Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition

Scott A. Lujan,[1,2] Anders R. Clausen,[1,2] Alan B. Clark,[1,2] Heather K. MacAlpine,[3] David M. MacAlpine,[3] Ewa P. Malc,[4] Piotr A. Mieczkowski,[4] Adam B. Burkholder,[5] David C. Fargo,[5] Dmitry A. Gordenin,[1] and Thomas A. Kunkel[1,2]

[1]Laboratory of Molecular Genetics, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina 27709, USA; [2]Laboratory of Structural Biology, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina 27709, USA; [3]Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710, USA; [4]Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA; [5]Integrative Bioinformatics, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina 27709, USA

Mutational heterogeneity must be taken into account when reconstructing evolutionary histories, calibrating molecular clocks, and predicting links between genes and disease. Selective pressures and various DNA transactions have been invoked to explain the heterogeneous distribution of genetic variation between species, within populations, and in tissue-specific tumors. To examine relationships between such heterogeneity and variations in leading- and lagging-strand replication fidelity and mismatch repair, we accumulated 40,000 spontaneous mutations in eight diploid yeast strains in the absence of selective pressure. We found that replicase error rates vary by fork direction, coding state, nucleosome proximity, and sequence context. Further, error rates and DNA mismatch repair efficiency both vary by mismatch type, responsible polymerase, replication time, and replication origin proximity. Mutation patterns implicate replication infidelity as one driver of variation in somatic and germline evolution, suggest mechanisms of mutual modulation of genome stability and composition, and predict future observations in specific cancers.

[Supplemental material is available for this article.]

DNA synthesis errors are a dual-edged sword. At a population level, accurate DNA replication maintains species identity, yet a small fraction of replication errors creates mutations that improve fitness and fuel evolution. At an individual level, DNA synthesis errors can be beneficial, e.g., by allowing a virus or microbe to survive in an adverse environment or by promoting affinity maturation of antibodies. Replication errors can also result in mutations that have deleterious consequences, cell death, or carcinogenesis. Because replication fidelity underpins so much biology, it has been intensively studied. These studies reveal that—in the absence of stress—replication fidelity is largely determined by nucleotide selectivity, proofreading, and mismatch repair (MMR), with considerable heterogeneity in each process (for review, see Kunkel 2009). Mutation rate heterogeneity is a feature of evolution (Sasaki et al. 2009; Prendergast and Semple 2011; Tolstorukov et al. 2011), including somatic evolution, i.e., tumorigenesis (for review, see Salk et al. 2010). This heterogeneity complicates the identification of genes responsible for the initiation and progression of cancer (Lawrence et al. 2013). Our understanding of the origins of heterogeneous replication fidelity is limited because most studies only monitor a tiny fraction of large, highly organized genomes. Whole-genome studies are required for a complete picture of variations in replication fidelity, the underlying mechanisms, and the consequences for evolution and disease.

One way to interrogate global replication fidelity is to allow mutations to accumulate through many cell divisions with minimal selection against deleterious mutations, and then to sequence the genome to identify the types, numbers, and locations of the mutations that arise (Nishant et al. 2009). To focus on replication errors per se, rather than on other sources of spontaneous mutations, mutation accumulation can be studied in cells defective in nucleotide selectivity, proofreading, or MMR. Such studies have been done in *Saccharomyces cerevisiae*, whose haploid nuclear genome contains 16 chromosomes and 12 million base pairs (bp). Studies of strains with complete or partial defects in MMR reported the accumulation of 76 to 140 mutations, mostly deletions in homonucleotide runs (Zanders et al. 2010; Ma et al. 2012; Lang et al. 2013). Another study of MMR-deficient haploid yeast (Serero et al. 2014) reported 1679 mutations, mostly substitutions. We (Larrea et al. 2010) previously used an MMR-defective haploid strain encoding a mutator variant of DNA polymerase delta (Pol delta), one of three major nuclear replicases. From the genome-wide distribution of 1099 transitions that accumulated and from similar studies using a reporter gene (for review, see Kunkel and Burgers 2008; Lujan et al. 2013), we proposed a model wherein DNA polymerase alpha (Pol alpha) and Pol delta are primarily lagging-strand replicases, whereas polymerase epsilon (Pol epsilon) is primarily a leading-strand replicase.

In these studies, small data sets and/or selective pressures precluded correlation of mutations with other key features of ge-

Corresponding author: kunkel@niehs.nih.gov
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.178335.114.

nomic structure. Here we report a study based on more than 40,000 mutations, accumulated in the absence of selective pressure, in diploid yeast encoding wild-type replicases or mutator variants of Pol alpha, delta, or epsilon, each either proficient or defective in MMR. The results allow calculations of single-base error rates per base pair per generation for replication across the yeast nuclear genome. They also permit genome-wide estimates of the efficiency of MMR for different mismatches. We find that fidelity varies with DNA sequence context, and establish relationships between fidelity and replication origins, replication timing, nucleosome positions, and protein coding potential.

Results

Collecting mutations and determining mutation rates

The rate and distribution of mutations were determined in eight *S. cerevisiae* strains. Diploid strains were used to minimize the effects of purifying selection. The strains (Supplemental Table S1) encode either wild-type replicases or homozygous mutator alleles of the catalytic subunits of Pol alpha (*POL1*; *pol1-L868M*), Pol delta (*POL3*; *pol3-L612M*), or Pol epsilon (*POL2*; *pol2-M644G*). In each case, we compared a strain that was wild type for MMR to one deleted for *MSH2* (or in a few clones, both *MSH3* and *MSH6*). Multiple clonal isolates were passaged on solid, complete media for up to 30 passages, about 900 generations (Supplemental Fig. S1), and their genomes were sequenced (see Methods). Mutations were identified by comparison to "zero passage" genomes for each strain and were filtered by coverage, allelic fraction and false-positive risk due to high internal homology (Supplemental Fig. S1). Sequencing of MMR-deficient *pol2-M644G* genomes at different passage numbers confirmed that mutation counts increased linearly with passage number (Supplemental Fig. S2), indicating that no suppressor or additional mutator phenotypes were acquired. Nonsynonymous substitution rates slightly exceeded synonymous rates (by no more than 15%, less than one standard deviation), indicating a lack of purifying selection against the majority of mutations (Supplemental Methods). Mutations from terminal passage genomes were pooled by strain. Depending on the number of sequenced genomes and passages, large numbers of mutations accumulated during ~2700–7200 total generations (Table 1), yielding high statistical power (see Methods). Most mutations were single-base events distributed across all chromosomes (Fig. 1A). Small gaps (Fig. 1A, black boxes, marked by ‡) are regions where mutations could not be identified with confidence due to high internal homology. The data were used to calculate mutation rates per base pair per generation (μ_{bp}) for each type of single-base change (after dividing by 0.38 or 0.62 for GC or AT templated mismatches, respectively) (Table 1; Fig. 1B,C; see Supplemental Methods). Rates in mmr$^-$ strains provide an estimate of the accuracy of replication, and the ratios of rates in mmr$^-$ to MMR$^+$ strains provide minimum estimates of MMR efficiency (some mutations may result from mismatches not generated by or subject to MMR).

Replication fidelity in strains with wild-type replicases

Mutation rates are higher in mmr$^-$ strains (Fig. 1B) as compared to MMR$^+$ strains (Fig. 1C), indicating that the vast majority of mutations in mmr$^-$ strains result from unrepaired replication errors. In the mmr$^-$ strain encoding wild-type polymerases, 1637 mutations were observed, yielding a μ_{bp} of 1.6×10^{-8} (Table 1). Deletions of AT pairs occur at the highest average rate, followed by transitions, with a twofold bias for CG-to-TA, and then transversions, with CG-to-AT having the highest rate. On average, substitutions from G or C occur at higher rates than substitutions from A or T. With functional MMR, the mutation rate per diploid genome per generation is 0.004, and the average substitution rate per base pair per generation is 1.7×10^{-10}.

Replication fidelity near and distal to origins

Using confirmed functional origins of replication (*S. cerevisiae* OriDB, version 2.1.0) (Supplemental Table S1; Siow et al. 2012) mapped onto the reference genome and the substitutions in the mmr$^-$ strains (Table 1), we calculated substitution (Supplemental Fig. S3A) and indel rates (Supplemental Fig. S3B) as a function of the distance traversed by each replication fork between adjacent origins. Rates are per base pair replicated at each time, accounting for the proportion of the genome at each inter-origin distance (Supplemental Fig. S3C). In all strains (Supplemental Fig. S3A,B), error rates are similar across inter-origin space, with small but statistically significant variations observed in four cases (noted by asterisks; $P \leq 0.0011$). In the strain with wild-type polymerases, substitution rates double near inter-origin midpoints. Slightly higher indel rates near origins (Supplemental Fig. S3B) are due to four- to eightfold higher rates at the origins (autonomously replicating sequence [ARS]; consensus sequences [ACSs]) (Supplemental Fig. S3J). When substitution rates for MMR$^+$ strains were compared to rates in mmr$^-$ strains, the apparent MMR efficiencies were all high, exceeding 99% in all but one case (Supplemental Fig. S3G). In strains with lagging-strand replicase variants, MMR efficiencies were not different for substitutions near and distal to origins (after correcting for multiple hypothesis testing; see Methods). However, MMR of substitutions in the *pol2-M644G* background was about twice as efficient near origins ($P < 10^{-9}$) (Supplemental Fig. S3G, blue bars).

Polymerase and strand specificity of replication errors across the genome

Studies with *URA3* and *CAN1* reporter genes showed that substitution rates are elevated in *pol1-L868M*, *pol3-L612M*, and *pol2-M644G* strains compared to strains with wild-type replicases. This is also true at the whole-genome level for all six substitutions (Table 1; Fig. 1A–C). Thus the substitutions in the mmr$^-$ strains primarily reflect mismatches made by Pol alpha, delta, or epsilon, providing the first opportunity to compare the roles of all three replicases in genome-wide replication. Consider the AT-to-GC transitions observed in the *pol3-L612M msh2Δ* genome. As depicted at the top of Figure 2A, L612M Pol delta preferentially generates T-dG as compared to A-dC mismatches (Nick McElhinny et al. 2008). If Pol delta preferentially synthesizes the nascent lagging strand, the highest proportion of T-to-C substitutions should be immediately to the right of replication origins, and the highest proportion of A-to-G substitutions should be immediately to the left. When the locations of the 5164 AT-to-GC transitions in the *pol3-L612M msh2Δ* genome were mapped relative to origins, the strand bias closely matched the prediction (Fig. 2B). Strand biases in the *pol3-L612M msh2Δ* genome are also seen for the other five types of substitutions (Fig. 2B), including transversions too sparse to analyze in the previous study (Larrea et al. 2010). Similar biases were observed for five of the six substitutions in the *pol1-L868M msh2Δ* genome. The results are consistent with primary roles for Pol alpha and Pol delta in lagging-strand replication. Biases are also observed

Table 1. Mutation counts and rates in eight *S. cerevisiae* strains

	msh2Δ		*pol1-L868M msh2Δ*		*pol3-L612M msh2Δ*		*pol2-M644G mmr-*		*WT*		*pol1-L868M*		*pol3-L612M*		*pol2-M644G*	
Isolates	5		7		4		6		8		6		7		8	
Passages	142	(22–30)	178	(22–30)	90	(16–30)	158	(16–30)	240	(30)	180	(30)	210	(30)	240	(30)
Generations	4260	(660–900)	5340	(660–900)	2700	(480–900)	4740	(480–900)	7200	(900)	5400	(900)	6300	(900)	7200	(900)
Mutations	Count	μ_{bp} [$\times 10^{-11}$]	Count	μ_{bp} [$\times 10^{-11}$]	Count	μ_{bp} [$\times 10^{-11}$]	Count	μ_{bp} [$\times 10^{-11}$]	Count	μ_{bp} [$\times 10^{-11}$]	Count	μ_{bp} [$\times 10^{-11}$]	Count	μ_{bp} [$\times 10^{-11}$]	Count	μ_{bp} [$\times 10^{-11}$]
AT→GC	151	240	1979	2500	5164	13,000	2889	4100	5	4.7	24	30	135	140	33	31
CG→TA	167	430	5513	11,000	5652	23,000	6704	16,000	6	9.1	23	47	79	140	30	46
AT→TA	30	47	237	300	219	550	735	1000	4	3.7	18	22	13	14	193	180
TA→GC	26	41	429	540	189	470	262	370	5	4.7	4	5.0	9	10	13	12
CG→AT	86	220	3206	6600	575	2300	1737	4000	6	9.1	13	26	18	31	22	34
CG→GC	17	44	258	530	15	61	117	270	3	4.6	4	8.1	1	1.7	8	12
–A or –T	955	1500	934	1200	1382	3400	1375	1900	0		0		2	2.1	3	2.8
–C or –G	11	28	56	110	195	790	60	140	0		0		0		2	3.0
+A or +T	46	73	112	140	244	610	168	240	0		0		0		1	0.9
+C or +G	5	13	6	12	49	200	17	39	0		0		0		0	
>1-bp delete	135	130	55	43	42	65	142	120	0		0		0		0	
>1-bp insert	8	7.8	10	7.8	3	4.6	11	9.7	0		0		0		0	
Total	1637	1600	12,795	9900	13,729	21,000	14,217	12,000	29	17	86	67	257	170	305	170
μ_g [diploid]	0.38	(0.20–0.51)	2.4	(0.88–5.4)	5.1	(3.7–6.5)	3.0	(2.5–3.5)	0.0040	(0.0011–0.0078)	0.016	(0.012–0.022)	0.042	(0.033–0.051)	0.042	(0.024–0.064)

Ranges are shown in parentheses.

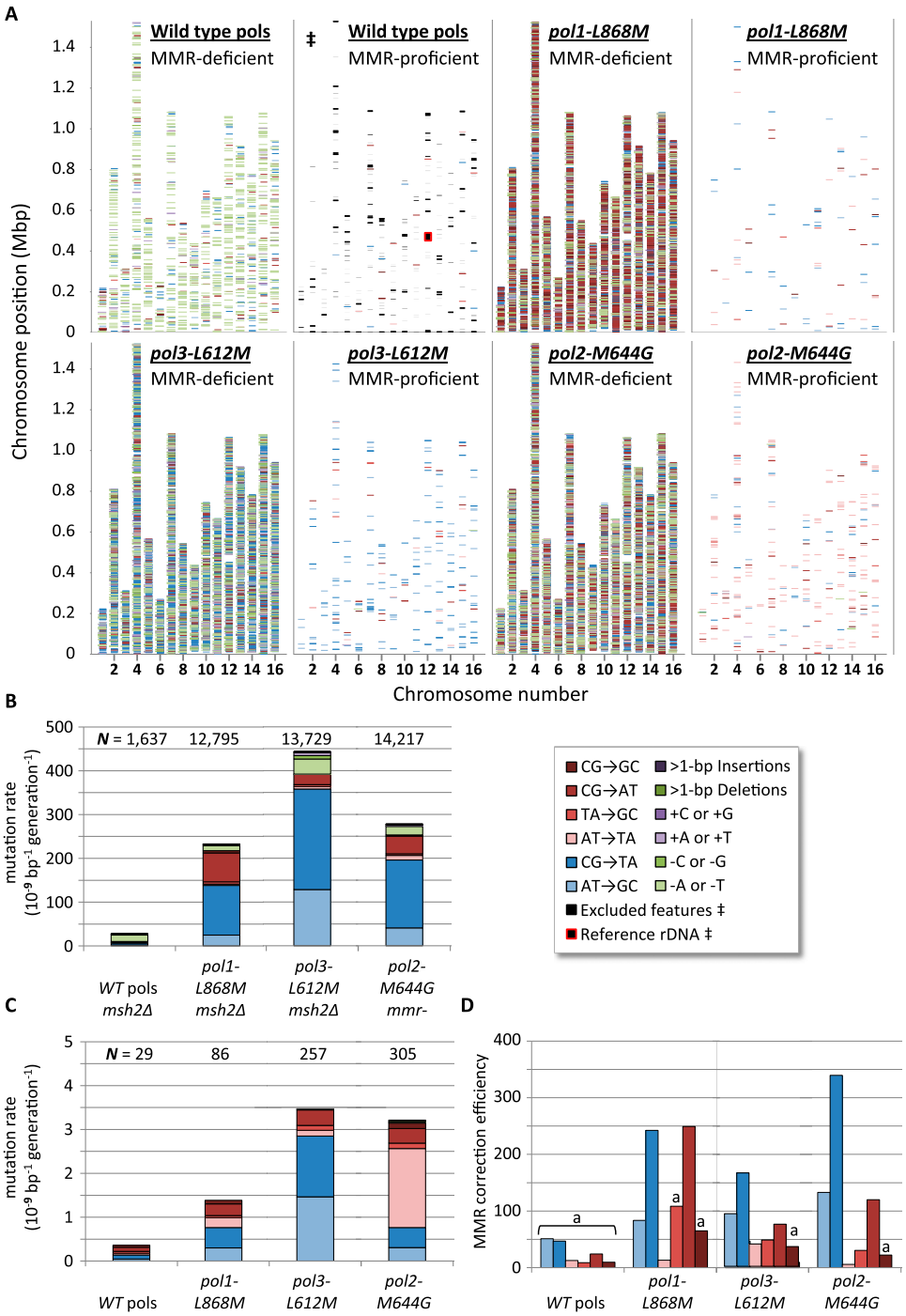

Figure 1. Genome-wide replication error positions, rates, and MMR efficiencies. Transitions are indicated by blue shades (light to dark: AT→GC, CG→TA), transversions by reds (light to dark: AT→TA, TA→GC, CG→AT, CG→GC), deletions by greens (light to dark: −A/T, −G/C, multibase), and insertions by purples (light to dark: +A/T, +G/C, multibase). (*A*) Positions of more than 43,000 mutations, from all eight strains used in this study, plotted along the 16 *S. cerevisiae* chromosomes. ([‡] Overlaid black boxes are regions excluded from mutation calling; see Supplemental Methods.) (*B*) Mutation rates, corrected for genomic GC content from MMR-deficient strains. ([*N*] Mutation count pooled by strain.) (*C*) As per *B*, but for MMR-proficient strains. (*D*) MMR correction efficiencies for substitution errors in four polymerase allelic backgrounds are ratios of MMR-deficient rates to MMR-proficient rates. ([a] Calculated from <10 observed mutations in MMR-proficient strains; see also Supplemental Fig. S2.)

in the *pol2-M644G msh2Δ* genome for five of the six substitutions. Four of these biases are "complementary" to those in the Pol alpha and Pol delta variant strains (Fig. 2B, opposite patterns of red and blue lines), as predicted by a model wherein Pol epsilon primarily replicates the leading strand. AT-to-TA substitution patterns are an exception, with a similar bias seen in the *pol2-M644G msh2Δ* and *pol3-L612M msh2Δ* strains. This exception is actually predicted (Fig. 2B, schematic in left column) by the fact that Pol epsilon

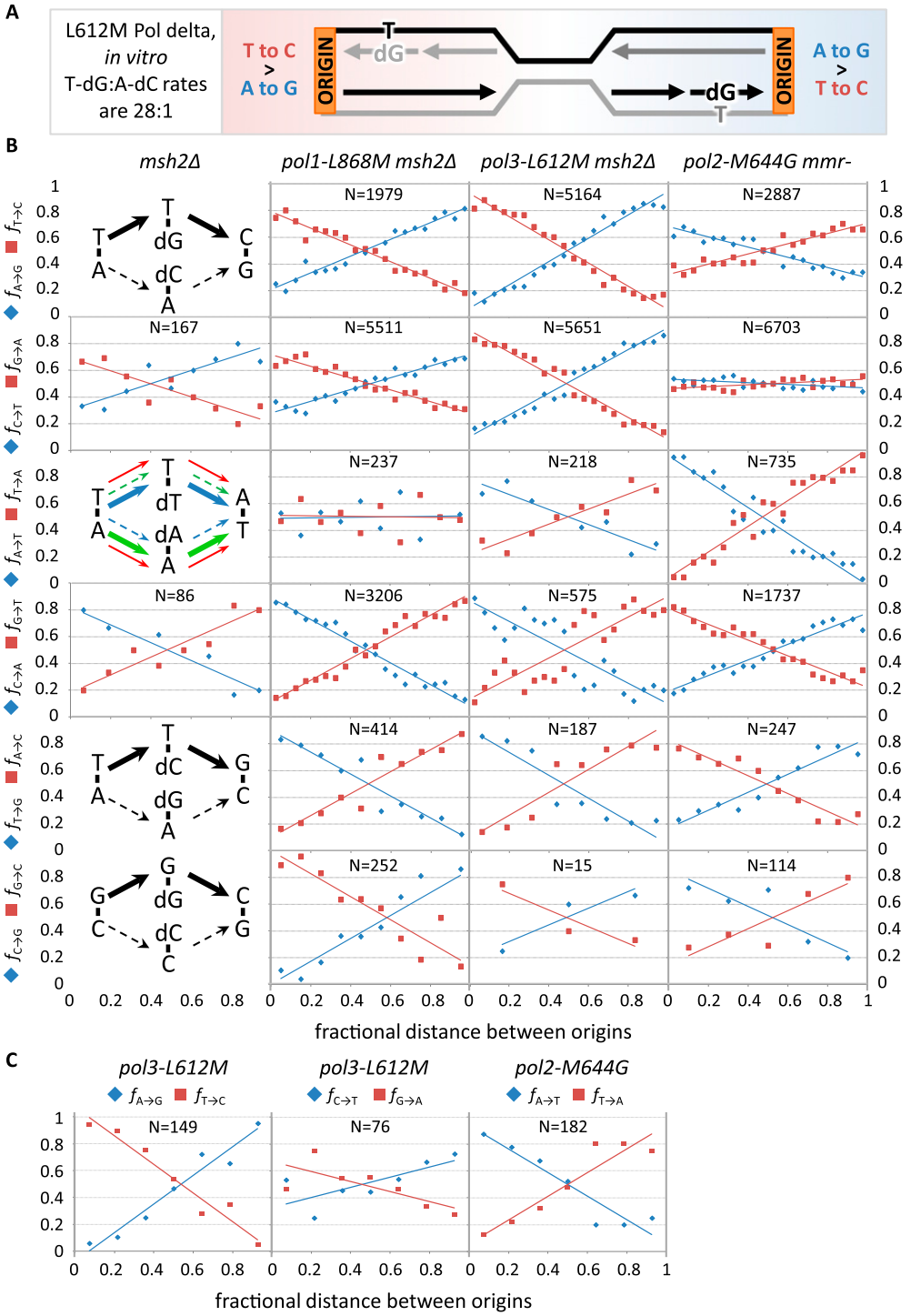

Figure 2. Polymerase and strand specificity of replication errors. Select polymerase-biased complementary mismatch pairs and mismatch motifs. (*A*) Schematic example of adjacent replication origins and their effects on lagging-strand–biased mutagenesis. The T-dG:A-dC ratio in vitro is from Nick McElhinny et al. (2008). (*B*) Diagrams are example preferences for complementary mutation pathways. In most cases, the three variant polymerases have the same preference (black arrows). Disagreements are color-coded by polymerase variant: Pol alpha (*pol1-L868M*), red; Pol delta (*pol3-L612M*), green; and Pol epsilon (*pol2-M644G*), blue. Plots are the fraction of each substitution mutation (f_i) paired with its complement as a function of relative distance between adjacent replication origins. ([N] Mutation count pooled by strain, excluding mutations in origins.) (*C*) As for *B*, but for those mutation types observed >50 times in individual MMR-proficient strains. See also Supplemental Figures S3 and S4.

preferentially forms T-dT mismatches (Shcherbakova et al. 2003; Pursell et al. 2007), whereas Pol delta preferentially forms A-dA mismatches (Fortune et al. 2005, 2006). Thus, among three variant replicases and six substitutions (Fig. 2B), 16 of 18 comparisons reveal strand biases supporting the interpretation that Pols alpha and delta are the primary lagging-strand replicases and Pol epsilon

is primarily a leading-strand replicase. Strand biases were not observed for CG-to-TA in *pol2-M644G msh2Δ* and for AT-to-TA in *pol1-L868M msh2Δ* (Fig. 2B), suggesting roughly equivalent probabilities of generating C-dA/G-dT and A-dA/T-dT mispairs, respectively.

The *msh2Δ* strain with wild-type replicases generated sufficient mutations to analyze strand biases for AT-to-GC, CG-to-TA and CG-to-AT substitutions. The first substitution was unbiased, while the latter two had biases matching those in the Pol alpha and Pol delta variants (Fig. 2B, left column) and opposite to those of the Pol epsilon variant. Moreover, strand-specific patterns for AT-to-GC and CG-to-TA transitions in the *pol3-L612M* background were the same in the MMR^+ and mmr^- strains (Fig. 2B), and the strand-specific pattern of AT-to-TA transversions in the *pol2-M644G* strain (Fig. 2C) matches that in the corresponding mmr^- strain (Fig. 2B). Rates in the mmr^- and MMR^+ strains with variant replicases indicate that all six substitutions are corrected by MMR (Fig. 1D), but with variable efficiencies. For example, correction in the *pol2-M644G* strain varies from sixfold for AT-to-TA mismatches to more than 300-fold for GC-to-AT mismatches. Average correction factors (Fig. 1D) are generally higher for mismatches generated at higher rates (transitions and CG-to-AT) (Fig. 1B).

A preferred sequence motif for generating replication errors

Motif detection algorithms were used to determine if replication errors are generated in preferred sequence contexts. We focused on two abundant substitutions that show strong strand biases, CG-to-AT and CG-to-TA. To infer the direction of replication and the responsible mismatch, we only used the subset of these events that is adjacent to origins (relative inter-origin distance < 0.1). The results reveal motifs for generating replication errors (Fig. 3A,B) that can be rationalized as discussed below (Fig. 3C).

Replication error rates and replication timing

We compared substitution rates as a function of time after release from alpha factor arrest (Supplemental Fig. S3D, top; Muller and Nieduszynski 2012). All rates were corrected for differences in the number of base pairs replicated at each time (Supplemental Fig. S3F). In all replicase backgrounds, rates were nonuniformly distributed ($P \leq 0.00123$) as a function of replication time (Supplemental Fig. S3D). These differences are small but significant (13% higher with L612M Pol delta before 22 min, $P \leq 0.007$; 8.9%–12% higher with L868M Pol alpha and M644G Pol epsilon after 30 min, $P < 0.0002$). Comparing rates in mmr^- and MMR^+ strains with variant replicases indicates that MMR corrects the vast majority of both early and late replication errors (Supplemental Fig. S3H). Nonetheless, in strains encoding the Pol delta and Pol epsilon variants, MMR efficiency is about twofold higher during early replication ($P < 10^{-5}$) (Supplemental Fig. S3H).

Substitution rates at nucleosome positions

To search for relationships between replication errors and nucleosome positions, we mapped the positions of 60,098 nucleosomes across the yeast genome (see Methods) and calculated mutation rates as a function of distance from the nearest nucleosome dyad (Fig. 4A). All rates account for target size (Fig. 4A, top, black plot), with most of the genome within 100 bp of the nearest dyad. MMR-deficient strains show a transition bias (Table 1) and no variation in transition (blue) to transversion (red) ratio with respect to absolute distance

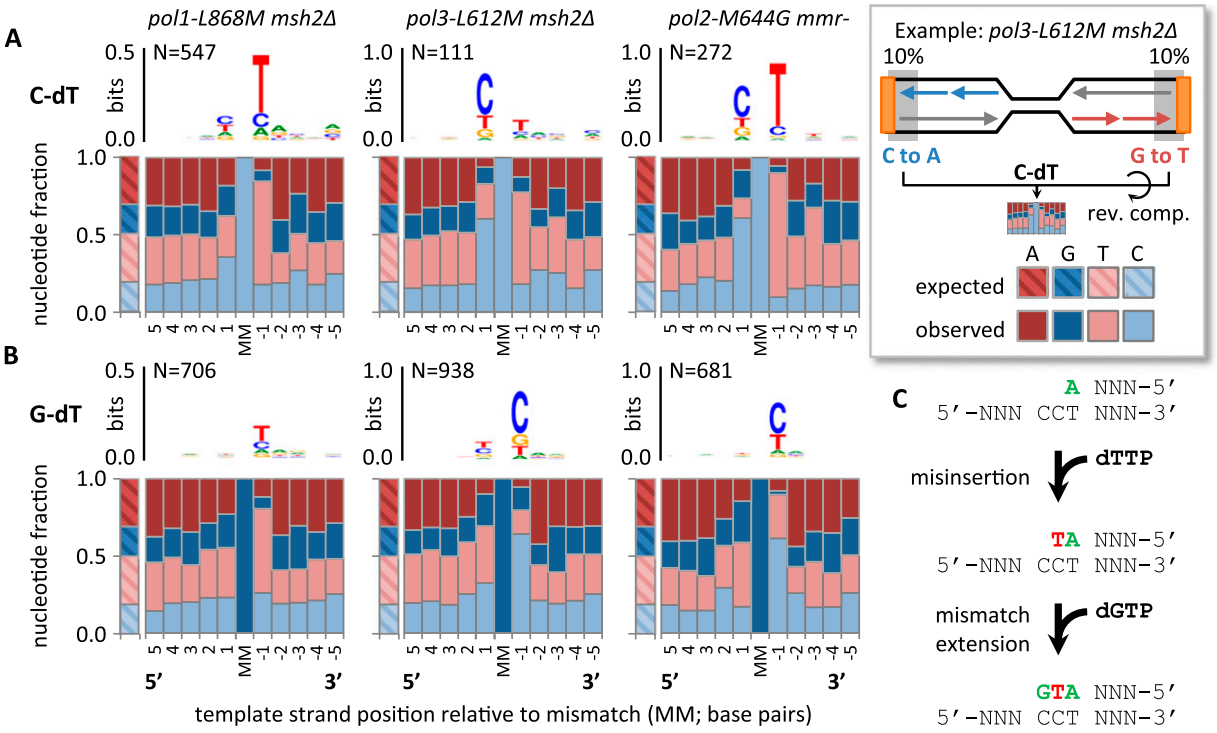

Figure 3. Sequence specificity of replication errors in the absence of MMR. (*A*) Nucleotide fractions and sequence logos (Schneider and Stephens 1990; Crooks et al. 2004) for five bases upstream of and downstream from mutations resulting from presumed C-dT mispairs, as calculated from sequences flanking mutations near replication origins (see example schematic). Expected fractions assume 38% G + C content. ([MM] Mismatch position; [N] mutation count pooled by strain.) (*B*) As per *A*, but for mutations resulting from presumed G-dT mispairs. (*C*) Example: An incoming mismatched nucleotide (red) stacks with adjacent pyrimidines (green) in the nascent strand, as indicated by logos in *A*.

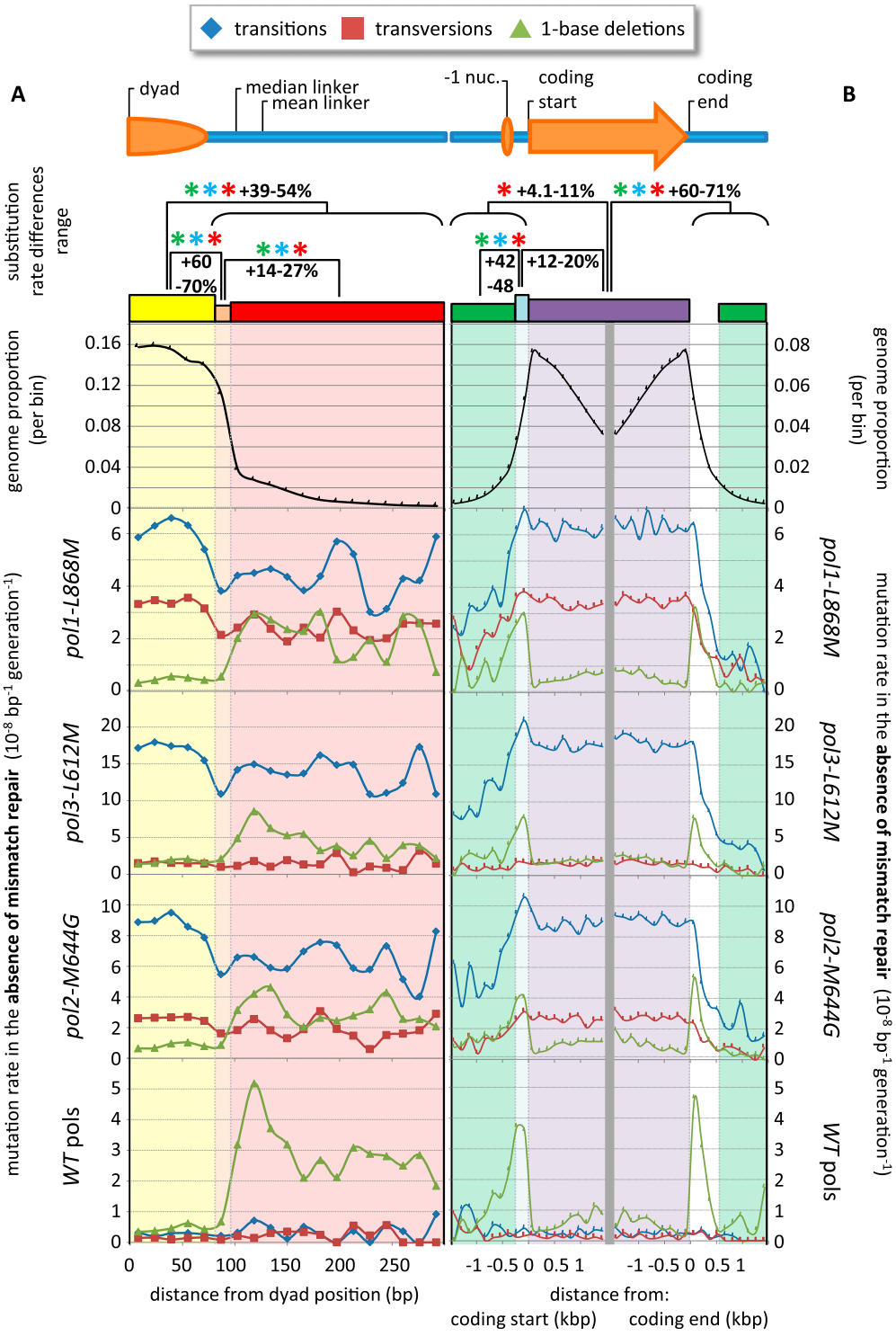

Figure 4. Variation in mutation rates near nucleosome positions and genes. Mutation rates (blue indicates transitions; red, transversions; green, one-base deletions) plotted versus either (*A*) the distance from either the nearest nucleosome dyad (in base pairs) or (*B*) from the nearest coding start (*left*) or end site (*right*; in kilobase pairs). Asterisks denote significantly different substitution rates between indicated regions (Pol alpha, red; Pol delta, green; Pol epsilon, blue). Percentages denote the magnitude of substitution excesses. Shaded areas are DNA regions: nucleosome-bound (yellow), shorter and longer than average linkers (orange and red, respectively), intergenic (green), 5′ nucleosome-free (blue), and coding (purple).

from or relative distance between adjacent dyads (Fig. 4A). Substitution rates in all three MMR-deficient polymerase variant strains are nonuniform with respect to dyad proximity ($P \leq 1.7 \times 10^{-49}$), being higher in nucleosome-bound regions (Fig. 4A, yellow area) than in linker DNA (orange and red areas). For example, the transition rate in the *pol1-L868M msh2Δ* strain is 70% higher, on average,

in nucleosome-bound DNA (6.6×10^{-8} vs. 3.9×10^{-8} per base pair per generation) (Fig. 4A). When all substitutions are considered, rates near dyad positions are significantly elevated ($P \leq 0.0035$), and substitution rates are higher at GC than at AT base pairs. The greatest difference is between the center of the nucleosome (<47 bp from the dyad) and the midpoint of the median linker region (78–94 bp from the dyad). In strains with variant replicases, rates in the core region exceed rates near the median midpoint by 67%–86% for substitutions at GC pairs, but only by 33%–50% for substitutions at AT pairs (data not shown). The difference is greater in the strain with wild-type replicases, where the excess at GC is similar to variant polymerase strains (76%), but rates at AT are similar at nucleosome core and median linker positions. Based on substitution rates in the presence and absence of MMR, repair efficiency is uniform with respect to dyad proximity (nonuniformity $P > 0.2$). MMR does not alter the underlying nucleosome-based mutation pattern.

Replication fidelity in and around genes

Substitution rates (Fig. 4B, blue transitions, red transversions) inside and outside of coding regions are nearly indistinguishable (coding rates ≤11% higher than noncoding), but rates are not evenly distributed with respect to open reading frames (ORFs). Substitution rates in mmr⁻ replicase variant strains are uniform across open reading frames (purple area) but peak upstream of coding sequences in the 5′-nucleosome-free region where transcription initiates ($P < 10^{-12}$) (Fig. 4B, light blue area), and decrease 2.5- to fourfold within a kilobase in the 5′ and 3′ directions away from ORF ($P < 10^{-56}$) (Fig. 4B, green areas).

Indel rates in nucleosome-free regions

Indels accumulated in all polymerase backgrounds (Table 1). The vast majority are loss of an AT pair from a homonucleotide run. Rates for these events were higher (1) in DNA between nucleosomes as compared to nucleosome-bound DNA (Fig. 4A, green), (2) in nucleosome-free DNA immediately upstream of and downstream from coding sequences as compared to both coding sequences and more distant DNA (Figs. 4B, 5A), and (3) within 200 bp of origin consensus motifs at origins of replication (Supplemental Fig. S3J). Rates are highest in sequences enriched in AT runs and depleted in nucleosomes (e.g., immediately upstream of coding sequences) (Supplemental Fig. 5A,B).

Discussion

This study provides new information on the rates by which the yeast nuclear DNA replication machinery generates errors and on the efficiency with which these errors are corrected by MMR.

Replication is incredibly accurate even without MMR

The single-base mutation rate per diploid genome per generation (μ_g) in the MMR-defective strain encoding wild-type replicases is 0.38 (Table 1). This value is similar to those from studies involving fewer mutations (Zanders et al. 2010; Ma et al. 2012; Lang et al. 2013; Serero et al. 2014) and indicates that even without MMR, replication is so accurate that most cells in a population will not contain even one mismatch generated during replication. Indeed, given that two forks emerge from each of about 400 replication origins, only one out of every ~2000 replication forks is likely to generate a mismatch for later correction by MMR.

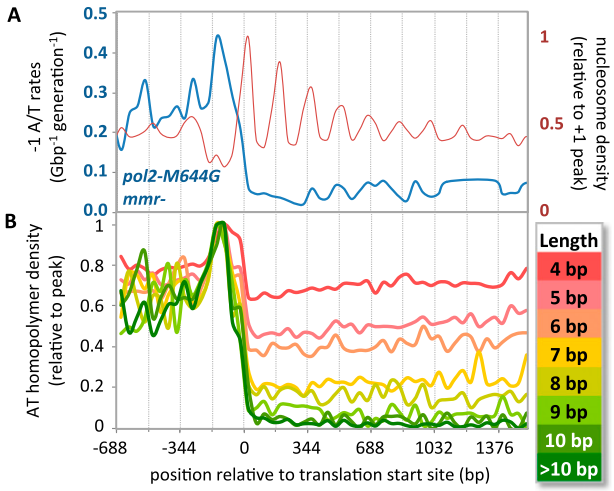

Figure 5. Indel rates in homopolymers in the absence of MMR. (*A*) A comparison of nucleosome density (red; relative to the +1 peak) and *pol2-M644G mmr-* AT deletion rates (blue) as a function of distance from translation start sites. (*B*) Homopolymer densities (relative to maximum density) for various homopolymer lengths as a function of distance from translation start sites.

Replication fidelity is heterogeneous by base pair identity and mismatch composition

Substitutions from G or C occur at higher rates than from A or T (Table 1). Previous studies in yeast (Lynch et al. 2008; Zanders et al. 2010; Lang et al. 2013), *Caenorhabditis elegans* (Denver et al. 2009), and *Drosophila melanogaster* (Keightley et al. 2009) reported a similar bias, suggesting that the error specificity of replication by wild-type replicases may be evolutionarily conserved. Among the six substitution types, CG-to-TA is generated at the highest rate in each mmr⁻ strain (Table 1). Because these are efficiently corrected by MMR (Fig. 1D, dark blue bars) and their rates are strongly elevated in the replicase variant strains, it is likely that they primarily result from G-dT rather than C-dA mismatches. AT-to-GC transitions are generated at a lesser, but still high, rate, likely more through T-dG than A-dC mismatches. Among transversions, CG-to-AT is consistently generated at a higher rate than the other three. While some CG-to-AT transversions could reflect insertion of adenine opposite 8-oxoguanine due to oxidative stress, in this study of replication infidelity in unstressed yeast, we favor the hypothesis that they mostly result from misinserting dTTP opposite template C. If so, this pyrimidine–pyrimidine mismatch would be among the most common mismatches generated during normal replication. This is surprising in light of previously suggested mispairing schemes and the hypothesis that pyrimidine–pyrimidine mismatches are rarely generated because they are hydrated (Kool 2001 and references therein).

Replication fidelity is heterogeneous during strand-specific replication

The strand-specific substitution patterns in Figure 2 support a model wherein Pol alpha and Pol delta are primarily lagging-strand replicases and Pol epsilon is primarily a leading-strand replicase. For the first time, our study applies this interpretation to replication of the whole-yeast nuclear genome by all three major replicases. Similar biases are observed in the mmr⁻ strain encoding

wild-type replicases (Fig. 2B) and also in MMR$^+$ strains (Fig. 2C). Thus, strand biases due to replication fidelity occur naturally, not merely with engineered replicases or only in the absence of MMR.

In the strain encoding wild-type replicases, the biases match those seen in the Pol alpha and Pol delta variant strains and are opposite to the biases in the Pol epsilon variant strain (Fig. 2B). This implies that lagging-strand replication by wild-type Pol alpha and/or Pol delta is less accurate than leading-strand replication by wild-type Pol epsilon, accounting for ~66% of base substitutions in the wild-type replicase background (rates in Table 1; biases in Fig. 2B). On an evolutionary time scale, strand-biased replication infidelity could influence the base composition of genomes. Strand-biased nucleotide excesses seen in bacteria, yeasts (Gierlik et al. 2000; Koren et al. 2010), and mammals, including humans, have been associated with transcription (Green et al. 2003; Touchon et al. 2003; Polak and Arndt 2008; Mugal et al. 2009) and replication (Touchon et al. 2005; Rocha et al. 2006; Chen et al. 2011). In the *S. cerevisiae* genome (excluding subtelomeric DNA) (Gierlik et al. 2000), there are excesses of C and A in lagging-strand templates and of G and T in leading-strand templates (Agier and Fischer 2012). This is nicely explained by the accumulation over an evolutionary time scale of substitutions of template G and C with A, with G-to-A substitutions about twice as frequent, as seen in the MMR-deficient strain with wild-type polymerases (Fig. 2B).

Replication fidelity is heterogeneous due to local sequence context

The results in Figure 3 suggest the existence of a preferred sequence context for two common base substitutions by all three major replicases. This preference can be rationalized by the effects of base stacking on stable misincorporation, which requires misinsertion followed by mismatch extension without proofreading. As one example (Fig. 3C), the alignment of sequences surrounding CG-to-AT transversions in the MMR-deficient *pol3-L612M msh2Δ* strain indicates that after correctly incorporating dATP opposite a template T (green A), L644M Pol delta (POL3) misinserts dTTP (red T) opposite template C and then correctly incorporates dGTP (green G) opposite the next C. This motif suggests that misinsertion of dTTP, as well as subsequent extension of the C-dT mismatch without proofreading, are favored by stacking of the misinserted dT with flanking purines, which are known to stack more strongly than pyrimidines (e.g., see Goodman et al. 1993; Hunter 1993; Kool 2001).

Substitution fidelity at the replication fork in relation to replication timing

Replication times for each 1-kb section of the genome were estimated by converting published relative copy number maps (Muller and Nieduszynski 2012) into replication timing units. Briefly, that study sorted S-phase and G2-phase cells from asynchronous diploid *S. cerevisiae* cultures and used quantitative deep sequencing to measure the relative copy number of each genomic section. The first sequences replicated in S phase had relative copy numbers twice as high as the last sequences replicated in S phase. Since they showed that the relative copy number is proportional to the mean replication time, we used published origin firing times, measured in minutes after release from alpha factor-induced G1 growth arrest (Yabuki et al. 2002), to transform the data from relative copy number into replication time. In all replicase backgrounds, substitution rates in the absence of MMR are not constant with replication time ($P \leq 0.0012$) (Supplemental Fig. S3D). In strains with variant replicases, substitution rates are marginally higher at the latest time points and, on the lagging strand, at the earliest time points as well (Supplemental Fig. S3D). Possible explanations include differences in the sequences being replicated, differences in chromatin status, and slight variations in dNTP concentrations during S phase that could modulate misinsertion and/or proofreading. Substitution rates increase with replication time in wild-type yeast (Lang and Murray 2011; Agier and Fischer 2012), in contemporary human diversity (Koren et al. 2012), and across taxa in the evolutionary record (Sharp et al. 1989; Wolfe et al. 1989; Stamatoyannopoulos et al. 2009; Flynn et al. 2010; Pink and Hurst 2010; Chen et al. 2011; Weber et al. 2012). A leading explanation has been that error-prone repair becomes more prominent later in replication (Lang and Murray 2011). Our data suggest that replication fidelity and MMR biases (below) may also contribute to these patterns.

Replication fidelity is heterogeneous relative to nucleosome positions

At the megabase level, variations between primate lineages (Prendergast et al. 2007; Ananda et al. 2011) are correlated with chromatin openness. At higher resolutions, comparative genomics studies within and between species have correlated nucleosome positions with the accumulation of genetic variation (Washietl et al. 2008; Sasaki et al. 2009; Prendergast and Semple 2011; Tolstorukov et al. 2011; Kenigsberg and Tanay 2013). On evolutionary time scales, substitutions accumulate near stable nucleosome positions, while indels accumulate in linker regions. Explanations for these patterns have included variations in replication fidelity, MMR, DNA damage, DNA repair, and purifying selection (Kenigsberg et al. 2010; Ying et al. 2010; Tolstorukov et al. 2011). In MMR-deficient yeast strains, we likewise find higher substitution rates near nucleosome dyads and higher deletion rates in linker regions (Fig. 4A). This does not necessarily suggest that nucleosomes themselves directly alter replication fidelity, as they are presumably displaced by the replicative helicase ahead of each fork. The fact that the mutation rates in question are elevated in all three mutator replicase backgrounds and in the absence of MMR suggests that they are not due to DNA damage. The results are consistent with the idea that nucleosomes prefer to bind to DNA sequence contexts that are more mutable than average but are normally protected by purifying selection, such that when the purifying selection is not operative, as in this study, these sequences are at higher than average risk of mutation. This could be due to a higher GC content in nucleosome binding sites and a higher AT-homopolymer density in linkers. Regardless of the explanation, the observations suggest that replication infidelity contributes to the evolutionary pattern of variation relative to nucleosome positions.

Genomic and locus-based comparative studies show a bias for transitions over transversions that has been attributed to the underlying rates of DNA mutation and/or repair (Li et al. 1984; Rosenberg et al. 2003; Lynch et al. 2008; Denver et al. 2009; Ossowski et al. 2010; Babbitt and Cotter 2011, and references therein). Other comparative studies (Keightley et al. 2009) and previous mutation accumulation studies in model organisms found no transition bias. Here we see a strong transition bias in all mmr$^-$ strains (Fig. 4A). The fact that the transition-to-transversion ratio correlates with nucleosome position led to the suggestion

that the transition bias was driven by selective pressure against transversions in stable nucleosome cores, where they might cause greater disruption to nucleosome assembly and localization (Babbitt and Cotter 2011). In our mmr⁻ strains, the transition-to-transversion ratio is stable relative to nucleosome proximity, suggesting that replication infidelity may indeed contribute to any overall transition bias seen in the evolutionary record but not to the variance at nucleosome dyads. The overall MMR efficiency for substitutions does not vary with regard to dyad proximity (nonuniformity test $P \geq 0.22$). MMR efficiency varies by mutation type (Fig. 1D), thus complicating its possible influence on a transition bias (Fig. 1C), but neither replication infidelity nor MMR causes anything like the nucleosome-dependent variation in the transition-to-transversion ratio seen in the evolutionary record.

Indel rates were highest in nucleosome-free DNA immediately upstream of and downstream from coding sequences (Figs. 4B, 5A) and within 200 bp of ACS motifs at origins of replication (Supplemental Fig. S3J). Previous studies found that indel rates increase with AT homopolymer length (Lynch et al. 2008; Lang et al. 2013). These correlations are not independent, because AT runs exclude nucleosomes (Field et al. 2008; for review, see Kaplan et al. 2009; Radman-Livaja and Rando 2010) and because long AT homopolymers are concentrated at conserved, nucleosome-free locations like those upstream of coding sequences (Fig. 5B). One implication is that regions that have been *selected* to be nucleosome-free will also be indel hotspots in the absence of mismatch repair. This means that important nucleosome-free areas, such as ORC binding sites and untranslated regions around genes, are among the most vulnerable to even a transient lapse in MMR activity.

Replication fidelity is heterogeneous in coding and noncoding DNA sequences

In mmr⁻ replicase variant strains, substitution rates for both transitions and transversions are higher in sequences that code for proteins (Fig. 4B, red and blue). Higher substitution rates in coding sequences ($P < 10^{-56}$), and perhaps even higher rates in immediate 5′-flanking regions where transcription initiates (*P*-values inconclusive), contrast with the evolutionary record, where substitutional variation is lower in genes and lowest in the nucleosome-free regions 5′ to genes, as compared to distant intergenic DNA (Sasaki et al. 2009; Tolstorukov et al. 2011). This inverse relationship implies that sequences that are normally protected by purifying selection are hypermutable in the absence of such selection. This extends to replication errors an effect that was previously shown for primate CpG sites (Subramanian and Kumar 2003; Schmidt et al. 2008). Likewise, our study reveals variations in AT base pair deletion rate with respect to coding sequences. Deletion rates are higher in intergenic DNA and highest in the nucleosome-free regions upstream of and downstream from coding sequences, as compared to coding sequences (Figs. 4B, 5A; green). Deletion rates in the evolutionary record also reach a local maximum in 3′-untranslated regions (Tolstorukov et al. 2011), but otherwise, the relationships are inverted: In the evolutionary record, variation due to indels reaches a local minimum in 5′ nucleosome-free regions, is higher in coding sequences, and is higher still in distant intergenic DNA (Sasaki et al. 2009; Tolstorukov et al. 2011). It is theoretically possible that collisions between replication forks and transcription complexes, and/or spontaneous damage to single-stranded DNA in transcription bubbles, might contribute to the higher mutation rate in coding sequences. Studies have been initiated to determine if the higher mutation rates in coding sequences observed here correlate with levels of gene expression.

MMR is very efficient

In the MMR⁺ strain with wild-type replicases, the average genome-wide base mutation rate per base pair (μ_{bp}) is 1.7×10^{-10}, similar to values from earlier studies based on fewer mutations (Lynch et al. 2008; Nishant et al. 2009; Lang et al. 2013). The average mutation rate per diploid genome of 0.004 (Table 1) is similar to the haploid μ_g of 0.003 extrapolated from data for the *CAN1* locus (Drake 1991; Drake et al. 1998). Our genome-wide μ_g in the corresponding MMR-deficient strain (0.38) is 100-fold higher, indicating that on average, MMR corrects at least 99% of all mismatches generated by the approximately 1600 replication forks in diploid yeast, each of which replicates a different genomic sequence. As a consequence, only one in 250 unstressed, wild-type yeast cells in a population will suffer from replication infidelity when MMR is operative.

Strand-specific variations in MMR efficiency

We previously suggested that 8-oxoguanine–adenine mismatches generated during lagging-strand replication may be corrected more efficiently than those made during leading-strand replication (Pavlov et al. 2003). We proposed that the 5′ ends of Okazaki fragments, possibly in conjunction with PCNA, are used as strand discrimination signals during lagging-strand MMR. This possibility was supported by a study indicating more efficient repair of mismatches generated by Pol alpha near the 5′ ends of Okazaki fragments than more internal mismatches made by Pol delta (Nick McElhinny et al. 2010) or leading-strand mismatches made by Pol epsilon (Lujan et al. 2012). Moreover, exonuclease 1, which digests DNA in a 5′ to 3′ direction to excise mismatches during MMR, contributes more to the repair of mismatches generated by Pol alpha (Liberti et al. 2013) and Pol delta (Hombauer et al. 2011; Liberti et al. 2013) than to repairing mismatches generated by Pol epsilon. In this study, comparing μ_g in mmr⁻ and MMR⁺ replicase variant strains (Table 1) reveals that MMR efficiency is higher for lagging-strand replicases Pol alpha (150-fold) and Pol delta (120-fold) than for the leading-strand replicase Pol epsilon (70-fold). This is the first genome-wide evidence for more efficient MMR of lagging-strand replication errors. This result, and evidence that lagging-strand replication is less accurate than leading-strand replication (Fig. 2B), supports the hypothesis that there may be a complementary relationship between generating and repairing replication errors (for review, see Kunkel 2011; Lujan et al. 2012), wherein mismatches generated at the highest rates are those that are most efficiently corrected by MMR, thus protecting the integrity of both DNA strands against a variety of replication errors made at different rates. A caveat to this hypothesis is that the average MMR efficiency for errors generated by M644G Pol epsilon is similar to that for lagging-strand replicases if one excludes AT-to-TA transversions, which are the least efficiently repaired substitutions (Fig. 1D, see next section).

Heterogeneity by mismatch composition and sequence context

All six types of substitutions are corrected by MMR (Fig. 1D), but correction factors vary widely, e.g., from 350-fold for G-dT mismatches in the *pol2-M644G* strain to less than sixfold for T-dT

mismatches in the same strain. Nonexclusive explanations for such variations include differences in the composition of the mismatch and the effect of flanking sequence context. A previous study (Lujan et al. 2012) reported that a flanking ATT triplet repeat sequence partially suppresses repair of a T-dT mismatch at base pair 686 in the *URA3* gene. This and other studies (Jones et al. 1987; Marsischky and Kolodner 1999) reveal that sequences flanking a mismatch can modulate its repair. At the same time, our study now shows that AT-to-TA transversions occur throughout the genome in many different sequence contexts, and the mismatches likely to explain these substitutions (A-dA, as in *pol3-L612M*, and especially T-dT, as in *pol2-M644G*) (Fig. 2B) are repaired less efficiently (Fig. 1D, pink bars) than are other mismatches. This implies additional explanations for variations in MMR efficiency beyond local sequence context. Finally, MMR correction factors (Fig. 1D) are generally higher for mismatches generated at higher rates (e.g., transitions and CG-to-AT transversions) (Fig. 1B). This genome-wide result agrees with *URA3* reporter studies (Lujan et al. 2012) and is again consistent with the concept of a complementary relationship between generating and repairing replication errors.

Leading-strand MMR decreases as replication proceeds from origins

MMR of substitution mismatches generated by Pol epsilon, but not by Pol alpha or Pol delta, is 2.4-fold more efficient near origins as compared to inter-origin midpoints ($P = 2.8 \times 10^{-10}$). It may be that leading-strand MMR is less efficient near fork collision points, or the MMR machinery may become uncoupled from the leading-strand replication machinery with increasing distance from origins.

MMR is less efficient during late replication

MMR operates very efficiently during both early and late replication. Nonetheless, MMR of substitutions is on average about twofold less efficient during late replication in strains encoding variants of Pol delta and Pol epsilon ($P < 10^{-5}$) (Supplemental Fig. S3H). These differences are consistent with a previous study of indels in a reporter gene (Hawk et al. 2005) that first suggested that MMR might be less efficient late in S phase. Because genomic regions replicated early and late differ in sequence composition, gene content, and chromatin status, there are several possible explanations for variations in MMR related to replication timing. Even small differences may explain the higher variation in late-replicated regions in yeast, human cancers, the evolutionary record, and contemporary human diversity.

Polymerase fidelity and MMR drive heterogeneous mutation rates

We have shown that replication fidelity is heterogeneous across the yeast genome in the absence of stress and purifying selection. We suspect that other mutation sources are similarly heterogeneous. This must be taken into account when using mutation spectra to reconstruct evolutionary histories or to calibrate molecular clocks.

We have tested a variety of extant hypotheses as to the origins of heterogeneous evolutionary variation. We have shown that the heterogeneity of replication fidelity can explain patterns like those seen in the evolutionary record with regards to the composition of variation (mismatch type), strand-biased genome composition, the correlation between variation and replication time, higher variation at nucleosomes based on substitutions, and in linker regions based on indels. On the evolutionary time scale, heterogeneity of replication fidelity can explain neither the nucleosome-dependent variation in the transition-to-transversion ratio nor the complex variation patterns regarding coding character (gene/UTR/intergenic). In fact, the inverse relationship between long-term evolutionary variation (Sasaki et al. 2009; Tolstorukov et al. 2011) and short-term replication error rates, depending on coding character, reveals regional differences in mutability that must be maintained over the long haul by differential selective pressure.

Tumorigenesis is driven by somatic evolutionary processes characterized by cyclical mutation and selection; one could anticipate the same patterns of variation in tumor genomes as in the evolutionary record. The same underlying processes should hold sway and thus may be explained in part by heterogeneous replication fidelity. It is important to note that despite similarities, patterns of cancer mutations do differ substantially from patterns in the germline evolutionary record (e.g., vs. replication time), and these differences themselves may be instructive. We already know that in human tumor genomes, mutation frequencies vary with replication time and correlate with chromatin openness (Hodgkinson et al. 2012; Schuster-Bockler and Lehner 2012; Woo and Li 2012; Lawrence et al. 2013). Substitution types observed at the highest rates here also predominate in tumors, with the majority of substitutions being G/C targeted (even excluding CpG motifs [Alexandrov et al. 2013], though not necessarily other tissue-specific, G/C-targeted mechanisms [Roberts et al. 2013]). Differences in leading- versus lagging-strand replication infidelity, as seen here even with wild-type polymerases and even in the presence of MMR, should also be relevant to the clonal evolution of tumors and to differences in tissue-specific tumorigenesis as observed in Pol delta or Pol epsilon proofreading-defective mice (Albertson et al. 2009).

As the analysis of variation in tumor genomes approaches the level of detail seen in evolutionary comparative genomics, we predict that the patterns seen here may be observed in human tumors. For example, a very careful and comprehensive study recently reported that microsatellite-stable, hypermutated endometrial carcinomas (ECs) bearing proofreading-defective Pol epsilon (exonuclease domain mutation [EDM]) variants are proportionally enriched for GC-to-TA transversions as compared to non-EDM ECs (Church et al. 2013). Proofreading efficiency depends on the balance between mismatch excision and extension (for review, see Kunkel 2009). L612M Pol delta (POL3) and M644G Pol epsilon (POL2) both have increased mismatch extension and are therefore proofreading defective (Pursell et al. 2007; Nick McElhinny et al. 2008). Church et al. say that >80% of their GC-to-TA transversions are flanked on either side by AT base pairs (A flanking each mutated G) and that non-EDM samples lacked this signature. For EDM ECs, which represent <10% of ECs (Kandoth et al. 2013), we can explain ~90% of GC-to-TA transversions as the result of C-dT mispairs on the nascent leading strand (as in Fig. 2A,B) with the template C flanked by pyrimidines (as in Fig. 3A). Further, for the >90% of ECs that are non-EDM, ~70% of GC-to-TA transversions also resemble pyrimidine-flanked C-dT mispairs, implicating replication infidelity in most of the GC-to-TA transversions in all 228 EC exomes in that study. Pyrimidine flanks were found for nearly all substitution types in this study, but the GC-to-TA class has an added diagnostic advantage in that it is both common in our spectra and distinct from the TCW to TTW or TGW patterns due to APOBEC cytidine deaminase in many human cancers (Roberts et al. 2013). The motifs in Figure 3 also explain the bulk of substitutions in highly mutated

colorectal cancers (Alexandrov et al. 2013). Future comparison with our other patterns would strengthen these explanations. This also suggests further questions. Do other patterns in ECs also suggest replication infidelity as a major source of variation? Perhaps this is unsurprising given the prevalence to MMR defects in EC tumors, but what of other forms of cancer? Do the positions or contexts of tumor suppressor genes make them more or less vulnerable to replication infidelity? How important is replication infidelity to tumorigenesis in general?

Methods

Yeast strains and methods

All sequenced *S. cerevisiae* strains are diploids descended from Δ(-2)-7B-YUNI300 (Pavlov et al. 2001) and are homozygous for the following markers: *CAN1*, *his7-2*, *leu2-Δ::kanMX*, *ura3-Δ::*, *trp1-289*, *ade2-1*, *lys2-ΔGG2899-2900*, and *agp1::URA3* (orientations vary between strains) (Supplemental Table S1).

Mutation accumulation

As per Supplemental Figure S1, yeast cells were subjected to up to 30 single-cell bottleneck passages on solid media. Samples were retained periodically for glycerol stocks and phenotype testing. Synonymous (K_s) and nonsynonymous (K_a) substitutions of each type were counted for each strain and the mutation rates calculated. K_a/K_s ratios exceed unity by less than one standard deviation, indicating no significant selective pressures during the experiment.

Library preparation and genome sequencing

Genomic DNA (isolated from saturated cultures) was fragmented to between 200 and 800 bp according to the Illumina TruSeq DNA protocol, and libraries were prepared using Illumina TruSeq DNA sample prep kits on a Tecan Freedom EVO 150 automated liquid handling system. Libraries were size-selected for insert fragments ~300 bp (Pippin prep system from Sage Science). Libraries were analyzed and quantified using a Qubit (fluorometric detection; Invitrogen) and Experion automated electrophoresis system (Bio-Rad). Quantified libraries were diluted to a 15-nM concentration and pooled for sequencing. The paired-end sequencing (2 × 100 cycles) was performed on HiSeq 2000 sequencers (Illumina).

Reference assembly

Assembly of the master reference sequence, L03, was described previously (see Data Access) (Larrea et al. 2010). L03 was annotated with gene, retrotransposon long terminal repeats, and repeat regions from the Saccharomyces Genome Database S288c genome version R64-1-1 (Engel et al. 2014), nucleosomes positions from MNase-seq experiments (see below), and origin consensus sequences (ACSs; derived from the *S. cerevisiae* OriDB version 2.1.0) (Supplemental Table S1; Siow et al. 2012). Replication times for each 1-kb section of the genome were estimated by converting published relative copy number maps (Muller and Nieduszynski 2012) into replication timing units via the linear correlation between relative copy number and mean replication time, using published origin firing times (Yabuki et al. 2002).

Calling variant base pairs from Illumina sequences

Sequencing reads were mapped to the L03 master reference, and variant base pairs were called using CLC bio Genomics Workbench version 5.1.5 with parameters set as per Supplemental Figure S1. Relative chromosomal coverage indicated aneuploidy. Variants were filtered as per Supplemental Figure S1 and pooled by genotype (for rate calculations).

Finding nucleosome locations via MNase-seq

MNase-digested Illumina paired end reads were aligned to the L03 reference via Bowtie 0.12.7 (filtered for quality, mismatches, and unique alignment) (Langmead et al. 2009). Positions of mononucleosome sized fragments were examined via NOrMAL (Polishko et al. 2012).

Calculating mutation rates

Each mutation rate, per base pair per generation, $\mu_{bp,i}$, for any mutation type, i, in any section of the genome (bin, b), is calculated as

$$\mu_{bp,i} = \frac{N_{i,b}}{N_{bp,b} \times gen_{tot}},$$

where $N_{i,b}$ is the number of mutations of type i in bin b; $N_{bp,b}$ is the number of base pairs in bin b, accounting for ploidy; and gen_{tot} is the total number of generations of mutation accumulation for all isolates of the selected genotype. For $N_{bp,b}$ estimation relative to genomic landmarks, see the Supplemental Methods.

The MMR correction efficiency, *cf*, for a given mutation type i in genomic bin b, is the ratio of the MMR-deficient and MMR-proficient mutation rates where all else is genetically equal:

$$cf = \frac{\mu_{bp,i,mmr-}}{\mu_{bp,i,MMR+}}.$$

Details on statistics, including Bonferroni and Dunn–Šidák (Dunn 1961; Šidák 1967) corrections for multiple hypothesis testing, may be found in the Supplemental Methods.

Mutable motif detection

Sequence motifs were detected via a custom Excel tool. Twenty template bases bracketing each variant were oriented and aligned, and initial hidden Markov, log-likelihood models were built. Sequence logos were created via WebLogo 3 (weblogo.threeplusone.com) (Schneider and Stephens 1990; Crooks et al. 2004).

Additional details regarding yeast strains and methods, mutation accumulation, library preparation and genome sequencing, assembly and feature selection, calling variant base pairs, finding nucleosome locations, calculating mutation rates, mutable motif detection, and hypothesis testing can be found in the Supplemental Material.

Data access

DNA-seq and MNase-seq data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE56939.

Acknowledgments

We thank Matthew Longley, Shayamal Peddada, Marta Garbacz, and Matthew Young for helpful comments on the manuscript and Tianyuan Wang for computational support. This work was supported by Project Z01 ES065070 to T.A.K. and Z01 ES065073 to

Michael A. Resnick (whom we also thank), both from the Division of Intramural Research of the NIH, NIEHS, and by 2R01GM052319-16A1 to P.A.M. and 1R01GM104097 to D.M.M., both from the NIH.

Author contributions: S.A.L and A.B.C. constructed yeast strains. S.A.L, A.B.C., D.A.G., and T.A.K. designed mutation accumulation experiments, which S.A.L and A.B.C. performed. D.M.M., H.K.M., and A.R.C. designed MNase-seq library preparations, which H.K.M. and A.R.C. performed. A.B.B. and D.C.F. performed nucleosome position calling. E.P.M. and P.A.M. designed and performed DNA-seq and MNase-seq procedures. S.A.L. and A.B.C mapped sequences and called mutations. S.A.L. designed and performed mutation filtering and meta-analyses. S.A.L. and D.A.G. devised statistical measures, which S.A.L. performed. S.A.L. and T.A.K. wrote the manuscript. All authors contributed to the manuscript.

References

Agier N, Fischer G. 2012. The mutational profile of the yeast genome is shaped by replication. *Mol Biol Evol* **29:** 905–913.

Albertson TM, Ogawa M, Bugni JM, Hays LE, Chen Y, Wang Y, Treuting PM, Heddle JA, Goldsby RE, Preston BD. 2009. DNA polymerase ε and δ proofreading suppress discrete mutator and cancer phenotypes in mice. *Proc Natl Acad Sci* **106:** 17101–17104.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500:** 415–421.

Ananda G, Chiaromonte F, Makova KD. 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biol* **12:** R27.

Babbitt GA, Cotter CR. 2011. Functional conservation of nucleosome formation selectively biases presumably neutral molecular variation in yeast genomes. *Genome Biol Evol* **3:** 15–22.

Chen CL, Duquenne L, Audit B, Guilbaud G, Rappailles A, Baker A, Huvet M, d'Aubenton-Carafa Y, Hyrien O, Arneodo A, et al. 2011. Replication-associated mutational asymmetry in the human genome. *Mol Biol Evol* **28:** 2327–2337.

Church DN, Briggs SE, Palles C, Domingo E, Kearsey SJ, Grimes JM, Gorman M, Martin L, Howarth KM, Hodgson SV, et al. 2013. DNA polymerase ε and δ exonuclease domain mutations in endometrial cancer. *Hum Mol Genet* **22:** 2820–2828.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188–1190.

Denver DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledo JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M, et al. 2009. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc Natl Acad Sci* **106:** 16310–16314.

Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci* **88:** 7160–7164.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. *Genetics* **148:** 1667–1686.

Dunn OJ. 1961. Multiple comparisons among means. *J Am Stat Assoc* **56:** 52.

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, et al. 2014. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3* **4:** 389–398.

Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol* **4:** e1000216.

Flynn KM, Vohr SH, Hatcher PJ, Cooper VS. 2010. Evolutionary rates and gene dispensability associate with replication timing in the archaeon *Sulfolobus islandicus*. *Genome Biol Evol* **2:** 859–869.

Fortune JM, Pavlov YI, Welch CM, Johansson E, Burgers PM, Kunkel TA. 2005. *Saccharomyces cerevisiae* DNA polymerase δ: high fidelity for base substitutions but lower fidelity for single- and multi-base deletions. *J Biol Chem* **280:** 29980–29987.

Fortune JM, Stith CM, Kissling GE, Burgers PM, Kunkel TA. 2006. RPA and PCNA suppress formation of large deletion errors by yeast DNA polymerase δ. *Nucleic Acids Res* **34:** 4335–4341.

Gierlik A, Kowalczuk M, Mackiewicz P, Dudek MR, Cebrat S. 2000. Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? *J Theor Biol* **202:** 305–314.

Goodman MF, Creighton S, Bloom LB, Petruska J. 1993. Biochemical basis of DNA replication fidelity. *Crit Rev Biochem Mol Biol* **28:** 83–126.

Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33:** 514–517.

Hawk JD, Stefanovic L, Boyer JC, Petes TD, Farber RA. 2005. Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. *Proc Natl Acad Sci* **102:** 8639–8643.

Hodgkinson A, Chen Y, Eyre-Walker A. 2012. The large-scale distribution of somatic mutations in cancer genomes. *Hum Mutat* **33:** 136–143.

Hombauer H, Campbell CS, Smith CE, Desai A, Kolodner RD. 2011. Visualization of eukaryotic DNA mismatch repair reveals distinct recognition and repair intermediates. *Cell* **147:** 1040–1053.

Hunter CA. 1993. Sequence-dependent DNA structure: the role of base stacking interactions. *J Mol Biol* **230:** 1025–1054.

Jones M, Wagner R, Radman M. 1987. Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. *Genetics* **115:** 605–610.

Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, Benz CC, et al. 2013. Integrated genomic characterization of endometrial carcinoma. *Nature* **497:** 67–73.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458:** 362–366.

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

Kenigsberg E, Tanay A. 2013. *Drosophila* functional elements are embedded in structurally constrained sequences. *PLoS Genet* **9:** e1003512.

Kenigsberg E, Bar A, Segal E, Tanay A. 2010. Widespread compensatory evolution conserves DNA-encoded nucleosome organization in yeast. *PLoS Comput Biol* **6:** e1001039.

Kool ET. 2001. Hydrogen bonding, base stacking, and steric effects in DNA replication. *Annu Rev Biophys Biomol Struct* **30:** 1–22.

Koren A, Tsai HJ, Tirosh I, Burrack LS, Barkai N, Berman J. 2010. Epigenetically-inherited centromere and neocentromere DNA replicates earliest in S-phase. *PLoS Genet* **6:** e1001068.

Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* **91:** 1033–1040.

Kunkel TA. 2009. Evolving views of DNA replication (in)fidelity. *Cold Spring Harb Symp Quant Biol* **74:** 91–101.

Kunkel TA. 2011. Balancing eukaryotic replication asymmetry with replication fidelity. *Curr Opin Chem Biol* **15:** 620–626.

Kunkel TA, Burgers PM. 2008. Dividing the workload at a eukaryotic replication fork. *Trends Cell Biol* **18:** 521–527.

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

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

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.

Larrea AA, Lujan SA, Nick McElhinny SA, Mieczkowski PA, Resnick MA, Gordenin DA, Kunkel TA. 2010. Genome-wide model for the normal eukaryotic DNA replication fork. *Proc Natl Acad Sci* **107:** 17674–17679.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499:** 214–218.

Li WH, Wu CI, Luo CC. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol* **21:** 58–71.

Liberti SE, Larrea AA, Kunkel TA. 2013. Exonuclease 1 preferentially repairs mismatches generated by DNA polymerase α. *DNA Repair (Amst)* **12:** 92–96.

Lujan SA, Williams JS, Pursell ZF, Abdulovic-Cui AA, Clark AB, Nick McElhinny SA, Kunkel TA. 2012. Mismatch repair balances leading and lagging strand DNA replication fidelity. *PLoS Genet* **8:** e1003016.

Lujan SA, Williams JS, Clausen AR, Clark AB, Kunkel TA. 2013. Ribonucleotides are signals for mismatch repair of leading-strand replication errors. *Mol Cell* **50:** 437–443.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci* **105:** 9272–9277.

Ma X, Rogacheva MV, Nishant KT, Zanders S, Bustamante CD, Alani E. 2012. Mutation hot spots in yeast caused by long-range clustering of homopolymeric sequences. *Cell Rep* **1:** 36–42.
Marsischky GT, Kolodner RD. 1999. Biochemical characterization of the interaction between the *Saccharomyces cerevisiae* MSH2-MSH6 complex and mispaired bases in DNA. *J Biol Chem* **274:** 26668–26682.
Mugal CF, von Grunberg HH, Peifer M. 2009. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol Biol Evol* **26:** 131–142.
Muller CA, Nieduszynski CA. 2012. Conservation of replication timing reveals global and local regulation of replication origin activity. *Genome Res* **22:** 1953–1962.
Nick McElhinny SA, Gordenin DA, Stith CM, Burgers PM, Kunkel TA. 2008. Division of labor at the eukaryotic replication fork. *Mol Cell* **30:** 137–144.
Nick McElhinny SA, Kissling GE, Kunkel TA. 2010. Differential correction of lagging-strand replication errors made by DNA polymerases α and δ. *Proc Natl Acad Sci* **107:** 21070–21075.
Nishant KT, Singh ND, Alani E. 2009. Genomic mutation rates: what high-throughput methods can tell us. *BioEssays* **31:** 912–920.
Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327:** 92–94.
Pavlov YI, Nguyen D, Kunkel TA. 2001. Mutator effects of overproducing DNA polymerase eta (Rad30) and its catalytically inactive variant in yeast. *Mutat Res* **478:** 129–139.
Pavlov YI, Mian IM, Kunkel TA. 2003. Evidence for preferential mismatch repair of lagging strand DNA replication errors in yeast. *Curr Biol* **13:** 744–748.
Pink CJ, Hurst LD. 2010. Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents. *Mol Biol Evol* **27:** 1077–1086.
Polak P, Arndt PF. 2008. Transcription induces strand-specific mutations at the 5′ end of human genes. *Genome Res* **18:** 1216–1223.
Polishko A, Ponts N, Le Roch KG, Lonardi S. 2012. NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model. *Bioinformatics* **28:** i242–i249.
Prendergast JG, Semple CA. 2011. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res* **21:** 1777–1787.
Prendergast JG, Campbell H, Gilbert N, Dunlop MG, Bickmore WA, Semple CA. 2007. Chromatin structure and evolution in the human genome. *BMC Evol Biol* **7:** 72.
Pursell ZF, Isoz I, Lundstrom EB, Johansson E, Kunkel TA. 2007. Yeast DNA polymerase ε participates in leading-strand DNA replication. *Science* **317:** 127–130.
Radman-Livaja M, Rando OJ. 2010. Nucleosome positioning: how is it established, and why does it matter? *Dev Biol* **339:** 258–266.
Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, et al. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45:** 970–976.
Rocha EP, Touchon M, Feil EJ. 2006. Similar compositional biases are caused by very different mutational effects. *Genome Res* **16:** 1537–1547.
Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. *Mol Biol Evol* **20:** 988–993.
Salk JJ, Fox EJ, Loeb LA. 2010. Mutational heterogeneity in human cancers: origin and consequences. *Annu Rev Pathol* **5:** 51–75.
Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, Matsushima K, Gu SG, Kasahara M, Ahsan B, et al. 2009. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* **323:** 401–404.
Schmidt S, Gerasimova A, Kondrashov FA, Adzhubei IA, Kondrashov AS, Sunyaev S. 2008. Hypermutable non-synonymous sites are under stronger negative selection. *PLoS Genet* **4:** e1000281.
Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18:** 6097–6100.
Schuster-Bockler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488:** 504–507.
Serero A, Jubin C, Loeillet S, Legoix-Ne P, Nicolas AG. 2014. Mutational landscape of yeast mutator strains. *Proc Natl Acad Sci* **111:** 1897–1902.
Sharp PM, Shields DC, Wolfe KH, Li WH. 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* **246:** 808–810.
Shcherbakova PV, Pavlov YI, Chilkova O, Rogozin IB, Johansson E, Kunkel TA. 2003. Unique error signature of the four-subunit yeast DNA polymerase ε. *J Biol Chem* **278:** 43770–43780.
Šidák Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc* **62:** 626–633.
Siow CC, Nieduszynska SR, Muller CA, Nieduszynski CA. 2012. OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Res* **40:** D682–D686.
Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41:** 393–395.
Subramanian S, Kumar S. 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res* **13:** 838–844.
Tolstorukov MY, Volfovsky N, Stephens RM, Park PJ. 2011. Impact of chromatin structure on sequence variability in the human genome. *Nat Struct Mol Biol* **18:** 510–515.
Touchon M, Nicolay S, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett* **555:** 579–582.
Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc Natl Acad Sci* **102:** 9836–9841.
Washietl S, Machne R, Goldman N. 2008. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet* **24:** 583–587.
Weber CC, Pink CJ, Hurst LD. 2012. Late-replicating domains have higher divergence and diversity in *Drosophila melanogaster*. *Mol Biol Evol* **29:** 873–882.
Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.
Woo YH, Li WH. 2012. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat Commun* **3:** 1004.
Yabuki N, Terashima H, Kitada K. 2002. Mapping of early firing origins on a replication profile of budding yeast. *Genes Cells* **7:** 781–789.
Ying H, Epps J, Williams R, Huttley G. 2010. Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. *Mol Biol Evol* **27:** 637–649.
Zanders S, Ma X, Roychoudhury A, Hernandez RD, Demogines A, Barker B, Gu Z, Bustamante CD, Alani E. 2010. Detection of heterozygous mutations in the genome of mismatch repair defective diploid yeast using a Bayesian approach. *Genetics* **186:** 493–503.

Received May 13, 2014; accepted in revised form September 5, 2014.