7160–7164 (1991).
37. Ochman, H., Elwyn, S. & Moran, N. A. Calibrating bacterial evolution. *Proc. Natl Acad. Sci. USA* **96,** 12638–12643 (1999).
38. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am. Nat.* **138,** 1315–1341 (1991).
39. Lenski, R. E. Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium *Escherichia coli*. *Plant Breed. Rev.* **24,** 225–265 (2004).
40. Jeong, H. *et al.* Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J. Mol. Biol.* doi:10.1016/j.jmb.2009.09.052 (26 September 2009).
41. Albert, T. J. *et al.* Mutation discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. *Nature Methods* **2,** 951–953 (2005).
42. Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A. & Lenski, R. E. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **103,** 9107–9112 (2006).
43. Cooper, V. S., Schneider, D., Blot, M. & Lenski, R. E. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol.* **183,** 2834–2841 (2001).
44. Pelosi, L. *et al.* Parallel changes in global protein profiles during long-term experimental evolution in *Escherichia coli*. *Genetics* **173,** 1851–1869 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank collaborators in the Lenski and Schneider laboratories for sharing unpublished data; N. Hajela and J. I. Lee for isolation of genomic DNA; C. T. Brown, C. Epstein, C. H. Lee and J. Plotkin for discussion; N. Hajela, L. Ekunwe and S. Simpson for years of technical assistance with the long-term lines; and W. J. Dittmar for assistance with fluctuation tests. We acknowledge support from the DARPA 'Fun Bio' Program (to R.E.L.); the US National Science Foundation (to J.E.B. and R.E.L.); the Agence Nationale de la Recherche Programme 'Génomique Microbienne à Grande Echelle', Centre National de la Recherche Scientifique, and Université Joseph Fourier (to D.S.); and the 21C Frontier Microbial Genomics and Applications Center Program, Ministry of Education, Science and Technology, Republic of Korea (to J.F.K.).

Author Contributions R.E.L., D.S. and J.F.K. conceived the project and its components. D.S.Y., J.E.B., S.H.Y., H.J., T.K.O. and J.F.K. performed the genome sequencing and confirmatory analyses. D.S. sequenced specific genes in other populations and performed additional molecular procedures. J.E.B. developed code for data analyses and statistical simulations. R.E.L. directs the long-term experiment while J.F.K. directed the genomics work. R.E.L., J.E.B. and J.F.K. analysed the data and wrote the paper. J.E.B., D.S.Y., S.H.Y., R.E.L., D.S. and J.F.K. prepared figures and tables.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.E.L. (lenski@msu.edu) or J.F.K. (jfk@kribb.re.kr).

METHODS

Genomic DNA isolation. Clones 2K (strain REL1164A), 5K (REL2179A), 10K (REL4536A), 15K (REL7177A), 20K (REL8593A) and 40K (REL10938) from the Ara-1 population of the *E. coli* long-term evolution experiment[38,39] were revived from stocks kept at –80 °C in 15% glycerol by growth overnight in LB medium. DNA was harvested and purified from several millilitres of each culture with the Qiagen Genomic-tip 100/G kit.

NimbleGen comparative genome sequencing (CGS). Genomic DNA from the 2K and 20K clones was sent to NimbleGen Systems for comparative genome sequencing. Tiling arrays with 29-mer probes optimized for length, melting temperature, and mismatch position based on the REL606 ancestral genome sequence[40] were used to detect mutational differences between the ancestor and each evolved clone. Data were visualized using SignalMap version 1.8. All putative SNP sites and candidates for DIP variations (deletions, insertions and inversions) were checked by capillary Sanger sequencing of PCR products amplified directly from the genome. The NimbleGen CGS approach did not find several mutations that had been previously discovered in evolved clones (see later), indicating that a high false-negative rate was an impediment to this approach.

Illumina whole-genome shotgun (WGS) re-sequencing. Genomic DNA samples from all six Ara-1 clones were sequenced on a 1G Genome Analyzer (Illumina) by Macrogen. Standard procedures produced data sets of opposite-strand, mated 36-bp read pairs with inserts averaging ~135 bp and calibrated base quality scores. Mutations in each genome were then identified using BRESEQ, a pipeline implemented in Perl for analysing bacterial WGS re-sequencing data (J. E. Barrick, unpublished algorithm). The average coverage at positions with only unique read alignments in each of the six sequenced genomes was between 55× and 65×. More than 99.9% of the positions outside of repeated regions in each genome had at least tenfold coverage. These levels are sufficient to ensure with great confidence that SNP mutations at virtually all positions in the reference genome would be discovered if present. For additional verification of mutations predicted by BRESEQ, we also used the software MAQ[45] (version 0.7.1), which predicted the same set of base substitutions in each WGS re-sequencing data set.

BRESEQ pipeline for mutation discovery. BRESEQ analyses gapped matches between each read sequence and the reference genome produced by MUMmer[46] (version 3.20) with minimal exact match and extension requirements of 14 bp. For each read, it first determines the set of best alignments that are not contained within other matches and that have fewer mismatches than other alignments with the same endpoints. Of these matches, those that do not contain at least three of the four possible DNA bases or where 80% or more of the length of the match is a single base are eliminated. These nearly homopolymeric matches are typically spurious read sequences. For remaining matches that have a unique best match in the reference genome, internal ambiguities in gapped alignments are systematically re-aligned to consistent reference genome positions by using a Needleman–Wunsch algorithm[47]. The ends of each alignment are also trimmed whenever it is possible that mutations introducing small indels would be compatible with an alternative, equally valid alignment to the reference sequence.

On the basis of unique alignments, BRESEQ predicts base changes and short indels from the calibrated base quality scores that support and contradict each candidate mutation. New sequence junctions (such as those produced by deletions) are predicted from hybrid reads consisting of two regions with best matches to discontinuous sites in the reference genome. The junction prediction procedure involves first assembling a list of candidate junction sequences compatible with these matches, then re-querying hybrid reads against these candidates with MUMmer, and finally predicting consensus junctions from reads that match the new candidates better than they match any portion of the original genome. Additionally, deletions too large to be identified as gaps in individual read alignments are predicted by taking all reference positions with zero unique coverage and propagating outward until the unique coverage at a position exceeds an arbitrary cutoff.

BRESEQ generates HTML output files with the genomic contexts of putative mutations annotated using BioPerl[48]. These tables are linked to alignments of read sequences so that information supporting and contradicting each predicted mutation can be examined in more detail by the user. Additionally, coverage histograms are generated for the entire genome and for each large deletion with the R statistics package[49]. These output files were used to determine the precise extent of deletions and the locations of new sequence junctions. BRESEQ performs a reference-based comparison, rather than *de novo* assembly and therefore this procedure may not reliably detect point mutations, indels and genomic rearrangements involving repeated sequences that occur at several locations in the genome, or short tandem repeats that approach or exceed the 36-bp size of the input reads.

Mutation identification. Supplementary Table 1 shows the precise genomic locations and other details for all 34 SNP mutations found in the genomes of the Ara-1 2K, 5K, 10K, 15K, and 20K clones. Supplementary Table 2 provides details for the 21 DIP mutations identified in these genomes, which include a large inversion, three 1-bp insertions, three 1-bp deletions, four more extensive deletions, and ten IS element insertions.

The following seven SNP mutations were discovered before genome re-sequencing: *mrdA* (ref. 42), *ompF*/*asnS* (D. Schneider, unpublished), *topA* (ref. 32), *yhdG*/*fis* (ref. 32), *malT* (ref. 44), *spoT* (ref. 31) and 20K *nadR* (ref. 42). Four of these known SNP mutations (*mrdA*, *yhdG*/*fis*, *spoT* and 20K *nadR*) were found independently using NimbleGen CGS, but the other three were not. Of the remaining 23 SNP mutations present in either the 2K or 20K genomes, NimbleGen CGS identified all except five (*insL*-2/*lon*, *nagC*, 2K *pykF*, *insB*-15 and *hypF*). By contrast, analysis of Illumina WGS data identified all seven of the known mutations and 33 of the 34 total SNP mutations reported. The final SNP mutation (*insB*-15) is within a multicopy IS element and was found in this study during additional Sanger sequencing. We also verified that the *kup*/*insJ*-5 1-bp insertion on the border of an IS*150* element shows an aberrant homoplastic distribution, being present in clones 10K and 20K but not 15K, with additional targeted sequencing.

Five DIP mutations were discovered before genome re-sequencing: Δ(*nmpC*-ECB_00513) (ref. 50), inv(*citC*-*gatZ*) (ref. 50), *pykF*::IS*150* (ref. 50), *glmU*/*atpC* (M. Stanek and R.E.L., unpublished), and Δ(*kup*-*yieO*) (ref. 43). Four of these mutations were found independently using NimbleGen CGS, with the exception being the 1-bp insertion in the *glmU*/*atpC* intergenic region. Five of the twelve other DIP mutations present in the 2K and 20K genomes were first discovered using NimbleGen CGS (*ynjI*::IS*150*, Δ(*manB*-*cpsG*), *kpsD*::IS*150*, Δ*gltB*, *fimA*::IS*186*), and the 1-bp insertion in the *kup*/*insJ*-5 intergenic region was found during additional Sanger sequencing. Analysis of Illumina WGS data identified 20 of the 21 DIP mutations reported, all except inv(*citC*-*gatZ*). New IS-insertions predicted by WGS re-sequencing were confirmed as size changes in PCR-amplified fragments.

We identified a total of 627 SNP and 26 DIP mutations in the 40K clone genome (Supplementary Tables 3 and 4). In addition to the 15 DIP mutations in the 20K clone that are on the line of descent, the 11 other DIP mutations include four IS-insertions, two 1-bp insertions, one 6-bp insertion, one 1-bp deletion, one 61-bp deletion, and two larger deletions of roughly 7 and 22 kb. Only two mutations that are present in the 20K clone (*tdcR*/*yhaB* and *nrdE*) are not found at 40K. One of the additional IS-insertions in the 40K clone is an IS*186* element in the *nupC*/*yfeA* intergenic region at the exact site where one occurs in the 5K clone. Because this mutation is missing in all of the sequenced clones from intervening generations, this insertion is the second example of a homoplastic change that evidently originated independently in two sub-lineages. Also, one of the new 1-bp insertions adds another G directly adjacent to the previous *kup*/*insJ*-5 1-bp insertion. We did not further confirm most of the predicted mutations in the 40K genome owing to the large number, but their quality is on par with the mutations that were correctly predicted in the earlier clones from WGS data.

Parallel mutations. We PCR-amplified and sequenced the *infB*, *pcnB*, *ompF*, *rpsD* and *rpsM* genes of three clones isolated at 20,000 generations from each of the 11 other experimental populations. Mutations were counted as parallel changes if they were within the protein coding sequence or upstream promoter elements, but not if they were downstream of the reading frame.

Fluctuation tests. We performed Luria–Delbrück fluctuation tests[33] to confirm that the Ara-1 population evolved an elevated mutation rate. Bacteria were revived from frozen stocks by growth overnight in LB medium. After dilution and 24 h of re-growth in Davis minimal medium supplemented with 25 mg l^{-1} glucose, we inoculated 24 replicate 10-ml cultures of Davis minimal medium with 250 mg l^{-1} glucose with 100–1,000 cells. After 24 h of growth to stationary phase, these cultures were concentrated by centrifugation and plated on LB agar containing 20 μg ml^{-1} nalidixic acid. The mutation rates to resistance of the mixed populations archived at 20,000, 30,000 and 40,000 generations were estimated as 5.8×10^{-10}, 1.7×10^{-8} and 6.3×10^{-9} per cell division by the maximum likelihood method using a custom Perl script. These values are roughly 4, 120 and 45 times the mutation rate of 1.4×10^{-10} per cell division estimated in the same experimental block for the ancestral strain.

Tests of constant rate of genomic evolution. We examined three data sets consisting of the total number of SNP and DIP mutations in the 2K, 5K, 10K, 15K and 20K genomes (TOT: 6, 15, 28, 36, 45), only those changes on the line of descent to the 40K clone (LOD: 4, 12, 22, 36, 43), and only the SNP mutations in each genome (SNP: 3, 9, 16, 22, 29). The LOD subset represents the estimated rate at which mutations fixed in the population; it excludes the effects of sampling within-population variation, which are unknown and may fluctuate over time given the effect of selective sweeps in purging variation. The SNP subset

represents those mutations that are most commonly modelled in theoretical studies of evolutionary biology. We tested the hypothesis that the rate of mutation accumulation was linear in two ways.

Cumulative distribution test. We used a randomization test to determine whether there was evidence that the number of mutations observed in the 2K, 5K, 10K, 15K and 20K genomes was inconsistent with an underlying probability distribution giving a linear increase over time. For each data set, 10 million simulated data sets were generated by randomly redistributing 20K mutations with a uniform probability at times between 0 and 20,000 generations. The confidence intervals shown in Fig. 1 and also *P*-values for deviations from linearity at a given generation were determined directly from these simulated distributions of mutation number over time. This test is similar in design to the Kolmogorov–Smirnov test, but it takes into account the uneven spacing of the observations of mutation number with respect to time and enforces stricter confidence limits on observations near the ends of the cumulative distribution. If any point from the observed genomic data set fell outside of the 95% confidence interval then the null hypothesis of linearity would be rejected. The most extreme deviations are not significant even by a one-tailed test for observing more mutations than would be expected from the uniform distribution for any of the three data sets: TOT (10K, $P = 0.052$), LOD (15K, $P = 0.092$), SNP (5K, $P = 0.227$).

Time-weighted test statistic. We tested the apparent skew towards excess mutations in early generations using a time-weighted test statistic equal to the summation of the number of new mutations observed in each subsequent genome multiplied by the generation at which that genome was sampled. For example, the value of this test statistic for the TOT data set is 487,000 (equal to $6 \times 2{,}000 + 9 \times 5{,}000 + 13 \times 10{,}000 + 8 \times 15{,}000 + 9 \times 20{,}000$). We determined the significance of the observed value of the test statistic for each data set using a randomization test implemented with the STATISTICS101 resampling program (http://www.statistics101.net/). The mutations at 20K were redistributed randomly with a uniform probability distribution with respect to time to generate randomized data sets consisting of counts of mutations occurring by 2,000, 5,000, 10,000, 15,000 and 20,000 generations. Then, one-tailed *P*-values were assigned to each data set by comparing the observed value of the test statistic to the distribution of simulated test-statistic values obtained from performing the randomization procedure 10 million times. The skew of excess mutations towards early generations, as measured by this test statistic, is not significant for any of the three data sets: TOT ($P = 0.066$), LOD ($P = 0.256$), or SNP ($P = 0.299$).

Lack of early synonymous substitutions. In the 2K, 5K, 10K, 15K and 20K clone genomes, 26 different point mutations within protein reading frames were observed, and all of these altered the encoded amino acid. Using the total codon frequencies in the ancestral genome sequence tabulated using a custom Perl script, we calculated the chances that each of the 12 possible nucleotide substitutions would result in a non-synonymous substitution if it occurred at a random coding position (A→C: 0.893; A→G: 0.758; A→T: 0.898; C→A: 0.762; C→G: 0.808; C→T: 0.559; G→A: 0.698; G→C: 0.785; G→T: 0.785; T→A: 0.797; T→C: 0.600; T→G: 0.840). From these probabilities, it is straightforward to calculate the 0.07% probability of observing zero synonymous mutations, by chance alone, given the observed distribution of the 26 coding base substitutions (A→C: 2; A→G: 2; A→T: 3; C→A: 2; C→T: 3; G→A: 7; G→C: 1; G→T: 1; T→A: 1; T→G: 4). The observed excess of non-synonymous substitutions is highly significant.

Synonymous substitution rates. Reading frames annotated in the ancestral genome sequence were extracted and analysed to calculate the probability that a random mutation would be synonymous using a custom Perl script. These calculations assume that the relative rates of different base changes are equal, except as indicated otherwise for the *mutT* mutator.

45. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).
46. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
47. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).
48. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12,** 1611–1618 (2002).
49. R Development Core Team. R: A language and environment for statistical computing Pages ⟨http://www.R-project.org⟩. (R Foundation for Statistical Computing, 2007).
50. Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. & Blot, M. Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics* **156,** 477–488 (2000).