Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing

Heewook Lee[a], Ellen Popodi[b], Haixu Tang[a], and Patricia L. Foster[b,1]

[a]School of Informatics and Computing and [b]Department of Biology, Indiana University, Bloomington, IN 47405

Edited by Paul Modrich, Duke University Medical Center, Durham, NC, and approved August 24, 2012 (received for review June 18, 2012)

Knowledge of the rate and nature of spontaneous mutation is fundamental to understanding evolutionary and molecular processes. In this report, we analyze spontaneous mutations accumulated over thousands of generations by wild-type *Escherichia coli* and a derivative defective in mismatch repair (MMR), the primary pathway for correcting replication errors. The major conclusions are (*i*) the mutation rate of a wild-type *E. coli* strain is $\sim 1 \times 10^{-3}$ per genome per generation; (*ii*) mutations in the wild-type strain have the expected mutational bias for G:C > A:T mutations, but the bias changes to A:T > G:C mutations in the absence of MMR; (*iii*) during replication, A:T > G:C transitions preferentially occur with A templating the lagging strand and T templating the leading strand, whereas G:C > A:T transitions preferentially occur with C templating the lagging strand and G templating the leading strand; (*iv*) there is a strong bias for transition mutations to occur at 5′ApC3′/3′TpG5′ sites (where bases 5′A and 3′T are mutated) and, to a lesser extent, at 5′GpC3′/3′CpG5′ sites (where bases 5′G and 3′C are mutated); (*v*) although the rate of small (≤4 nt) insertions and deletions is high at repeat sequences, these events occur at only 1/10th the genomic rate of base-pair substitutions. MMR activity is genetically regulated, and bacteria isolated from nature often lack MMR capacity, suggesting that modulation of MMR can be adaptive. Thus, comparing results from the wild-type and MMR-defective strains may lead to a deeper understanding of factors that determine mutation rates and spectra, how these factors may differ among organisms, and how they may be shaped by environmental conditions.

evolution | mutation accumulation | neutral mutation | mutational hotspots | indels

Mutations are the source of variation upon which natural selection acts; thus, a complete understanding of evolutionary processes must include an accurate assessment of mutation rates and of the molecular spectrum of mutational events. In addition, we need to know whether, and how, intrinsic and extrinsic factors influence mutational processes. This understanding must be founded on baseline parameters established by analyzing mutations that accumulate in a neutral fashion, unbiased by selective pressures. Much of our knowledge of spontaneous mutation is based on mutations that occur in nonessential reporter genes during short-term laboratory culture of microorganisms (1). An alternative approach is to compare presumably neutral mutations that have accumulated over evolutionary time periods in diverged species (2). Both methods have substantial uncertainties. The experimental approach may use reporter loci that are not representative of the whole genome and necessarily incorporates assumptions about the expression and neutrality of the mutant phenotypes. The historical approach relies on estimated divergence times and the absence of selective pressure on synonymous sequence changes. High-throughput whole-genome sequencing allows some of these limitations to be overcome. For example, a recent study analyzed the base-pair substitutions (BPSs) that arose in *Escherichia coli* strains that had been subjected to long-term evolution studies (3). Although the contribution of selection was minimized by considering only synonymous BPSs, selection based on codon usage nonetheless may have biased the results (4).

The mutation-accumulation (MA) strategy combined with whole-genome sequencing overcomes many of these limitations. The MA protocol is designed to allow mutations to occur in a neutral manner, devoid of selective pressure (5). The general strategy is to establish a number of clonal populations from a founder individual and then to take each population through repeated single-individual bottlenecks for thousands of generations. Because the effective population size of each line is one, genetic drift prevents selection from eliminating all but the most deleterious mutations, which typically are less than 1% of mutations (6). For microorganisms, streaking for single colonies on agar medium accomplishes the bottleneck, allowing around 30 generations of growth between passages. Selection within a colony is minimal because most new mutations arise during the last few generations of growth when the population is large. Furthermore, the great majority of mutations have no fitness effects, and of those that do, the effects are small (6–9). The application of whole-genome sequencing to MA lines has made this protocol an extremely valuable way to determine a complete and nearly unbiased picture of mutation profiles. Recent results from applying this protocol to microbes include the genome-wide spontaneous mutational spectrum in *Saccharomyces cerevisiae* (10) and the mutational consequences of loss of DNA oxidative damage repair in *Salmonella typhimurium* (11).

Because of its decades of use as a model for genetic and physiological research, *Escherichia coli* is the ideal microorganism for investigations at the genomic level. However, there still are questions about mutational processes in this best-known-of-all microbe. For example, in the papers cited above (1, 3) estimates of the mutation rate vary by an order of magnitude. The experiments reported here were designed to yield a highly accurate estimate of the spontaneous mutation rate of *E. coli*. In addition, by obtaining a large number of neutral mutations, we anticipated determining the spectrum of mutagenic changes across the whole genome at a greater density than has been possible previously.

The remarkably low rates at which mutations occur result from both the high intrinsic accuracy of DNA replication and various enzymatic activities that survey and repair DNA. Chief among these is mismatch repair (MMR), which surveys newly replicated DNA, detects mismatched bases, recruits enzymes to destroy the new DNA strand, and forces repolymerization using the old DNA strand as the template (reviewed in ref. 12). MMR is highly

Author contributions: P.L.F. designed research; E.P. performed research; H.L. and H.T. contributed new reagents/analytic tools; H.L., E.P., H.T., and P.L.F. analyzed data; and P.L.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences and mutations reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive, http://www.ncbi.nlm.nih.gov/sra (accession nos. SRA054030 and SRA054031).

[1]To whom correspondence should be addressed. E-mail: plfoster@indiana.edu.

See Author Summary on page 16416 (volume 109, number 41).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1210309109/-/DCSupplemental.

conserved and found in all domains of life. However, bacterial strains isolated from nature often lack MMR capacity (13, 14), and MMR-defective strains frequently become dominant in long-term evolution experiments (15). Indeed, the presence or absence of various DNA repair pathways can be quite sporadic among microbial species (16–18). In addition, MMR proteins are subject to regulation (19, 20), and thus the activity of MMR is responsive to both intrinsic and extrinsic factors. For these reasons, we have included in this report the results of an MA experiment with an isogenic MMR-defective strain. The resulting large pool of mutations gives a picture of genomic mutagenic processes hitherto unseen.

Results

Spontaneous Mutation Rate of Wild-Type *E. coli* Is Lower than Expected. The wild-type strain we used for the MA protocol is a prototrophic derivative of the reference *E. coli* strain, MG1655. The parameters of a two-part MA experiment with this strain are given in Table 1. In the first part (designated "wild-type 3K"), MA lines were passaged for about 3,000 generations before sequencing; in the second part (designated "wild-type 6K"), unsequenced lines continued to be passaged for an additional 3,000 generations and then were sequenced. Surprisingly, the 6K lines accumulated more mutations per generation than did the 3K lines. Thus, we have two estimates of the spontaneous mutation rate that differ by 30% but are within 95% confidence limits (CL) of each other. This result illustrates the inherent uncertainty of estimating mutation rates from sparse data. Nonetheless, because the MA protocol minimizes selection (see below), and because we have a fairly accurate measure of the number of generations that the lines experienced, we believe that these two values put accurate limits on the spontaneous mutation rate of *E. coli.* Both parts of the experiment gave results that appear to be random and unbiased (see below); thus, the best estimate of the spontaneous mutation rate is the midpoint of these limits, 2.2×10^{-10} mutations per nucleotide per generation or 1.0×10^{-3} mutations per genome per generation (Table 1). This rate is about threefold lower than Drake's value of $3–4 \times 10^{-3}$ per genome per generation for all DNA-based microbes (1, 21). Possible reasons for this difference are considered in the *Discussion*.

Table 2 gives mutation rates for the wild-type strain estimated using classical fluctuation tests, scoring for resistance to rifampicin (Rif^R) and to nalidixic acid (Nal^R). When normalized to the number of nucleotides that, when mutated, give the resistance phenotypes (22–24), the two mutation rates are 0.33 and 0.21×10^{-10} mutations per nucleotide per generation, respectively. The mean rate for BPS mutations from the MA experiment was 1.99×10^{-10} per nucleotide per generation, six- to nine-fold higher than that obtained from fluctuation tests. Low values obtained in the fluctuation tests probably are responsible for this difference; several cell generations may be required for drug-sensitive molecules to be replaced by resistant ones before newly arisen Rif^R and Nal^R mutations are expressed. Such phenotypic lag will result in low apparent mutation rates (25).

Selection is Minimal During MA Experiments. There are several ways to evaluate the degree to which selective pressures may have biased the mutational profile obtained from the MA experiments. Assuming that selective pressure on noncoding regions is minimal, if mutations accumulate in a neutral manner, the ratio of the number of mutations occurring in coding versus noncoding DNA should reflect the ratio of the number of nucleotides that are in coding versus noncoding DNA, which is 5.74 for the MG1655 genome (3.95×10^6 nucleotides in protein-coding sequences and 6.88×10^5 in noncoding sequences). The observed ratio from the wild-type data was 3.31 (179/54), which is significantly less than 5.74 ($\chi^2 = 5.0, P = 0.03$). However, because of possible differences in nucleotide content in coding and noncoding DNA, the ratio of mutations in these regions could be biased by the types of mutational events that occur. To address this issue, we performed Monte Carlo simulations using the actual spectrum of BPSs observed from the wild-type results; the results from 1,000 simulations gave a ratio of 5.93 ± 1.04 (mean ± SD), close to the theoretical ratio. Thus it appears that the actual ratio is less than expected for an unbiased distribution of mutations. This result may reflect the inherent variability of sparse data or, more interestingly, may indicate that coding DNA is less susceptible to mutation than noncoding DNA (see below).

Another commonly used measure of bias is the ratio of nonsynonymous to synonymous BPSs, under the assumption that synonymous mutations are relatively neutral. Given the codon usage in MG1655, the expected ratio is 3.25. The ratio observed from the MA experiments with the wild-type strain was 2.25 (55/124), which is not significantly different from expected ($\chi^2 = 2.3$, $P = 0.13$). One thousand Monte Carlo simulations using the observed spectrum of BPSs from the wild-type results gave a ratio of 2.71± 0.41, which is not significantly different from the actual ratio ($\chi^2 = 0.5, P = 0.47$). Thus, the MA protocol resulted in little or no apparent bias against nonsynonymous mutations.

Even synonymous BPSs could be selected against if they result in less favorable codons. Based on the codon usage in MG1655, of the 55 codon changes resulting from synonymous BPSs in the wild-type strain, 16 resulted in a more commonly used codon, four changes were neutral (<10% difference in usage), 11 changes resulted in a less commonly used codon (10–30% difference in usage), and 24 resulted in a relatively rarely used codon (>30% difference in usage) (Table S1). Whether these codon changes would affect fitness depends on whether the genes are highly expressed; of the 55 genes affected, three are considered to be highly expressed (26). In one of these, *guaA*, the mutation resulted in a more commonly used codon, and in two, *aspS* and *metQ*, the mutation resulted in a less commonly used codon. Thus, synonymous BPSs creating less favorable codons occurred both in highly expressed genes and in less highly expressed genes, indicating that these mutations were not purged by selection during the MA experiment.

Finally, most mutations creating chain-terminating (nonsense) codons are expected to have deleterious consequences (21). Of the 1.2×10^7 possible codon-changing BPSs in the *E. coli* ge-

Table 1. Parameters of MA experiments

Strain	No. of BPSs	No. of indels*	No. of lines	Generations per line	Total no. of generations	BPSs per line	Indels per line	Mutation rate per nucleotide (× 10^{10})	95% CL†	Mutation rate per genome‡ (× 10^3)	95% CL†
Wild-type 3K	93	9	38	3,080	117,040	2.45	0.24	1.88	±0.46	0.87	±0.21
Wild-type 6K	140	12	21	6,356	133,476	6.67	0.57	2.45	±0.49	1.14	±0.23
MutL$^-$	1625	306	34	375	12,750	47.8	9.00	326	±153	151	±71

*Insertion or deletion of ≤4 nt.
†Critical values of the t distribution were used to calculate 95% CLs for the means of the Poisson distributions shown in Fig. 1. These values then were used to compute the CLs for the mutation rates.
‡Genome = 4,639,675 nt.

Table 2. BPS mutation rates

	Phenotype	Mutation rate per locus ($\times 10^{09}$)	Mutation rate per nucleotide ($\times 10^{10}$)	95% CL	Mutation rate per genome ($\times 10^{3}$)	95% CL
Fluctuation test results*						
Wild type	Rif^R	2.6	0.33	0.22–0.46	0.15	0.10–0.21
	Nal^R	0.43	0.21	0.11–0.34	0.10	0.05–0.16
$MutL^-$	Nal^R	99	49	44–55	23	21–25
Increase caused by loss of $MutL^-$			233			
MA results						
Wild-type 3K			1.71	$\pm 0.44^{\dagger}$	0.80	$\pm 0.20^{\dagger}$
Wild-type 6K			2.26	$\pm 0.45^{\dagger}$	1.05	$\pm 0.21^{\dagger}$
Wild type mean			1.99		0.92	
$MutL^-$			275	$\pm 14^{\dagger}$	127	$\pm 6^{\dagger}$
Increase caused by loss of $MutL^-$			138			

*Mutation rates and 95% CL were calculated as described in *SI Materials and Methods*. Values were normalized assuming Rif^R is conferred by 79 BPSs in the *rpoB* gene (22) and Nal^R is conferred by 18 BPSs in the *gyrA* gene (23) plus two BPSs in the *gyrB* gene (24).
†Critical values of the t distribution were used to calculate 95% CLs for the means of the Poisson distributions of the numbers of mutations per MA line.These values then were used to compute the CLs for the mutation rates.

nome, 423,094 will create a chain-terminating codon. Thus, genome-wide, 3% {423,094 ÷ [(3 × (4.6 × 10^6)]} of all BPS mutations should be nonsense mutations. This calculation predicts that seven of the 233 BPSs observed in the wild-type strain should have been nonsense; eight were recovered.

Previous studies estimated the rate of deleterious mutations in wild-type *E. coli* to be 0.05–0.2 × 10^{-3} per genome per generation (7, 27), which is, at most, only a fourth of the total mutation rate observed here. The rate of beneficial mutations is even lower, and, in addition, most beneficial and deleterious mutations have only small (<3%) effects on fitness (7, 27). Taken together, our results with the wild-type strain are consistent with these findings and support the basic premise that the MA protocol minimizes selection and allows mutations to accumulate in a nearly neutral fashion.

Loss of MMR Increases the BPS Mutation Rate and Changes the Mutational Bias. Loss of MMR typically increases mutation rates 100- to 200-fold (12). The MMR-defective strain we used for the MA experiment has a deletion of the *mutL* gene, which encodes MutL, one of the key enzymes required for MMR. As shown in Table 2, the $MutL^-$ strain had a 233-fold increase in mutation rate to Nal^R as measured by a fluctuation test. The 375-generation MA experiment resulted in 1,625 BPS mutations in 34 lines, giving a BPS mutation rate of 2.75 × 10^{-8} per nucleotide per generation, a 138-fold increase relative to the wild-type strain.

The observed ratio of BPSs in coding versus noncoding regions in the $MutL^-$ strain was 6.63 (1,412/213), which is greater than but not significantly different from the theoretical ratio of 5.74 (χ^2 = 2.13, P = 0.14). One thousand Monte Carlo simulations using the observed spectrum of BPSs from the $MutL^-$ strain gave a ratio of 5.29 ± 0.36; a value as large as 6.63 was not recovered. Thus, it appears that in the absence of MMR, BPSs are more likely to occur in coding regions than in noncoding regions, suggesting that MMR preferentially prevents errors in coding DNA.

In contrast to the wild-type strain, the ratio of nonsynonymous to synonymous BPSs was only 1.92 (589/307) in the $MutL^-$ strain, significantly less than the expected 3.25 (χ^2 = 39; P < 0.0001). Loss of MMR changes the spectrum of BPSs (see below), which change accounts for most, if not all, of this relative increase in synonymous BPSs. One thousand Monte Carlo simulations using the observed spectrum of BPSs from the $MutL^-$ strain gave a ratio of 1.66 ± 0.09; values of 1.92 or higher were obtained twice. Thus it appears that the $MutL^-$ results are at the low end of the expected result for an unbiased distribution of nonsynonymous and synonymous BPSs.

Distribution of Mutations Is Random among Wild-Type Lines but Not Among $MutL^-$ Lines. If the mutation rate were constant throughout the MA experiment, then the numbers of mutations among the lines should have a Poisson distribution. As shown in Fig. 1*A*, this hypothesis was well supported for the wild-type strain; the two mutation datasets fit very well to Poisson distributions with means of 2.68 and 7.24 mutations per line (χ^2 = 1.3, P = 0.97 and χ^2 = 3.7, P = 0.89, respectively). However, the distribution of mutations per line in the $MutL^-$ strain clearly deviated from expected (χ^2 = 299, P = <0.0001) (Fig. 1*B*), suggesting that some lines may have accumulated mutation-rate modifiers. Indeed, the high and low sections of the $MutL^-$ data for BPSs can be fit to two Poisson distributions with means of 37 and 54 mutations per line (χ^2 = 1.6, P = 0.99 and χ^2 = 2.9, P = 0.99, respectively). However, fluctuation tests performed at the end of the MA protocol with $MutL^-$ lines whose numbers of accumulated mutations spanned the distribution shown in Fig. 1*B* showed no consistent differences in mutation rates. Thus, if mutation modifiers appeared, they had been eliminated by the end of the experiment.

Spectrum of BPSs Is Shifted by Loss of MMR. The spectrum of BPSs in the wild-type strain is given in Table 3, and the rates of each type of BPSs are shown in Fig. 2*A*. The wild-type 3K and 6K datasets resulted in slightly different proportions of the different types of BPSs, but none of the differences are statistically significant at the 95% level. In addition, in 1,000 Monte Carlo simulations of the two spectra the proportions of most of the mutational events from the two datasets fell within one SD of each other; the exceptions were A:T > C:G and G:C > T:A, which fell within the 95% CL. Transitions, particularly G:C > A:T mutations, dominated the wild-type spectrum, comprising 56% of the BPSs; of the transversions, G:C > C:G and A:T > T:A were the rarest. This spectrum is the expected one for wild-type strains based on both locus-specific and genome-wide experiments (see refs. 2, 3, and 28 and references therein).

The spectrum of the $MutL^-$ strain also was strongly biased toward transitions, which made up 98% of the BPSs. However, in contrast to the wild-type strain, A:T > G:C transitions were by far the dominant class, accounting for 70% of all of the BPSs (Table 3 and Fig. 2*B*). Overall, the effect of loss of MMR was to shift the BPS bias from changing G:C to A:T base pairs to changing A:T to G:C base pairs. The increase in the proportion of A:T > G:C transitions in the $MutL^-$ strain was not the result of decreases in the rates of other mutational events. As shown in Table 3 and Fig. 2*B*, the mutation rates of all types of BPSs were increased in the $MutL^-$ strain, but A:T > G:C events rose disproportionally. This

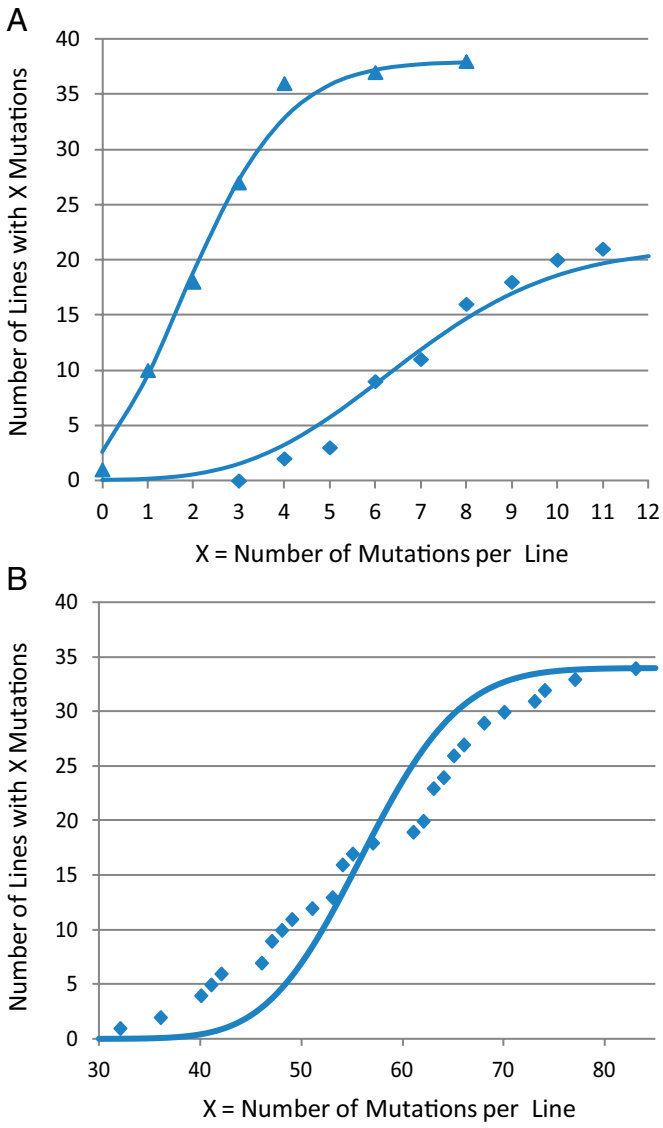

Fig. 1. Distribution of mutations among MA lines. The number of lines with a given number of mutations is plotted against the number of mutations per line and compared with the Poisson distribution expected for the mean number of mutations per line (solid traces). (*A*) Wild-type 3K lines (38 lines; triangles) and wild-type 6K lines (21 lines; diamonds). (*B*) MutL$^-$ lines (34 lines).

change in mutational bias has been seen in some, but not all, previous studies of MMR defective strains (*Discussion*).

Occurrence of BPSs Has a Strong DNA-Strand Bias. Assuming that the BPSs observed are the result of replication errors, A:T > G:C mutations result from C mispaired with a template A or G mispaired with a template T. The *E. coli* genome is divided into two replichores, each replicated in opposite directions starting at the origin of replication centered on nucleotide 3,923,883 and ending at the terminus approximately half-way around the chromosome [nucleotide 1,604,045 is the exact half-way point, but termination is imprecise, usually occurring between TerA at nucleotide 1,339,769 and TerC at nucleotide 1,607,200 (29)]. Thus, the leading and lagging strands are switched in the two replichores relative to the conventional 5′-to-3′ nucleotide numbering system defining the "top" strand. In the right replichore A:T > G:C transitions in the MutL$^-$ strain were twice as likely to occur with A rather than T in the top strand (390 As versus 175 Ts); in the left replichore this bias was reversed: A:T > G:C transitions were twice as likely to occur with T rather than A in the top strand (394 Ts versus 182 As). This bias held true when the mutations were normalized to the A and T content of the strands. Thus, in both the right (clockwise) and left (counterclockwise) replichores, A:T > G:C transitions were twice as likely to occur with A templating the lagging strand and T templating the leading strand during replication than in the opposite orientation. In contrast, G:C > A:T transitions were twice as likely to occur (305/142) with C templating the lagging strand and G templating the leading strand during replication than in the opposite orientation (also true when normalized to G and C content of the strands). The bias for transitions occurring with C templating the lagging strand also was evident in the wild-type strain (58/24), but the bias for A templating the lagging strand was not as strong (29/20).

BPSs in the MutL$^-$ Strain Are Strongly Biased by the Local Sequence Context. Of the BPSs that occurred at A:T sites in the MutL$^-$ strain, 77% (899/1,165) occurred in the sequence 5′**A**pC3′ or the equivalent 5′Gp**T**3′ (the two bases indicated are on the same DNA strand with the mutated base in bold; the "p" represents the phosphate linking the two nucleosides). The influence of the neighboring C or G was particularly prominent at the A:T sites that gave rise to A:T > G:C mutations: of these, 79%, (897/1,141) were at 5′**A**pC3′ or 5′Gp**T**3′ (Table S2). G:C > A:T transitions in the MutL$^-$ strain also were biased by the local sequence context: 53% (238/447) were at 5′**G**pC3′ or 5′Gp**C**3′ sites (Table S2). Transition mutations in the wild-type strain were not as strongly biased by neighboring Gs or Cs, suggesting that MMR is relatively efficient at preventing mutations at these sites (Table S2).

Methylated Bases Are Mutational Hotspots. The internal Cs in the sequences C**C**AGG and C**C**TGG are methylated at the 5 position by the Dcm methylase (30). Because 5meC can deaminate, creating thymine in the DNA, Dcm sites are hotspots for G:C > A:T transitions (31). Of the 82 G:C > A:T transitions that occurred in the wild-type strain, eight (about 10%) occurred at Dcm sites. There are 12,045 Dcm sites in the *E. coli* genome, so only 2% of the Cs potentially are methylated. Thus, our data confirm that Dcm sites are mutational hotspots. However, only one of the 447 G:C > A:T transitions in the MutL$^-$ strain occurred at a Dcm site; because the rate of this event is only threefold higher than in the wild-type strain, MMR appears to be poor at preventing these mutations (*Discussion*).

The A in GATC sites is methylated at the 6 position by the Dam methylase (30); 6meA is prone to depurination, producing transversions (32, 33). Of the 24 A:T transversions that occurred in the MutL$^-$ strain, 17 occurred at Dam sites; 13 of those were A:T > T:A, and four were A:T > C:G transversions. Four A:T > G:C transitions also occurred at Dam sites. A similar but less prominent pattern was observed in the wild-type strain: of 55 A:T transversions, 11 occurred at Dam sites; four were A:T > T:A, and seven were A:T > C:G. Two A:T > G:C transitions also occurred at Dam sites in the wild-type strain. There are 19,120 Dam sites in the *E. coli* genome, so only about 3% of the As potentially are methylated. Therefore, the number of A:T transversions occurring at Dam sites far exceeds the expected number, indicating that these sites are mutational hotspots.

The Dam methylase has some activity at noncanonical target sites, particularly GACC sequences but also CATC, TATC, AATC, and GATT sequences (34), but these do not appear to be hotspots for transversions (see *SI Text* for further discussion).

BPSs Are Not Biased Toward Highly Expressed Genes. Of *E. coli*'s 4,146 protein-coding genes, 253 are considered to be highly expressed (26). These genes account for 6% of the nucleotides in the genome and 7% of the nucleotides in the coding DNA. In the wild-type strain 16 BPSs were in highly expressed genes, representing

Table 3. Spectra of BPS in wild-type and MutL⁻ strains

	Wild type		MutL⁻		
	Number	Fraction	Number	Fraction	Increase caused by MutL⁻*
Type of substitution					
Transitions	131	0.56	1,588	0.97	240
A:T > G:C	49	0.21	1,141	0.70	465
G:C > A:T	82	0.35	447	0.28	108
Transversions	102	0.44	37	0.02	7
A:T > T:A	17	0.07	14	0.009	5
A:T > C:G	38	0.16	10	0.006	5
G:C > T:A	30	0.13	10	0.006	6
G:C > C:G	17	0.07	3	0.002	4
A:T sites	104	0.45	1,165	0.71	224
G:C sites	129	0.55	460	0.28	70
Total	233		1,625		138
Consequences of substitutions					
Position					
Noncoding	54	0.23	213	0.13	77
Coding	179	0.77	1,412	0.87	157
Within coding sequences					
Synonymous	55	0.31	482	0.34	176
Nonsynonymous	124	0.69	930	0.66	149
Amino acid changes					
Conservative	57	0.46	644	0.69	224
Nonconservative	67	0.54	286	0.31	85

*Mutation rates normalized to the relevant genomic nucleotide composition, compared with the average of the rates for wild-type 3K and 6K.

6% of all of the BPSs and 9% of BPSs in coding DNA. In the MutL⁻ strain 102 BPSs (6% of all of the BPSs and 7% of the BPSs in coding DNA) were in highly expressed genes. None of these values is significantly different from expected ($P \geq 0.42$ in every case). Thus, our results do not support previous observations that highly expressed genes have either high (11, 35, 36) or low (37–39) mutation rates. Unlike a previous report (39), we found no bias for any type of base change occurring preferentially on either the transcribed or the nontranscribed DNA strand (Table S3).

Sequence Repeats are Hotspots for Indels. Insertions and deletions ≤4 nt accounted for the vast majority of the indels in both the wild-type and the MutL⁻ strain; the few larger indels and insertion-sequence element movements will be treated in a separate analysis. When normalized to the whole genome, small indels occurred at about 1/10th the rate of BPSs in both the wild-type and the MutL⁻ strain (Table 4). Loss of MMR resulted in a 288-fold increase in the rate of indel formation, with the increase biased toward additions of G:C base pairs. The majority of indels were gain or loss of a single nucleotide in runs of identical nucleotides; the few multiple-nucleotide indels were gains or losses of 2- or 3-nt units in runs of identical units.

The rate at which indels occurred normalized to the whole genome obscures the true mutation rate at sequence repeats, which are known hotspots for indel formation (40). As shown in Fig. 3*A*, the rate at which indels occurred as a function of the length of the run increased exponentially in both the wild-type and the MutL⁻ strain. The slope of the least-squares line fitted to the MutL⁻ data is 25% greater than that fitted to the wild-type data, suggesting that the ability of MMR to prevent indels increases with increasing run length. Alternatively, the wild-type data can be fit to two lines, with the rate of increase of mutations in runs ≥5 nt being about 20% greater than that of the MutL⁻ strain. In either case, although the increase in the rate of indel formation caused by the loss of MMR is high overall, the increase in runs of a given length is at most 70-fold.

To determine the probability of an indel occurring at a run of a given length, we multiplied the total number of indels observed by the fraction of nucleotides that are in runs of each length in the genome. Compared with the actual number of indels observed, any run ≥4 nt was a hotspot (Fig. 3*B*). However, certain runs may be more likely to sustain an indel than other runs of the same length. To identify these potential hotspots, we looked for runs in which more than one indel occurred, and these are given in Table S4.

Discussion

The major conclusions of the study presented here are (*i*) the best estimate of the spontaneous mutation rate of a wild-type *E. coli* strain growing on rich medium is 1×10^{-3} per genome per generation; (*ii*) in the absence of MMR, mutations are dominated by A:T > G:C transitions; (*iii*) A:T > G:C transitions occur preferentially with A templating the lagging strand and T templating the leading strand, whereas G:C > A:T transitions occur preferentially with C templating the lagging strand and G templating the leading strand; and (*iv*) there is a strong bias for transitions to occur at 5′**A**pC3′/3′**T**pG5′ sites and, to a lesser degree, at 5′**G**pC3′/3′**C**pG5′ sites.

In addition to these major points, our results identify GATC sites as hotspots for A:T transversion mutations. We extend to the whole genome the previous findings that CCAGG and CCTGG sites are hotspots for G:C transitions (41) and that indels occur preferentially at sequence repeats (40). However, although the rate of indels at repeats is high and increases exponentially with the length of a sequence run, the overall genomic rate of indels is only 1/10th that of BPSs.

Our finding that the mutation rate of this model prokaryote is one third of Drake's estimate based on analyses of specific-locus experiments (1, 21) was unexpected. As pointed out by Drake (4), different estimates obtained from MA experiments and specific-locus methods may reflect selection against the mutations that accumulate during the former protocol, which, by the use of a carefully chosen reporter gene, are minimized in the latter. However, by several criteria the MA protocol followed here minimized selection:

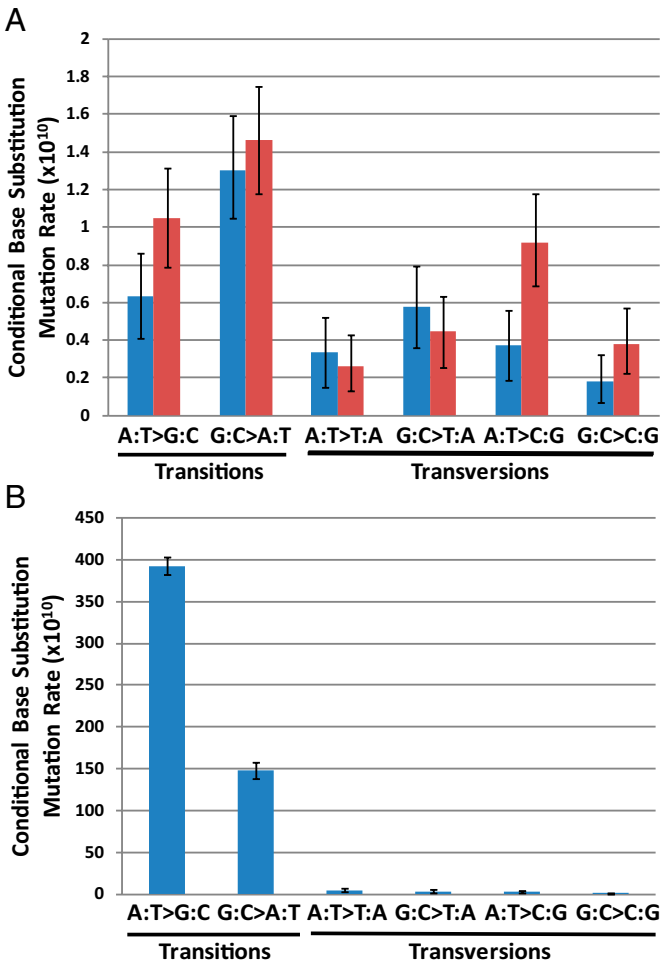

Fig. 2. Mutation rates of each of the six BPSs. Bars represent the mutation rate of each type of BPS normalized to the number of AT or GC base pairs in the genome. (*A*) Wild-type 3K dataset (93 mutations; blue) and wild-type 6K dataset (140 mutations; red). (*B*) MutL$^-$ dataset (1,625 mutations). Error bars represent the fifth percentile and 95th percentile values from 1,000 Monte Carlo simulations of a random distribution with the mutational spectra observed for each dataset.

(*i*) the ratio of synonymous to nonsynonymous BPSs in the wild-type strain did not differ significantly from that expected by chance; (*ii*) the wild-type mutation rate based solely on synonymous BPSs differs little from that calculated based on all BPSs (see below); (*iii*) most synonymous BPSs resulted in less commonly used codons, indicating that these mutations were not selected against; and (*iv*) nonsense mutations were recovered at the expected frequency, indicating that these mutations also were not selected against.

Alternative explanations for the discrepancy are that the specific loci used by Drake (1, 21) to estimate the genomic mutation rate were not representative of the genome as a whole or that some of the approximations used were not appropriate for *E. coli*. In addition, the growth conditions used in the two types of experiments are very different. Most obviously, during MA experiments the cells grow in colonies, which can become microaerobic, whereas for specific-locus experiments cells typically are grown in well-aerated liquid cultures. Given these differences, perhaps it is remarkable that the two estimates are within threefold.

Recently Wielgoss et al. (3) estimated the *E. coli* genomic mutation rate based on synonymous BPSs occurring during long-term evolution experiments. Their value, 0.41×10^{-3} per genome per generation, is less than half the rate that we observed. The mutation rate based solely on the 55 synonymous BPSs in our wild-type dataset is 1.1×10^{-3} per genome per generation. This value is close to the rate we obtained considering all the BPSs, and is nearly threefold higher than the value calculated by Wielgoss et al. (3). Drake (4) has argued that selection against synonymous codon changes in the long-term evolution experiments was sufficient to account for the low rate obtained by Wielgoss et al. We believe that the higher mutation rate that we obtained reflects the fact that the MA protocol truly minimizes selective pressure against mutations. The lowest mutation rate estimate for *E. coli*, $0.1–0.2 \times 10^{-3}$ per genome per generation, comes from comparative genomics (42). Although the various contributions of selection, mutational bias, and experimental conditions appear sufficient to explain differences in experimentally determined mutation rates, there is as yet no consistent explanation for the difference between estimates based on comparative genomics and those derived from experimental evidence (2).

Although whole-genome sequencing can give an unparalleled view of mutational events across the entire genome, it cannot supersede the results obtained with more traditional methods. For example, except for highly frequent events, such as indels at repeat sequences, sequencing unselected genomes reveals only single mutational events. Cataloging events at similar sites in the genome allows aspects of mutagenic specificity to be deduced (see below) but does not identify true hotspots, i.e., specific sites that are particularly mutation-prone. Thus, the mutational topography revealed by single-locus studies is a necessary component of our understanding of mutational processes.

A striking result from our MA experiments is that, in the absence of MMR, the BPS spectrum of cells growing on rich medium is dominated by A:T > G:C transitions (Table 3 and Fig. 2). Because of individual idiosyncrasies of each mutational target, this bias had not been seen in many previous studies of MMR-defective microbes. Indeed, the observed bias in MMR-defective *E. coli* often was toward G:C > A:T mutations (43–45). Only when collections of mutations were sequenced, as in the *mnt* gene (46), the *lacI* gene (47), the *rpoB* gene (22), and the *gyrA* gene (23), was an A:T > G:C bias in MMR-defective strains revealed. Our data confirm and extend these results to the entire genome.

Table 4. Indel mutation rates

Strain	Mutation rate per genome ($\times 10^5$)	Single-nucleotide indels				Multiple-nucleotide indels*	
		Plus	Minus	A:T	G:C	Plus	Minus
Wild-type 3K	7.69	5	3	6	2	0	1
Wild-type 6k	8.99	1	10	4	7	0	1
MutL$^-$	2,400	177	120	97	200	3	6
Increase caused by loss of MutL†	288	553	187	187	451	—	59

*2–4 nt.
†Mutation rates for the MutL$^-$ strain compared with the average of the rates for wild-type 3K and 6K.

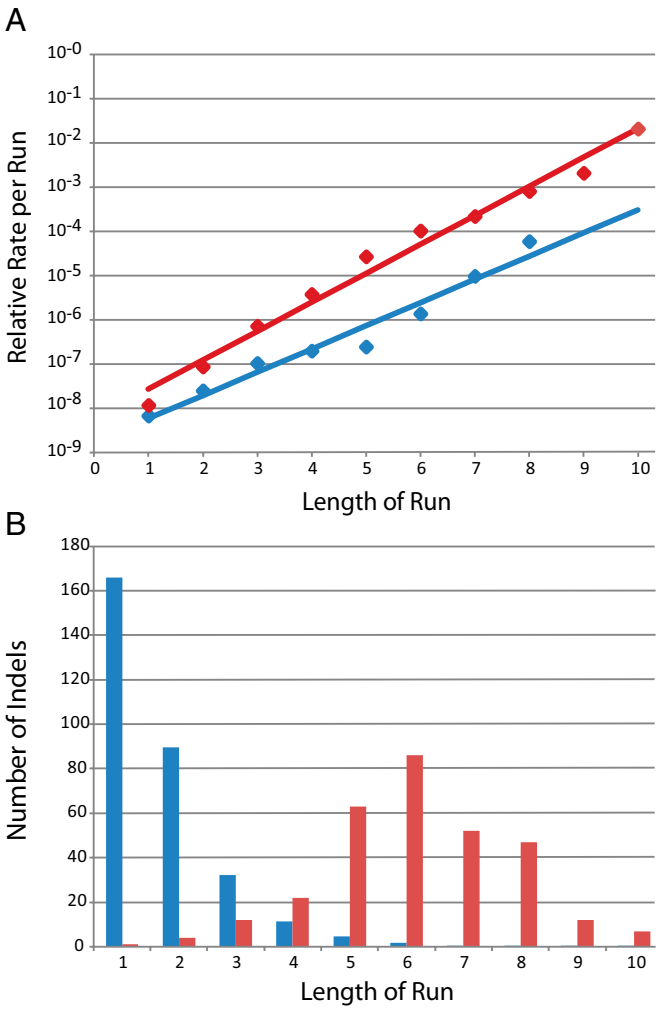

Fig. 3. Indel formation at runs. (*A*) The relative mutation rate of indels in a run of a given length is plotted against the length of the run. The relative mutation rate of indels in each run length is the number of observed indels divided by the total number of target nucleotides (= nucleotides in the run × the number of runs of that length in the genome × the number of MA lines in the analysis). Blue diamonds represent the combined wild-type 3K and 6K datasets (21 indels); red diamonds represent the MutL⁻ dataset (306 indels). Lines are the least-squared fits to the data. (*B*) Values for the MutL⁻ dataset (306 indels). Blue bars represent the expected number of indels in a run of a given length calculated as the total number of indels obtained × the fraction of runs in the genome of that length. Red bars represent the actual number of indels observed in each run length.

Mutational data from a variety of sources have confirmed that in almost all organisms spontaneous mutations are dominated by G:C > A:T transitions, a bias that tends to drive genomes toward greater A:T content (28). However, the G:C content of genomes varies widely, so some selective pressure or nonadaptive mechanism must drive genomes back toward G:C-richness. Because MMR corrects replication errors, the mutational spectrum in its absence reveals that errors made during replication would, if uncorrected, increase the genomic G:C content. Thus, the MMR system serves not only to keep mutation rates low but also to prevent genomes from drifting toward ever-higher G:C content. This observation suggests that the G:C/A:T balance of the genomes of different organisms might be determined, at least in part, by the functioning of their MMR system. In a recent survey of 699 bacterial genomes, about 20% lacked one or more genes in the MMR pathway, but there was no correlation with the G:C content of the genomes (48). Two caveats apply to these results: (*i*) other repair activities in the cell may compensate for the absence of MMR; and (*ii*), the presence of the MMR genes does not mean that the system is functioning to the same extent in each organism. Our results also have a forensic implication. Because the activity of the MMR varies by growth phase and is influenced by environmental conditions (19, 20), an organism's mutational spectrum might be a fingerprint of its recent history.

Loss of MMR caused a larger increase in mutations in coding than in noncoding DNA (Table 3), suggesting that the MMR system is more active on the DNA that encodes genes. This difference could reflect the participation of the MMR proteins, including MutL, in transcription-coupled repair (49). However, contrary to a recent report (39), we found no bias for any specific mutational event to occur on the transcribed versus the nontranscribed DNA strand (Table S3); such a bias could reflect the activity of transcription-coupled repair. Nor did our data support previous conclusions that highly expressed genes are either vulnerable or resistant to mutation (37–39).

Our results shed light on the origins of spontaneous mutations. Deamination of cytosine to uracil frequently is cited as a cause of G:C > A:T transitions. The mutation rate of G:C > A:T transitions in the MutL⁻ strain was 0.035 per genome per generation, which is 60 times the estimated deamination rate of cytosine in dsDNA and 10–20 times the estimated deamination rate of cytosine in ssDNA (Table S5). Adding to the low level of deamination are the activities of uracil glycosylase, which removes uracil from DNA, and mismatch uracil DNA glycosylase, which removes uracils paired with guanines. Thus, unless the rate of cytosine deamination is greatly underestimated, it cannot be responsible for the number of G:C > A:T mutations observed. However, the creation of thymines by deamination of 5meC is important. In the wild-type strain, 10% of the G:C > A:T transitions occurred at the C that is methylated by the Dcm methylase. Thus, our results support and extend to the whole genome the prescient finding that Dcm sites are mutational hotspots (41).

Mutations at Dcm sites can be prevented by very short patch (VSP) repair, an enzyme system that removes the T mispaired with C at these sites (reviewed in ref. 50). MutL participates in VSP repair, so it was surprising that the rate of G:C > A:T mutations at Dcm sites was increased only threefold in the MutL⁻ strain. However, MutL facilitates but is not required for the relatively inefficient VSP repair (51, 52). Because G:C >A:T mutations were increased 100-fold overall in the MutL⁻ strain, other sources of G:C > A:T mutations must occur at much higher frequencies than deamination of 5meC. Because MMR normally prevents these G:C > A:T transitions, mutations at Dcm sites become prevalent in wild-type cells.

Transitions arise from A:C and G:T mismatches, both of which are well corrected by the *E. coli* MMR system in vitro (53) and in vivo (54, 55). If both these mismatches occurred with equal frequency and irrespective of DNA strand, then loss of MMR would elevate both A:T > G:C and G:C > A:T transitions to the same extent. Therefore, the observed bias for A:T transitions in the MutL⁻ strain implies either that the mismatches that lead to A:T transitions are replication errors that occur more frequently than those that lead to G:C transitions or that a separate error-correcting pathway efficiently and preferentially prevents G:C transitions. One candidate for this error-correcting function is the proofreading activity of the *E. coli* replicase DNA polymerase III, which resides in a separate subunit, epsilon. Whether error correction by epsilon is biased has been debated for years, but a recent study (56) showed a bias toward correcting G:C > A:T mutations that would account for our results. Another candidate is MutY, which can remove an A mispaired with C, preventing G:C > A:T mutations if C is the template (and promoting A:T > G:C mutations if A is the template) (57). Future MA studies will reveal if these enzymes contribute to mutational specificity in the absence of MMR.

Transitions arise when a purine mispairs with the wrong pyrimidine during replication. Long ago, such mispairings were postulated to result from tautomeric shifts in the bases (58, 59). Others have proposed mispairs involving ionized bases (60). Crucial to these schemes is that the postulated mispair match the geometry of the normal base pair, allowing it to fit within the DNA polymerase-active site for catalysis. Recently both A:C and the G:T mispairs have been captured in the active site of a crystallized polymerase (61, 62); the former involved amino:imino tautomers, and the latter probably involved ionized bases. In both cases the mismatched base pair had the canonical Watson–Crick geometry of the cognate base pair. One of the striking aspects of the MutL$^-$ data is that 79% of the A:T transitions occurred at sites with the sequence 5′**A**pC3′/3′**T**pG5′ (Table S2), a preference that has been observed in previous studies (22, 63). We favor the hypothesis that these transitions are templated by the A for the following reason: because polymerase approaches the template base from the 3′ side, a C:G base pair would be established first and, because of its strong base-pairing and stacking interactions, could stabilize the A:C mispair. In the alternative configuration, requiring a T:G mispair, the stabilizing G:C base pair would not be established until after the T:G mispair formed. The strong preference for C, not G, 3′ to the A, further suggests that the stabilizing base pair must have the correct orientation.

Another striking aspect of the MutL$^-$ data is the 2:1 bias for A:T transitions occurring with A in the top strand on the right replichore and T in the top strand in the left replichore. The purine:pyrimidine strand-bias for G:C transitions was the reverse: they were twice as likely to occur with C in the top strand on the right replichore and G in the top strand in the left replichore. This mutational pattern would tend to produce the well-known "keto-skew" of the *E. coli* chromosome (64); i.e., in the first half of the chromosome the top strand has more Gs and Ts, the bases with keto groups, and in the second half of the chromosome, the top strand has more As and Cs, the bases with amino groups. A similar mutational bias and keto-skew near replication origins has been described recently in *S. cerevisiae* (65). However, the bias in itself does not tell us which base gives rise to the mutation.

If A is the templating base for A:T transitions, as hypothesized above, the A/T strand bias would require that the mutational event be A templating the insertion of a C during lagging-strand synthesis. If the G/C strand bias reflects the same molecular constraints, then the G:C transitions would be produced by C templating insertion of an A during lagging-strand synthesis. Arguing against this hypothesis is the fact that G:C transitions did not show a strong bias for a particular base 3′ to the C. However, 74% of the G:C transitions occurred at sites with either C or G 3′ to the C (Table S2), suggesting that the C:A mispair could be stabilized by the formation of a 3′G:C base pair regardless of its strand orientation.

Although we favor the hypothesis that A and C are the templating bases and that the A:C mispair forms during lagging-strand synthesis, the alternative is that T and G template the G:T mispair during leading-strand synthesis. Then the mismatch would be stabilized by the G:C base pair formed after the mismatch. In support of this hypothesis, extension past a mismatch by only 4 nucleotides protects the mismatch from exonucleolytic proofreading (66). The hypothesis that the mismatch occurs during leading-strand synthesis also would agree with the often-stated conclusion that leading-strand synthesis is less accurate than lagging-strand DNA synthesis (67). However, although the fidelity of synthesis of the two strands may not be equal, the conclusion that synthesis of the leading strand is the less accurate one is based on primer-extension reactions in vitro and may not pertain in vivo.

Transversions made up 44% of the mutations in the wild-type strain. Because loss of MMR increased transversions only about 10-fold, compared with 200-fold for transitions, transversions were only a minor component of the MutL$^-$ spectrum. This result means that MMR is relatively inefficient in correcting the mismatches that lead to transversions and/or that other repair pathways are active in preventing transversions. For example, mutations that inactivate the enzymes that deal with oxidized guanines are powerful mutators, greatly increasing both A:T > C:G and G:C > T:A mutations (68). These two transversions accounted for 29% of the BPSs recovered in the wild-type strain, but their rates were increased only sixfold in the MutL$^-$ strain, suggesting that MMR is relatively blind to the mismatches that cause these mutations. As mentioned above, A:T transversions occurred preferentially at GATC sites where the A is methylated at the 6 position by the Dam methylase. GATC sites also were hotspots in the mutational spectrum of the *rpoB* gene (69). Depurinated bases lead to transversion mutations because for many DNA polymerases the preferred order of base insertion opposite a missing base is A followed by G, although this order is far from strict (33, 70). The estimated rate of depurination of 6meA is about 3×10^{-3} per *E. coli* genome per generation (Table S5), which is sufficient to account for the observed rate of A:T transversions at GATC sites in both the wild-type and the MutL$^-$ strains (0.09×10^{-3} and 2.4×10^{-3} per genome per generation, respectively).

Our data show that the rate of indel formation at mononucleotide runs increases exponentially with the length of the run and that the ability of MMR to correct these indels may increase with run length (Fig. 3*A*). Although we identified indel hotspots based on repeated occurrence, any run of 4 nucleotides or more is a potential hotspot (Fig. 3*B*). Because an indel is likely to inactivate a gene, there should be selection against runs in genes encoding important cell functions. On the other hand, genes that have significant runs could be dispensable; indeed, such genes may be "contingency loci," able to be activated and inactivated by frequently occurring mutations (71). We inspected the list of genes in which indels were recovered to find those whose activation or inactivation might have a selective advantage under stressful conditions. Interestingly, we recovered two indels in *dinB* and one in *umuC*, the genes that encode *E. coli*'s two error-prone polymerases. The two runs in the *dinB* gene, of four and six Gs respectively, occur early in the coding sequence and thus indels in these runs certainly inactivate the gene. These runs are conserved in most *E. coli* and *Shigella* genomes in the database but are not conserved in *Salmonella*. The mutated run in the *umuC* gene, consisting of five As late in the coding sequence, likewise is conserved in most *E. coli* and *Shigella* genomes but not in *Salmonella*. Surprisingly, an indel was isolated midway in the *mutT* gene in a run of six Cs that is not well conserved even among *E. coli* strains. The phenotype of *mutT* mutants is an increase in G:C > TA mutations (68); because the MA strain carrying this mutation did not accumulate any G:C > TA mutations, the indel must have occurred late in the MA experiment. Other mutated genes of interest are *xthA*, which encodes an exonuclease involved in base-excision repair, and *alkB*, which encodes an enzyme involved in the repair of alkylation DNA damage. A recent survey of bacterial genomes found that hundreds of species had repeat sequences in MMR genes in which indels could act as genetic switches to change the mutation rates of these microbes (72). Thus, in addition to mechanisms that modulate the activity of DNA repair enzymes, microbes also may have adaptive or serendipitous mechanisms that genetically alter their DNA repair capabilities, and thus their mutation rates, allowing them to respond to changes in environmental conditions.

Materials and Methods

Bacterial Strains and Media. The wild-type strain PFM2 is a prototrophic derivative of the *E. coli* K12 reference strain MG1655 (73, 74). The MutL strain PFM5 is a derivative of PFM2 with the *mutL* gene replaced by an in-frame scar sequence (75). Details of the genetic techniques and media used are given in *SI Materials and Methods*.

Estimation of Mutation Rates by Fluctuation Tests. Mutation rates to Nal^R or Rif^R were estimated using fluctuation tests as described (76). Further details are given in *SI Materials and Methods*.

MA Procedure. MA lines originated from single colonies isolated on LB agar plates; each day each line was streaked for single colonies on an LB agar plate, two lines per plate, and incubated at 37 °C for 24 h. To avoid bias, the colony chosen for passage was a well-isolated one closest to a line drawn down the center of the plate. Originally 100 lines of the wild-type strain and 70 lines of the $MutL^-$ strain were started, but losses occurred during passage (when a single colony was not obtained), DNA preparation, library construction, and because of lineage ambiguities (see *Shared Mutation Analysis* below). Each of these steps accounted for a reduction of about 20%, resulting in the number of lines given in Table 1. The losses were random and so should not affect the results. In no case was a line lost because it went extinct. Further details are given in *SI Materials and Methods*.

Estimation of Generations and Cell Viability in Colonies. The number of generations per passage, estimated from the number of cells in colonies of a given diameter, usually was 28 generations. The fraction of dead cells in resuspended colonies determined with the Live/Dead BacLight Bacterial Viability Kit (Invitrogen, Inc) was 0.05 ± 0.01 (mean ± SEM). Further details are given in *SI Materials and Methods*.

Genomic DNA Preparation. The PureLink Genomic DNA purification kit (Invitrogen Corp.) was used to purify genomic DNA from 0.5–1 mL of overnight LB cultures inoculated from the freezer stocks. DNA concentration and purity were assessed using a NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific, Inc.).

Sequencing and Quality Control. Sequencing was performed at Beijing Genome Institute (BGI) using the Illumina HiSeq2000 platform. For each MA line, ~101 (470 Mbp) paired-end reads (2 × 90 bp) were retained after discarding reads that did not meet BGI's quality control. Reads with any one of the following characteristics were discarded: (*i*) ≥10% unreadable bases; (*ii*) ≥20% low-quality ($\leq Q_{20}$) bases; (*iii*) adapter contamination (≥15-bp overlap allowing up to 3-bp mismatch); or (*iv*) duplicate read-pairs. After such filtering an average of 91.1% of reads were retained.

SNP and Short Indel Calling. The reference genome sequence was NCBI Reference Sequence NC_000913.2. For each sample, Illumina reads were aligned to the *E. coli* K12 (strain MG1655) genome with the short read alignment tool, BWA (ver. 0.5.9) (77). Short indels (≤4 bp) were called based on the read mapping of the SNP calling procedures. Further details are given in *SI Materials and Methods*.

Mutation Confirmation by Conventional Sequencing. We randomly selected 19 BPSs from the wild-type 3K MA lines (~20% of the total) for confirmation. Because indels generally are more challenging to call accurately than SNPs (78), we checked all 10 of the small indels called from the wild-type 3K MA line. There were no false calls in the BPS dataset. One indel that had been called as −G actually was a deletion of 24 nt in a repeat element; all other indel calls were correct. Further details are given in *SI Materials and Methods*.

Mutation Annotation. Several custom scripts were written to annotate candidate variants. The coordinates of protein-coding genes were obtained from the GenBank page of NCBI reference sequence NC_000913.2. BPSs were determined to be synonymous, nonsynonymous, or noncoding based on the genetic code. Nonsynonymous BPSs were designated as conservative or nonconservative based on the BLOSUM62 matrix (79); a value ≥0 was considered conservative.

Shared Mutation Analysis. The wild-type and $MutL^-$ strains experienced ~80 and ~75 generations of growth, respectively, as their phenotypes were checked and as they were frozen and regrown before the first bottleneck for the MA protocol. Thus, MA lines could share mutations that occurred during this period. Sequence alignment and lineage analysis were used to distinguish mutations that arose independently from those that shared a common ancestor. Further details are given in *SI Materials and Methods*.

Monte Carlo Simulations. A custom script was written to simulate a random distribution of BPSs corresponding to the observed mutational spectra. For simplicity, the number of mutations was fixed at the observed numbers; 1,000 trials were simulated.

Statistical Analyses. Standard statistical analyses were used (80, 81).

ACKNOWLEDGMENTS. We thank D. Osiecki, D. Simon, K. Storvik, J. Townes, N. Yahaya, and A. Ying Yi Tang for valuable technical assistance; members of the Lynch laboratory for useful discussions; and E. Eisenstadt, S. Finkel, A. Hanson, M. Lynch, and the anonymous reviewers for helpful comments on this manuscript. This research was supported by Multidisciplinary University Research Initiative Award W911NF-09-1-0444 from the US Army Research Office (to P.L.F.).

1. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* 88:7160–7164.
2. Ochman H (2003) Neutral mutations and neutral substitutions in bacterial genomes. *Mol Biol Evol* 20:2091–2096.
3. Wielgoss S, et al. (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda)* 1:183–186.
4. Drake JW (2012) Contrasting mutation rates from specific-locus and long-term mutation-accumulation procedures. *G3 (Bethesda)* 2:483–485.
5. Keightley PD, Halligan DL (2009) Analysis and implications of mutational variation. *Genetica* 136:359–369.
6. Lynch M, Walsh JB (1998) *Genetics and Analysis of Quantitative Traits* (Sinauer Associates., Inc, Sunderland, MA).
7. Kibota TT, Lynch M (1996) Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. *Nature* 381:694–696.
8. Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nat Rev Genet* 4:457–469.
9. Schultz ST, Lynch M, Willis JH (1999) Spontaneous deleterious mutation in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 96:11393–11398.
10. Lynch M, et al. (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA* 105:9272–9277.
11. Lind PA, Andersson DI (2008) Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci USA* 105:17878–17883.
12. Marinus MG (2010) DNA methylation and mutator genes in *Escherichia coli* K-12. *Mutat Res* 705:71–76.
13. LeClerc JE, Li B, Payne WL, Cebula TA (1996) High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science* 274:1208–1211.
14. Matic I, et al. (1997) Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. *Science* 277:1833–1834.
15. Sniegowski PD, Gerrish PJ, Lenski RE (1997) Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387:703–705.
16. Eisen JA, Hanawalt PC (1999) A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat Res* 435:171–213.
17. Denver DR, Swenson SL, Lynch M (2003) An evolutionary analysis of the helix-hairpin-helix superfamily of DNA repair glycosylases. *Mol Biol Evol* 20:1603–1611.
18. Lucas-Lledó JI, Lynch M (2009) Evolution of mutation rates: Phylogenomic analysis of the photolyase/cryptochrome family. *Mol Biol Evol* 26:1143–1153.
19. Tsui H-CT, Feng G, Winkler ME (1997) Negative regulation of *mutS* and *mutH* repair gene expression by the Hfq and RpoS global regulators of *Escherichia coli* K-12. *J Bacteriol* 179:7476–7487.
20. Feng G, Tsui H-CT, Winkler ME (1996) Depletion of the cellular amounts of the MutS and MutH methyl-directed mismatch repair proteins in stationary-phase *Escherichia coli* K-12 cells. *J Bacteriol* 178:2388–2396.
21. Drake JW (2009) Avoiding dangerous missense: Thermophiles display especially low mutation rates. *PLoS Genet* 5:e1000520.
22. Garibyan L, et al. (2003) Use of the *rpoB* gene to determine the specificity of base substitution mutations on the *Escherichia coli* chromosome. *DNA Repair (Amst)* 2:593–608.
23. Becket E, Tse L, Yung M, Cosico A, Miller JH (2012) Polynucleotide phosphorylase plays an important role in the generation of spontaneous mutations in *Escherichia coli*. *J Bacteriol* 194:5613–5620.
24. Yamagishi J, Yoshida H, Yamayoshi M, Nakamura S (1986) Nalidixic acid-resistant mutations of the *gyrB* gene of *Escherichia coli*. *Mol Gen Genet* 204:367–373.
25. Armitage P (1952) The statistical theory of bacterial populations subject to mutation. *J R Stat Soc, B* 14:1–40.
26. Puigbò P, Romeu A, Garcia-Vallvé S (2008) HEG-DB: A database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. *Nucleic Acids Res* 36(Database issue):D524–D527.
27. Trindade S, Perfeito L, Gordo I (2010) Rate and effects of spontaneous mutations that affect fitness in mutator *Escherichia coli*. *Philos Trans R Soc Lond B Biol Sci* 365:1177–1186.
28. Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* 107:961–968.
29. Duggin IG, Bell SD (2009) Termination structures in the *Escherichia coli* chromosome replication fork trap. *J Mol Biol* 387:532–539.
30. Marinus MG (1996) Methylation of DNA. *Escherichia coli and Salmonella Cellular and Molecular Biology*, eds Neidhardt FC, et al. (ASM, Washington, DC), pp 782–791.
31. Farabaugh PJ, Schmeissner U, Hofer M, Miller JH (1978) Genetic studies of the *lac* repressor. VII. On the molecular nature of spontaneous hotspots in the *lacI* gene of *Escherichia coli*. *J Mol Biol* 126:847–857.

32. Lindahl T, Nyberg B (1972) Rate of depurination of native deoxyribonucleic acid. *Biochemistry* 11:3610–3618.
33. Loeb LA, Preston BD (1986) Mutagenesis by apurinic/apyrimidinic sites. *Annu Rev Genet* 20:201–230.
34. Clark TA, et al. (2012) Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res* 40:e29.
35. Klapacz J, Bhagwat AS (2002) Transcription-dependent increase in multiple classes of base substitution mutations in *Escherichia coli*. *J Bacteriol* 184:6866–6872.
36. Hudson RE, Bergthorsson U, Ochman H (2003) Transcription increases multiple spontaneous point mutations in *Salmonella enterica*. *Nucleic Acids Res* 31:4517–4522.
37. Eyre-Walker A, Bulmer M (1995) Synonymous substitution rates in enterobacteria. *Genetics* 140:1407–1412.
38. Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* 4:222–230.
39. Martincorena I, Seshasayee AS, Luscombe NM (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485: 95–98.
40. Streisinger G, et al. (1966) Frameshift mutations and the genetic code. *Cold Spring Harb Symp Quant Biol* 31:77–84.
41. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* 274:775–780.
42. Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. *Proc Natl Acad Sci USA* 96:12638–12643.
43. Coulondre C, Miller JH (1977) Genetic studies of the *lac* repressor. IV. Mutagenic specificity in the *lacI* gene of *Escherichia coli*. *J Mol Biol* 117:577–606.
44. Leong PM, Hsia HC, Miller JH (1986) Analysis of spontaneous base substitutions generated in mismatch-repair-deficient strains of Escherichia coli. *J Bacteriol* 168: 412–416.
45. Cupples CG, Miller JH (1989) A set of *lacZ* mutations in *Escherichia coli* that allow rapid detection of each of the six base substitutions. *Proc Natl Acad Sci USA* 86: 5345–5349.
46. Rewinski C, Marinus MG (1987) Mutation spectrum in *Escherichia coli* DNA mismatch repair deficient (*mutH*) strain. *Nucleic Acids Res* 15:8205–8215.
47. Schaaper RM, Dunn RL (1991) Spontaneous mutation in the *Escherichia coli lacI* gene. *Genetics* 129:317–326.
48. Garcia-Gonzalez A, Rivera-Rivera RJ, Massey SE (2012) The presence of the DNA repair genes *mutM, mutY, mutL*, and *mutS* is related to proteome size in bacterial genomes. *Front Genet* 3:3.
49. Mellon I, Champe GN (1996) Products of DNA mismatch repair genes *mutS* and *mutL* are required for transcription-coupled nucleotide-excision repair of the lactose operon in *Escherichia coli*. *Proc Natl Acad Sci USA* 93:1292–1297.
50. Bhagwat AS, Lieb M (2002) Cooperation and competition in mismatch repair: Very short-patch repair and methyl-directed mismatch repair in *Escherichia coli*. *Mol Microbiol* 44:1421–1428.
51. Lieb M (1987) Bacterial genes *mutL, mutS*, and *dcm* participate in repair of mismatches at 5-methylcytosine sites. *J Bacteriol* 169:5241–5246.
52. Robertson AB, Matson SW (2012) Reconstitution of the Very Short Patch repair pathway from *Escherichia coli*. *J Biol Chem* 287:32953–32966.
53. Su SS, Lahue RS, Au KG, Modrich P (1988) Mispair specificity of methyl-directed DNA mismatch correction *in vitro*. *J Biol Chem* 263:6829–6835.
54. Kramer B, Kramer W, Fritz HJ (1984) Different base/base mismatches are corrected with different efficiencies by the methyl-directed DNA mismatch-repair system of *E. coli*. *Cell* 38:879–887.
55. Dohet C, Wagner R, Radman M (1985) Repair of defined single base-pair mismatches in *Escherichia coli*. *Proc Natl Acad Sci USA* 82:503–505.
56. Nowosielska A, Wrzesiński M, Nieminuszczy J, Janion C, Grzesiuk E (2005) Mutator activity and specificity of *Escherichia coli dnaQ49* allele—effect of *umuDC* products. *Mutat Res* 572:113–122.
57. Kim M, Huang T, Miller JH (2003) Competition between MutY and mismatch repair at A x C mispairs *In vivo*. *J Bacteriol* 185:4626–4629.
58. Watson JD, Crick FHC (1953) The structure of DNA. *Cold Spring Harb Symp Quant Biol* 18:123–131.
59. Topal MD, Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. *Nature* 263:285–289.
60. Yu H, Eritja R, Bloom LB, Goodman MF (1993) Ionization of bromouracil and fluorouracil stimulates base mispairing frequencies with guanine. *J Biol Chem* 268: 15935–15943.
61. Wang W, Hellinga HW, Beese LS (2011) Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proc Natl Acad Sci USA* 108:17644–17648.
62. Bebenek K, Pedersen LC, Kunkel TA (2011) Replication infidelity via a mismatch with Watson-Crick geometry. *Proc Natl Acad Sci USA* 108:1862–1867.
63. Schaaper RM, Dunn RL (1987) Spectra of spontaneous mutations in *Escherichia coli* strains defective in mismatch correction: The nature of *in vivo* DNA replication errors. *Proc Natl Acad Sci USA* 84:6220–6224.
64. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* 13:660–665.
65. Agier N, Fischer G (2012) The mutational profile of the yeast genome is shaped by replication. *Mol Biol Evol* 29:905–913.
66. Johnson KA (1993) Conformational coupling in DNA polymerase fidelity. *Annu Rev Biochem* 62:685–713.
67. Fijalkowska IJ, Jonczyk P, Tkaczyk MM, Bialoskorska M, Schaaper RM (1998) Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *Proc Natl Acad Sci USA* 95:10020–10025.
68. Michaels ML, Miller JH (1992) The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine). *J Bacteriol* 174:6321–6325.
69. Wolff E, Kim M, Hu K, Yang H, Miller JH (2004) Polymerases leave fingerprints: Analysis of the mutational spectrum in *Escherichia coli rpoB* to assess the role of polymerase IV in spontaneous mutation. *J Bacteriol* 186:2900–2905.
70. Taylor JS (2002) New structural and mechanistic insight into the A-rule and the instructional and non-instructional behavior of DNA photoproducts and other lesions. *Mutat Res* 510:55–70.
71. Moxon R, Bayliss C, Hood D (2006) Bacterial contingency loci: The role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* 40:307–333.
72. Chen F, et al. (2010) Multiple genetic switches spontaneously modulating bacterial mutability. *BMC Evol Biol* 10:277.
73. Blattner FR, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–1462.
74. Riley M, et al. (2006) *Escherichia coli* K-12: A cooperatively developed annotation snapshot—2005. *Nucleic Acids Res* 34:1–9.
75. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA* 97:6640–6645.
76. Foster PL (2006) Methods for determining spontaneous mutation rates. *Methods Enzymol* 409:195–213.
77. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
78. Albers CA, et al. (2011) Dindel: Accurate indel calls from short-read data. *Genome Res* 21:961–973.
79. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919.
80. Zar JH (1984) *Biostatistical Analysis* (Prentice Hall, Englewood Cliffs, NJ).
81. Rice JA (1995) *Mathematical Statistics and Data Analysis* (Wadsworth, Belmont, CA).