Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in *Escherichia coli*

Ashok S. Bhagwat[a,b,1], Weilong Hao[c], Jesse P. Townes[d], Heewook Lee[e], Haixu Tang[e], and Patricia L. Foster[d]

[a]Department of Chemistry, Wayne State University, Detroit, MI 48202; [b]Department of Immunology and Microbiology, Wayne State University School of Medicine, Detroit, MI 48201; [c]Department of Biological Sciences, Wayne State University, Detroit, MI 48202; [d]Department of Biology, Indiana University, Bloomington, IN 47405; and [e]School of Informatics and Computing, Indiana University, Bloomington, IN 47405

Edited by Graham C. Walker, Massachusetts Institute of Technology, Cambridge, MA, and approved January 5, 2016 (received for review November 12, 2015)

The rate of cytosine deamination is much higher in single-stranded DNA (ssDNA) than in double-stranded DNA, and copying the resulting uracils causes C to T mutations. To study this phenomenon, the catalytic domain of APOBEC3G (A3G-CTD), an ssDNA-specific cytosine deaminase, was expressed in an *Escherichia coli* strain defective in uracil repair (*ung* mutant), and the mutations that accumulated over thousands of generations were determined by whole-genome sequencing. C:G to T:A transitions dominated, with significantly more cytosines mutated to thymine in the lagging-strand template (LGST) than in the leading-strand template (LDST). This strand bias was present in both repair-defective and repair-proficient cells and was strongest and highly significant in cells expressing A3G-CTD. These results show that the LGST is accessible to cellular cytosine deaminating agents, explains the well-known GC skew in microbial genomes, and suggests the APOBEC3 family of mutators may target the LGST in the human genome.

uracil-DNA glycosylase | APOBEC3A | APOBEC3B | kataegis | cancer genome mutations

Pairing of complementary DNA strands protects the DNA bases against modification by a number of hydrolytic, oxidizing, and alkylating chemicals (1–4). For example, water reacts with cytosine, creating uracil, and the rate of this reaction in single-stranded DNA (ssDNA) is more than 100-fold the rate in double-stranded DNA [dsDNA (5–7)]. Uracil-DNA glycosylase (Ung) excises uracils created by cytosine deamination in both ssDNA and dsDNA, resulting in abasic (AP) sites. In dsDNA, the AP sites are replaced with cytosines as a result of copying of the guanine in the complementary strand during repair by the base-excision repair (BER) pathway (8). In contrast, cytosine deaminations occurring in ssDNA are problematic because the complementary strand is not available to the BER pathway. Uracils that escape repair create C:G to T:A mutations, and incomplete repair of uracils can result in persistent AP sites and strand breaks that can destabilize the genome. Hence, identifying ssDNA regions that are susceptible to damage will increase our understanding of causes of mutations and genome instability.

The AID/APOBEC (activation-induced deaminase/apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) family of DNA–cytosine deaminases are specific for ssDNA (9, 10). They are found only in vertebrates and are good probes of ssDNA in cells because of their relatively small size (about 190-amino acid catalytic domain). They are active in heterologous hosts such as *Escherichia coli* (11, 12) and yeast (13–15) and cause mutations in the same sequence context as in their known targets, such as Ig genes and the DNA copy of HIV-1 genome (11, 16, 17). In particular, the catalytic domain of human APOBEC3G (A3G-CTD) was expressed in an engineered yeast strain lacking the *UNG* gene and was shown to target ssDNA generated through aberrant resection of telomeric ends (13).

To similarly probe ssDNA in *E. coli*, we expressed A3G-CTD on a plasmid (pA3G-CTD) in an *ung* mutant *E. coli* strain, and the resulting mutations were determined by whole-genome sequencing. These results were compared with results from similar experiments performed using *ung* mutant cells without a plasmid, *ung* mutant cells expressing catalytically inactive A3G-CTD (pA3G-CTDmut), and published results from the ung^+ wild-type (WT) *E. coli* strain (18, 19).

Results and Discussion

C:G to T:A Mutations in the Absence of Ung. About 200 independent lines of the four *E. coli* strains, PFM2, PFM16 (=PFM2 Δ*ung*), PFM275 (=PFM16 with plasmid pA3G-CTD), and PFM277 (=PFM16 with plasmid pA3G-CTDmut), were grown for more than 1,000 generations each, and the accumulated mutations in each line were determined by whole-genome sequencing (Table 1). More than 1,100 base-substitution mutations were identified in these lines and were classified according to the mutation type and whether they were in coding or noncoding DNA. These data are summarized in Table S1. A small number of the mutations (<10%) in each strain were insertion/deletions (Table S2), and these were not analyzed further.

In the absence of pA3G-CTD, the *ung* mutant strain had a twofold higher mutation rate compared with the WT strain (Table 1). This weak mutator phenotype contrasts with the 160-fold increase in the overall mutation rate that occurred when a pathway for the repair of 8-oxoguanines was blocked in the same genetic background (18) and shows that cytosine deamination is not the major form of endogenous DNA damage. The mutator phenotype of the *ung* mutant strain was almost entirely the result

Significance

C:G to T:A mutations constitute the largest class of spontaneous base substitutions in all organisms. These mutations are thought to be a result of cytosine deaminations, but what promotes these deaminations is unclear. We confirm here the hypothesis that they occur predominantly in single-stranded DNA (ssDNA) and identify the ssDNA in the lagging strand template as the preferred site of C:G to T:A mutations. As a consequence, replication creates a strand bias in these mutations, and this overwhelms any strand bias resulting from transcription. These results explain a long-recognized bias in base composition of microbial genomes called GC skew and predicts that C:G to T:A mutations created by the APOBEC3 family deaminases in cancer genomes should occur with the same strand bias.

Author contributions: A.S.B. and P.L.F. conceived the project; A.S.B. and P.L.F. designed research; J.P.T. performed research; A.S.B., W.H., J.P.T., H.L., H.T., and P.L.F. analyzed data; and A.S.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: axb@chem.wayne.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1522325113/-/DCSupplemental.

Table 1. Base-substitution mutations

Strain	WT, no plasmid*		*ung*, no plasmid		*ung*, pA3G-CTD		*ung*, pA3G-CTDmut	
Independent lines	61		50		48		46	
Generations per line	4,230		3,372		1,137		1,139	
Mutations		Rate ($\times 10^{10}$)†		Rate ($\times 10^{10}$)†		Rate ($\times 10^{10}$)†		Rate ($\times 10^{10}$)†
Total (*n*, %)	246 (100)	2.1	358 (100)	4.6	411 (100)	16.2	171 (100)	7.0
Transition								
Total (*n*, %)	136 (55)	1.1	280 (78)	3.6	368 (90)	14.5	127 (74)	5.2
C:G to T:A (*n*, %)	86 (35)	1.4	247 (69)	6.2	349 (85)	27.1	104 (61)	8.4
Transversion	110 (45)	0.9	78 (22)	1.0	43 (10)	1.7	44 (26)	1.8

*Data from refs. 18 and 19.

†Mutations per nucleotide per generation.

of a fourfold increase in the rate of C:G to T:A mutations, which increased from 35% of all of the mutations in the WT strain to 69% in the *ung* mutant strain, consistent with the lack of repair of U•G mispairs in the *ung* mutant strain. These mutation rates obtained using whole-genome sequencing are somewhat lower but are still consistent with previous studies using reporter genes that found loss of Ung in *E. coli* increased the frequencies of spontaneous mutations 1.5- to fivefold overall, and C:G to T:A mutations seven- to 50-fold (12, 20–24). When the plasmid pA3G-CTD was introduced in the *ung* mutant strain, the overall mutation rate increased a further 3.5-fold, principally as a result of a fourfold increase in C:G to T:A mutations, which constituted 85% of all of the mutations obtained (Table 1).

In all the strains tested, mutated cytosines were not localized to specific genomic regions, such as the origin or terminus of replication, but were distributed across the genome (Fig. 1*A*). This apparent lack of clustering may be because of the relatively small number of mutations obtained (several hundred), and a much larger number of mutations could reveal regional clustering (25). In the *ung* mutant strain both without a plasmid and carrying pA3G-CTDmut plasmid, there was no strong preference for flanking nucleotides (Fig. 1*B* and Fig. S1). However, when catalytically active A3G-CTD was expressed in the *ung* mutant strain, C:G to T:A mutations occurred preferentially in 5′CC**C**R3′/5′Y**G**GG3′ sequences (target cytosine in bold and underlined; R is purine, Y is pyrimidine; Fig. 1*B*). Twenty-eight percent of the mutations occurred in this sequence, whereas only 2.5% were expected on the basis of the composition of *E. coli* genome ($P \sim 0$ for the difference). This sequence preference of APOBEC3G in *E. coli* has been reported before (11) and is similar to that found in mammalian targets (12, 16, 26, 27).

Replication Strand Bias in Mutations Caused by APOBEC3G. C:G to T:A mutations obtained in each *E. coli* strain were mapped to the two *E. coli* replichores and then separated into two groups on the basis of whether the mutated cytosine was in the lagging strand template (LGST) or in the leading strand template (LDST) for replication (Fig. S2). The numbers of C to T mutations in the two groups are presented in Table 2. Although in all four strains there were more mutations in the LGST than in the LDST, the LGST to LDST mutation ratio was highest, 3 to 4, when pA3G-CTD was present in the *ung* mutant strain (Table 2). This strong mutational strand bias was present in both *E. coli* replichores and was statistically highly significant ($P << 0.0001$). The presence of pA3G-CTD in the *ung* mutant strain increased the rate of C to T mutation in both LGST and LDST; the increase was sixfold when C was in the LGST and twofold when C was in the LDST (Tables 1 and 2). Thus, most of these strand-biased mutations must be a result of cytosine deaminations catalyzed by A3G-CTD.

There are two reasons for the observed excess of C to T mutations in the LGST. First, A3G-CTD acts on ssDNA (28), and during replication, the LGST has longer stretches of this substrate than does the LDST (Fig. 2) (29). Second, once a cytosine in the LGST is deaminated to uracil (class I uracils in Fig. 2), it will be immediately copied by the lagging strand polymerase, creating a C:G to T:A mutation. This is likely to happen with or

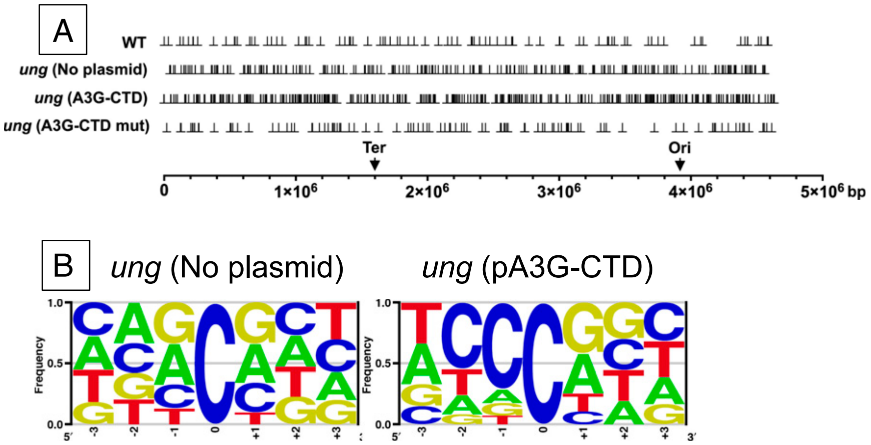

Fig. 1. Distribution and sequence context of C:G to T:A mutations. (*A*) Distribution of mutations. The positions of C:G to T:A mutations were plotted using the Graph Prism 6 for Mac software and are shown using the perpendicularity symbol (⊥); the strains in which the mutations occurred are indicated on the left. A straight line at the bottom is used to represent the *E. coli* genome, and downward arrows mark the positions of the replication origin (Ori) and terminus (Ter). (*B*) Sequence context of mutations. The three nucleotides on either side of the C:G pair that was mutated to T:A were used to create LOGOS plots using enoLOGOS software (63).

Table 2. Replication strand bias in C:G to T:A mutations

Replichore	C in the LGST scored mutation		C in the LDST scored mutation		Total	Ratio of C in LGST/LDST	*P* value*	Ratio normalized for C†
WT								
Right	C to T	30	G to A	14	44	2.1	0.06	2.3
Left	G to A	31	C to T	11	42	2.8	0.02	3.0
Total	C:G to T:A	61	C:G to T:A	25	86	2.4	0.003	2.6
ung								
Right	C to T	64	G to A	55	119	1.2	0.41	1.2
Left	G to A	77	C to T	51	128	1.5	0.06	1.6
Total	C:G to T:A	141	C:G to T:A	106	247	1.3	0.05	1.4
ung (pA3G-CTD)								
Right	C to T	151	G to A	35	186	4.3	4x10^{-11}	4.6
Left	G to A	126	C to T	37	163	3.4	6x10^{-08}	3.6
Total	C:G to T:A	277	C:G to T:A	72	349	3.8	2x10^{-17}	4.1
ung (pA3G-CTDmut)								
Right	C to T	37	G to A	17	54	2.2	0.03	2.3
Left	G to A	26	C to T	24	50	1.1	0.72	1.2
Total	C:G to T:A	63	C:G to T:A	41	104	1.5	0.08	1.6

**P* value is based on χ^2 test of observed versus expected values. Expected values were calculated from the ratio of cytosines in LGST/LDST in each replichore.

†Normalized for the number of cytosines in the LGST versus the LDST (ratio = 0.941 for the right replichore and 0.936 for the left replichore).

without A3G-CTD in cells in both *ung* mutant and *ung*$^+$ genetic backgrounds (see following). Hence, the strand bias in mutations when A3G-CTD was expressed in the *ung* mutant strain strongly suggests that cytosines in LGST were much more accessible to the deaminase than cytosines in the LDST.

Replication Strand Bias in Mutations in WT *E. coli*. The WT strain lacking A3G-CTD also acquired twofold more C:G to T:A mutations when C was in the LGST than when it was in the LDST (P = 0.003; Table 2). This strand bias was reported previously, but it was not attributed to cytosine deaminations in that study (19). In light of the conclusion presented here that the A3G-CTD can access LGST, it is attractive to suggest that water, other small molecules, or proteins can access the LGST and deaminate cytosines in WT cells lacking A3G-CTD. Most uracils are likely to be copied immediately by DNA polymerase III, creating C:G to T:A mutations. However, occasionally, Ung may find the uracil before the polymerase and excise it, creating an AP site. This AP site cannot be replaced with cytosines by BER because of the lack of the complementary strand (Fig. 2). If the AP site is copied by one of the translesion synthesis DNA polymerases, it will generate a C:G to T:A mutation in about 50–70% of the cases, according to the "A" rule (30–32). Thus, both of these replicative pathways in *ung*$^+$ cells will create more C:G to T:A mutations in the LGST than in the LDST.

Alternately, the AP site may be processed by an AP endonuclease, creating a double-strand break that will stop DNA replication (Fig. 2). If the break is not repaired through recombination, this will lead to replication fork collapse (29). This model further predicts that the strand breaks generated by AID/APOBEC enzymes should lead to cell death unless they are repaired through homology-directed repair or nonhomologous end-joining (in eukaryotes). These mutagenic and genome destabilizing effects of uracils created in the LGST contrast with the uracils created by deamination in dsDNA (class II uracils; Fig. 2). In WT cells, these uracils will be excised by Ung and repaired efficiently by BER, restoring the C:G pair.

Implications for Mutations in Cancer Genomes. Many of the mutations in human tumors display "signatures" of APOBEC3 family

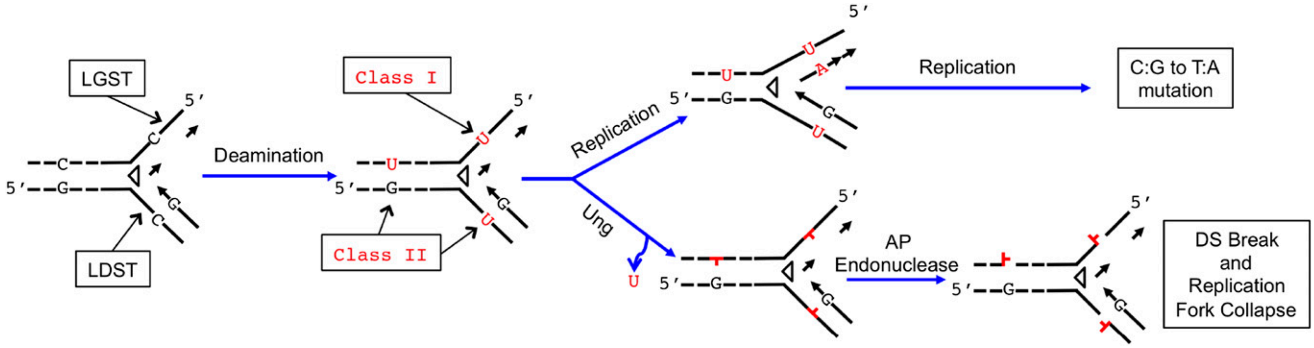

Fig. 2. Consequences of cytosine deaminations at the replication fork. Deamination of three cytosines and the consequences of processing of the resulting uracils through replication or repair pathways are shown. Copying of the LGST is discontinuous, and two Okazaki fragments are shown. The open arrowhead represents the helicase DnaB and is pointed in the overall direction of replication. Class 1 and class 2 refer to two classes of uracils generated through cytosine deamination and have different biochemical consequences. The class 2 uracils are likely to be replaced with cytosines through BER, whereas the class 1 uracils lead to C:G to T:A mutations or double-strand (DS) breaks. The latter may be repaired through recombinational repair or result in the collapse of the replication fork.

deaminases (33, 34). In particular, a signature defined by C to T or C to G mutations in TCW context (W is A or T) is found in the genomes of a number of different cancers and is attributed to mutations caused by APOBEC3 enzymes (33, 35, 36). As in *E. coli*, C:G to T:A mutations in tumors would be caused when uracils created in LGST by one of the APOBEC3s are copied by the replicative DNA polymerases δ, or when AP sites created by the excision of uracils by UNG2 are copied by polymerase η, resulting in an insertion of adenines (37). Alternately, the AP sites may be copied by Rev1, creating C:G to G:C transversions (38, 39). One characteristic of cancer genome mutations is that they are found in clusters, suggesting that they occur in genomic regions that contain stretches of ssDNA (34). Although a number of cellular processes, including replication, transcription, and recombination, have been proposed as sources of ssDNA targets for APOBEC3s (13, 34, 40–42), experimental evidence for these ideas is limited. The data presented here suggest that the LGST at the replication forks of rapidly dividing cancer cells would be accessible to these enzymes, and hence the cytosines mutated in cancer genomes should be found preferentially in the LGST compared with the LDST. Recent results from whole-genome sequencing of mutations in yeast expressing APOBEC3A or APOBEC3B (43) and in 590 human tumors (44) found a strand bias in mutations consistent with this prediction.

Implications for the Microbial GC Skew. Bacterial genomes contain a bias in base composition that is connected with replication. With few exceptions, these genomes have an excess of guanines over cytosines in the LGST [when normalized for the total G+C content (45)]. The hypotheses proposed to explain this "GC skew" include: the two DNA strands are replicated with different accuracy, the two strands are repaired with different frequency, and/or cytosines in the two strands deaminate at different rates (46–49). The data presented here strongly support the last of these hypotheses. As the LGST accumulates C to T mutations at least twice as frequently as the LDST, in the absence of any selection, there would be a progressive loss of cytosines from LGST creating the GC skew.

Mutations in *ung* Mutant Strain Lacking Active A3G-CTD. The LGST/LDST ratio for C:G to T:A transitions in *ung* mutant cells without A3G-CTD or with inactive A3G-CTD was smaller than the ratio in the WT strain (Table 2). This was despite the fact that the overall C:G to T:A mutation rate was fourfold higher in the *ung* mutant strain without a plasmid and sixfold higher in the strain containing pA3G-CTDmut compared with the WT strain (Table 1). In the absence of Ung, class II U•G pairs will not be repaired, and both class I and class II U•G pairs will be replicated by DNA polymerases, increasing the overall mutation rate (Fig. 2). The class II uracils occur in dsDNA and have no strand bias. As a consequence, the addition of the resulting mutations to the overall C:G to T:A mutations reduces the LGST/LDST mutation ratio. Both the overall C:G to T:A mutation rate and LGST/LDST bias is slightly higher in the strain containing pA3G-CTDmut compared with the strain without a plasmid, suggesting the E259A substitution in A3G-CTD may have a residual ability to deaminate cytosines.

Lack of Strand Bias with Respect to Transcription. As observed previously with the WT strain (18, 19), when normalized to the number of nucleotides in coding versus noncoding DNA, all three *ung* mutant strains showed a twofold bias against mutations occurring in coding sequences (Table S1). In *ung*$^+$ strains, this bias was shown to be a result of better mismatch repair of the coding regions than of noncoding regions (19). Considering all the genes together, there was no significant bias in the frequency of C to T mutations in the transcribed versus the nontranscribed (coding) strand (Table S3). It is possible that a larger mutational harvest would reveal a significant transcription strand bias, especially in highly transcribed genes, but in our data, any such bias was overwhelmed by the observed replication strand bias. In contrast, Lada et al. found that most mutations caused by the lamprey cytosine deaminase, PmCDA1, in nondividing yeast were correlated with transcription (40). Together, these observations suggest that in dividing cells, the greatest source of C:G to T:A mutations is the deamination of cytosines in the template for the lagging strand synthesis, but transcription may play a larger role in nondividing cells.

Concluding Remarks

LGST is protected against digestion by nucleases by the ssDNA-binding protein (SSB) in bacteria and the replication protein A (RPA) in eukaryotes (50, 51). However, structural studies of both the proteins have found that both SSB and RPA wrap DNA around themselves in a conserved structure called an OB-fold, and the DNA bases largely lie on the outside of these complexes (52, 53). Thus, SSB or RPA are unlikely to prevent access to DNA bases by reactive small molecules or enzymes. This leaves the bases in LGST in both bacterial and eukaryotic replication forks highly susceptible to chemical and enzymatic damage. Therefore, the LGST is a chink in the armor with which the cell protects its DNA.

Experimental Procedures

Bacterial Media, Strains, and Plasmids. Cultures were grown in liquid Miller Luria broth (LB), or on Miller LB agar plates (54). When appropriate, antibiotics were added to the growth media at the following concentrations: carbenicillin (Carb), 100 μg/mL; kanamycin (Kn), 50 μg/mL; and rifampicin (Rif), 100 μg/mL.

The *ung* mutant *E. coli* K12 strain used in this study, PFM16, was derived from WT strain PFM2 (18, 19). The Δ*ung*::Kn allele was moved into PFM2 from the Keio strain JW2564 (55) via P1 bacteriophage transduction, and the kanamycin-resistance gene was then removed using FLP recombination (56), leaving an in-frame scar sequence that encodes a 34-amino acid peptide.

The plasmid, pA3G-CTD, has been described previously (57), and the E259A mutant of A3G-CTD was constructed using the QuikChange site-directed mutagenesis strategy (Agilent Technologies). To remove restriction barriers, plasmid DNA was first transformed into *E. coli* strain DH5α, selecting for resistance to Carb (58). Plasmid DNA was then purified from DH5α, using Zyppy Plasmid Miniprep Kit (Zymo Research), and it was introduced into PFM16 via TSS transformation (59). PFM16/pA3G-CTD was designated PFM275, and PFM16/pA3G-CTDmut was designated PFM277.

Estimation of Mutation Rates by Fluctuation Tests. Mutation rates to rifampicin-resistance (Rif^R) were estimated using fluctuation tests as described (60). Cultures of PFM275 and PFM277 were grown in LB broth containing Carb to maintain the presence of the plasmids. Because both A3G-CTD and A3G-CTDmut are under the control of *lac* promoter and inducible by IPTG (isopropyl β-D-1-thiogalactopyranoside), fluctuation assays with PFM275 and PFM277 were done both in the absence and presence of 1mM isopropyl β-D-1-thiogalactopyranoside. The addition of isopropyl β-D-1-thiogalactopyranoside made little difference to the mutation rates, so it was not added to the plates for the mutation accumulation (MA) procedure. Mutation rates from fluctuation tests were calculated using the Ma-Sandri-Sarkar maximum likelihood method (61) implemented using the FALCOR web tool found at www.mitochondria.org/protocols/FALCOR.html (62).

MA Protocol. For PFM2 and PFM16, MA-line founders were generated by inoculating LB broth from a freezer stock, growing the culture at 37 °C overnight, and plating an appropriate dilution on LB agar plates (19). For PFM275 and PFM277, founders were generated by streaking from the freezer stock onto LB agar plates containing Carb and incubating the plates at 37 °C overnight. Two well-isolated colonies were then excised, soaked for 30 min in 0.01% gelatin solution in 0.85% NaCl, and vortexed for 60 s. Appropriate dilutions were then plated onto LB agar plates containing Carb to obtain at least 30 well-isolated colonies from each of the two parental colonies. These separate lines were then propagated for the duration of the MA procedure.

Each MA line was streaked for single colonies each day on an LB agar plate (PFM2 and PFM16) or LB agar plate with Carb (PFM275 and PFM277) and incubated overnight at 37 °C for 23–25 h. A well-isolated colony of each line was then picked and restreaked. After passage, plates were stored at 4 °C for a maximum of 2 d, to be used again if the original streaking did not yield well-isolated colonies. The number of required passages was determined from the mutation rates, as determined by fluctuation tests. The propagation of PFM2 was previously described (19). Both PFM275 and PFM277 were propagated for 40 passages. PFM16 was initially streaked for 55 passages, and then frozen stocks were made. Sequencing of the MA lines at that point revealed that the mutation rate had been overestimated. Lines were then reinoculated from the frozen stocks and streaked for another 65 passages, giving a total of 120 passages.

Estimation of Generations. The number of generations between passages was estimated from the diameter of the colonies and the number of cells in colonies of a given diameter for each strain, as described (19). The average generation count was ∼27.5 per day. Generation counts for each line for the course of the MA experiment were totaled, and the average of this number of all of the lines for each strain multiplied by the number of lines was used to calculate the mutation rates.

Genomic DNA Preparation, Library Construction, and Whole-Genome Sequencing. Genomic DNA was purified from 0.8 mL of overnight cultures in LB (PFM2 and PFM16) or LB broth with Carb (PFM275 and PFM277). DNA concentration and purity were assessed using an Epoch Microplate Spectrophotometer (BioTek Instruments, Inc.).

Before library construction, the *ung* deletion in each line was confirmed by visualizing the appropriate-sized PCR product of the genomic DNA, using primers ungFW (5′TGTCCAGCAGCCAGAAAGAG3′) and ungRV (5′ATAAAT-CAGCCGGGTGGCAA3′). To ensure presence of the appropriate plasmid, diagnostic restriction digestion were performed with each of the PFM275 and PFM277 MA lines. Plasmid DNA containing WT A3G-CTD has a unique Bgl1 restriction site, whereas plasmid containing A3G-CTDmut has a unique HaeII restriction site.

Libraries for PFM2 and PFM16 were made by the Beijing Genomics Institute and sequenced using the Illumina HiSeq2000 platform. Libraries for PFM275 and PFM277 were made by the Indiana University Center for Genomics and Bioinformatics and sequenced at the University of New Hampshire Hubbard Center for Genomic Studies, using the Illumina HiSeq2500 platform.

For quality control purposes, reads with any of the following characteristics were discarded, as described (19): $\geq$10% unreadable bases, $\geq$20% low-quality ($\leq Q_{20}$) bases, adapter contamination ($\geq$15 bp overlap allowing up to 3 bp mismatch), and duplicate read-pairs. After this filtering, retained reads averaged 91.1%. Two MA lines from the PFM277 set had less than 30× sequence coverage and were eliminated.

Single-Nucleotide Polymorphism. Procedures for single-nucleotide polymorphism calling were as described (18, 19). National Center for Biotechnology Information reference sequence NC_000913.2 was used as the reference genome sequence because it more accurately matches the sequence of PFM2 than the subsequent release. The sequences and mutations reported in this article have been deposited at the National Center for Biotechnology Information Sequence Read Archive (BioProject accession no. SPR013707) and in the IUScholarWorks Repository (URI TBA).

Some MA lines carried shared mutations arising either from mutations that occurred during the initial growth of founders or from cross-contamination during streaking. Such mutations were assigned to one MA line based on deduced lineage, if possible; otherwise, mutations were assigned randomly. Two MA lines of PFM275 that had many shared mutations but only one or zero unique mutations were eliminated.

Mutation Annotation. Variants were annotated using custom scripts, as described (19).

ACKNOWLEDGMENTS. A.S.B. thanks Gad Getz and Michael Lawrence (Broad Institute), and Steve Roberts (Washington State University) for sharing unpublished results and manuscripts. We thank the following current and past members of the P.L.F. laboratory for technical assistance: C. Coplen, J. Ferlmann, N. Ivers, D. Osiecki, S. Riffert, B. Souders, L. Whitson, N. Yahaya, A. Ying Yi Tan, and Ellen M. Popodi for useful discussions and B. Wanner and the National BioResource Project-E. coli at the National Institute of Genetics for providing bacterial strains and plasmids. This research was supported by US Army Research Office Multidisciplinary University Research Initiative (MURI) Award W911NF-09-1-0444 to P.L.F., H.T., M. Lynch, and S. E. Finkel; NIH Grant GM 57200 (to A.S.B.); and funds from Wayne State University.

1. Hayatsu H (1976) Bisulfite modification of nucleic acids and their constituents. *Prog Nucleic Acid Res Mol Biol* 16:75–124.
2. Paleček E, Bartošík M (2012) Electrochemistry of nucleic acids. *Chem Rev* 112(6): 3427–3481.
3. Sinden RR (1994) *DNA Structure and Function* (Academic Press, San Diego, CA), p 398.
4. Singer B, Grunberger D (1983) *Molecular Biology of Mutagens and Carcinogens* (Plenum Press, New York, NY), p 347.
5. Frederico LA, Kunkel TA, Shaw BR (1990) A sensitive genetic assay for the detection of cytosine deamination: Determination of rate constants and the activation energy. *Biochemistry* 29(10):2532–2537.
6. Lindahl T, Nyberg B (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* 13(16):3405–3410.
7. Shen JC, Rideout WM, 3rd, Jones PA (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res* 22(6):972–976.
8. Lindahl T (1982) DNA repair enzymes. *Annu Rev Biochem* 51:61–87.
9. Chiu YL, Greene WC (2008) The APOBEC3 cytidine deaminases: An innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu Rev Immunol* 26:317–353.
10. Conticello SG, Langlois MA, Yang Z, Neuberger MS (2007) DNA deamination in immunity: AID in the context of its APOBEC relatives. *Adv Immunol* 94:37–73.
11. Harris RS, Petersen-Mahrt SK, Neuberger MS (2002) RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol Cell* 10(5):1247–1253.
12. Petersen-Mahrt SK, Harris RS, Neuberger MS (2002) AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* 418(6893):99–103.
13. Chan K, et al. (2012) Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS Genet* 8(12): e1003149.
14. Lada AG, et al. (2013) Genome-wide mutation avalanches induced in diploid yeast cells by a base analog or an APOBEC deaminase. *PLoS Genet* 9(9):e1003736.
15. Mayorov VI, et al. (2005) Expression of human AID in yeast induces mutations in context similar to the context of somatic hypermutation at G-C pairs in immunoglobulin genes. *BMC Immunol* 6:10.
16. Beale RC, et al. (2004) Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: Correlation with mutation spectra in vivo. *J Mol Biol* 337(3):585–596.
17. Rogozin IB, Pavlov YI (2006) The cytidine deaminase AID exhibits similar functional properties in yeast and mammals. *Mol Immunol* 43(9):1481–1484.
18. Foster PL, Lee H, Popodi E, Townes JP, Tang H (2015) Determinants of spontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing. *Proc Natl Acad Sci USA* 112(44):E5990–E5999.
19. Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* 109(41):E2774–E2783.
20. Duncan BK, Rockstroh PA, Warner HR (1978) *Escherichia coli* K-12 mutants deficient in uracil-DNA glycosylase. *J Bacteriol* 134(3):1039–1045.
21. Duncan BK, Weiss B (1982) Specific mutator effects of ung (uracil-DNA glycosylase) mutations in *Escherichia coli*. *J Bacteriol* 151(2):750–755.
22. Foster PL (1990) *Escherichia coli* strains with multiple DNA repair defects are hyperinduced for the SOS response. *J Bacteriol* 172(8):4719–4720.
23. Lutsenko E, Bhagwat AS (1999) Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells. A model, its experimental support and implications. *Mutat Res* 437(1):11–20.
24. Nordman J, Wright A (2008) The relationship between dNTP pool levels and mutagenesis in an *Escherichia coli* NDP kinase mutant. *Proc Natl Acad Sci USA* 105(29): 10197–10202.
25. Foster PL, Hanson AJ, Lee H, Popodi EM, Tang H (2013) On the mutational topology of the bacterial genome. *G3 (Bethesda)* 3(3):399–407.
26. Lecossier D, Bouchonnet F, Clavel F, Hance AJ (2003) Hypermutation of HIV-1 DNA in the absence of the Vif protein. *Science* 300(5622):1112.
27. Zhang H, et al. (2003) The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature* 424(6944):94–98.
28. Harris RS, et al. (2003) DNA deamination mediates innate immunity to retroviral infection. *Cell* 113(6):803–809.
29. Langston LD, O'Donnell M (2006) DNA replication: Keep moving and don't mind the gap. *Mol Cell* 23(2):155–160.
30. Lawrence CW, Borden A, Banerjee SK, LeClerc JE (1990) Mutation frequency and spectrum resulting from a single abasic site in a single-stranded vector. *Nucleic Acids Res* 18(8):2153–2157.
31. Reuven NB, Arad G, Maor-Shoshani A, Livneh Z (1999) The mutagenesis protein UmuC is a DNA polymerase activated by UmuD′, RecA, and SSB and is specialized for translesion replication. *J Biol Chem* 274(45):31763–31766.
32. Tang M, et al. (1999) UmuD′(2)C is an error-prone DNA polymerase, Escherichia coli pol V. *Proc Natl Acad Sci USA* 96(16):8919–8924.
33. Alexandrov LB, et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013) Signatures of mutational processes in human cancer. *Nature* 500(7463):415–421.
34. Nik-Zainal S, et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* 149(5):979–993.
35. Roberts SA, Gordenin DA (2014) Hypermutation in human cancer genomes: Footprints and mechanisms. *Nat Rev Cancer* 14(12):786–800.

36. Swanton C, McGranahan N, Starrett GJ, Harris RS (2015) APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. *Cancer Discov* 5(7):704–712.
37. Masutani C, Kusumoto R, Iwai S, Hanaoka F (2000) Mechanisms of accurate translesion synthesis by human DNA polymerase eta. *EMBO J* 19(12):3100–3109.
38. Lawrence CW, Hinkle DC (1996) DNA polymerase zeta and the control of DNA damage induced mutagenesis in eukaryotes. *Cancer Surv* 28:21–31.
39. Prakash S, Johnson RE, Prakash L (2005) Eukaryotic translesion synthesis DNA polymerases: Specificity of structure and function. *Annu Rev Biochem* 74:317–353.
40. Lada AG, et al. (2015) Disruption of Transcriptional Coactivator Sub1 Leads to Genome-Wide Re-distribution of Clustered Mutations Induced by APOBEC in Active Yeast Genes. *PLoS Genet* 11(5):e1005217.
41. Roberts SA, et al. (2012) Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell* 46(4):424–435.
42. Taylor BJ, et al. (2013) DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife* 2:e00534.
43. Hoopes J, et al. (2016) APOBEC3A and APOBEC3B deaminate the lagging strand template during DNA replication. *Cell Reports*, 10.1016/j.celrep.2016.01.021.
44. Haradhvala NJ, et al. (2016) Mutational strand asymmetries across cancer reveal mechanisms of DNA damage and repair. *Cell*, 10.1016/j.cell.2015.12.050.
45. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* 13(5):660–665.
46. Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. *Trends Genet* 13(6):240–245.
47. Frank AC, Lobry JR (1999) Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. *Gene* 238(1):65–77.
48. Rocha EP (2004) The replication-related organization of bacterial genomes. *Microbiology* 150(Pt 6):1609–1627.
49. Karlin S (1999) Bacterial DNA strand compositional asymmetry. *Trends Microbiol* 7(8): 305–308.
50. Alani E, Thresher R, Griffith JD, Kolodner RD (1992) Characterization of DNA-binding and strand-exchange stimulation properties of y-RPA, a yeast single-strand-DNA-binding protein. *J Mol Biol* 227(1):54–71.
51. Molineux IJ, Gefter ML (1975) Properties of the *Escherichia coli* DNA-binding (unwinding) protein interaction with nucleolytic enzymes and DNA. *J Mol Biol* 98(4): 811–825.
52. Bochkareva E, Korolev S, Lees-Miller SP, Bochkarev A (2002) Structure of the RPA trimerization core and its role in the multistep DNA-binding mechanism of RPA. *EMBO J* 21(7):1855–1863.
53. Raghunathan S, Kozlov AG, Lohman TM, Waksman G (2000) Structure of the DNA binding domain of *E. coli* SSB bound to ssDNA. *Nat Struct Biol* 7(8):648–652.
54. Miller JH (1992) *A short course in bacterial genetics: A laboratory manual and handbook for Escherichia coli and related bacteria* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).
55. Baba T, et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol Syst Biol* 2:0008.
56. 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(12):6640–6645.
57. Carpenter MA, Rajagurubandara E, Wijesinghe P, Bhagwat AS (2010) Determinants of sequence-specificity within human AID and APOBEC3G. *DNA Repair (Amst)* 9(5): 579–587.
58. Grant SG, Jessee J, Bloom FR, Hanahan D (1990) Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants. *Proc Natl Acad Sci USA* 87(12):4645–4649.
59. Chung CT, Niemela SL, Miller RH (1989) One-step preparation of competent *Escherichia coli*: Transformation and storage of bacterial cells in the same solution. *Proc Natl Acad Sci USA* 86(7):2172–2175.
60. Foster PL (2006) Methods for determining spontaneous mutation rates. *Methods Enzymol* 409:195–213.
61. Sarkar S, Ma WT, Sandri GH (1992) On fluctuation analysis: A new, simple and efficient method for computing the expected number of mutants. *Genetica* 85(2): 173–179.
62. Hall BM, Ma CX, Liang P, Singh KK (2009) Fluctuation analysis CalculatOR: A web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics* 25(12):1564–1565.
63. Workman CT, et al. (2005) enoLOGOS: A versatile web tool for energy normalized sequence logos. *Nucleic Acids Res* 33(Web Server issue):W389–W392.