Supporting Information

Wielgoss et al. 10.1073/pnas.1219574110

SI Text

We used the ratio of the rates of nonsynonymous (dN) to synonymous (dS) substitutions to estimate the fraction of nonsynonymous changes that are deleterious (f_d). Here, we examine several confounding factors that might, in principle, affect the observed dN/dS ratio including codon bias and selection, GC mutational skew, selection for higher GC content, and bottleneck effects. As explained below, these other factors would have had negligible effects in the long-term experiment and left no signatures in the evolved genomes. We also discuss whether the *mutY* gene is a special case in certain respects.

Codon Bias and Selection. The rate of accumulation of synonymous substitutions provides an estimate of the genomic mutation rate, based on the reasoning that synonymous mutations are neutral and thus should accumulate at a rate that depends only on the mutation rate and not on selection (1, 2). However, organisms differ in their use of synonymous codons, and these differences indicate not only mutational biases but also selection for preferred codons that may reflect translational efficiency, accuracy, or both (3–6). Our analyses examined the effects of the *mutT* and *mutT mutY* backgrounds on the mutational spectra, but we did not adjust for selection on codon use because its fitness effects are extremely small (4–6). According to one recent study (6), the strength of selection for optimal codon use for the 40 most highly expressed genes in *E. coli* is roughly equal to the reciprocal of its effective population size, and selection is even weaker for genes with lower expression levels. The effective population size for *E. coli* in nature is unknown, but a recent estimate (7) based on sequence diversity across hundreds of genes put the size at $\sim 2.5 \times 10^7$. The reciprocal of that value implies that the selection coefficient against a suboptimal synonymous change would be $<10^{-6}$ per generation, and thus several orders of magnitude weaker than the selection to reduce the load of deleterious nonsynonymous mutations (Table 2). However, a recent study of two highly expressed genes in *Salmonella typhimurium* reported that synonymous changes had unexpectedly large effects on fitness (8). These effects were not correlated with suboptimal codon use, but they may indicate changes in mRNA stability or structure (8). In any case, one would expect highly expressed genes to be subject to stronger selection, on average, and thus exhibit greater conservation of synonymous sites than typical genes. Therefore, we asked whether the synonymous mutations in our study were less likely to occur in the 40 highly expressed genes (6). These genes comprise 1.3% of the coding sequences, and 4/235 (1.7%) of the unique synonymous mutations in our study occurred in those genes (binomial test, $P = 0.55$). Taken together, there is no indication that selection on codon use or transcript integrity had any appreciable effect on our estimates of mutation rates and genetic loads.

GC Mutational Skew. We examined whether mutational skew affected the accuracy of our mutation rate estimates. In many organisms, the complementary DNA strands have somewhat different G–C and A–T ratios, a property called GC skew (9, 10). This skew may be caused, at least in part, by deamination of C to T in single-stranded DNA, which leads to C-to-T changes on the leading strand. However, none of the 414 changes in the *mutT* background, and only a few in the *mutT mutY* backgrounds (2/274 for *mutT mutY*-E and 2/319 for *mutT mutY*-L), were C:G to T:A changes, and these included only a single synonymous mutation in each *mutT mutY* background. Instead, the mutational spectra (Table 1) were dominated by the transversion biases typical for *mutT* (414/414 changes) and *mutT mutY* backgrounds (271/274 for *mutT mutY*-E and 316/319 for *mutT mutY*-L).

Selection for Higher GC Content. A recent study found evidence that *E. coli* strains carrying highly expressed genes with artificially increased GC content at synonymous sites grew faster than their native counterparts, and this effect was independent of codon use (11). This selection opposes the mutational bias toward increased AT content seen in many species (11) including the ancestral strain in our study (2). All 48 synonymous changes that we observed in the *mutT* background were A:T→C:G mutations (Table 2) and thus increased GC content. By contrast, both *mutY* mutations reduced the rate of A:T→C:G mutations but raised the rate of C:G→A:T mutations. Thus, GC content would have increased faster in the *mutT*-only background than in the *mutT mutY* backgrounds, and therefore, any selection for increased GC content per se would have opposed (and cannot explain) the parallel rise of the two *mutY* alleles.

Bottleneck Effects. The daily serial transfers during the long-term evolution experiment caused demographic bottlenecks. In principle, bottlenecks reduce the efficacy of selection, allowing the accumulation of deleterious mutations that selection would eliminate in an infinite population. However, the minimum population size in the experiment is $\sim 3 \times 10^6$, and the effective size taking the bottlenecks into account is $>10^7$ (12); thus, the fixation of deleterious mutations with selection coefficients $>1/10^7$ by pure drift is very unlikely and, moreover, would require millions of generations (1). Selection for beneficial mutations also reduces the effective population size, and this effect could allow more deleterious mutations to spread by hitchhiking (rather than by simple drift) on a shorter timescale. As a consequence of hitchhiking, some nonsynonymous mutations in the sequenced genomes might be deleterious rather than neutral or beneficial. This effect would lead us to underestimate the fraction of deleterious mutations (f_d), and hence, we would underestimate the genetic load and the strength of selection to reduce it. However, neither the demographic nor selection bottleneck effects influence the mutation rate we estimated from the accumulation of neutral mutations (using synonymous mutations as a proxy thereof), because this calculation depends on the number of generations between two sequenced genomes but not on the population size (1).

Is *mutY* a Special Case? One way that the population we studied could have evolved a lower mutation rate would have been by reverting the mutation in *mutT*. The resulting reduction in the genetic load would have been larger than for the *mutY* mutations that reduced the mutation rate by about half, and the fitness gain would have been correspondingly greater. One factor that contributed to the emergence of the *mutY* alleles is that they were loss-of-function mutations, and thus, the *mutY* gene presented a larger target for mutations to compensate for the hypermutability caused by the earlier *mutT* mutation.

However, the target size for mutations might not be the only relevant factor. The *mutT mutY* backgrounds had mutation rates that were about half that of the *mutT*-only background, but they were still ~50-fold higher than the corresponding rate for a revertant to the ancestral *mutT* state. As a consequence, a *mutY* mutant was ~50 times more likely to gain an additional boost from a subsequent beneficial mutation than would have been a revertant, and the *mutY* mutants would still have enjoyed half the per capita rate of beneficial mutations as the *mutT*-only background. Thus, the *mutT mutY* genotype might be closer to the “sweet spot” with respect to the tension between adaptation and load reduction when the pace of adaptation has slowed but not stopped. Further theoretical work or numerical simulations might shed new light on this hypothesis.

The mutations in *mutT* and *mutY* interact epistatically with respect to their effects on the mutation rate; mutations in either gene alone cause an increase in the overall point-mutation rate, but a mutation in *mutY* reduces the rate if a *mutT* mutation is already present. This interaction has the interesting consequence that it might trap an asexual lineage in a hypermutable state. If the mutation rate for the *mutY*-only genotype was higher than that for the *mutT mutY* genotype (as is true for *mutT*-only genotypes), then the double mutant would occupy a local minimum for genetic load, and the trap would be difficult to escape. In fact, however, the *mutY*-only mutation rate is lower than the *mutT mutY* rate (13), and thus, the ancestral rate can re-evolve by successive reversions in *mutT* and *mutY*, each of which reduces the genetic load.

1. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
2. Wielgoss S, et al. (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* 1(3): 183–186.
3. Sharp PM, Li WH (1987) The codon Adaptation Index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15(3): 1281–1295.
4. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res* 33(4): 1141–1153.
5. Henry I, Sharp PM (2007) Predicting gene expression level from codon usage bias. *Mol Biol Evol* 24(1):10–12.
6. Sharp PM, Emery LR, Zeng K (2010) Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci* 365(1544):1203–1212.
7. Charlesworth B (2009) Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10(3):195–205.
8. Lind PA, Berg OG, Andersson DI (2010) Mutational robustness of ribosomal protein genes. *Science* 330(6005):825–827.
9. Lobry JR, Sueoka N (2002) Asymmetric directional mutation pressures in bacteria. *Genome Biol* 3(10):research0058.
10. Lobry JR, Louarn J-M (2003) Polarisation of prokaryotic chromosomes. *Curr Opin Microbiol* 6(2):101–108.
11. Raghavan R, Kelkar YD, Ochman H (2012) A selective force favoring increased G+C content in bacterial genes. *Proc Natl Acad Sci USA* 109(36):14504–14507.
12. 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(6):1315–1341.
13. Fowler RG, et al. (2003) Interactions among the *Escherichia coli mutT*, *mutM*, and *mutY* damage prevention pathways. *DNA Repair (Amst)* 2(2):159–173.

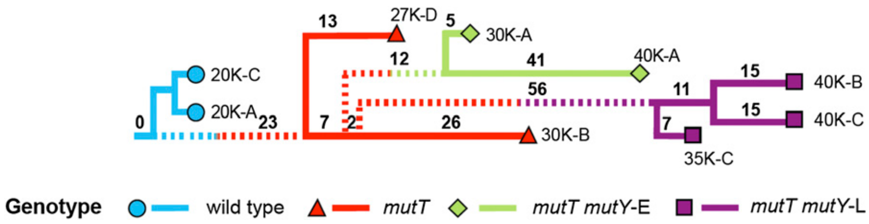

Fig. S1. Inferred history of accumulation of synonymous mutations after 20,000 generations. The phylogeny corresponds to the later part of the tree shown in Fig. 1*A*. The sequenced clones and branches are colored according to mutator genotypes. The bold numbers along each branch show the number of synonymous mutations.

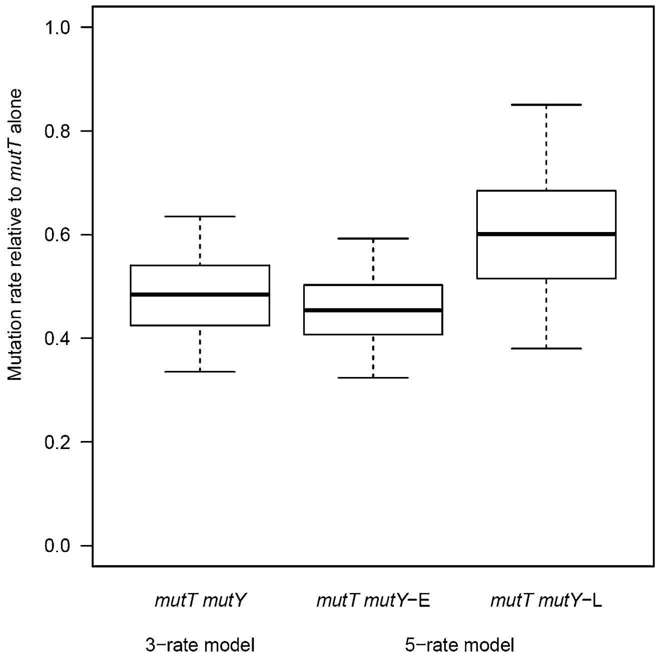

Fig. S2. Effects of evolved *mutY* alleles on mutation rates based on two maximum likelihood models. Rates are expressed relative to the *mutT* background, and they are calculated using synonymous mutations only. The box plot on the left is based on the three-rate model, in which the *mutT mutY*-E and *mutT mutY*-L backgrounds have identical mutation rates. The two box plots on the right are based on the five-rate model, in which these two backgrounds have different mutation rates. Each plot summarizes the probability distribution of the relative mutation rate, where the box indicates the upper and lower quartiles, the heavy line the median, and the whiskers the 95% confidence interval.

Table S1. *E. coli* strains sequenced in this study

Strain	Description*	Mutator alleles[†]	Accession no.[‡]	Coverage[¶]
REL1164A	2K-A		SRS007214	52.8
REL1164B	2K-B		ERS068522	37.5
REL1164C	2K-C		ERS068520	43.9
REL2179A	5K-A		SRS007215	62.0
REL2179B	5K-B		ERS068533	107.0
REL2179C	5K-C		ERS068524	75.2
REL4536A	10K-A		SRS007216	59.4
REL4536B	10K-B		ERS068527	70.8
REL4536C	10K-C		ERS068525	77.8
REL7177A	15K-A		SRS007217	59.7
REL7177B	15K-B		ERS068528	76.3
REL7177C	15K-C		ERS068532	65.4
REL8593A	20K-A		SRS007218	54.1
REL8593B	20K-B		ERS068523	35.0
REL8593C	20K-C		ERS068521	43.5
REL11395	27K-D	*mutT*	ERS068534	222.6
REL10391	30K-A	*mutT mutY*-E	ERS068531	76.9
REL10392	30K-B	*mutT*	ERS068530	70.2
REL10707	35K-C	*mutT mutY*-L	ERS068535	301.9
REL10938	40K-A	*mutT mutY*-E	SRS007219	60.3
REL10939	40K-B	*mutT mutY*-L	ERS068526	74.9
REL10940	40K-C	*mutT mutY*-L	ERS068529	62.7

All the strains are evolved clones sampled from population Ara–1 of the long-term evolution experiment.

*Generation number and letter identifying particular clones (e.g., 20K-A and 20K-B are two clones sampled at generation 20,000).

[†]Known mutations affecting mutation rates are *mutT* (insertion of one C at genome position 114,034), *mutY*-E (T→G mutation at position 2,988,792 causing Leu-to-Trp substitution at amino acid 40), and *mutY*-L (T→G mutation at position 2,989,164 causing Leu to Stop at amino acid 164).

[‡]Previously published (ref. 1) short-read data were deposited in the National Center for Biotechnology Information Sequence Read Archive (SRS numbers); new sequence data obtained during this work were deposited in the European Nucleotide Archive Sequence Read Archive maintained by the European Bioinformatics Institute (ERS numbers).

[¶]Average depth of sequencing coverage at positions in the ancestral genome based on uniquely mapped reads only. The average coverage was at least 35× for each clone; this level provides very high confidence in the discovery of point mutations in the 97.5% of the ancestral genome that excludes multicopy repeat sequences.

1. Barrick JE, et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461(7268):1243–1247.

Table S2. Mutation-rate models and their likelihoods using synonymous mutations only

Genetic background							
mutT		*mutT mutY*-E		*mutT mutY*-L			
A:T→C:G	C:G→A:T	A:T→C:G	C:G→A:T	A:T→C:G	C:G→A:T	No. of parameters	Log(Lk) and parameter penalty
x	0	x	y	x	y	2	−51.9
x1	0	x2	y	x2	y	3	−42.6–2
x1	0	x2	y1	x3	y2	5	−40.7–6

In all models, the lower limit for the origin of the *mutT* allele was set at 20,000 generations.

Table S3. Mutation-rate models and their likelihoods using all point mutations

Genetic background							
mutT		*mutT mutY*-E		*mutT mutY*-L			
A:T→C:G	C:G→A:T	A:T→C:G	C:G→A:T	A:T→C:G	C:G→A:T	No. of parameters	Log(Lk) and parameter penalty
x	0	x	y	x	y	2	−135
x1	0	x2	y	x2	y	3	−66.4–2
x1	0	x2	y1	x3	y2	5	−60.6–6

In all models, the lower limit for the origin of the *mutT* allele was set at 20,000 generations.