Published online 24 November 2015

Nucleic Acids Research, 2016, Vol. 44, No. 4 ***1669–1680***
doi: 10.1093/nar/gkv1298

Genome-wide analysis of the specificity and mechanisms of replication infidelity driven by imbalanced dNTP pools

Danielle L. Watt[1,2,†], Robert J. Buckland[1,3,†], Scott A. Lujan[2,†], Thomas A. Kunkel[2] and Andrei Chabes[1,3,*]

[1]Department of Medical Biochemistry and Biophysics, Umeå University, SE-901 87, Umeå, Sweden, [2]Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, NC 27709, USA and [3]Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå University, SE-901 87, Umeå, Sweden

Received July 21, 2015; Revised November 5, 2015; Accepted November 6, 2015

ABSTRACT

The absolute and relative concentrations of the four dNTPs are key determinants of DNA replication fidelity, yet the consequences of altered dNTP pools on replication fidelity have not previously been investigated on a genome-wide scale. Here, we use deep sequencing to determine the types, rates and locations of uncorrected replication errors that accumulate in the nuclear genome of a mismatch repair-deficient diploid yeast strain with elevated dCTP and dTTP concentrations. These imbalanced dNTP pools promote replication errors in specific DNA sequence motifs suggesting increased misinsertion and increased mismatch extension at the expense of proofreading. Interestingly, substitution rates are similar for leading and lagging strand replication, but are higher in regions replicated late in S phase. Remarkably, the rate of single base deletions is preferentially increased in coding sequences and in short rather than long mononucleotides runs. Based on DNA sequence motifs, we propose two distinct mechanisms for generating single base deletions *in vivo*. Collectively, the results indicate that elevated dCTP and dTTP pools increase mismatch formation and decrease error correction across the nuclear genome, and most strongly increases mutation rates in coding and late replicating sequences.

INTRODUCTION

The three main safety systems that determine DNA replication fidelity are DNA polymerase selectivity, proofreading and mismatch repair. The major replicative DNA polymerases (Pols) α, δ and ϵ insert correct nucleotides during DNA replication with high accuracy that partly depends on the correct absolute and relative concentrations of the four dNTPs (1,2). The correct concentrations of dNTPs are primarily maintained by the enzyme ribonucleotide reductase (RNR) (3,4). Occasional nucleotide misinsertions can be excised by the 3′-5′ exonuclease activities intrinsic to the catalytic subunits of Pols δ and ϵ, and rare replication errors that escape this proofreading can be repaired by Msh2-dependent mismatch repair (MMR, recently reviewed in (5). We have previously shown that mutations in the allosteric specificity site of the budding yeast large RNR subunit Rnr1 result in imbalanced dNTP pools and can strongly reduce replication fidelity (6,7). In an *rnr1-Y285A* mutant strain, with elevated dCTP and dTTP concentrations, the mutation rate at the *CAN1* locus on chromosome 5 increased ∼14-fold. Loss of MMR in this strain (*rnr1-Y285A msh2Δ*) raised the mutation rate additionally ∼40-fold. The fidelity of both leading and lagging strand replication was reduced, and the types and locations of mutations were consistent with increased misinsertion of dCTP and dTTP followed by efficient mismatch extension that reduced exonucleolytic proofreading by Pols δ and ϵ (8). That study determined the effects of the *rnr1-Y285A* mutation on the fidelity of replicating the *CAN1* locus on chromosome 5, representing only 0.014% of the yeast genome. Here, we expand this view by performing a mutation accumulation experiment in the absence of purifying selection using a homozygous diploid *rnr1-Y285A msh2Δ* mutant yeast strain, allowing comprehensive analysis of the effects of the *rnr1-Y285A* mutation on replication fidelity across the genome. When compared to a recent analysis of an *msh2Δ* single mutant strain (9), the results indicate that elevated dCTP and dTTP in the *rnr1-Y285A* strain increase base substitution and base deletion rates on both DNA strands across

[*]To whom correspondence should be addressed. Tel: +46 90 786 5937; Fax: +46 90 786 9795; Email: andrei.chabes@umu.se
[†]These authors contributed equally to the paper as first authors.

the entire yeast nuclear genome, and do so preferentially in coding and late replicating sequences when purifying selection is not operating. Replication errors are observed in signature DNA sequence motifs indicating increased misinsertion and increased mismatch extension at the expense of proofreading, in patterns that differ for deleting iterated and non-iterated bases, implying that deletions are generated by two distinct misalignment mechanisms.

MATERIALS AND METHODS

Yeast strains and culture conditions

A homozygous diploid strain of *rnr1-Y285A-TRP msh2Δ*::Hyg was created by crossing two *rnr1-Y285A msh2Δ* haploid strains created as described in (8). Zygotes were selected on Hygromycin and −Trp (Tryptophan) media to give *rnr1-Y285A/rnr1-Y285A msh2Δ/msh2Δ*. All culturing was carried out at 30°C in YPAD (1% yeast extract, 2% bacto-peptone, 20mg/l adenine, 2% agar for plates) liquid cultures in a shaking incubator at 160 rpm.

dNTP pool measurements

dNTP pools were measured in asynchronous cultures as described in (10). Briefly, cells were harvested by filtration at a density of 0.4×10^7 to 0.5×10^7 cells/ml, and NTPs and dNTPs were extracted in trichloroacetic acid and $MgCl_2$ followed by extraction with a Freon-trioctylamine mix. dNTPs were separated from NTPs using boronate columns (Affigel 601, BioRad) and analyzed by HPLC on a LaChrom Elite UV detector (Hitachi) using a Partisphere SAX column (Hichrom, UK).

Flow cytometry

Cells in an asynchronous growing culture were analyzed for cell cycle progression and ploidy by staining of DNA with SYBR Green (Molecular Probes) as per (11) in a Becton Dickinson FC500.

Whole genome sequence analysis

Experiments were done as described in the workflow diagram of the Supplementary Figure S1 in Lujan *et al.* (9). Briefly, cells were subjected to 28 single-cell bottleneck passages on complete media (YPAD), which equates to ∼810 generations (12). Samples were retained periodically for glycerol stocks and phenotype testing. Genomic DNA from 10 ml cultures was isolated via the Epicentre MasterPure Yeast DNA Purification Kit (MPY80200), including the optional RNase A treatment step. Genome sequencing and analysis was completed as described previously (9). Briefly, genomic DNA was fragmented to between 200 and 800 bp and libraries were prepared using Illumina TruSeqTM DNA Sample Prep Kits. The quantified libraries were diluted to 15 nM and pooled for paired-end sequencing (2 × 100 bp read length) performed on a HiSeq 2500 sequencer (Illumina). The generation 0 sequence was assembled and used as the reference to map sequence reads and call variant base pairs using CLC bio Genomics Workbench version 5.1.5. The variants were filtered and pooled for mutation rate calculations for any mutation type in any section of the genome (bins). Sequence motifs were detected and logos with the nucleotide frequency and the relative information content measured in bits (13) were created using custom Excel tools.

RESULTS

Cell cycle progression and dNTP pools

To determine the effects of imbalanced dNTP pools on genome-wide replication fidelity, we generated and characterized an *rnr1-Y285A msh2Δ* yeast strain. In comparison to our earlier study that measured mutation rates at a specific locus in a haploid strain (8), this new strain was made diploid in order to reduce the effect of purifying selection on mutations expected to accumulate during many generations. Initially (Figure 1A), and after one passage on solid medium representing ∼30 generations (Figure 1B, top), clonal isolates of this diploid strain had normal cell cycle progression (three of seven clonal isolates are shown in Figure 1B). By passage 28, the isolates grew noticeably slower than at passage 1 (Supplementary Figure S1) and one had an abnormal flow cytometry profile (Figure 1B, bottom left), while cells of the other isolates were closer to normal but did accumulate more in G2/M. Differences among clonal isolates are to be expected because, as shown below, these isolates stochastically accumulate large numbers of sequence changes during outgrowth that are not shared among isolates. Despite accumulation of many mutations, the dNTP pool imbalances in the diploid mutant isolates at passage 1 and passage 28 were comparable to those in the haploid strain, with ∼20- and 16-fold higher than wild-type (WT) concentrations of dCTP and dTTP, respectively, and slightly higher than WT concentrations of dATP and dGTP (Figure 1C).

Spontaneous mutation rates and distribution of mutations across the genome

We sequenced the nuclear genomes of the seven clonal isolates of the *rnr1-Y285A msh2Δ* strain after 28 passages comprising about 800 generations each (Table 1). Alignments to the ‘zero passage’ control genome revealed that a total of 18 664 single base mutations accumulated (Table 1). The vast majority of these were single base changes that were broadly distributed across all 16 chromosomes (Figure 2A). Among them, 15 678 (84%) were base substitutions, with a 7:3 ratio of transversions to transitions. The remaining 2986 sequence changes were insertions-deletions, 2682 (96%) of which were single base deletions. From these data, we calculated mutation rates per diploid genome per generation (μ_g) and per base pair per generation (μ_{bp}) (Table 1). These rates were compared to previously published rates for an *msh2Δ* strain (9), which are listed again here for comparison.

When considering all sequence changes in the double mutant *rnr1-Y285A msh2Δ* strain, μ_g and μ_{bp} were 3.3 and 140×10^{-9}, respectively (Table 1, right side). Both values are about 9-fold higher than the corresponding rates in the *msh2Δ* strain (0.38 and 17×10^{-9}, respectively). Thus, the formation of replication errors that remain uncorrected in *msh2Δ* strains due to the defect in MMR is strongly increased by the *rnr1-Y285A* mutation that selectively elevates

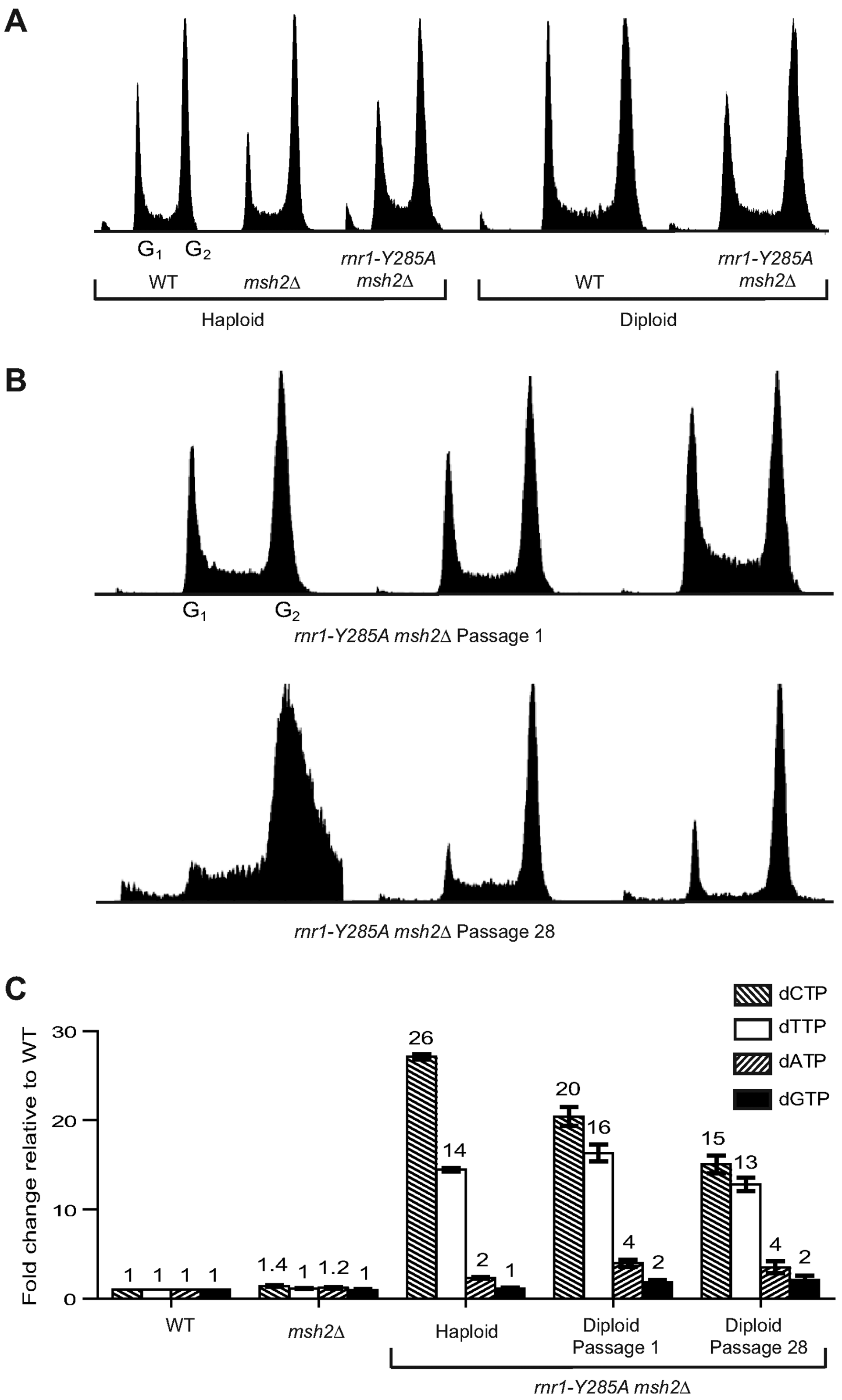


Figure 1. Cell cycle progression and dNTP pools. (**A**) Flow cytometry histograms showing DNA content of haploid and diploid cells stained with SYBR Green. The first small peak prior to G1 is cell debris and between the two peaks are cells in S phase. (**B**) Flow cytometry histograms for three *rnr1-Y285A msh2* isolates at passage 1 and 28. (**C**) dNTP pools as measured by HPLC, average of 8–10 cultures (4 cultures for the diploid at Pass 28) with the Standard Error of Mean (SEM) and fold change relative to WT above the data bars. The average WT values were dCTP 48, dTTP 131, dATP 66 and dGTP 29 pmol per 10^8 cells.

Table 1. Classification of mutation types and rates

Strain	*msh2Δ* [a]		*rnr1-Y285A msh2Δ*		
Isolates	5		7		
Elapsed passages	142 (22–30)		189 (27)		
Elapsed generations	4260 (660–900)		5670 (810)		
	Count	μ_{bp} x 10^{-9}	Count	μ_{bp} x 10^{-9}	μ_{RNR1}:$\mu_{rnr1\text{-}Y285A}$
Base substitutions	477	4.9	15 678	120	25
T/A → C/G [b]	151	2.5	1067	13	5.3
C/G → T/A [b]	167	4.5	3694	74	17
T/A → A/T [b]	30	0.49	517	6.4	13
T/A → G/C [b]	26	0.43	892	11	26
C/G → A/T [b]	86	2.3	9499	191	83
C/G → G/C [b]	17	0.46	9	0.18	0.40
Insertions/Deletions	1017	12	2986	23	1.9
−T/A [b]	955	16	2015	25	1.6
−C/G [b]	11	0.29	667	13	46
+T/A [b]	46	0.47	186	1.4	3.0
+C/G [b]	5	0.051	5	0.038	0.75
>1 bp deletions	135	1.4	93	0.71	0.52
>1 bp insertions	8	0.081	20	0.15	1.9
Total	1637	17	18 664	140	8.6
μ_g (diploid)	0.38 (0.20–0.51)		3.3 (2.8–4.0)		8.6

[a] Data published previously (9).

[b] Corrected for target size based on the percentage of A + T and G + C in the genome as previously described (9).

The number in parenthesis is the range for the individual isolates. All values are rounded to two significant digits

dCTP and dTTP concentrations. The extent of this increase varied over a wide range depending on the type of mutations. For total base substitutions, the average substitution rate per base pair increased by 25-fold (4.9 × 10^{-9} in *msh2Δ* versus 120 × 10^{-9} in *rnr1-Y285A msh2Δ*), whereas the average single base deletion rate per base pair increased by only 1.9-fold (12 × 10^{-9} in *msh2Δ* versus 23 × 10^{-9} in *rnr1-Y285A msh2Δ*). Variations further extend to subclasses of substitutions and single base deletions. For example, the mutator effect of the *rnr1-Y285A* mutation was greatest for C/G to A/T (83-fold), T/A to G/C (26-fold) and C/G to T/A (17-fold) substitutions (Figure 2B). These substitutions are inferred to result from C-dTTP, T-dCTP and G-dTTP replication errors driven by excess dTTP and dCTP in the *rnr1-Y285A msh2Δ* strain. In contrast, the rate of C/G to G/C substitutions that would result from C-dCTP or G-dGTP mismatches was not elevated, but instead was slightly reduced. This result is anticipated because C-dCTP mismatch are rarely generated during replication, and because formation of G-dGTP mismatches should diminish due to the high ratio of correct dCTP to incorrect dGTP (Figure 1C). Variable effects were likewise seen among single base deletion rates, with the average rate of deleting a C/G base pair increased by 46-fold (Figure 2C), with smaller or no changes in average rates observed for other classes of single base deletions. Large changes in rates for specific subsets of single base deletions are described and discussed further below.

The rnr1-Y285A mutator effect on leading and lagging strand replication

The above results indicate that most base substitutions in the *rnr1-Y285A msh2Δ* strain can be explained by misincorporation of the two dNTPs present in excess, dCTP and dTTP. This circumstance provides the opportunity to determine whether the *rnr1-Y285A* mutation drives replication infidelity to similar or different extents during leading and lagging strand replication (primarily catalyzed by Pol ϵ versus Pols α and δ, respectively). To examine this, we performed a meta-analysis of the distribution of mismatches generated immediately to the left and right of replication origins across the genome as described in (9). For example, a C/G to A/T substitution to the right of a replication origin would result from misincorporation of dTTP opposite template C during leading strand replication, whereas the same mismatch generated to the left of a replication origin would occur during lagging strand replication (see schematic in Figure 3A). When the 9499 C/G to A/T substitutions observed here were mapped relative to replication origins, the distribution of C to A versus G to T substitutions was constant between origins (Figure 3B). Similar results were obtained for the other base substitutions driven by the *rnr1-Y285* mutation (Supplementary Figure S2). Moreover, base substitution (and base deletion) rates did not change in the *rnr1-Y285A msh2Δ* strain as a function of the distance traversed by each replication fork between adjacent origins (data not shown). This last result is similar to what was observed earlier in the *msh2Δ* single mutant strain (9).

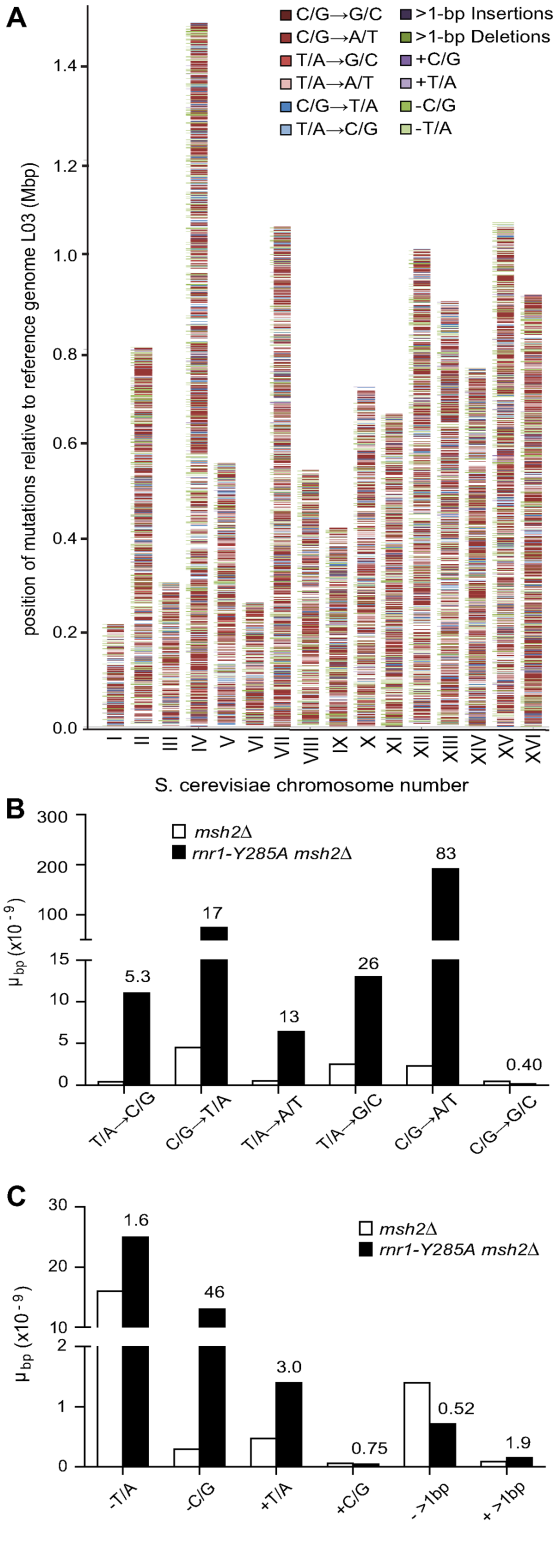


Figure 2. Genome-wide mutations and rates. (**A**) Distribution of mutations across 16 chromosomes. Transversions are indicated by red shades (dark to light: C/G→G/C, C/G→A/T, T/A→G/C, T/A→A/T), transitions by blue shades (dark to light: C/G→T/A, T/A→C/G), deletions by green shades (dark to light: >1 bp, −C/G, −T/A) and insertions by purple shades (dark to light: >1 bp, +C/G, +T/A). (**B**) Replication error rates for single base substitutions. (**C**) Replication error rates for single and multi-base indels. Numbers above the bars are the fold change compared to the *msh2*Δ strain. *msh2*Δ data previously published in (9).

Preferential substitution mutator effects in late S phase

The above data imply that the *rnr1-Y285A* mutation promotes replication infidelity to similar extents during leading and lagging strand replication across the nuclear genome. To examine if this mutator effect was constant throughout S phase, we calculated the transversion and transition rates in the *rnr1-Y285A msh2*Δ mutant for replication occurring late as compared to early after release from α-factor as described in (9) (Figure 3C). These rates steadily increased as S phase progressed, ultimately increasing in late S phase by more than 2-fold. These increases are preferential for the *rnr1-Y285A msh2*Δ mutant, i.e. they are greater than those seen in the *msh2*Δ single mutant strain (Figure 3C, reproduced from (9)). We also calculated substitution rates in coding sequences as compared to 5′ and 3′ flanking sequences. Compared to the *msh2Δ* single mutant strain analyzed earlier (Figure 3D, middle panel, reproduced from (9)), the new results demonstrate that the *rnr1-Y285A* mutation strongly reduces fidelity during replication of coding and non-coding DNA.

A preferred sequence motif for generating base substitutions

A motif detection algorithm was used to determine if single base–base mismatches were generated in preferred sequence contexts (9). For this purpose, we analyzed the most common base substitution, the C/G to A/T mutation inferred to result from a C-dTTP mismatch. As in an earlier study (9), to increase confidence in assigning the direction of replication during formation of the C-dTTP mismatch, we focused on the subset of C/G to A/T substitutions occurring closest to replication origins (relative inter-origin distance <0.1, gray shade in Figure 3A). During both leading and lagging strand replication, the most common template strand sequence motif for generating the C-dTTP mismatch was 5′-AAAG$\underline{\text{C}}$T-3′ (Figure 4A), where $\underline{\text{C}}$ is the position of the substitution and the direction of synthesis of the template strand is from right to left. The newly synthesized nascent strand is thus 3′-TTTC$\overline{\text{T}}$A-5′. Explanations for this motif (Figure 4B) are discussed below.

Preferential effects on single base deletion rates in repeat sequences

Taking into account differences in the number of T/A and C/G base pairs in mononucleotide runs of various lengths in the yeast genome (Supplementary Table S1), we calculated the single base deletion rates per base pair replicated per generation as a function of increasing mononucleotide run length. The results (Figure 5A and B, Supplementary Table S2) reveal that, compared to our earlier analysis of the *msh2Δ* single mutant strain (Figure 5A/B, gray curves), some but not all deletion rates were higher in the *rnr1-Y285A msh2Δ* strain (black curves). Specifically, the mutator effects of the *rnr1-Y285A* mutation (Figure 5C and D) were greater for short mononucleotide runs as compared to long mononucleotide runs. For example, the *rnr1-Y285A* mutation has no effect on the rate of deleting a C/G pair from a 10-base run, but elevated the rate of deleting a C/G pair from a 4-base run by 175-fold. Based on these preferential mutator effects and using the approach described above

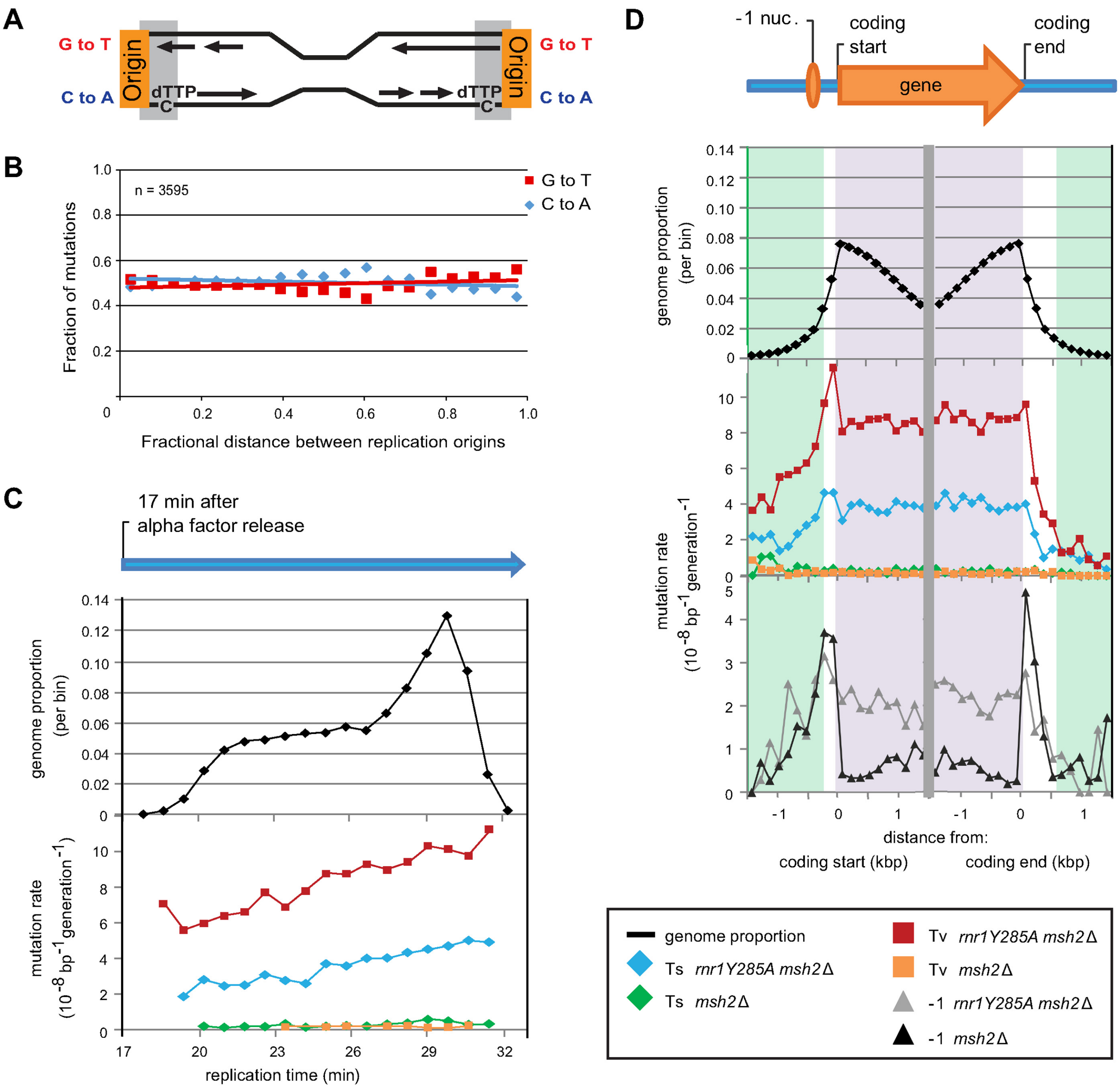


Figure 3. Mutation rates relative to genomic landmarks. (**A**) Schematic of adjacent replication origins and strand specific mutagenesis. Gray boxes indicate the closest 10% of the inter-origin distance. The long arrow represents continuous leading strand replication by Pol ϵ and the short arrows represent discontinuous lagging strand replication by Pol α/δ. For simplicity, misincorporation events are shown only on the bottom strands, but can occur on both the top and bottom strands. (**B**) Distribution of C-G to A-T base substitutions plotted as a function of relative distance between adjacent origins. Mutation rates plotted versus (**C**) replication timing across the genome after α-factor release and (**D**) the nearest gene coding start (left) or coding end (right) site in kilobase pairs (kbp). The top black plot is the target size, blue and green are transitions, red and orange are transversions, and black and gray are 1-base deletions. Gene regions shaded green are intergenic, white are 5′-UTR and 3′-UTR, and purple is coding.

for substitutions, we generated sequence logos for deleting a C/G pair from runs of different lengths (Figure 6). These motifs differ, with mechanistic (Figure 6, top) and biological implications discussed below.

Preferential mutator effects in coding as compared to non-coding DNA

When calculated as described earlier (9), average single base deletion rates across the open reading frames of genes (shaded purple in Figure 3D, lower panel) were substantially higher in the *rnr1-Y285A msh2Δ* as compared to the *msh2Δ* strain. This contrasts with single base deletion rates

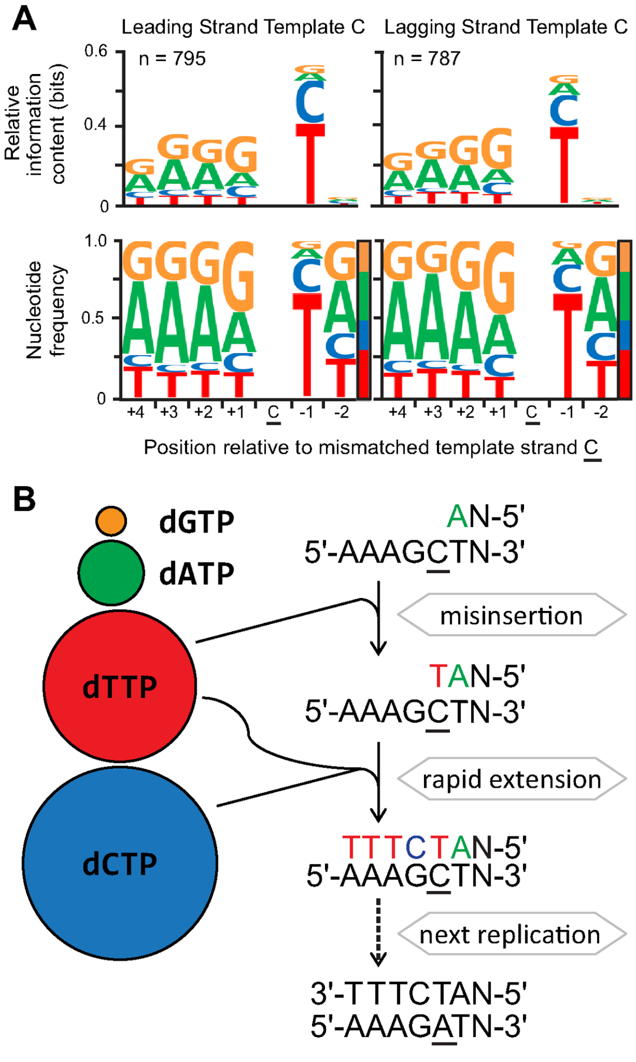


Figure 4. Context specificity of substitutions promoted by a dNTP pool imbalance. (**A**) Sequence logos of relative information content and nucleotide frequency for template strand C-dTTP mispairs. 'n' is the mutation count for the inter-origin distance used (Figure 3A, gray boxes). (**B**) Mechanism for a C–A mutation. The colored circles are the relative dNTP pools, the underlined character is the mutational event and N represents an undefined nucleotide.

in non-coding sequences that flank genes, which are not strongly increased in the *rnr1-Y285A msh2Δ* as compared to the *msh2Δ* mutant strain.

DISCUSSION

This study provides a genome-wide view of the consequences of a dNTP pool imbalance on nuclear DNA replication fidelity. The results offer insights into the mechanisms underlying formation of single base–base and deletion mismatches during leading and lagging strand replication, and they have several biological implications.

Base–base mismatches driven by the dNTP pool imbalance

The vast majority of base substitutions in the *rnr1-Y285A msh2Δ* strain (Table 1, Figure 2) can be explained by misincorporation of the two dNTPs present in excess, dCTP and dTTP (Figure 1C). This fact is key to mechanistic interpretations, because it allows the mismatch composition and the identity of the template and nascent strands involved in the replication error to be deduced. In a recent study of the *rnr-1-Y285A msh2Δ* strain that monitored replication fidelity at a specific locus (*CAN1*) that comprises only 0.014% of the yeast genome (8), the rates and spectra of base substitutions were similar when each DNA strand was replicated as the leading or the lagging strand. This fact implied that the *rnr1-Y285A* mutation promotes infidelity to similar extents during both leading and lagging strand replication. The current results (Figure 3A and B and Supplementary Figure S2) strongly support this interpretation and expand it to a variety of single base mismatches generated across the whole genome. This similarity in the consequences of a dNTP pool imbalance on replication of the two strands is remarkable given that replication of the leading and lagging strand templates is catalyzed by DNA polymerases that differ in structure, subunit composition, partnerships with replication accessory proteins, proofreading potential, processivity and the fidelity of DNA synthesis. Although one recent study proposed that Pol δ is the major polymerase for both the leading and lagging strands (14), substantial evidence from many different laboratories ((9) and references therein and (15–18)) supports the view that the leading and lagging strands are primarily replicated by Pol ϵ and Pol δ, respectively.

The genome-wide view provided by this study suggests that the dNTP pool imbalance resulting from the *rnr1-Y285A* mutation reduces replication fidelity as S phase progresses (Figure 3C). Given that substitution rates in the *rnr1-Y285A msh2Δ* strain are substantially higher at C/G as compared to T/A base pairs (Table 1, Figure 2), the higher mutation rate in late S phase suggests that the dNTP pool imbalance driven by the *rnr1-Y285A* mutation increases further as S phase progresses, a possibility yet to be examined. The results in Figure 3D further show that the dNTP pool imbalance resulting from the *rnr1-Y285A* mutation strongly reduces the fidelity of replication of both coding and non-coding DNA sequences, and that base substitution rates in the *rnr1-Y285A msh2Δ* strain are higher in coding sequences (shaded purple) and their immediately flanking 5′- and 3′-untranslated regions (white) than in intergenic regions (shaded green). This latter observation was also made in our earlier study of *msh2Δ* strains encoding WT *RNR1* and WT or variant replicases (9). As mentioned in that study, these elevated rates contrast with the evolutionary record, wherein nucleotide variation is lower in genes and lowest in the nucleosome free regions 5′ to genes compared to intergenic DNA regions (19,20). This inverse relationship suggests that sequences in which mutations are normally deleterious are highly mutable in the absence of purifying selection, as in these experimental conditions. The present observations in the *rnr1-Y285A msh2Δ* mutant suggest this same interpretation applies when dNTP pools are imbalanced. In our earlier study, we speculated that collisions between transcription and replication machineries, and/or spontaneous damage to single-stranded DNA in transcription bubbles, might contribute to the higher mutation rate in coding sequences (9). We are currently inves-

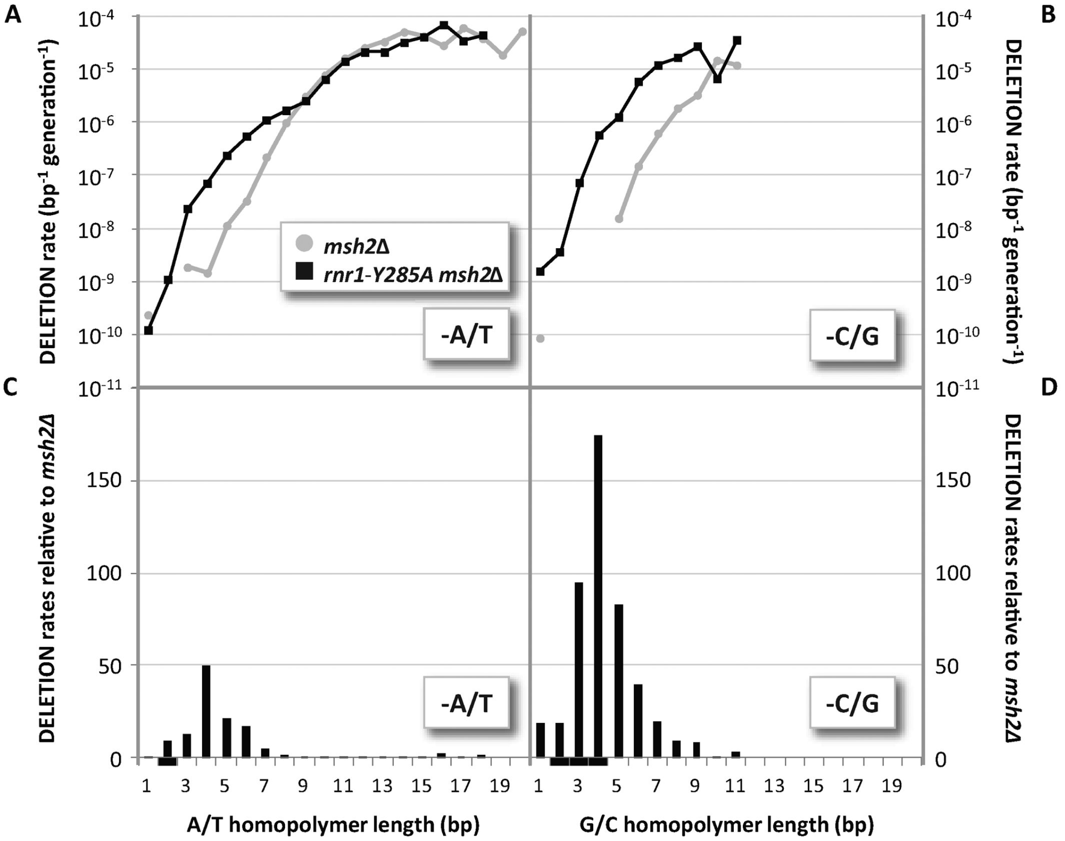


Figure 5. Deletion rates as a function of homopolymer length. Rates for *msh2*Δ and *rnr1-Y285A msh2*Δ strains for (**A**) deletion of T/A bp and (**B**) deletion of C/G bp. Deletion rates for given homopolymer length for *rnr1-Y285A msh2*Δ relative to *msh2*Δ for (**C**) T/A bp and (**D**) C/G bp. Black horizontal bar on X-axis indicate regions where the indel counts in the *msh2*Δ strain equaled 0 and therefore, the histograms in these regions are minimum estimates. *msh2*Δ data published in (28).

tigating whether higher mutation rates in coding sequences correlate with levels of gene expression.

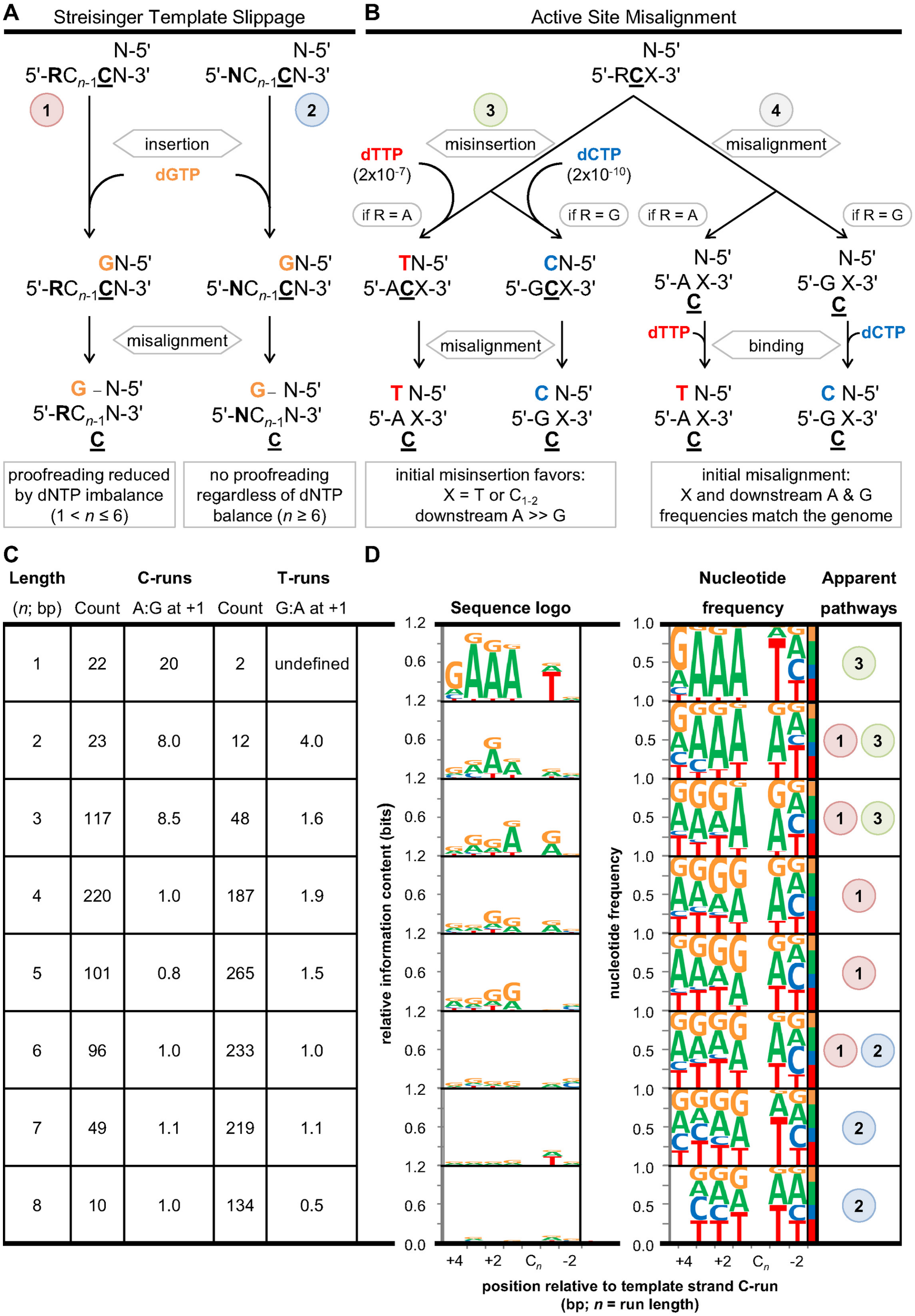
The likely mechanisms by which excess dTTP and dCTP result in base substitution errors during replication in the *rnr1-Y285A msh2Δ* strain can be inferred by considering the preferred DNA sequence motif for the substitution observed at the highest rate, i.e. C/G to A/T (Figures 2 and 4A). Theoretically, this substitution could result from misincorporation of dTTP opposite template C on one strand, or from misincorporation of dATP opposite G on the other strand. However, as depicted in the schematic diagram in Figure 4B, the first possibility is much more likely because in the *rnr1-Y295A* mutant, the concentration of incorrect dTTP is much higher than the concentration of correct dGTP, whereas the concentrations of incorrect dATP and correct dGTP are more similar. This implies that the excess of incorrect dTTP over correct dGTP fuels misinsertion of dTTP opposite template C. On average across the genome, this misinsertion is preferentially preceded by a template pyrimidine, implying that the incoming incorrect dTTP is stabilized by stacking with a previously incorporated purine, more often adenine opposite the template T (position −1 in Figure 4A), but sometimes guanine opposite template C (underlined in the same logos). The preferential presence of several consecutive template purines on the 5′ side of the substitution further suggests that misinsertion of dTTP is followed by efficient correct incorporation of dCTP and dTTP, both of which are present in much higher than normal concentrations, opposite four consecutive template purines. Efficient mismatch extension diminishes the opportunity for fraying of the primer terminus, thereby reducing formation of single-stranded DNA that must move to the exonuclease active site for excision of the incorrect dTTP. These results are in agreement with the concept of the next nucleotide effect (21–23), first predicted by Ninio (24). The observation that the motifs for leading and lagging strand replication are remarkably similar (Figure 4A) implies that the above explanations are about equally relevant to the leading and lagging strand replicases.

Single base deletion mismatches driven by the dNTP pool imbalance

The *rnr1-Y285A* mutation also elevates the rates of single base deletions. The sequence contexts in which these deletions occur differ (Figure 6), implying that different misalignment mechanisms give rise to single base deletion mutations during DNA replication *in vivo*.

Strand slippage

The classical explanation for the origin of insertions-deletions (indels) during DNA synthesis is strand slippage



Length	C-runs		T-runs	
(n; bp)	Count	A:G at +1	Count	G:A at +1
1	22	20	2	undefined
2	23	8.0	12	4.0
3	117	8.5	48	1.6
4	220	1.0	187	1.9
5	101	0.8	265	1.5
6	96	1.0	233	1.0
7	49	1.1	219	1.1
8	10	1.0	134	0.5

Figure 6. Misalignment mechanisms for single base deletions in mononucleotide runs. Predicted context preferences for the deletion of a single base in a poly-C run due to either (**A**) Streisinger template slippage or to (**B**) template misalignment in the polymerase active site. The context of Streisinger template slippage is predicted to differ between (1) regimes where dNTP imbalance promotes extension over proofreading ($C_{1 < n \leq 6}$) and (2) regimes where proofreading is poor regardless of nucleotide pools ($C_{n \geq 6}$). The predicted context is also expected to differ when active site misalignment immediately (3) follows a misinsertion or (4) precedes next-nucleotide binding. (**C**) Table comparing deletions in C- and T-run of lengths 1–8 and template A:G or G:A ratios in the +1 position. (**D**) Sequence logos (relative information content), nucleotide frequency and apparent causative pathways of deletions in C-runs of lengths 1–8. Pathways are numbered as per A and B. **R** indicates any purine, **N** indicates any nucleotide and *n* is the number of iterated nucleotides. **X** stands for specific nucleotides or sequences as indicated in B.

in repetitive DNA sequences (25) and reviewed in (26,27). This mechanism explains the >100 000-fold increase in single base deletion rates across the yeast genome as a function of increasing homonucleotide run length. We recently observed this 'rate versus length' relationship in a genome wide study of the *msh2Δ* single mutant strain (gray curves in Figure 5A and B (28)). We now see a generally similar pattern in the *rnr1-Y285A msh2Δ* strain (black curves), implying that strand slippage during replication underlies the majority of the single base deletion mutagenesis in both strains. Interestingly, deletion rates for runs of about seven or more base pairs are similar in the two strains (Figure 5C and D), suggesting that the excess dCTP and dTTP does not increase the probability that slippage occurs during replication of long homonucleotide runs (Figure 6A, pathway 2). No effect on slippage when dNTP concentrations are *increased* contrasts with the effect of decreasing dNTP concentrations, where, at least *in vitro*, the probability of slippage is *increased* by *decreasing* the dNTP complementary to the template run being replicated (29). In the future, it will be interesting to determine if, unlike the current situation, decreasing dNTP levels below normal, e.g. in an *rnr1-R293A* or *rnr1-Q288A* (6), increases the rate of indels in long runs.

In marked contrast to no effect in long runs, the *rnr1-Y285A* mutation does elevate the rate of deletions in shorter runs, by factors up to 175-fold (Figure 5C and D). We suggest that these increases result from an event downstream of initial strand slippage, namely efficient extension of the resulting misaligned substrates driven by high dCTP and dTTP concentration. Just as for base–base mismatches, efficient extension of misalignment mismatches in short runs will diminish the probability of proofreading, or alternatively, the opportunity for the two strands to simply re-align. In support of this interpretation, extensive biochemical and genetic data (reviewed in (27)) have previously shown that proofreading is ineffective at correcting misalignments in long runs because the misaligned base can be so deeply embedded in the duplex DNA upstream of the polymerase active site as to avoid the fraying needed for proofreading. In contrast, a misaligned base in a shorter run is closer to the primer terminus as the bulge is not able to migrate further away from the active site than the end of the homopolymer run and is presumably trapped in the DNA polymerase channel. Therefore a misaligned base in a shorter run is less protected against fraying and can still be proofread (28,30). If so, we would expect the downstream template sequences to be purine-rich when excess dTTP and dCTP drives extension (Figure 6A, pathway 1), as compared to upstream sequences that closely match genomic nucleotide frequencies when proofreading is poor regardless of dNTP pools (Figure 6A, pathway 2). This is precisely the result of the present genome-wide study of the *rnr1-Y285A msh2Δ* strain: the rate of deletions in runs of six or fewer base pairs (Figure 5C and D) is preferentially increased in sequences enriched for template purines on the 5′-side of the run (Figure 6D). However, this is not the case for runs of seven or more base pairs (Figure 6D, bottom), in which the misaligned template base can be six correct base pairs upstream of the polymerase active site. Beyond the implications for replication fidelity *per se*, the pattern in Figure 6D is, to our knowledge, the first genome-wide 'footprint' for functional interactions of replicases with the upstream duplex DNA during replication *in vivo*, a footprint that nicely correlates with protein—duplex DNA interactions observed in the crystal structures of the yeast replicases (31,32). Recent functional studies of yeast Pol epsilon showed that contacts between R988 and bases at position n-4/n-5 are important to stabilize the 3′-end of the nascent strand in the polymerase site. Thus, replication errors that lead to a loss of interaction between the R988 and duplex DNA cause a switch to the exonuclease site and removal of misincorporated nucleotides (33). The position of a sensor (R988 in Pol ϵ and R839 in Pol δ) at n-4/n-5 correlates well with our observed length dependence.

Misinsertion-primer relocation and dNTP stabilized misalignment

The *rnr1-Y285A* mutation also increases by 10-fold or more the rate of deleting non-iterated bases and bases in runs of only two or three bases (Figure 5A and B). Importantly, we expect the preferred sequence motif for deleting a non-iterated template C to differ somewhat from the motif for deletion-containing template C homonucleotide runs. The greatest difference is in the identity of the +1 template base, 5′ to the deleted C. Here an A is strongly preferred for non-iterated base deletions, and only slightly less preferred for deletions from 2–3 base runs, but is not preferred for longer runs (Figure 6C and D). In light of earlier studies of indel error formation *in vitro*, this pattern implies a mechanism for initial misalignment that is distinct from classical Streisinger strand slippage. In the absence of the correct base pairing possible in iterated sequences, a misaligned, non-iterated base was proposed to reside at or near the polymerase active site itself, rather than in the upstream duplex (34). Two models were proposed to give rise to misalignment at the active site. In one model supported by kinetic data using lesion-containing substrates (35–37), initial slippage is followed by binding of the next correct dNTP to stabilize the misaligned base (Figure 6B, pathway 4). In the *rnr1-Y285A msh2Δ* strain, dCTP or dTTP are present at high and similar concentrations (Figure 1C), predicting that A and G should be about equally represented at the +1 template position. However, this is not the case for non-iterated base deletions and deletions from 2–3 base runs, where A is strongly preferred over G at the +1 position (Figure 6C and D). Thus, the observed sequence motif disfavors the dNTP-stabilized misalignment model for the deletions observed here, unless dCTP pairs less avidly with template G than dTTP pairs with template A when template C is misaligned, which seems unlikely.

A different model for active site misalignment (38) that is also supported by evidence *in vitro* (39) proposed that a misalignment can be initiated by misinsertion of a base. In the present study for example, this could be dTTP or dCTP misinserted opposite template C (Figure 6B, pathway 3). If this misinsertion is followed by primer relocation to the +1 template base, the substrate contains a misaligned C and a correct terminal A-dTTP or G-dCTP base pair that could be extended by additional incorporation of correct dTTP and dCTP. The sequence motif observed here for loss of non-iterated C *in vivo* supports this model, based on the following observations. Just as seen in the motif for C-dTTP mismatches that result in base substitutions (Figure 4), the template base adjacent to the site of the deletion is a T (position −1 in Figure 6D), such that a misinserted T can strongly stack with an adjacent adenine. Among 22 observed non-iterated C deletions (Figure 6C), 20 are flanked by 5′-template A to which misinserted T could pair. This preference correlates well with the fact that the *rnr1-Y285A* mutation strongly promotes misinsertion of dTTP opposite C (Figure 2B, rate of 2×10^{-7}, also shown in Figure 6B pathway 3, left branch). In contrast, among 22 total non-iterated C deletions (Figure 6C), only one is flanked by 5′-template G to which misinserted C could pair and this correlates well with the fact that the *rnr1-Y285A* mutation very rarely promotes misinsertion of dCTP opposite C (Figure 2B, rate of 2×10^{-10}, Figure 6B pathway 3, right branch). Moreover, the A:G ratios at the +1 position are also high in the sequence motifs for loss of a C from 2–3 base runs and the G:A is high for loss of a T from TT runs (Figure 6C). These facts provide the *in vivo* evidence that, in addition to classical Streisinger strand slippage, indels in runs may be initiated by misinsertion during DNA replication. Collectively then, the preferred sequence motifs for single base deletions suggest that the *rnr1-Y285A* mutation promotes deletion mutagenesis by two different and non-exclusive mechanisms. One mechanism involves, but is not limited by, strand slippage in repetitive DNA. The other mechanism involves active site misalignment initiated by misinsertion followed by primer relocation.

Biological implications

The preferred sequence motifs for both substitutions and deletions in the *rnr1-Y285A msh2Δ* strain are all consistent with efficient mismatch extension that reduces proofreading. Because defects in proofreading by Pol ϵ and Pol δ have been implicated in tumor development in mice and humans (40–43), one implication of the present study is that mutations in *RNR1* or in other genes that elevate dNTP pools may also increase cancer susceptibility by reducing proofreading, even when there is no genetic defect in a replicase. Interestingly, RRM2 coding for the rate-limiting small RNR subunit in human cells is among the top 10% most overexpressed genes in 73 out of the 168 cancer analyses (44). Relevant here is also a recent study in yeast reporting that an extraordinary mutator effect of a cancer-associated Pol δ mutation depends on high dNTP levels (45). As in the current study, this effect was proposed to partly be due to enhanced mismatch extension. Additional suggestive evidence for a relationship between altered dNTP pools and cancer is the progressive increase in substitution rates through S phase promoted here by the *rnr1-Y285A* mutation (Figure 3C), which is consistent with recent evidence for an increased frequency of substitutions in late replicating regions of the genome in tumors (46). Also possibly relevant to disease etiology is the observation that excess dCTP and dTTP place coding sequences at high risk of mutations (Figure 3D), including highly deleterious single base deletions in short homonucleotide runs that are abundant in coding sequences.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank Paulina Wanrooij, Mercedes Arana and Jessica Williams for thoughtful comments on the manuscript.

FUNDING

Swedish Cancer Society (to A.C.); Knut and Alice Wallenberg Foundation (to A.C.); Swedish Research Council (to A.C.); Division of Intramural Research of the National Institutes of Health, National Institute of Environmental Health Sciences [Project Z01 ES065070 to T.A.K.]. Funding for open access charge: Swedish Research Council.
Conflict of interest statement. None declared.

REFERENCES

1. Yao,N.Y., Schroeder,J.W., Yurieva,O., Simmons,L.A. and O'Donnell,M.E. (2013) Cost of rNTP/dNTP pool imbalance at the replication fork. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 12942–12947.
2. Kunz,B.A., Kohalmi,S.E., Kunkel,T.A., Mathews,C.K., McIntosh,E.M. and Reidy,J.A. (1994) International commission for protection against environmental mutagens and carcinogens. Deoxyribonucleoside triphosphate levels: a critical factor in the maintenance of genetic stability. *Mutat. Res.*, **318**, 1–64.
3. Mathews,C.K. (2006) DNA precursor metabolism and genomic stability. *FASEB*, **20**, 1300–1314.
4. Mathews,C.K. (2014) Deoxyribonucleotides as genetic and metabolic regulators. *FASEB J.*, **28**, 3832–3840.
5. Jiricny,J. (2013) Postreplicative mismatch repair. *Cold Spring Harb. Perspect. Biol.*, **5**, 1–23.
6. Kumar,D., Abdulovic,A.L., Viberg,J., Nilsson,A.K., Kunkel,T.A. and Chabes,A. (2011) Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. *Nucleic Acids Res.*, **39**, 1360–1371.
7. Kumar,D., Viberg,J., Nilsson,A.K. and Chabes,A. (2010) Highly mutagenic and severely imbalanced dNTP pools can escape detection by the S-phase checkpoint. *Nucleic Acids Res.*, **38**, 3975–3983.
8. Buckland,R.J., Watt,D.L., Chittoor,B., Nilsson,A.K., Kunkel,T.A. and Chabes,A. (2014) Increased and imbalanced dNTP pools symmetrically promote both leading and lagging strand replication infidelity. *PLoS Genet.*, **10**, e1004846.
9. Lujan,S.A., Clausen,A.R., Clark,A.B., MacAlpine,H.K., MacAlpine,D.M., Malc,E.P., Mieczkowski,P.A., Burkholder,A.B., Fargo,D.C., Gordenin,D.A. *et al.* (2014) Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res.*, **24**, 1751–1764.
10. Jia,S., Marjavaara,L., Buckland,R., Sharma,S. and Chabes,A. (2015) Determination of deoxyribonucleoside triphosphate concentrations in yeast cells by strong anion-exchange high-performance liquid chromatography coupled with ultraviolet detection. *Methods Mol. Biol.*, **1300**, 113–121.
11. Sabouri,N., Viberg,J., Goyal,D.K., Johansson,E. and Chabes,A. (2008) Evidence for lesion bypass by yeast replicative DNA polymerases during DNA damage. *Nucleic Acids Res.*, **36**, 5660–5667.
12. Larrea,A.A., Lujan,S.A., McElhinny,S.A.N., Mieczkowski,P.A., Resnick,M.A., Gordenin,D.A. and Kunkel,T.A. (2010) Genome-wide model for the normal eukaryotic DNA replication fork. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 17674–17679.
13. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
14. Johnson,R.E., Klassen,R., Prakash,L. and Prakash,S. (2015) A major role of DNA polymerase delta in replication of both the leading and lagging DNA strands. *Mol. Cell*, **59**, 163–175.
15. Clausen,A.R., Lujan,S.A., Burkholder,A.B., Orebaugh,C.D., Williams,J.S., Clausen,M.F., Malc,E.P., Mieczkowski,P.A., Fargo,D.C., Smith,D.J. *et al.* (2015) Tracking replication enzymology

in vivo by genome-wide mapping of ribonucleotide incorporation. *Nat. Struct. Mol. Biol.*, **22**, 185–191.

16. Daigaku,Y., Keszthelyi,A., Müller,C.A., Miyabe,I., Brooks,T., Retkute,R., Hubank,M., Nieduszynski,C.A. and Carr,A.M. (2015) A global profile of replicative polymerase usage. *Nat. Struct. Mol. Biol.*, **22**, 192–198.
17. Koh,K.D., Balachander,S., Hesselberth,J.R. and Storici,F. (2015) Ribose-seq: global mapping of ribonucleotides embedded in genomic DNA. *Nat. Methods*, **12**, 251–257.
18. Reijns,M.A., Kemp,H., Ding,J., de Proce,S.M., Jackson,A.P. and Taylor,M.S. (2015) Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, **518**, 502–506.
19. Sasaki,S., Mello,C.C., Shimada,A., Nakatani,Y., Hashimoto,S., Ogawa,M., Matsushima,K., Gu,S.G., Kasahara,M., Ahsan,B. *et al.* (2009) Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science*, **323**, 401–404.
20. Tolstorukov,M.Y., Volfovsky,N., Stephens,R.M. and Park,P.J. (2011) Impact of chromatin structure on sequence variability in the human genome. *Nat. Struct. Mol. Biol.*, **18**, 510–515.
21. Phear,G., Nalbantoglu,J. and Meuth,M. (1987) Next-nucleotide effects in mutations driven by DNA precursor pool imbalances at the aprt locus of Chinese hamster ovary cells. *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 4450–4454.
22. Kunkel,T.A., Schaaper,R.M., Beckman,R.A. and Loeb,L.A. (1981) On the fidelity of DNA replication. Effect of the next nucleotide on proofreading. *J. Biol. Chem.*, **256**, 9883–9889.
23. Fersht,A.R. (1979) Fidelity of replication of phage phi X174 DNA by DNA polymerase III holoenzyme: spontaneous mutation by misincorporation. *Proc. Natl. Acad. Sci. U.S.A.*, **76**, 4946–4950.
24. Ninio,J. (1975) Kinetic amplification of enzyme discrimination. *Biochimie*, **57**, 587–595.
25. Streisinger,G., Okada,Y., Emrich,J., Newton,J., Tsugita,A., Terzaghi,E. and Inouye,M. (1966) Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. *Cold Spring Harb. Symp. Quant. Biol.*, **31**, 77–84.
26. Kunkel,T.A. (1990) Misalignment-mediated DNA synthesis errors. *Biochemistry*, **29**, 8003–8011.
27. Garcia-Diaz,M. and Kunkel,T.A. (2006) Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem. Sci.*, **31**, 206–214.
28. Lujan,S.A., Clark,A.B. and Kunkel,T.A. (2015) Differences in genome-wide repeat sequence instability conferred by proofreading and mismatch repair defects. *Nucleic Acids Res.*, **43**, 4067–4074.
29. Bebenek,K., Roberts,J.D. and Kunkel,T.A. (1992) The effects of dNTP pool imbalances on frameshift fidelity during DNA replication. *J. Biol. Chem.*, **267**, 3589–3596.
30. Kroutil,L.C., Register,K., Bebenek,K. and Kunkel,T.A. (1996) Exonucleolytic proofreading during replication of repetitive DNA. *Biochemistry*, **35**, 1046–1053.
31. Swan,M.K., Johnson,R.E., Prakash,L., Prakash,S. and Aggarwal,A.K. (2009) Structural basis of high-fidelity DNA synthesis by yeast DNA polymerase delta. *Nat. Struct. Mol. Biol.*, **16**, 979–986.
32. Hogg,M., Osterman,P., Bylund,G.O., Ganai,R.A., Lundstrom,E.B., Sauer-Eriksson,A.E. and Johansson,E. (2014) Structural basis for processive DNA synthesis by yeast DNA polymerase epsilon. *Nat. Struct. Mol. Biol.*, **21**, 49–55.
33. Ganai,R.A., Bylund,G.O. and Johansson,E. (2015) Switching between polymerase and exonuclease sites in DNA polymerase epsilon. *Nucleic Acids Res.*, **43**, 932–942.
34. Kunkel,T.A. (1986) Frameshift mutagenesis by eucaryotic DNA polymerases in vitro. *J. Biol. Chem.*, **261**, 13581–13587.
35. Efrati,E., Tocco,G., Eritja,R., Wilson,S.H. and Goodman,M.F. (1997) Abasic translesion synthesis by DNA polymerase beta violates the 'A-rule'. Novel types of nucleotide incorporation by human DNA polymerase beta at an abasic lesion in different sequence contexts. *J. Biol. Chem.*, **272**, 2559–2569.
36. Hashim,M.F., Schnetz-Boutaud,N. and Marnett,L.J. (1997) Replication of template-primers containing propanodeoxyguanosine by DNA polymerase beta. Induction of base pair substitution and frameshift mutations by template slippage and deoxynucleoside triphosphate stabilization. *J. Biol. Chem.*, **272**, 20205–20212.
37. Tippin,B., Kobayashi,S., Bertram,J.G. and Goodman,M.F. (2004) To slip or skip, visualizing frameshift mutation dynamics for error-prone DNA polymerases. *J. Biol. Chem.*, **279**, 45360–45368.
38. Kunkel,T.A. and Soni,A. (1988) Mutagenesis by transient misalignment. *J. Biol. Chem.*, **263**, 14784–14789.
39. Bebenek,K. and Kunkel,T.A. (1990) Frameshift errors initiated by nucleotide misincorporation. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 4946–4950.
40. Palles,C., Cazier,J.B., Howarth,K.M., Domingo,E., Jones,A.M., Broderick,P., Kemp,Z., Spain,S.L., Guarino,E., Salguero,I. *et al.* (2013) Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.*, **45**, 136–144.
41. Church,D.N., Briggs,S.E., Palles,C., Domingo,E., Kearsey,S.J., Grimes,J.M., Gorman,M., Martin,L., Howarth,K.M., Hodgson,S.V. *et al.* (2013) DNA polymerase epsilon and delta exonuclease domain mutations in endometrial cancer. *Hum. Mol. Genet.*, **22**, 2820–2828.
42. Goldsby,R.E., Hays,L.E., Chen,X., Olmsted,E.A., Slayton,W.B., Spangrude,G.J. and Preston,B.D. (2002) High incidence of epithelial cancers in mice deficient for DNA polymerase delta proofreading. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 15560–15565.
43. Albertson,T.M., Ogawa,M., Bugni,J.M., Hays,L.E., Chen,Y., Wang,Y., Treuting,P.M., Heddle,J.A., Goldsby,R.E. and Preston,B.D. (2009) DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 17101–17104.
44. Aye,Y., Li,M., Long,M.J. and Weiss,R.S. (2015) Ribonucleotide reductase and cancer: biological mechanisms and targeted therapies. *Oncogene*, **34**, 2011–2021.
45. Mertz,T.M., Sharma,S., Chabes,A. and Shcherbakova,P.V. (2015) Colon cancer-associated mutator DNA polymerase delta variant causes expansion of dNTP pools increasing its own infidelity. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E2467–E2476.
46. Donley,N. and Thayer,M.J. (2013) DNA replication timing, genome stability and cancer: late and/or delayed DNA replication timing is associated with increased genomic instability. *Semin. Cancer Biol.*, **23**, 80–89.