Targeted Next Generation Sequencing in *Saccharomyces cerevisiae* Reveals Complex Mutation Profiles in the Presence of Mismatch Repair and Ribonucleotide Reductase Mutations that Compromise the Fidelity of DNA Replication

Natalie A. Lamb[1], Jonathan Bard[2,,3], Michael J. Buck[1,3], Jennifer A. Surtees[1,3*]

[1] Department of Biochemistry, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo (SUNY), Buffalo, NY, 14203

[2] University at Buffalo Genomics and Bioinformatics Core, Buffalo, NY, 14203

[3] Genetics, Genomics and Bioinformatics Graduate Program, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo (SUNY), Buffalo, NY, 14203

[*] Corresponding author:

Jennifer A. Surtees

Department of Biochemistry

Jacobs School of Medicine and Biomedical Sciences

University at Buffalo, State University of New York

Rm 4215 - 955 Main Street, Buffalo, NY, 14203

Phone: (716) 829-6083; Fax, (716) 829-2725; Email: jsurtees@buffalo.edu

ABSTRACT

Determining mutation signatures is becoming the standard for understanding the etiology of human tumors and informing cancer treatment. Multiple levels of replication fidelity prevent mutagenesis that leads to carcinogenesis, including regulation of free deoxyribonucleoside triphosphate (dNTP) pools and the mismatch repair (MMR) pathway. We developed a targeted deep-sequencing approach on the *CAN1* gene in *Saccharomyces cerevisiae* to determine mutational signatures associated with both skewed and balanced increases in dNTP levels and/or defects in MMR substrate recognition. Notably, even small changes in dNTP pools led to distinct mutation profiles, as did deletions of each MMR recognition subunit *MSH2*, *MSH3* or *MSH6*. Altered dNTP levels increased and altered the mutational load, therefore *combining rnr1* and *msh* mutations allowed us to observe previously unreported specificities of Msh2-Msh3 and Msh2-Msh6. Msh2-Msh3 is uniquely able to direct repair of G/C single base deletions in GC runs, while Msh2-Msh6 specifically directs repair of substitutions at G/C dinucleotides. We also identified broader sequence contexts that influence variant profiles in different genetic backgrounds and found that there was not necessarily a simple additive relationship of mutation profiles in double mutants. Our results have implications for interpreting mutation signatures from human tumors.

INTRODUCTION

Cancer is a genetic disease caused by mutation accumulation; DNA replication is an important source of mutation. Replicative polymerases, Polε and Polδ, minimize errors via a highly selective nucleotide binding domain to prevent deoxynucleotide triphosphate (dNTP) misincorporation and an exonuclease domain to proofread and remove errors (1-4). Together, these functions lead to error rates on the order of 10^{-7} (3,4). Appropriate levels and ratios of the four dNTPs is also essential for maintaining high fidelity polymerase function. This has been demonstrated in yeast using allosteric mutations in *RNR1* that alter dNTP pools in different ways. *RNR1* encodes the large subunit of the enzyme that catalyzes the rate-limiting step in dNTP synthesis, ribonucleotide reductase (RNR). Even a modest 2-fold increase above normal levels in *rnr1D57N*, increased nucleotide misincorporation by DNA polymerases and elevated mutation rates (5,6) (**Figure 1A)**. More pronounced and skewed elevations in dNTP pools are generated in *rnr1Y285F* and *rnr1Y285A*, which increase dCTP and dTTP 3-fold and 20-fold, respectively, without inducing a cell cycle defect (7). These increases in dNTP pool levels further compromise replication fidelity (7-10).

Altered dNTP levels that have been generated by mutations in other biosynthetic pathways, such as yeast *SMH2* (11) or mammalian *SAMHD1* (12), also result in increased mutation rates in both yeast and mammals (5,6,8,13,14). Elevated dNTP pools are thought to: 1) increase the frequency of misinsertion and misalignment events and 2) shift replicative DNA polymerases toward synthesis at the expense of proofreading, promoting extension beyond a mispair (7,15,16).

Nucleotide misincorporation and DNA slippage events in DNA replication result in DNA structures that are substrates for the mismatch repair (MMR) system, which functions as a spell-check, recognizing and directing the repair of errors in replication (17), thereby increasing the fidelity of replication by an additional 10-1,000 fold (2,3). Once recognized by a MutS homolog (Msh) complex, errors are targeted for excision. In most eukaryotes, two heterodimeric complexes bind errors at the

replication fork: Msh2-Msh6 and Msh2-Msh3, which recognize a broad spectrum of mismatches and insertion/deletion loops (IDLs) with different specificities. The current model of post-replicative repair posits that Msh2-Msh3 recognizes, binds and directs repair of IDLs up to 17 nucleotides long, (18,19)while Msh2-Msh6 targets mispairs and IDLs of 1-2 nucleotides. Msh2-Msh3 also has affinity for some mispairs, especially C-C, A-A and (possibly) G-G mispairs (20,21). There was a synergistic effect on mutation rate in *rnr1D57N msh2Δ* and *rnr1D57N msh6Δ* (6) and in *rnr1Y285A msh2Δ* strains, (8,10) indicating that mismatch repair normally corrects the errors generated by DNA polymerases when dNTP pools are elevated (**Figure 1B)**.

Overexpression or mutation of *RNR1*, encoding the large subunit of RNR, elevates free dNTP levels and increases mutagenesis in yeast and mammalian cells (22 2003,23-25). Overexpression of *Rrm2*, encoding the small subunit of RNR, induced lung neoplasms in a mouse model (6). Mutations in SAMHD1, a dNTP triphosphohydrolase that modulates dNTP pools in mammalian systems, lead to oncogenesis likely driven by increases in dNTP pool levels (13,14). And nucleobase analogs are frequently used as chemotherapeutics, commonly targeting RNR (26,27). Similarly, MMR defects promote tumorigenesis. MMR is deficient in ~25% of sporadic cancers caused by increased rate of mutagenesis (28). Deficiencies in MMR genes *MSH2*, *MSH6*, *PMS2* and *MLH1* cause hereditary Lynch syndrome, which leads to a strong predisposition to cancers of the gastrointestinal (GI) tract, endometrial cancer and lymphomas, and are defined by microsatellite instability (MSI) (29,30). MMR mutations are also implicated in breast and ovarian cancer (31,32). While Msh2-Msh3 has not been directly linked to Lynch syndrome, mutations in *MSH3* lead to cancer predisposition within and outside the GI tract(33-38) as well as chemoresistant tumors.(39-41) Loss of *MSH3* function leads to Elevated Microsatellite Alterations of Selected Tetranucleotide repeats (EMAST), which is distinct from MSI and has been associated with a number of different cancers, including up to ~60% of colorectal cancers.(42) Therefore, while *MSH3* also plays a role in tumorigenesis, its role is distinct from that of Msh2-Msh6.

In mammalian cells, characterizing mutation signatures is a powerful tool to identify potential pathways that are implicated in tumorigenesis. However, this analysis rarely considers the likely presence of altered dNTP pools as a mitigating factor. Cancer cells have increased proliferative nature and thus elevated dNTP pools may be necessary to support accelerated DNA replication (43,44), which may, in turn, increase mutagenesis, promote molecular evolution and provide a selective advantage to the tumor. Notably, combining defects in both MMR and free dNTP pool synthesis leads to synergistic increases in mutation rates and uncharacterized mutagenic potential (6). A major goal of this study was to determine how altered dNTP pools might influence the mutation signature or "mutation profile" with and without functional MMR, with an eye to illuminating the interpretation of mutation signatures observed in tumors, i.e. to what extent do altered dNTP pools influence the mutation profile in MMR mutants.

For this analysis, we paired selection for mutations at *CAN1* with next-generation sequencing (NGS), allowing higher throughput and increased sequencing depth compared with previous studies. We built mutation profiles from first principles, characterizing single mutants and then evaluating combinatorial effects on the mutation profiles of double mutants. In this study we: 1) characterized mutagenesis/mutation load as a result of altered dNTP pools, with and without MMR (*msh2Δ*), and in the presence of only Msh2-Msh6 (*msh3Δ*) or Msh2-Msh3 (*msh6Δ*), 2) determined novel discrete substrates for Msh2-Msh6 and Msh2-Msh3 error recognition and 3) assessed mutation profiles and the combinatorial effect of elevated dNTPs and defined defects in MMR. We observed novel, distinct profiles in *rnr1* and *msh* backgrounds, by analyzing mutation spectra and sequence context effects. By combining deletions of MMR factors with altered dNTP pools, we increased and altered mutation rates which allowed us to observe rare variant types and positions that are typically repaired by MMR and to determine Msh2-Msh6 versus Msh2-Msh3 specificity for recognizing variant classes in distinct sequence context.

MATERIALS AND METHODS

Strains and plasmids

All strains in this study were derived from the W303 *RAD5+* background (**Table S1**). Strains containing *rnr1Y28F/A* in the *pGAL-RNR1* background and integration plasmids to re-create these strains (7) were kindly provided by Andrei Chabes. To integrate *rnr1* alleles at the endogenous *RNR1* locus in the absence of *pGAL-RNR1*, we created a new set of *rnr1* integration plasmids. First, we amplified pRS416 (45) with SO261 and SO262, which both contain an *AatII* recognition site (**Table S2)**. The resulting PCR product, which consists of pRS416 without the *ARS/CEN* region, was digested with *AatII* and ligated to generate the yeast integration plasmid, pEM3, which carries a counter-selectable *URA3* marker (**Table S3**). *RNR1* was amplified from JSY13 with SO263 (**Table S2**; encodes a *Xho*I site) and SO265 (**Table S2**; encodes a *Spe*I site) to amplify a product that extends ~ 1 kb upstream of the endogenous *RNR1* start site to ~3kb downstream of the *RNR1* start site, to capture the full *RNR1* gene. This fragment was digested with *Xho*I and *Spe*I and ligated into pEM3 to generate the *RNR1* integration plasmid pNL1 (**Table S3**). To generate a truncated version of *RNR1* in the same plasmid, *RNR1* was amplified from JSY13 with SO263 (encodes a *Xho*I site) and SO264 (**Table S2**; encodes a *Spe*I site) to amplify a product that extends ~ 1 kb upstream of the endogenous *RNR1* start site to ~2kb into the *RNR1* open reading frame. The PCR product was digested with *Xho*I and *Spe*I and ligated into pEM3 to generate pNL2 (**Table S3**).

We generated *rnr1D57N*, *rnr1Y285F* and *rnr1Y285A* versions of pNL1 and pNL2 using the Q5 site-directed mutagenesis kit (New England Biolabs). Oligonucleotides used for the mutagenesis are described in **Table S2**. Plasmids were sequenced to confirm the presence of the desired mutations and absence of secondary mutations. To integrate each *rnr1* allele into the endogenous *RNR1* locus, both the full-length and truncated integration plasmid for that allele were linearized with Sph1 and used to co-transform JSY13 using the standard LiAC approach (46). Transformants were selected on synthetic

complete medium lacking uracil (SC-URA). Single colonies were patched onto YPD to allow for homologous recombination and transplacement of one copy of *rnr1*. After 2 days growth at 30°C patches were struck out to generate single colonies on plates containing 5-FOA, selecting for a loss of *URA3*. Plasmid integration was confirmed by PCR using SO266 (**Table S2**; specific to the integration plasmid) and SO230 (**Table S2**; specific to the genomic locus). Sanger sequencing was used to confirm retention of a mutant *rnr1* allele and a loss of *RNR1*. *rnr1* was amplified in two fragments for sequencing, using the SO32/SO21 and SO25/SO22 primer pairs (**Table S2**).

Mismatch repair factors were deleted by amplifying *msh2Δ::KanMX*, *msh6Δ::KanMX*, and *msh3Δ::KanMX* chromosomal fragments from deletion collection strains utilizing primers A and D specific to each locus (**Table S2**). PCR products from these strains were used for transformation to replace the endogenous *MSH* gene with the *KanMX* cassette, which confers resistance to the drug, G418. Transformants were selected on YPD plates containing G418 and deletions were confirmed by PCR.

Pooling Canavanine resistant colonies

Strains were patched on SC-ARG plates and grown at 30°C for 3 days, ~30-35 generations. A quarter of the patch was used to inoculate a 25 mL SC-ARG liquid culture. A patch was used to inoculate instead of a single colony to reduce the potential effects of dominant mutations in liquid culture. Cells were grown for approximately 3-4 generations and then plated on SC-ARG + canavanine plates, to select for approximately 1,000-2,000 canavanine resistant (Can^R) colonies. It requires another ~30 generations to generate colonies, for a total of ~65-70 generations of accumulated mutations. Canavanine resistant colonies were selected in at least four independent experiments per genotype. For each genotype, at least two independent isolates were used (**Table S4**).

Colonies were counted, and collected in TE (10 mM Tris-HCl, pH 7.5; 1 mM EDTA) using a sterile glass spreader. Colonies from multiple plates were pooled and resuspended to a final volume of

approximately 10-12 mL. One mL of the colony suspension was used to extract genomic DNA (gDNA). Briefly, cells were lysed by vortexing in 200 µl 1:1 phenol: chloroform, 200 µl chromosome preparation buffer (10 mM Tris,-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 1% SDS, 2% Triton X-100) and 0.3 grams acid-washed glass beads (Sigma; 425-600 microns). 200 µl TE pH 8.0 was added and reactions were centrifuged for 5 minutes at 16,000 x g. The resulting supernatant was collected. Three additional phenol/chloroform extractions were performed to increase purity. gDNA was precipitated by addition of ammonium acetate to a final concentration of 100 mM followed by 2 volumes of 95% ethanol. The gDNA was collected by centrifugation at 16,000 x g for 10 minutes, followed by a wash with 70% ethanol. The gDNA pellets were resuspended in 50 µl nuclease free water with RNase A (final concentration of 50 µg/ml) and incubated at room temperature for at least 1 hour after which gDNA was stored at -20°C.

Pooling unselected samples

In addition to canavanine resistant colonies, we also pooled ~1000-2000 unselected colonies from *RNR1*, *rnr1D57N*, *rnr1Y285F* and *rnr1Y285A* backgrounds as permissive controls. These colonies were grown as described above, except that in the final step, cells were grown on SC-ARG in the absence of canavanine. Colonies were pooled and genomic DNA extracted as described above.

Library Preparation and Sequencing

CAN1/can1 was amplified from gDNA in 6 overlapping, 349-350 base pair fragments using primers listed in **Table S2**. KAPA HiFi ReadyMix (Roche) was used to amplify these fragments in 25 µL reactions for each *can1* region for each of the 150 samples (total of 906 reactions). Two-µL of DNA was added to each reaction, in the range of 50-500 ng. Two-µL of each PCR reaction was electrophoresed on a 1% agarose gel to confirm amplification. For each sample, 20 µL of PCR product

from each of the 6 regions were pooled in a 96-well plate and purified using the Zymo ZR-96 DNA Clean-up Kit. A total of 150 pooled sample sets (**Table S4**) were generated for paired-end sequencing (2x300). Technical replicates were also included in the 150 pooled samples described above. PCR products from the same genomic preparation of pooled samples were independently generated and sequenced.

Library barcoding and QC

Amplicons containing Nextera barcoding adapters were minimally PCR amplified (8 cycles) for attachment of Illumina Nextera XT index primers set A (Illumina). After PCR, excess adapters were removed using Ampure XP beads (Beckman Coulter) and samples were eluted into EB buffer. Barcoded amplicons were checked for quality using an Advanced Analytical Fragment Analyzer and Qubit Fluorescence (Invitrogen). Amplicons were pooled to 10 nM in EB buffer and the final concentration was determined using the Illumina Universal qPCR Amplification kit from Kapa Biosystems. All pooled samples were diluted to 4 nM, denatured using NaOH and loaded onto an Illumina MiSeq sequencing platform (PE300, V3) with 20% PhiX control. The sequencing was performed in two separate runs to increase coverage and as a check for reproducibility.

Upstream sequencing analysis

Reads were trimmed using a variable length trimmer (CutAdapt version 1.14) specifying a quality score of Q30. Trimmed reads were further processed using CLC Genomics Workbench Version 11. Paired-end reads were merged, primer locations were trimmed, and processed reads were aligned to the SacCer3 reference genome. Variants were then called using the CLC low frequency variant caller with required significance of 0.01%. Variant files were exported from CLC as VCF files and

downstream analysis was performed in RStudio (version 1.2.1335), paired with custom python scripting (**Fig. S1**).

Permissive variant analysis and filtering

We applied a filter based on our permissive samples to remove background mutations from Can^R samples. The permissive filter removes any variant that occurs below the average permissive sample variant frequency of 0.108% (**Fig. S3C)**. This represents approximately 1 variant being called in every 1,000 reads sequenced. We customized the permissive filter because we also observed position-specific variants at a frequency higher than the overall average permissive variant frequency of 0.108%. Thus, these systematically higher frequency variants were removed from the selected samples if they occurred at a lower frequency than the highest frequency variant that occurred at the same position in the permissive data set. **Fig. S3E-G** shows an example of the permissive filter applied, with a blanket filter of 0.108% across *CAN1* except for the four positions where variants occurred at frequencies higher than the permissive average, up to approximately 1%. This is a conservative filter, and undoubtedly low frequency events that are biologically relevant may be removed, but are below the sensitivity of this assay.

Hierarchical Cluster Analysis

Hierarchical Cluster Analysis was performed on data that was grouped by biological replicates and the number of times the different classes of variants occurred within a genotype. Data was combined on the average of unique counts of variants from all the biological replicates within a genotype. The different classes of variants include 6 classes of single nucleotide variants (SNVs), single base A/T or G/C insertions and deletions, complex insertions and deletions, as well as mononucleotide variants (MNVs) and replacements (Replac.). MNVs and replacements are complex errors identified in CLC.

MNVs are dinucleotide SNVs, where two neighboring nucleotides are both mutated, ex: CC> AT. Replacements are complex insertions or deletions, where the deleted or replaced base is a variant. Two examples include AAC > G and C > AT. Both MNVs and replacements are extremely low frequency events and rarely occur in our data set; neither had a significant impact on clustering.

The data was condensed based on genotype by combining biological replicates and adding the total number of times a variant was seen in the total number of replicates. This analysis does not take frequency into account and instead totals how many unique types of variants occur in a sample. If the same variant occurred in multiple biological replicates within a genotype it was counted as such. Spearman correlation coefficients were calculated across all samples. Spearman correlation was chosen because we did not consistently observe linear relationships between variant frequencies in our data set, although Pearson correlation yielded nearly identical results. Correlation analysis was plotted in RStudio using the ggplots and ggplot2 libraries.

Principal Components Analysis (PCA)

PCA was performed on the number of times a unique variant was observed within a genotype. PCA was plotted using the "factoextra" package in RStudio, with a singular value decomposition (SVD) approach using the prcomp() function.

Determining SNV in trinucleotide context

In our filtered data set there are a total of 13,031 SNVs, with 907 unique SNV changes. We cannot determine which strand incurred a mutation and therefore all SNVs were categorized as C or T changes for this analysis. The trinucleotide context surrounding the SNV was determined by taking a 3 bp window surrounding the reference position within *CAN1*. There are a total of 96 different possible SNV changes in unique sequence context (47). For a given sample, the number of SNVs in each of these

96 contexts were totaled. This analysis does not take frequency into account and only scores the presence or absence of a particular type of SNV. The frequency plotted is calculated based on unique counts of an SNV that occurred, divided by the total number of times that SNV in tri-context occurred. The data was further condensed by taking the average of each of the 96 different contexts for all the biological replicates in one genotype.

The number of trinucleotide sequence contexts in *CAN1* was calculated using a sliding window approach through custom python scripts. For each of the 96 different SNV changes in triplet context, the average number of SNVs in a genotype was divided by the number of times the triplet sequence context occurs in *CAN1*. This dataset was imported into R-studio and plotted via the barplot() function. The average number of times an SNV is seen in a genotype out of the total number of times the sequence context occurs across *CAN1* is plotted (**Fig. 5)**.

Cluster Analysis to look at correlation between genotypes

To identify variants specific to a particular genotype and to eliminate frequency bias from variants that occurred early on in the growth of a particular sample, we condensed unique variants based on the number of biological replicates sequenced for that genotype. If a particular variant occurred in 4 out of 4 biological replicates it was represented as 1, if it occurred in 3 out of 6 replicates it was represented as .5. This gives an unbiased approach to score for the probability that a particular variant was present in a particular genotype, without considering the absolute frequency at which that variant occurred. These data were clustered on rows (or unique variants), after applying a row sum cutoff of greater than 2. This cutoff eliminates low frequency variants which are less likely to be driving the differences in mutation spectra that we observe. By clustering the data only on unique variants, it allows us to see groups of different *types* of variants in specific sequence contexts that are potentially diagnostic for a particular genotype.

CC Dinucleotide cluster analysis

A python script using a sliding window approach was used to identify all reference positions containing CC dinucleotides within *CAN1*. Our dataset was then subset to include only variants that occurred in these dinucleotide CC positions. Of the 138 CC dinucleotide contexts across *CAN1*, 110 (~80%) were mutated, compared to 857/1,711 base pairs or ~50% of the base pairs in *CAN1*. Unique CC run variants were clustered based on the number of times that variant occurred in a given genotype, while accounting for (normalizing by) the number of biological replicates sequenced for each genotype, as described above. Heatmaps were plotted using the pheatmap library in RStudio and motif enrichment was performed using Berkely web logos (48).

RESULTS

Rates of canavanine resistance in different genetic backgrounds

In this study, we were interested in the combined effects of altered dNTP pools and deficient MMR on mutation profiles. We genetically altered dNTP pools, using previously characterized mutations in *RNR1*, the large subunit of RNR. The *rnr1D57N* mutation, in the activity site, leads to a balanced two-fold increase in each of the four dNTPs (5). Two mutations in the RNR specificity site, *rnr1Y285F* and *rnr1Y285A*, lead to three- and twenty-fold skewed increases in dCTP and dTTP, respectively (8). Both *rnr1Y285F* and *rnr1Y285A* also encode a wild-type copy of *RNR1* under control of the inducible *pGAL* promoter (*pGAL-RNR1*) (7). We also constructed strains that encoded only *rnr1Y285F* and *rnr1Y285A* for use in this study (**Table S1**). In parallel, we deleted each of the MSH complex subunits individually. By deleting *MSH2* (*msh2Δ*), we eliminated Msh2-Msh3 and Msh2-Msh6 activity, effectively eliminating MMR. To determine the relative roles of Msh2-Msh3 and Msh2-Msh6, we tested *msh3Δ* (only Msh2-Msh6 present) and *msh6Δ* (only Msh2-Msh3 present) (**Fig. 1B**).

We also constructed and characterized all possible double mutations with *rnr1* alleles and *msh* deletions (**Table S1**).

The relative mutation rates for *msh* deletions were consistent with previous observations (6,49-51), with *msh2Δ* exhibiting the highest mutation rate, *msh6Δ* exhibiting a lower but still significantly increased mutation rate and *msh3Δ* not exhibiting any mutator effect in this assay (**Table 1**). Similarly, the *rnr1* mutation rates were consistent with previous work (5-8); *rnr1D57N* and *rnr1Y285F* exhibited relatively low mutation rates (2-4 fold increase), while *rnr1Y285A* exhibited a larger ~10-fold increase in mutation rate, (7) similar to *msh2Δ*. The *pGAL-RNR1* construct had little effect on the mutation rates in either the single or double mutants. When *rnr1D57N* and *rnr1Y285F* were combined with *msh* deletions, the mutation rates exhibited the same hierarchy (*msh2Δ* > *msh6Δ* > *msh3Δ*). Notably, the *msh* hierarchy of mutation rate was reversed in the *rnr1Y285A* background; *rnr1Y285A msh3Δ* exhibited a strong, synergistic mutator phenotype, indicating that the mutations generated in the presence of *rnr1Y285A* are substrates for Msh2-Msh3-mediated MMR. This observation is also consistent with very high mutations rates in *msh2Δ* (and possibly *msh6Δ*) compromising cell viability (see discussion).

Selection at *CAN1* paired with next generation sequencing (NGS) to define mutation profiles

In *Saccharomyces cerevisiae,* mutation spectra have been determined via two general strategies. The first requires selection of mutants, most often at the *CAN1* locus (8,10), which encodes an arginine permease that also imports the toxic arginine analog, L-canavanine, leading to cell death. (52,53) Inactivating mutations in *CAN1* block canavanine uptake and toxicity, allowing selection for mutations in *CAN1*. The *can1* gene from individual colonies can be subjected to Sanger sequencing, revealing mutation spectra in different genetic backgrounds. Importantly, *CAN1* has been found to be representative of mutations occurring genome wide when both Sanger sequencing and whole genome sequencing (WGS) approaches were compared (8) (9,10). However, this approach is relatively low

throughput (5,6,8) and will miss low frequency variants that might exist within the colony. The second approach has been mutation accumulation (MA) experiments followed by WGS, which avoids potential bias associated with focusing on a single genetic locus (i.e. *CAN1*). However, the depth of sequencing is limited because the entire genome is sequenced, complicating statistically significant comparisons among genotypes (13,54,55). Increased depth requires increased numbers of MA lines and many passages are required to accumulate sufficient mutations.

We developed a combined *CAN1* selection-based next generation sequencing (NGS) approach to characterize genotype-specific mutation profiles. Our approach was to select 1,000-2,000 canavanine resistant colonies, each of which sustained at least one mutation within *CAN1*, pool them, and extract genomic DNA for amplification of *can1*. Amplicons from each pooled sample were barcoded and combined for NGS (see methods). NGS allowed deep sequencing of pooled samples at each position along *can1*, providing: 1) large sample sizes for each pooled group, 2) sequencing depth sufficient to uncover low frequency variants and 3) novel insight into mutation profiles and positional effects on mutations.

Correlation analyses revealed systematic, genotype-specific mutation profiles

To determine the mutation spectra in different genetic backgrounds, we sequenced pools of up to 2,000 Can^R colonies (**Fig. S1A**), i.e. "selected samples". Because mutations are stochastic and we were interested in identifying trends in mutation profiles, we sequenced at least four independent samples for each genotype, using at least two independent isolates. Genomic DNA was extracted from each pooled sample and the *can1* gene was PCR amplified in 6 regions and sequenced using 2x300 paired-end sequencing. Upstream analysis was performed in CLC Genomics Workbench 11, determining variants utilizing a low frequency variant caller, and downstream analysis was performed in R Studio (**Fig. S1B**). NGS of a single gene provided enough sequencing depth to determine mutation spectra for 150 unique

samples, sequencing *CAN1* at an average depth of approximately 16,000x coverage per sample (**Table S4**). By definition, the pooled colonies selected in the presence of canavanine should each harbor at least one mutation in *can1*. Consistent with this prediction, the total variant frequency for all selected samples ranged from ~80% to >100%, with an average of 99.35% (**Fig. S2A**). Frequencies above 100% likely indicate more than 1 mutation within *can1*, on average, and were observed in strains with higher mutation rates (i.e. *rnr1Y285A* genotypes) (**Table 1**).

In parallel, we sequenced pooled samples (1000 colonies each) grown in the *absence* of canavanine selection, which we dubbed "permissive" samples. These permissive controls allowed us to develop a threshold for background mutations at each position along *CAN1* that are due to low frequency mutations and sequencing and/or PCR polymerase bias.

All sequence data were scored for: 1) single nucleotide variants (SNVs), 2) single base (A/T or G/C) insertions or deletions, 3) complex (>1 base) insertions or deletions, 4) dinucleotide SNVs at adjacent nucleotides, i.e. mononucleotide variants (MNVs), and 5) complex replacements (see Materials and Methods). MNVs and replacements were extremely rare in our datasets. The total variant frequency for each sample was calculated by taking the sum of the number of variants and dividing by the total number of sequence reads for that sample. The variant frequencies for all biological replicates within a genotype were combined (see below). The average variant frequency for four different genetic backgrounds (wildtype, *rnr1D57N, rnr1Y285F* and *rnr1Y285A*) under permissive conditions was 8.7% compared to >90% in the samples selected in the presence of canavanine (**Fig. S2A**). While the lower variant frequencies in permissive controls were expected, it was higher than predicted given the known mutation rates of replicative polymerases, (56-58) indicating that background mutations were being introduced as part of the PCR/sequencing pipeline. Consistent with this prediction, the variant distribution in permissive controls from wildtype, *rnr1D57N, rnr1Y285F* and *rnr1Y285A* were virtually identical, despite differences in mutation rates and mutation spectra (5,6) (**Table 1, Fig. S3A & S3B)**.

Moreover, the single nucleotide variant (SNV) spectra from permissive samples (**Fig. S3A**) closely resembled the SNV spectrum observed for KAPA HiFi polymerase (59,60), with a bias toward GC>AT changes.

To correct for these effects, we developed a permissive variant filter as part of the analytical pipeline. At each position in *can1,* we determined the average frequency of variants observed in permissive samples, which was 0.108% compared to 0.407% in the selected samples (**Fig. S3C & S3D).** The permissive variant filter removed any variant that occurred below the average permissive variant frequency of 0.108%, which represents ~1 variant called for every 1,000 reads sequenced. Variants that occurred in permissive samples at a frequency above this average were position specific. The permissive filter set the threshold for these variants at the highest frequency for a given position-specific variant in a permissive sample. This conservative approach ensured background variants were removed from our analysis of mutation profiles. The absolute variant frequency of selected samples decreased on average by 7.5% after applying the permissive filter (**Fig. S2B**), consistent with the average absolute variant frequency in permissive samples being 8.7%. **Fig. S3E** illustrates the application of this filter, where the filter cutoff is 0.108%, except for positions 31658 and 31659, where variants occurred at position-specific higher average frequencies in the permissive samples. This reduced the noise in the selected data (**Fig. S3F, G**) with only minor changes to variant frequency (**Fig. S2B**) and to mutation spectra overall (**Fig. S4**).

After applying the permissive filter described above, we analyzed the mutation profiles observed in individual "selected" replicates for each genotype to compare mutation profiles from sample to sample within a genotype. We characterized mutation events in two ways (**Table 2**): 1) the presence of a particular variant above background, i.e., the number of different C>A changes along *can1* ("unique counts") (Counts in **Table 2**) and 2) the frequency at which each of these unique variants occurred, i.e., the combined frequency of all C>A observed changes along *can1* ("sum of frequencies") (Freq. in **Table**

2). These analyses allowed us to determine whether different *types* of mutations occurred in a genotype-dependent manner, independent of frequency, and whether variant frequencies were altered in a significant way (Counts/Freq. in **Table 2**). For example, a decreased number for "unique counts" combined with unchanged or increased "sum of frequencies" would indicate that variant type is more localized, possibly indicating a mutational hotspot (e.g. *rnr1Y285A msh3Δ* in **Table 2**). Independent biological replicates with the same genotype closely resembled each other (**Fig. S5**); variant frequencies were more variable than unique counts among biological replicates of the same genotype (**Fig. S5**), consistent with the stochastic nature of mutation accumulation during the growth of a culture. Distinct trends in mutagenic events were observed in multiple, *independent*, experiments and were consistent with systematic, genotype-specific changes in mutation profiles.

We used Spearman rank correlation and cluster analysis to determine the quantitative relationship between mutation profile and genotype. All unique variants for all genotypes were assessed in parallel, based on both the presence and frequency of unique variants, as described above. In general, biological replicates of the same genotype clustered because their mutational profiles were highly correlated (**Fig. S6**). Biological replicates from strains with lower mutation rates (e.g. wild-type, *msh3Δ*, *rnr1D57N*) were less well correlated than those from strains with higher mutation rates (e.g., *msh2Δ*, *msh6Δ*, *rnr1Y285A*), indicating that mutation events occurred more systematically (less stochastically) in these genetic backgrounds, consistent with specific mutation profiles. For example, the correlation values comparing "wildtype biological replicate 1" to all other samples ranged from 0.035 to 0.310 (**Table S5: column "DJ"**). The average correlation value of all 7 wildtype biological replicates with all other samples was 0.192, while the average correlation values within the 7 wildtype biological replicates was r_s = 0.191. In contrast, the *rnr1Y285A msh3Δ* biological replicates were much more highly correlated to one another (r_s = 0.527) than they were with all other samples (r_s = 0.264). Because of the observed positive correlations for biological replicates within a genotype from each genetic background,

we combined variants from all biological replicates within a genotype for the remainder of our analyses. Importantly, *rnr1Y285F* and *rnr1Y285F pGAL-RNR1* biological replicates were well-correlated, as were *rnr1Y285A* and *rnr1Y285A pGAL-RNR1*(**Fig. S6, box II, Fig. S7**).

The distribution and cumulative frequency of variant types differed significantly between genotypes, indicating unique mutations profiles (**Table 2**). For instance, *msh6Δ* exhibited the highest proportion of unique events contributing to the mutation spectrum (Counts/Freq. = 1.62; **Table 2**). In contrast, *rnr1Y285F msh3Δ* and *rnr1Y285A msh3Δ* exhibited the lowest proportion of unique variants; the mutation spectra were instead dominated by G/C single base deletions, which occur at high frequencies (**Table 2**). Hierarchical cluster analysis using Spearman rank correlations based on the profile of unique variants between genotypes was consistent with genotype-specific mutation profiles (**Fig. 2A, Fig. S6)**, as was principal component analysis (PCA) based on unique variants (**Fig. 2B**). Overall, these results indicated that it is possible to distinguish between genotypes based on unique variant profiles observed from *can1* deep sequencing.

Alterations in dNTP pools change the mutation profiles

We used "unique counts" and "sum of frequencies" to generate mutation profiles for each genotype. We compared mutation profiles in the presence of the *rnr1* alleles with the wildtype profile, derived from two independent isolates of wildtype W303 *RAD5+* (**Table 2**, **Fig. 3A**). The relative frequency of SNVs, insertions and deletions, normalized by the number of sequence reads, varied significantly by *rnr1* allele (**Table 2**, **Fig. 3A**). In the *rnr1D57N* background, deletions increased substantially, while insertions remained unchanged. The SNV profile of *rnr1D57N* was very similar to wildtype, even though the overall SNV frequency was reduced (**Table 2**, **Fig. 3B, C**), consistent with more stochastic events. In contrast, SNVs dominated the *rnr1Y285F* profile. Despite the low mutation rate of *rnr1Y285F*, the variant types observed were almost identical to those of *rnr1Y285A*, which

exhibited much higher mutation rates (**Table 1**). *rnr1Y285A* was more similar to wildtype, but overall deletions increased with a concomitant decrease in SNVs and insertions, and the *types* of variants observed were distinct (**Table 2**, **Figs. 3B, C**). In wildtype cells, the SNVs were distributed across all 6 possible variants, while in *rnr1Y285A* the profile was dominated by CG>AT changes, as was the SNV profile of *rnr1Y285F*. While the *rnr1Y285A* and *rnr1Y285A-pGAL RNR1* strains resulted in almost indistinguishable mutation profiles, *rnr1Y285F* and *rnr1Y285F-pGAL RNR1* showed more variation compared to one another (**Table 2**, **Fig. S7A, B**), likely because there are more stochastic mutations when mutation rates are lower. All three *rnr1* backgrounds exhibited a significant increase in G/C -1 bp deletions compared to wild-type, although the absolute frequency varied (**Table 2**, **Fig. 3D, E**). The frequency of G/C -1 bp deletions in *rnr1* backgrounds was significantly different than in other genotypes (Kruskal-Wallis chi-squared = 88.367, df = 22, p-value =6.489e-10). Overall, very few insertions were observed, but we noted that the G/C +1 bp insertions were extremely rare in *rnr1Y285A* cells compared to other genotypes (**Table 2**, **Fig. 3F, G**). These observations for *rnr1D57N*, *rnr1Y285F* and *rnr1Y285A* are in line with previous observations (6-10), although the depth of sequencing in the current study revealed more distinct and detailed mutation profiles in these *rnr1* backgrounds, with clear shifts in the types and frequency of mutations that accumulate in the presence of balanced versus skewed elevations in dNTP levels.

Deletion of *MSH* genes alters mutation profiles in distinct ways

We deleted *MSH2*, *MSH6* and *MSH3* genes, which encode subunits of the MMR recognition complexes (**Figure 1**, **Table S1**). Deletion of *MSH2* effectively eliminated all MMR activity, retaining all errors made by replicative polymerases. We also removed only Msh2-Msh6 (*msh6Δ*) or Msh2-Msh3 (*msh3Δ*) to learn more about substrate specificity and efficacy of each complex. In general, *msh2Δ*, *msh3Δ* and *msh6Δ* mutation profiles were consistent with previous studies performed in yeast

(6,19,20,49,51). Deleting *MSH2* resulted in increased deletions and insertions compared to wildtype, particularly A/T single base deletions (**Table 2**, **Fig. 3D, E)**. In *msh3Δ* cells there was an increase in G/C -1 bp deletions and complex deletions and insertions (> 1 base pair), consistent with Msh2-3 preference for binding larger IDLs (61-63). Nonetheless, over 25% of the mutations that accumulated in *msh3Δ* were SNVs (**Table 2**, **Fig. 3A)**, consistent with Msh2-Msh3 also playing a role in correcting misincorporation events (20,51). In contrast, the *msh6Δ* spectrum was dominated by SNVs, which represented >80% of the mutations (**Table 2**, **Fig. 3A)**. Although *msh6Δ* resulted in the highest accumulation of SNVs, the *types* of SNVs generated were similar for *msh2Δ* and *msh6Δ* (**Table 2**, **Fig. 3B, C).** The *msh3Δ* SNV spectrum consisted of higher levels of CG>GC and TA>AT errors than *msh2Δ* and *msh6Δ*, despite the fact that these errors should be efficiently repaired by Msh2-Msh6 (64,65); TA>AT errors in *msh3Δ* were also higher than observed in wildtype.

***rnr1* alleles combined with MMR deletion result in unique mutation profiles**

The *rnr1* alleles each exhibited a unique mutation profile, but these strains retained MMR, which modified that mutation profile based on MMR specificity and efficacy. We deleted *MSH2* in combination with the *rnr1* alleles, eliminating all MMR activity, to determine how altered dNTP pools affected replication fidelity at the level of the polymerases. When *msh2Δ* was combined with each *rnr1* allele, distinct mutation profiles were observed (**Table 2**, **Fig. 3A**). Although all mutation classes were observed in *rnr1Y285F* and *rnr1Y285A* alone, the profiles were biased toward G/C single base deletions, CG>TA and CG>AT SNVs, and A/T single base deletions. Notably, the profiles in the double mutants (*rnr1Y285F msh2Δ* and *rnr1Y285A msh2Δ*) were almost completely dominated by these four types of errors (**Table 2**, **Fig. 3B, C, D, E)**. This indicated that polymerases generated these errors at extremely high frequencies, under even only modestly skewed elevations in dCTP and dTTP (*rnr1Y285F*) and began to saturate MMR activity and/or are poorly repaired by MMR.

Deleting *MSH6* in combination with *rnr1* alleles, so that only Msh2-Msh3-mediated MMR was present, also resulted in unique shifts in mutagenesis across *can1* (**Table 2**, **Fig. 3A)**. The effect was most dramatic in *rnr1Y285F msh6Δ* with a profile that was almost completely dominated by CG>TA and CG>AT SNV errors (**Table 2**, **Fig. 3B, C)**, indicating that Msh2-Msh3 is not efficient in correcting these mismatches. This effect was less apparent in *rnr1Y285A msh6Δ*, where the majority of the variant classes were still observed. However, in both the *rnr1Y285F msh6Δ* and *rnr1Y285A msh6Δ* double mutants there was almost a complete loss of the G/C -1 bp deletions that were observed at an increased frequency in *rnr1Y285F* and *rnr1Y285A* backgrounds (**Table 2**, **Fig. 3D, E**), consistent with efficient repair of these slippage events by Msh2-Msh3-mediated MMR. Strikingly, the opposite was observed in *msh3Δ* genotypes. In both *rnr1Y285F msh3Δ* and *rnr1Y285A msh3Δ*, there was a significant increase in G/C -1 bp deletions compared to either single mutant. This indicated that Msh2-Msh6 was effectively unable to correct this type of error. Notably the mutation spectra of the *rnr1Y285F* and *rnr1Y285A* double mutants compared to their *pGAL-RNR1* counterparts were largely consistent (**Table 2**, **Fig. S7**). The most striking difference is an increased frequency in G/C -1 bp deletions in *rnr1Y285A-pGAL msh6Δ*, which is the opposite trend observed in the other *msh6Δ* double mutants. Unlike *rnr1Y285F msh3Δ* or *rnr1Y285A msh3Δ*, in which multiple independent locations across *CAN1* accumulated G/C -1 bp deletions at elevated frequencies, the mutation profile of all *rnr1Y285A-pGAL msh6Δ* replicates was dominated by two different G/C -1 bp deletions at a discrete positions (32658 and 32608). These variants occurred at frequencies greater than 10% in at least two *rnr1Y285A-pGAL msh6Δ* biological replicates and likely represent mutations that occurred early in the growth of liquid culture and were carried through subsequent rounds of replication.

Positions in *CAN1* that are susceptible to mutation vary by genotype

We identified positions within *CAN1* where particular variants occurred in a genotype-specific manner, with varying degrees of susceptibility. The most highly susceptible positions, defined as "hotspots", were positions where a variant occurred in at least 50% of replicates with a frequency 2-fold or greater above wildtype (**Tables S6-S11**). The susceptible *CAN1* positions that we observed in *msh2Δ* (**Table S9**) overlapped hotspots previously observed (9,10); we noted additional susceptible positions, particularly in repetitive sequences (see below). We also noted positions that were moderately susceptible to mutation in *msh6Δ* and *msh3Δ* backgrounds, which had not been studied in this way previously (**Tables S9, S10 and S11**). Moderately susceptible positions were defined as positions that occurred in more than 1 biological replicate with an average frequency above wildtype. Both *msh2Δ* (49 positions) and *msh6Δ* (96 positions) exhibited more susceptible positions than *msh3Δ* (18 positions), consistent with more stochastic mutations contributing to the mutation spectrum in the absence of *MSH3* (**Table S10**). The *msh6Δ* sustained 13 hotspots, all SNVs, which occurred at elevated variant frequencies in 3 or more biological replicates and were not observed in wildtype (**Table S11**). These hotspots occurred almost exclusively in repetitive sequence contexts, including at dinucleotides and in runs 3bp or greater in length. A similar trend was observed in *msh2Δ* (**Table S9)** which had a total of 11 hotspots, mostly deletions, which occurred in repetitive runs, with few spots overlapping with *msh6Δ*. This analysis revealed significant and distinct genotype-specific mutagenic fingerprints.

Previous work uncovered no hot spots in *rnr1Y285F* (8) and *rnr1D57N* was previously not assessed in this way (6). Here we did note susceptible and moderately susceptible positions in these backgrounds (**Tables S6, S7**). As noted with *msh* backgrounds, many of these susceptible positions in all *rnr1* backgrounds were associated with repetitive DNA sequences, particularly insertions and deletions.

The susceptible positions in *rnr1Y285A msh2Δ* identified in our study largely overlapped with those previously identified, which were defined as variants occurring at a rate >10-fold higher than wild-

type (8,10). The increased sample sizes and depth of sequencing in our study also revealed new susceptible positions. One noteworthy position is a G>T at 32,940, which occurs in A-rich sequence (CCAA**G**AAAA) and is susceptible to mutation in *msh2Δ*, *rnr1Y285A* and *rnr1Y285A-pGAL*. G>A SNVs also occurred in a variety of genotypes at this position, indicating decreased replication fidelity in this context. The variant frequency at this position increased synergistically in *rnr1Y285A-pGAL msh2Δ* and *rnr1Y285A msh2Δ* double mutants, occurring at a frequency at least 10-fold greater than most single mutants. The majority of susceptible positions in *rnr1Y285A msh2Δ* hotspots are also susceptible in *rnr1Y285A* but not *msh2Δ*,indicating that highly skewed/elevated dCTP and dTTP levels promoted specific errors to a level that were substrates for MMR and approached saturation of the MMR system. However, *rnr1Y285A msh6Δ* and *rnr1Y285A msh3Δ* deletions showed more unique mutational fingerprints.

Insertions and deletions happen in distinct sequence contexts depending on genotype

In general, susceptible *CAN1* positions tended to be in or near repetitive sequences and it is well established that insertion and deletion events in particular occur more frequently in repetitive runs (54,66); our study demonstrated similar results. Within *CAN1*, there are only short homopolymeric runs and A/T repeats are more abundant than G/C repeats: 266 AA/TT dinucleotides versus 160 GG/CC dinucleotides*;* 98 mononucleotide runs of 2 or more A/T bases versus only 32 G/C runs. These ratios are consistent with our observation that more A/T insertion and deletion events occurred in A/T repeats than in G/C repeats in wildtype (**Fig. 3**). 96.1% of all A/T -1 bp deletions occurred in a repetitive run of 3 bp or greater. 89.6% of G/C deletions occurred in a repetitive run of 2 or more nucleotides. Notably, some G/C deletions occurred adjacent to repetitive runs in specific *rnr1* backgrounds. For example, a GΔ at position 31971 occurred in 4 out of 7 *rnr1D57N* biological replicates, between an A dinucleotide and C dinucleotide (AT<u>AA</u>**G**CC<u>AA</u>). CΔ deletions at positions 32398 (G<u>AAA</u>**C**GTAG) and 32630

(CCTT**C**GTTTT) were specific to r*nr1Y285A*. A GΔ at position 33018 was specific to *msh3Δ* ; it is not flanked by a repetitive run but dinucleotides are nearby upstream and downstream (GGAT**G**TAAC).

Complex insertions and deletions (i.e. involving more than one nucleotide) were rare, but occurred at increased frequency in *msh3Δ* (**Fig. 4A, B**). The majority of these complex events, especially insertions, were observed in a single biological replicate. The complex insertion of CT in the repetitive run stretching from positions 33206- 33215 (CTTAAG**CT**CTCTC) is noteworthy. It was observed almost exclusively in *msh2Δ* genotypes, and more frequently when paired with *rnr1* alleles. This indicates that this run was particularly susceptible to mutation when dNTPs were elevated, even a small amount, e.g. *rnr1D57N*. However, this CT insertion was very efficiently repaired by MMR, via either Msh2-Msh3 or Msh2-Msh6 directed repair; it was not observed in either *msh3Δ* or *msh6Δ*.

SNVs in trinucleotide context reveal unique mutation signatures

Sequence context can influence both misincorporation and MMR efficiency. Therefore, we examined the occurrence of different SNVs as a function of the trinucleotide context by genotype, an increasingly common approach to extract mutation signatures, especially in human cancers (67,68). Each unique SNV was categorized by trinucleotide context, with respect to the nucleotide immediately 5' and 3' to the variant, with 96 possible triplet contexts. We determined the average number of times an SNV was observed in a particular triplet context per genotype, normalized by the number of times the triplet context occurs in *CAN1* (**Fig. 5)**. Spearman correlation coefficients of each SNV in unique trinucleotide context were calculated to characterize patterns of SNV mutagenesis (**Fig. S8**). Notably, in all genotypes, C→T changes (red bars, **Fig. 5**), particularly in G**C**C and G**C**G sequence contexts, dominated. The proportion of G**C**C and G**C**G changes increased with altered dNTPs, most dramatically in *rnr1Y285F* and *rnr1Y285A* samples (note change in y-axis), which were highly correlated (r_s = 0.7284). This sequence context has been previously shown to be mutated in *rnr1Y285A* via whole

genome sequencing (9) and was proposed to be a result of increased extension after misincorporation of dTTP, which is in excess, superseding proofreading (15).

Wildtype, *rnr1D57N*, and *msh3Δ* had broad ranges of mutation in the trinucleotide contexts, particularly C>G, T>A, T>C and T>G variants (black, gray, green and pink bars respectively), which were rarely seen in either *rnr1Y285F* or *rnr1Y285A.* The wildtype SNV trinucleotide pattern was highly correlated with that of *msh3Δ* (r_s = 0.7351), *rnr1D57N msh3Δ* (r_s = 0.6752) and *rnr1D57N* (r_s = 0.6682). We noted that no C>T SNVs occurred in G**C**T context in these four genotypes, while they were observed in all *rnr1Y285A* and *rnr1Y285F* double mutant backgrounds, still at relatively low frequencies. This example shows an error that was specific to skewed increases in dCTP and dTTP and was efficiently repaired by either Msh2-Msh3 or Msh2-Msh6.

We also observed a decrease in the number of unique SNVs which occurred in trinucleotide context when *rnr1Y285F* or *rnr1Y285A* alleles were paired with *msh2Δ, msh3Δ,* or *msh6Δ.* All of these genotypes are also highly correlated to one another, with *rnr1Y285F/A msh6Δ* or *msh2Δ* double mutants showing the highest correlation values (Spearman rank, **Table S8**). For example, the spectrum of SNVs in trinucleotide context of *rnr1Y285A msh2Δ* was most highly correlated to *rnr1Y285F-pGAL RNR1 msh2Δ* (0.8616), *rnr1Y285F msh6Δ* (0.8595), *rnr1Y285F msh2*Δ (0.8458), *rnr1Y285F pGAL-RNR1 msh6Δ* (0.8197), *rnr1D57N msh6Δ (0.7686), rnr1Y285A msh6Δ* (0.7662) and *rnr1Y285A pGAL msh2Δ* (0.7640), descending in value. We observed a complete loss of C>G SNVs in all trinucleotide contexts in *rnr1Y285A msh3Δ*, *rnr1Y285A msh6Δ*, and *rnr1Y285F msh6Δ* backgrounds, as noted above (**Fig. 3)**. The C>G variant rarely occurred in *msh2Δ* backgrounds which indicates that the replicative polymerases rarely generate these errors. It is noteworthy that double mutant profiles often showed distinct changes in the SNV signatures, compared to either single mutant. For example, *rnr1Y285A* sustained a number of C>G mutations in multiple trinucleotide contexts, which were also observed in *rnr1Y285A msh2Δ*, but

these were completely absent in *rnr1Y285A msh3Δ* and *rnr1Y285A msh6Δ*. This indicated that there were combinatorial effects on the mutation spectra of specific MMR deletions in the presence of *rnr1* alleles, and different *rnr1* alleles paired with different MMR deletions result in distinct and unique mutational fingerprints.

Extended sequence context reveals sequence specificity of Msh2-3 and Msh2-6 error recognition

The depth of sequence within *can1* in single and double mutants, allowed us to evaluate potential synergistic (combinatorial) effects of *rnr1* and *msh* alleles on mutation spectra. We performed hierarchical cluster analysis of all unique errors in our entire sample set, to evaluate potentially larger sequence contexts that could drive distinct signatures (**Fig. S9)**. Genotypes were clustered based on the types of variants that were differentially enriched as a function of genotype, using Pearson correlation (Spearman correlation analysis yielded similar results) (**Fig. S9**). It is worth noting that overall, clusters were similar to those observed in **Fig. 2**, based on variant frequencies, which suggests that unique variants were the main drivers of genotype-specific mutation profiles. We identified differentially enriched variants and performed motif analysis (12 base window) to determine whether broader sequence context influenced the occurrence of these variants, indicating the same biological mechanism. The 12 base window was chosen because the longest repetitive runs in *CAN1* are 6 nucleotides long and the majority of variants are associated with repetitive runs.

Differentially enriched, genotype-specific variants occurred in distinct sequence contexts. Some of the clusters with the most pronounced differential enrichment between genotypes, were SNVs that occurred in CC dinucleotide context (**Fig. S9**). CC/GG dinucleotides bordered by G on either the 5' or the 3' side were frequently mutated in the presence of *rnr1* alleles, particularly *rnr1Y285F* and *rnr1D57N,* but not with *msh* alleles (**Fig. 6)**. MMR status, (especially *msh3Δ* or *msh6Δ*) in combination with these *rnr1* alleles appeared to reverse this bias with negative enrichment. The majority of SNVs

that occurred in other CC dinucleotide contexts were positively enriched in *msh2Δ* and *msh6Δ* samples, but negatively enriched in *msh3Δ* (**Fig. 6- II, III, & IV)**. This indicated that Msh2-Msh6 was uniquely required to recognize and direct repair of misincorporation events in CC dinucleotide context. CC dinucleotides are frequently mutated in different types of cancer, including human lung cancer (69,70). A defect in Msh2-Msh6-mediated MMR or alteration in free dNTP levels in these cancers may result in synergistic increases in mutagenesis.

We also noted that *msh3Δ*, *rnr1Y285F* and *rnr1Y285A* single mutants showed an increase in G/C single base deletions within specific sequence contexts (**Fig. 7**). The frequency at which these errors occurred increased synergistically in the *rnr1Y285F msh3Δ* and *rnr1Y285A msh3Δ* double mutants (**Fig. 7).** This variant was neutral or negatively enriched in all other genotypes, consistent with apparent specificity for G/C single base deletions in *msh3Δ* backgrounds (**Fig. 3, Table 2**). Some G/C errors along *CAN1* were more specific to *msh3Δ* while others were more specific to *rnr1Y285F* and/or *rnr1Y285A*. Msh2-Msh3 was important for recognizing and directing repair of G/C single base deletions that occurred in homopolymeric runs bordered by G/C rich sequence on the 5' side of the run (**Fig. 7A-C**). G/C runs bordered by A/T nucleotides were more prone to mutagenesis when dCTP and dTTP were increased (**Fig. 7D-F)**. The frequency of these variants directly bordered by A/T increased synergistically when *MSH3* was deleted, but not when *MSH6* was deleted, indicating Msh2-Msh3 has specificity in directing repair of G/C -1 deletions in repetitive G/C context.

DISCUSSION

We developed a *CAN1* selection-based next generation sequencing approach, characterizing mutation spectra in over 30 genetic backgrounds with unparalleled depth. While *CAN1* is a widely used reporter gene in yeast, the application of deep-sequencing allowed streamlined sample preparation and more accurate determination of mutation spectra, by increasing both the number of colonies and biological

replicates analyzed at one time. This approach allowed us to 1) identify low frequency sequence variants, 2) develop mutational fingerprints that result from compromised replication fidelity and 3) determine how mutations in different molecular pathways interact to modulate mutation profiles. We expanded our understanding of the impact of altered dNTP pools, in the presence or absence of MMR, on replication fidelity. By using *rnr1* backgrounds that alter the *type* and *frequency* of mutations sustained, we revealed previously unrecognized specificities for the MMR initiation complexes, Msh2-Msh3 and Msh2-Msh6.

Mechanisms of mutagenesis from distinct elevations in dNTP levels

Despite a 2-fold increase in mutation rate above wildtype, the *rnr1D57N* mutation spectrum closely resembled wildtype, consistent with previous studies (6,8). In *rnr1D57N msh2Δ* the mutation rate increased 150-fold above wildtype and 3-fold above *msh2Δ*, yet the mutation spectrum of *rnr1D57N msh2Δ* is closely related to *msh2Δ* (**Figure 2**). The same is true of *rnr1D57N msh3Δ* and *rnr1D57N msh6Δ;* these genotypes are most closely related to *msh3Δ* and *msh6Δ*, respectively, despite high increases in mutation rates (**Figure 2, Table 1**). This indicates the 2-fold balanced increase in dNTPs causes errors to accumulate in a stochastic manner, consistent with wildtype, with a large majority of errors being efficiently repaired by MMR.

Similar to *rnr1D57N*, the *rnr1Y285F* allele by itself has a modest effect on mutation rate, but the mutation spectrum is distinct, especially when paired with MMR deletions. In fact, the *rnr1Y285F* spectrum closely resembles that of *rnr1Y285A*, despite this allele having a higher dNTP skew and higher mutation rate (~20-fold). This demonstrates that even modest-skewed increases in dNTPs result in distinct error accumulation, likely due to both a decrease in selectivity of the replicative polymerases, along with an increase in efficient mismatch extension at the expense of proofreading (8,9). **Figure 8A**

illustrates an example of two different positions in the *rnr1Y285A/F* background that are mutated at increased frequencies by this mechanism.

Interestingly the increase in G/C deletions in this background can be explained by limiting levels of dGTP (**Figure 8B)**. dGTP levels are limiting in both yeast and mammalian cells and notably led to distinct patterns of mutagenesis when the concentration decreases further, relative to the increase in the other 3 nucleotides in *rnr1Y285F/A*. Notably, *rnr1* alleles that cause increases in dGTP are also found to be extremely mutagenic (71), highlighting the importance of maintaining the proper level and relative abundance of dGTP.

New insights into MMR specificity

In assessing mutation spectra and profiles, we are limited by the errors generated by replicative (and other) DNA polymerases. By altering dNTP pools, we altered the frequency and types of replication errors that occurred, unveiling new specificities in the types of errors that Msh2-Msh3 and Msh2-Msh6 recognize. While G/C SNVs are common in wild-type backgrounds, G/C-1 deletions are relatively rare, making it difficult to determine the relative efficiencies of MSH complexes with respect to this type of error. However, elevating the dNTP pools to any extent increased the ratio of G/C-1:A/T-1 deletions and allowed us to assess MMR function at these sites. In particular, the meta-analysis of our dataset showed that the majority of mutations in *rnr1Y285A/F* and *msh3Δ* genetic backgrounds, specifically single base deletions, occurred in G/C rich context (**Fig. 7, 8B**). The double mutants, *rnr1Y285F msh3Δ* and *rnr1Y285A msh3Δ*, exhibited mutation profiles that were completely dominated by G/C-1 deletions. Therefore Msh2-Msh6 was not able to compensate for the loss of Msh2-Msh3 for repair of G/C-1 deletions, despite the fact that both complexes have been implicated in directing repair of single base deletions (72).

It has been previously shown that there are few strand specific effects from altering dNTP levels (10), but numerous studies have indicated that MMR is more efficient on the lagging strand (73,74). It is also known that polymerase δ mutants deficient in proofreading (*pol3 L612M*) accumulate significantly more G/C deletions (66). This leads us to hypothesize that Msh2-Msh3 may have greater specificity for lagging strand DNA repair. Our targeted sequencing approach could be applied on strains with reverse orientation *CAN1* to further explore leading versus lagging strand DNA repair specificity of Msh2-Msh3 and Msh2-Msh6, as well as in the presence of elevated dNTP pools.

Similarly, SNVs at GG/CC dinucleotides were more prevalent in *rnr1D57N* and *rnr1Y285F* backgrounds than in wild-type (**Fig. 6**). We demonstrated that mutations in these G/C-rich patches were enriched in *msh2Δ* and *msh6Δ* backgrounds, but negatively enriched in *msh3Δ*, indicating that Msh2-Msh6 directs more efficient repair of mutations in GG/CC dinucleotides than does Msh2-Msh3. GpG dinucleotides are mutated as a signature mutation in colorectal cancer (75); cancers which are defined by defects in MMR. Mutations at CC dinucleotides are also found in human lung cancer (69,70) and could be generated due to defects in MMR and elevations in dNTP levels in combination. Notably, RNR subunit overexpression and deletion of *MSH6* in combination caused a synergistic increase in lung carcinogenesis in a mouse model(6). Our sequence context data provide new insight into MMR substrate specificity and sequence context information that influence the relative efficiency of MMR directed by either Msh2-Msh3 and/or Msh2-Msh6.

Notably, pairing elevations in dNTP levels with MMR deletions led to increased mutation rates and distinct mutation spectra. However, when measuring the rate of canavanine resistance in *rnr1Y285A msh2Δ* and *rnr1Y285A msh6Δ* backgrounds, we observed substantially lower colony numbers under permissive conditions, indicating reduced fitness even under conditions when all cells should be able to grow. This presumably would also reduce the number of cells that are able to grow under selective conditions and likely accounts for mutation rates being lower than expected (**Table 1**). We have also

noted slow growth phenotypes and survival defects in these genetic backgrounds (our unpublished results), consistent with the phenomenon of error extinction (11,76,77). We plan to evaluate this compromised fitness further as this has important implications for MMR defects and dNTP imbalances in maintaining cellular fitness.

Implications for understanding human cancer mutation signatures

Mutation signatures of human tumors are used to identify molecular drivers of carcinogenesis (67,78-80). This has clear implications for diagnosis, prognosis and treatment options for patients (81). Tumor mutation signatures are typically extracted from a large sampling of human tumor information via mathematical approaches and molecular pathways are inferred. This has led investigators to determine spurious associations between mutation profiles and driver mutations, e.g. RefSig 3 and *BRCA1/BRCA2* (82). In contrast, we used the *S. cerevisiae* model system to develop mutation signatures from first principles, using multiple biological replicates of defined genetic backgrounds alongside wild-type controls. By layering genetic backgrounds, the data revealed the potential for genotype-specific combinatorial effects on mutation signatures. Importantly, up to this point elevations in dNTP levels have not been considered when evaluating mutation signatures from human tumors, although they almost certainly contribute to mutagenesis in cancer.

We observed distinct similarities between mutation spectra in specific sequence contexts to particular COSMIC signatures, most notably signatures 6 and 15. Signature 6 is most commonly occurring in colorectal and uterine cancers and is associated with defective MMR. It is possible that elevations, specifically elevations in dCTP and dTTP contribute to this mutation signature in human cancers leading to synergistic increases in these types of mutations in the absence of certain MMR genes. COSMIC signature 15 is characterized by C>T SNVs within NpCpG sequences in lung and stomach cancers within unknown origin. The origin of these cancers could be in part, due to defects in

MMR coupled with skewed increases in dCTP and dTTP, consistent with the increased frequency of these SNVs at dinucleotides observed in *rnr1Y285F/A msh6Δ* backgrounds.

While elevations in dNTP levels are implicated in supporting rapidly proliferating cancer cells, whether skewed or balanced increases occur remains to be determined. Quantificaton in HeLa cell lines show elevated levels of all four dNTPs with an approximate 2-fold increase in dATP and dCTP and a 4-fold increase in dTTP and dGTP (43). Different skewed elevations result in distinct mutation spectra (71) and thus more studies are necessary to determine what dNTP imbalances are relevant to different types of cancers, although certain indicator mutations, such as a high number of G/C single base deletions can point towards specific dNTP imbalances *i.e.* high dCTP and dTTP levels seen in *rnr1Y285A*.

The *msh2Δ, msh3Δ* and *msh6Δ* mutation spectra all had features of MMR deficient human tumor samples (79). Notably, the *msh6Δ* spectrum in our study closely resembles that of *Msh6*$^{-/-}$ in a HAP1 cell line (83). The percentage of substitutions and IDLs in the mutation spectrum is consistent with what is seen in a *Msh6*$^{-/-}$ cell line. While T>C variants did not dominate the *msh6Δ* mutation signature in yeast as it did in the *Msh6*$^{-/-}$ cell line, the overall frequency and proportion of T>C changes did increase significantly. These data indicate that mutation signatures developed through defined mutations and deletions in *S. cerevisiae* will be broadly applicable to mammalian systems. Mutation signatures from *C. elegans* also resemble human cancer signatures, albeit with some minor discrepancies (72,84). We note that *C. elegans* lacks a *MSH3* homolog, which is present in both yeast and humans (85). Our results highlight the importance of considering altered dNTP pools and combinatorial effects of genetic backgrounds in future analyses.

AVAILABILITY

All data is available upon request.

SUPPLEMENTARY DATA

ACKNOWLEDGEMNTS

We are grateful to Dr. Andrei Chabes for providing strains and plasmids. We are grateful to Dr. Don Yergeau for his advice in next generation sequencing. We are grateful to members of the Surtees lab (past and present) for helpful discussions and input. We are grateful to Dr. Mark Sutton and Dr. John Panepinto for input and helpful discussions.

FUNDING

N.A.L. is a University at Buffalo Presidential Scholar. This work was supported by the American Cancer Society (RSG-14-2350-01) to J.A.S. J.A.S. is an ACS Research Scholar. J.A.S. is also grateful for support from the University at Buffalo's Genome, Environment and Microbiome (GEM) Community of Excellence. Funding for open access charge: GEM.

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

REFERENCES

1. Kunkel, T.A. (2004) DNA Replication Fidelity. *J. Biol. Chem.*, **279**, 16895-16898.
2. McCulloch, S.D. and Kunkel, T.A. (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res*, **18**, 148-161.
3. Ganai, R. and Johansson, E. (2016) DNA Replication - A matter of fidelity. *Molecular Cell*, **62**, 745-755.
4. St. Charles, J.A., Liberti, S.E., Williams, J.S., Lujan, S.A. and Kunkel, T.A. (2015) Quantifying the Contributions of Base Selectivity, Proofreading and Mismatch Repair to Nuclear DNA Replication in Saccharomyces cerevisiae. *DNA Repair*, **31**, 41-51.
5. Chabes, A., Georgieva, B., Domkin, V., Zhao, X., Rothstein, R. and Thelander, L. (2003) Survival of DNA damage in yeast directly depends on increased dNTP levels allowed by relaxed feedback inhibition of ribonucleotide reductase. *Cell*, **112**, 391-401.
6. Xu, X., Page, J.L., Surtees, J.A., Liu, H., Lagedrost, S., Lu, Y., Bronson, R., Alani, E., Nikitin, A.Y. and Weiss, R.S. (2008) Broad Overexpression of Ribonucleotide Reductase Genes in Mice Specifically Induces Lung Neoplasms. *Cancer Res*, **68**, 2652-2660.
7. Kumar, D., Viberg, J.r., Nilsson, A.K. and Chabes, A. (2010) Highly mutagenic and severely imbalanced dNTP pools can escape detection by the S-phase checkpoint. *Nucleic Acids Research*, **38**, 3975-3983.
8. Kumar, D., Abdulovic, A.L., Viberg, J.r., Nilsson, A.K., Kunkel, T.A. and Chabes, A. (2011) Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. *Nucleic Acids Research*, **39**, 1360-1371.
9. Watt, D.L., Buckland, R.J., Lujan, S.A., Kunkel, T.A. and Chabes, A. (2016) Genome-wide analysis of the specificity and mechanisms of replication infidelity driven by imbalanced dNTP pools. *Nucleic Acids Research*, **44**, 1669-1680.

10. 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.

11. Schmidt, T.T., Reyes, G., Gries, K., Ceylan, C.Ü., Sharma, S., Meurer, M., Knop, M., Chabes, A. and Hombauer, H. (2017) Alterations in cellular metabolism triggered by URA7 or GLN3 inactivation cause imbalanced dNTP pools and increased mutagenesis. *Proceedings of the National Academy of Sciences*, **114**, E4442-E4451.

12. Franzolin, E., Giovanna, P., Rampazzo, C., Christina, M., Ferraro, P., Elisa, P., Peter, R. and Bianchi, V. (2013) The deoxynucleotide triphosphohydrolase SAMDH1 is a major regulator of DNA precursor pools in mammalian cells. *Proceedings of the National Academy of Sciences*, **110**, 14272-14277.

13. Rentoft, M., Lindell, K., Tran, P., Chabes, A.L., Buckland, R., Watt, D.L., Marjavaara, L., Nilsson, A.K., Melin, B., Johansson, E. *et al.* (2016) Heterozygous Colon Cancer-Associated Mutations of SAMHD1 Have Functional Significance. *Proceedings of the National Academy of Sciences*, **113**, 4723-4728.

14. Mauney, C. and Hollis, T. (2018) SANDH1: Recurring roles in cell cycle, viral restriction, cancer , and innate immunity. *Autoimmunity*, **51**, 96-110.

15. Kunkel, T.A. and A, S. (1988) Mutagenesis by Transient Misalignment. *Journal of Biological Chemistry*, **263**, 14784-14789.

16. Phear, G., Nalbantoglu, J. and Meuth, M. (1987) Next-nucleotide effcts in mutations driven by DNA precuros pool imbalances at the aprt locus of Chinese hamter overy cells. *PNAS*, **84**, 4450-4454.

17. Kunkel, T.A. and Erie, D.A. (2015) Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annual Review of Genetics*, **49**, 291-313.

18. Jensen, L.E., Jauert, P.A. and Kirkpatrick, D.T. (2005) The Large Loop Repair and Mismatch Repair Pathways of Saccharomyces cerevisiae Act on Distinct Substrates During Meiosis. *Genetics*, **170**, 1033-1043.

19. Sia, E., Kokoska, R., Dominska, M., Greenwell, P. and Petes, T. (1997) Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol. Cell. Biol.*, **17**, 2851-2858.

20. Harrington, J.M. and Kolodner, R.D. (2007) Saccharomyces cerevisiae Msh2-Msh3 Acts in Repair of Base-Base Mispairs. *Mol. Cell. Biol.*, **27**, 6546-6554.

21. Srivatsan, A., Bowen, N. and Kolodner, R.D. (2014) Mispair-specific Recruitment of the Mlh1-Pms1 Complex Identifies Repair Substrates of the Saccharomyces cerevisiae Msh2-Msh3 Complex. *Journal of Biological Chemistry*, **289**, 9352-9364.

22. Chabes, A. and Stillman, B. (2007) Constitutively high dNTP concentration inhibits cell cycle progression and the DNA damage checkpoint in yeast Saccharomyces cerevisiae 10.1073/pnas.0610585104. *PNAS*, **104**, 1183-1188.

23. Reichard, P., Eliasson, R., Ingemarson, R. and Thelander, L. (2000) Cross-talk between the Allosteric Effector-binding Sites in Mouse Ribonucleotide Reductase. *Journal of Biological Chemistry*, **275**, 33021-33026.

24. Caras, I.W. and Martin, D.W.J. (1988) Molecular cloning of the cDNA for a mutant mouse ribonucleotide reductase M1 that produces a dominant mutator phenotype in mammalian cells. *Mol. Cell. Biol.*, **8**, 2698-2704.

25. Weinberg, G., Ullman, B. and Martin Jr, D. (1981) Mutator phenotypes in mammalian cell mutants with ditinct biochemical defects and abnormal deoxyribonucleoside triphosphate pools`. *Proceedings of the National Academy of Sciences*, **78**, 2447-2451.

26. Jordheim, L.P., Durantel, D., Zoulim, F. and Dumontet, C. (2013) Advances in the Development of Nucleoside and Nucleotide Analogues for Cancer and Viral Diseases. *Nature Reviews in Drug Discovery*, **12**, 447-464.

27. Kohnken, R., Kodigepalli, K. and Wu, L. (2015) Regulation of deoxynucleotide metabloism in cancer: novel mechanisms and therapeutic implications. *Molecular Cancer Research*, **14**, 176.

28. Mastrocola, A.S. and Heinen, C.D. (2010) Lynch syndrome-associated mutations in MSH2 alter DNA repair and checkpoint response functions in vivo. *Human Mutation*, **31**, E1699-E1708.

29. Pino, M., Mino-Kenudson, M., Wildemer, B., Ganguly, A., Batten, J., Sperduti, I., Iafrate, A. and Chung, D. (2009) Deficient DNA mismatch Repair is Common in Lynch Syndrome-Associated Colorectal Adenomas. *Journal of Molecular Diagnostics*, **11**, 238-247.

30. Heinen, C.D. (2010) Genotype to phenotype: Analyzing the effects of inherited mutations in colorectal cancer families. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **693**, 32-45.

31. Davies, H., Morganella, S., Purdie, C., Jang, S., Borgen, E., Russnes, H., Glodzik, D., Zou, X., Viari, A., Richardson, A. *et al.* (2017) Whole-Genome Sequencing Reveals Breast Cancers with Mismatch Repair Deficiency. *Cancer Research*, **77**, 4755-4762.

32. Fusco, N., Lopez, G., Corti, C., Pesenti, C., Colapietro, P., Ercoli, G., Gaudioso, G., Faversani, A., Gambini, D., Michelotti, A. *et al.* (2018) Mismatch Repair Protein Loss as a Prognostic and Predictive Biomarker in Breast Cancers Regardless of Microsatellite Instability. *JNCI Cancer Spectrum*, **2**, pky056.

33. Edelmann, W., Umar, A., Yang, K., Heyer, J., Kucherlapati, M., Lia, M., Kneitz, B., Avdievich, E., Fan, K., Wong, E. *et al.* (2000) The DNA Mismatch Repair Genes Msh3 and Msh6 Cooperate in Intestinal Tumor Suppression. *Cancer Res*, **60**, 803-807.

34. van Oers, J.M.M., Edwards, Y., Chahwan, R., Zhang, W., Smith, C., Pechuan, X., Schaetzlein, S., Jin, B., Wang, Y., Bergman, A. *et al.* (2013) The MutS[beta] complex is a modulator of p53-driven tumorigenesis through its functions in both DNA double-strand break repair and mismatch repair. *Oncogene*.

35. Santos, L.S., Silva, S.N., Gil, O.M., Ferreira, T.C., Limbert, E. and Rueff, J. (2018) Mismatch repair single nucleotide polymorphisms and thyroid cancer susceptibility. *Oncology Letters*, **15**, 6715-6726.

36. Morak, M., Käsbauer, S., Kerscher, M., Laner, A., Nissen, A.M., Benet-Pagès, A., Schackert, H.K., Keller, G., Massdorf, T. and Holinski-Feder, E. (2017) Loss of MSH2 and MSH6 due to heterozygous germline defects in MSH3 and MSH6. *Familial Cancer*, **16**, 491-500.

37. Adam, R., Spier, I., Zhao, B., Kloth, M., Marquez, J., Hinrichsen, I., Kirfel, J., Tafazzoli, A., Horpaopan, S., Uhlhaas, S. *et al.* (2016) Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. *The American Journal of Human Genetics*, **99**, 337-351.

38. Valle, L., de Voer, R.M., Goldberg, Y., Sjursen, W., Försti, A., Ruiz-Ponte, C., Caldés, T., Garré, P., Olsen, M.F., Nordling, M. *et al.* (2019) Update on genetic predisposition to colorectal cancer and polyposis. *Molecular Aspects of Medicine*.

39. Takahashi, M., Koi, M., Balaguer, F., Boland, C.R. and Goel, A. (2011) MSH3 Mediates Sensitization of Colorectal Cancer Cells to Cisplatin, Oxaliplatin, and a Poly(ADP-ribose) Polymerase Inhibitor. *Journal of Biological Chemistry*, **286**, 12157-12165.

40. Park, J.M., Huang, S., Tougeron, D. and Sinicrope, F.A. (2013) MSH3 Mismatch Repair Protein Regulates Sensitivity to Cytotoxic Drugs and a Histone Deacetylase Inhibitor in Human Colon Carcinoma Cells. *PLoS ONE*, **8**, e65369.

41. Nogueira, G.A.S., Costa, E.F.D., Lopes-Aguiar, L., Lima, T.R.P., Visacri, M.B., Pincinato, E.C., Lourenço, G.J., Calonga, L., Mariano, F.V., Altemani, A.M.d.A.M. *et al.* (2018) Polymorphisms in DNA mismatch repair pathway genes predict toxicity and response to cisplatin chemoradiation in head and neck squamous cell carcinoma patients. *Oncotarget*, **9**, 29538-29547.

42. Carethers, J.M., Koi, M. and Tseng-Rogenski, S.S. (2015) EMAST is a Form of Microsatellite Instability That is Initiated by Inflammation and Modulates Colorectal Cancer Progression. *Genes*, **6**, 185-205.

43. Mathews, C.K. (2015) Deoxyribonucleotide metabolism, mutagenesis and cancer. *Nature Reviews Cancer*, **15**, 528.

44. Connor, A., Denroche, R., Jang, G., Timms, L., Kalimuthu, S., Selander, I., MsPherson, T., Wilson, G., Chan-Seng-Yue, M., Borozan, I. *et al.* (2017) Association of Distinct Mutational Signatures with Correlates of Increased Immune Activity in Pancreatic Ductal Adenocarcinoma. *JAMA Oncology*, **3**, 774-783.

45. Sikorski, R.S. and Hieter, P. (1989) A System of Shuttle Vectors and Yeast Host Strains Designed for Efficient Manipulation of DNA in Saccharomyces cerevisiae. *Genetics*, **122**, 19-27.

46. Gietz, D., Jean, A.S., Woods, R.A. and Schiestl, R.H. (1992) Improved method for high efficiency transformation of intact yeast cells. *Nucleic Acids Research*, **20**, 1425.

47. Alexandrov, L.B., Jones, P., Wedge, D., Sale, J., Campbell, P., Nik-Zainal, S. and Stratton, M.R. (2015) Clock-like Mutational Process in Human Somatic Cells. *Nature Genetics*, **47**, 1402-1407.

48. Crooks, G., Hion, G., Chandonia, J.-M. and Brenner, S. (2004) WebLogo: A Sequence Logo Generator. *Genome Research*, **15**, 1188-1190.

49. Lang, G.I., Parsons, L. and Gammie, A.E. (2013) Mutation Rates, Spectra, and Genome-Wide Distribution of Spontaneous Mutations in Mismatch Repair Deficient Yeast. *G3: Genes|Genomes|Genetics*, **3**, 1453-1465.

50. Marsischky, G.T., Filosi, N., Kane, M.F. and Kolodner, R.D. (1996) Redundancy of *Saccharomyces cerevisiae MSH3* and *MSH6* in *MSH2*-dependent mismatch repair. *Genes & Dev.*, **10**, 407-420.

51. Kumar, C., Piacente, S.C., Sibert, J., Bukata, A.R., O'Connor, J., Alani, E. and Surtees, J.A. (2011) Multiple Factors Insulate Msh2–Msh6 Mismatch Repair Activity from Defects in Msh2 Domain I. *Journal of Molecular Biology*, **411**, 765-780.

52. Fantes, P. and Creanor, J. (1984) Canavanine resistance and the mechanism of arginie uptake in the fission yeast Schizosaccharomyces cerevisiae. *Journal of General Microbiology*, **130**, 3265-3273.

53. Hoffmann, W. (1985) Molecular Characterization of the CAN1 Locus in Saccharomyces Cerevisiae. A Transmembrane Protein Without N-Terminal Hydrophobic Signal Sequence. *Journal of Biological Chemistry*, **260**, 11831-11837.

54. Lujan, S.A., Clausen, A.R., Clark, A.B., MacAlpine, H., MacAlpine, D., 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 Research*, **24**, 1751-1764.

55. Zhu, Y.O., Siegal, M.L., Hall, D.W. and Petrov, D.A. (2014) Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences*, **111**, E2310-E2318.

56. Kunkel, T.A. (1992) DNA replication fidelity. *Journal of Biological Chemistry*, **267**, 18251-18254.

57. Bebenek, K., Roberts, J.D. and Kunkel, T.A. (1992) The effects of dNTP pool imbalances on frameshift fidelity during DNA replication. *Journal of Biological Chemistry*, **267**, 3589-3596.

58. Lee, D.F., Lu, J., Chang, S., Loparo, J.J. and Xie, X.S. (2016) Mapping DNA Polymerase Errors by Single-Molecule Sequencing. *Nucleic Acids Research*, **44**, e118.

59. Oyola, S., Otto, T., Gu, Y., Maslen, G., Maske, M., Campino, S., Turner, D., MacInnis, B., Kwiatkowski, D., Swerdlow, H. *et al.* (2012) Optimizing illumina next-generation sequencing library prepartion for extremelyAT-biased genomes. *BMC Genomics*, **13**.

60. Potapov, V. and Ong, J.L. (2017) Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLOS ONE*, **12**, e0169774.

61. Chatterjee, A., Dasgupta, S. and Sidransky, D. (2011) Mitochondrial Subversion in Cancer. *Cancer Prevention Research*, **4**, 638-654.

62. Habraken, Y., Sung, P., Prakash, L. and Prakash, S. (1996) Binding of insertion/deletion DNA mismatches by the heterodimer of yeast mismatch repair proteins MSH2 and MSH3. *Current Biology*, **6**, 1185-1187.

63. Surtees, J.A. and Alani, E. (2006) Mismatch Repair Factor MSH2-MSH3 Binds and Alters the Conformation of Branched DNA Structures Predicted to form During Genetic Recombination. *Journal of Molecular Biology*, **360**, 523-536.

64. Bowers, J., Sokolsky, T., Quach, T. and Alani, E. (1999) A Mutation in the MSH6 Subunit of the Saccharomyces cerevisiae MSH2-MSH6 Complex Disrupts Mismatch Recognition. *J. Biol. Chem.*, **274**, 16115-16125.

65. Genschel, J., Littman, S.J., Drummond, J.T. and Modrich, P. (1998) Isolation of MutSbeta from Human Cells and Comparison of the Mismatch Repair Specificities of MutSbeta and MutSalpha. *J. Biol. Chem.*, **273**, 19895-19901.

66. 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 Research*, **43**, 4067-4074.

67. Haradhvala, N.J., Kim, J., Maruvka, Y.E., Polak, P., Rosebrock, D., Livitz, D., Hess, J., Leshchiner, I., Kamburov, A., Mouw, K. *et al.* (2018) Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nature Communications*, **9**.

68. Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T. *et al.* (2016) Mutational Signatures associated with tobacco smoking in human cancer. *Science*, **354**, 618-622.

69. Greenman, C., Stephens, P., Smith, R.D., Dalgliesh, G., Hunter, C., Bignell, G., Davies, H., teague, J., Butler, A., Stevens, C. *et al.* (2007) Paterns od somatic mutation in human cancer genomes. *Nature*, **446**, 153-158.

70. Lee, W., Jiang, Z., Liu, J., Haverty, P.M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K., Bhatt, D. *et al.* (2010) The Mutation Spectrum Revelaed by Genome Sequences from a Lung Cancer Patient. *Nature*, **465**, 473-477.

71. Schmidt, T., Sharma, S., Reyes, G., Gries, K., Gross, M., Zhao, B., Yuan, J.-H., Wade, R., Chabes, A. and Hombauer, H. (2019) A Genetic Screen Pinpoints Ribonucleotide Reductase Residues that Sustain dNTP homeostasis and Specifies a Highly Mutagenic Type of dNTP Imbalance. *Nucleic Acids Research*, **47**, 237-252.

72. Meier, B., Volkova, N.V., Hong, Y., Schofield, P., Campbell, P.J., Gerstung, M. and Gartner, A. (2018) Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *Genome Research*, **28**, 666-675.

73. Lujan, S.A., Williams, J.S., Pursell, Z.F., Abdulovic-Cui, A.A., Clark, A.B., Nick McElhinny, S.A. and Kunkel, T.A. (2012) Mismatch Repair Balances Leading and Lagging Strand DNA Replication Fidelity. *PLoS Genetics*, **8**, e1003016.

74. Pavlov, Y.I., Mian, I.M. and Kunkel, T.A. (2003) Evidence for Preferential Mismatch Repair of Lagging Strand DNA Replication Errors in Yeast. *Current Biology*, **13**, 744-748.

75. Rubin, A. and Green, P. (2009) Mutation Patterns in Cancer Genomes. *Proceedings of the National Academy of Sciences*, **106**, 21766-21770.

76. Herr, A.J., Kennedy, S.R., Knowels, G.M., Schultz, E.M. and Preston, B.D. (2014) DNA Replication Error-Induced Extinction of Diploid Yeast. *Genetics*, **196**, 677-691.

77. Williams, L.N., Herr, A.J. and Preston, B.D. (2013) Emergence of DNA Polymerase ε Antimutators That Escape Error-Induced Extinction in Yeast. *Genetics*, **193**, 751-770.

78. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N. *et al.* (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94-101.

79. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G., Bolli, N., Borg, A., Borresen-Dale, A.-L. *et al.* (2013) Signatures of Mutational Preocess in Human Cancer. *Nature*, **500**, 415-421.

80. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **advance online publication**.

81. Van Hoeck, A., Tjoonk, N.H., van Boxtel, R. and Cuppen, E. (2019) Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer*, **19**, 457.

82. Degasperi, A., Amarante, T.D., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D., Morganella, S., Nanda, A.S., Badja, C., Koh, G. *et al.* (2020) A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nature Cancer*, **1**, 249-263.

83. Zou, X., Owusu, M., Harris, R., Jackson, S.P., Loizou, J.I. and Nik-Zainal, S. (2018) Validating the concept of mutational signatures with isogenic cell models. *Nature Communications*, **9**, 1744.

84. Meier, B., Cooke, S.L., Weiss, J., Bailly, A.P., Alexandrov, L.B., Marshall, J., Raine, K., Maddison, M., Anderson, E., Stratton, M.R. *et al.* (2014) C. elegans whole genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Research*.

85. Denver, D.R., Feinberg, S., Estes, S., Thomas, W.K. and Lynch, M. (2005) Mutation Rates, Spectra and Hotspots in Mismatch Repair-Deficient Caenorhabditis elegans. *Genetics*, **170**, 107-113.

Figure Legends

Figure 1. A cartoon depiction of decreased replication fidelity at the replication fork. (**A**) Elevated dNTP pool levels lead the replicative polymerases to favor synthesis over error-checking, leaving behind more mutations. Mutagenic replication is represented in red. (**B**) MMR repairs errors that are left behind during replication. Msh2-3 and Msh2-6 recognize and bind errors, initiating repair. Shown here, Msh2-3 binds a small deletion loop, while Msh2-6 recognizes a base-base mismatch based on known specificities.

Figure 2. Distinct genotypes share unique features. (**A**) Hierarchical cluster analysis using Spearman Rank correlation on all the distinct genotypes in our study. Data was clustered based on the unique counts of the 14 different classes of variants that occurred in each genotype. The histogram shows the distribution of correlation coefficients across samples. (**B**) Principal components analysis performed on all unique variants from biological replicates within a genotype.

Figure 3. Mutation Spectra are distinct across genotypes. (**A**) The relative distribution of SNVs, deletions, insertions, MNVs and replacements by genotype. **(B)** The SNV spectra normalized relative to total SNVs. **(C)** The SNV spectra normalized relative total variants. **(D)** The deletion spectra normalized relative to total deletions. **(E)** The deletion spectra normalized relative to total variants. **(F)** The insertion spectra normalized relative to total insertions. **(G)** The insertion spectra normalized relative to total variants.

Figure 4. Larger deletions are relatively rare across genotypes. The most unique > 1bp deletions and insertions occur in *msh3Δ*. The number of unique insertions greater than 1 base pair (peach) and the number of unique deletions greater than 1 base pair (turquoise) within a genotype are plotted.

Figure 5. The average number of each SNV as it occurs in unique triplet nucleotide context differs by genotype. Bars are colored according to the six different types of SNVs. (**A**) The 16 different triplet contexts are lettered for display purposes. The variant change (C>A, turquoise bar) occurs at the middle nucleotide marked X in each triplet context. The same triplet context is repeated for each possible variant in the following panels. **(B)** Wildtype and single *rnr1* alleles. **(C)** Genotypes with *MSH2* deleted **(D)** Genotypes with *MSH3* deleted. **(E)** Genotypes with *MSH6* deleted.

Figure 6. SNVs occur in C-C dinucleotide sequence contexts with differential enrichment between genotypes. (A) Hierarchical cluster analysis of all SNVs that occur at CC dinucleotides. Clusters of interest are boxed in black, labeled by roman numerals. (**B**) Motif enrichment of a 12 base window surrounding the mutated nucleotide was performed using Berkley web logos (48). The mutated base is at the 7th nucleotide position in all logos, indicated by the black arrow. (**C**) The most predominant type(s) of SNV in the cluster are displayed. (**D**) A summary of genotypes that show negative or positive correlation in a given cluster.

Figure 7. G/C single base deletions that show synergistic increases in variant frequency in *rnr1Y285A msh3Δ* genotypes. The deleted base is starred. (**A**) Motif enrichment of a 12 base window surrounding two G deletions (starred nucleotide) that are specific to genotypes with *msh3Δ*. (**B**) The variant frequency from biological replicates in wildtype, *rnr1Y285F/A* and *msh3Δ* genotypes for the single base deletion that occurred at position 31979 is plotted. (**C)** The variant frequencies across replicates for the G deletion at 32272. In both cases (**B** and **C**) there are very few events in the single mutants, but significant frequencies in double mutant backgrounds. (**D**) Motif enrichment for two C deletions that are specific to genotypes with *rnr1Y285F* and *rnr1Y285A*, but increase synergistically in

double mutants with *msh3Δ* in combination. (**E)** The variant frequencies across replicates for the G deletion at 32747. (**F**) The variant frequencies across replicates for the G deletion at 32658. In **E** and **F**, this event occurs in *rnr1Y285A*, but not *rnr1Y285F* or *msh3Δ*. It occurs at increased frequencies in double mutant backgrounds.

Figure 8. Mechanisms of mutagenesis for the incorporation of errors specific to either Msh2-Msh6 or Msh2-Msh3 repair. The mutated base is represented in red. (**A**)**.** Two examples of sequence context surrounding CC dinucleotides from **Fig. 7B**, where mis-insertion is due to the nucleotide in excess in *rnr1Y285F/A* backgrounds. These errors are efficiently repaired by Msh2-Msh6. This specificity becomes apparent when *msh6Δ* is paired with *rnr1Y285F/A* alleles. (**B**) Two examples of sequence context from **Figs. 8A** and **8D**, where mis-alignment events occur due to the severely limiting amount of dGTP in *rnr1Y285F/A* genetic backgrounds. The run where the deletion occurred is underlined in black. These single base G/C deletions are efficiently repaired by Msh2-Msh3, but not Msh2-Msh6, a previously unidentified specificity of the repair complex.

Table 1. Canavanine mutation rates		
Genotype	**Mutation Rate [95% Confidence Intervals]**	**Fold Change**
Wildtype (n=115)	2.7×10^{-7} [2.3×10^{-7} – 3.2×10^{-7}]	1
msh2Δ (n=24)	6.7×10^{-6} [5.8×10^{-6} – 7.8×10^{-6}]	24.8
msh3Δ (n=32)	7.8×10^{-7} [6.1×10^{-7} – 9.7×10^{-7}]	2.9
msh6Δ (n=24)	1.8×10^{-6} [1.4×10^{-6} – 2.2×10^{-6}]	6.7
rnr1D57N (n=91)	7.3×10^{-7} [6.3×10^{-7} – 8.3×10^{-7}]	2.7
rnr1D57N msh2Δ (n=24)	2.0×10^{-5} [1.8×10^{-5} – 2.3×10^{-5}]	74.1
rnr1D57N msh3Δ (n=24)	2.2×10^{-6} [1.7×10^{-6} – 2.7×10^{-6}]	8.2
rnr1D57N msh6Δ (n=24)	5.6×10^{-6} [4.7×10^{-6} – 6.6×10^{-6}]	20.7
rnr1Y285F (n=42)	7.5×10^{-7} [5.8×10^{-7} – 9.5×10^{-7}]	2.8
rnr1Y285F msh2Δ (n=24)	2.9×10^{-5} [2.6×10^{-5} – 3.2×10^{-5}]	107.4
rnr1Y285F msh3Δ (n=24)	3.7×10^{-6} [3.0×10^{-6} – 4.5×10^{-6}]	13.7
rnr1Y285F msh6Δ (n=24)	2.3×10^{-5} [2.0×10^{-5} – 2.6×10^{-5}]	85.2
rnr1Y285F pGAL-RNR1 (n=24)	7.8×10^{-7} [5.7×10^{-7} – 1.0×10^{-6}]	2.9
rnr1Y285F pGAL-RNR1 msh2Δ (n=24)	1.5×10^{-5} [1.3×10^{-5} – 1.8×10^{-5}]	55.6
rnr1Y285F pGAL-RNR1 msh6Δ (n=12)	1.6×10^{-5} [1.3×10^{-5} – 1.9×10^{-5}]	59.3
rnr1Y285A (n=46)	5.5×10^{-6} [4.7×10^{-6} – 6.2×10^{-6}]	20.4
rnr1Y285A msh2Δ (n=33)	2.5×10^{-5} [2.1×10^{-5} – 3.0×10^{-5}]	92.6
rnr1Y285A msh3Δ (n=49)	6.2×10^{-5} [5.7×10^{-5} – 6.6×10^{-5}]	229.6
rnr1Y285A msh6Δ (n=33)	1.0×10^{-5} [8.9×10^{-6} – 1.2×10^{-5}]	37.0
rnr1Y285A pGAL-RNR1 (n=22)	1.5×10^{-5} [1.2×10^{-5} – 1.7×10^{-5}]	55.6
rnr1Y285A pGAL-RNR1 msh2Δ (n=24)	1.3×10^{-5} [1.1×10^{-5} – 1.5×10^{-5}]	48.2
rnr1Y285A pGAL-RNR1 msh3Δ (n=36)	7.3×10^{-5} [6.8×10^{-5} – 7.9×10^{-5}]	270.4
rnr1Y285A pGAL-RNR1 msh6Δ (n=30)	1.0×10^{-5} [8.7×10^{-6} – 1.2×10^{-5}]	37.0

Table 2. Mutation spectra										
	WT (5)[a]		**WT_2424 (4)**		***msh2Δ* (4)**		***msh3Δ* (6)**		***msh6Δ* (4)**	
Variants	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**
CG>AT	22.20	15.94	13.00	19.27	19.00	11.13	15.83	5.20	45.50	26.44
CG>GC	8.80	6.51	5.75	3.63	1.00	0.79	8.00	2.06	4.00	1.03
CG>TA	35.20	13.41	27.75	32.03	36.75	15.18	28.50	12.81	66.50	49.32
TA>AT	6.20	2.60	2.75	0.62	0.75	-.13	4.50	1.66	3.00	0.66
TA>CG	4.60	1.19	2.50	0.58	6.75	1.83	4.00	-.71	16.00	6.79
TA>GC	5.60	13.22	3.50	0.74	3.50	2.26	4.17	1.20	6.75	2.91
SNV total[d]	**82.60**	**52.85**	**55.25**	**56.87**	**67.75**	**31.32**	**65.00**	**23.65**	**141.75**	**87.15**
% SNV[e]	**79.12**	**66.43**	**81.55**	**72.98**	**68.26**	**32.15**	**61.03**	**31.07**	**88.59**	**88.01**
A/T-1	10.60	18.29	5.75	6.64	18.75	51.69	10.33	11.89	9.00	6.45
G/G-1	3.20	0.60	1.75	12.83	5.00	8.00	12.17	20.03	3.00	1.77
>1 bp del	2.40	1.23	1.75	0.41	1.000	0.62	12.83	18.99	1.25	1.05
Deletion total[f]	**16.20**	**20.12**	**9.25**	**19.88**	**24.75**	**60.30**	**36.33**	**50.91**	**13.25**	**9.28**
% Deletion[g]	**15.50**	**25.29**	**13.65**	**25.51**	**24.94**	**61.91**	**33.18**	**66.86**	**8.28**	**9.38**
A/T+1	3.20	2.35	1.75	0.87	5.25	5.51	2.50	-.61	4.50	2.52
G/G+1	0.40	0.09	0	0	0.75	0.16	0.17	0.06	0	0
>1 bp insert	1.00	0.27	0.75	0.15	0.50	0.08	2.67	0.75	0.25	0.03
Insertion total[h]	**4.60**	**2.71**	**2.5**	**1.02**	**6.50**	**5.75**	**5.33**	**1.41**	**4.75**	**2.55**
% Insertion[i]	**4.40**	**3.40**	**3.69**	**1.30**	**6.55**	**5.91**	**5.01**	**1.85**	**2.97**	**2.58**
MNV	0.40	3.36	0.50	0.12	0.25	0.03	0.33	0.10	0.25	0.04
Replacement	0.60	0.52	0.25	0.04	0	0	0.50	0.07	0	0
Complex total[j]	**1.00**	**3.88**	**0.75**	**0.16**	**0.25**	**0.03**	**0.83**	**0.17**	**0.25**	**0.04**
% Complex[k]	**0.96**	**4.87**	**1.11**	**0.21**	**0.25**	**0.03**	**0.78**	**0.22**	**0.16**	**0.04**
Sum total[l]	**104.40**	**79.56**	**67.75**	**77.92**	**99.25**	**97.41**	**106.50**	**76.14**	**160**	**99.02**
Counts/Freq	**1.31**		**0.87**		**0.92**		**1.40**		**1.62**	

[a] Number of replicates

[b] Sum of unique variants observed within a genotype divided by the number of replicates

[c] Sum of variant frequencies observed within a genotype divided by the number of replicates

[d] Average total number of SNVs per replicate for each genotype

[e] Percentage of total variants that were SNVs

[f] Average total number of deletions per replicate for each genotype

[g] Percentage of total variants that were deletions

[h] Average total number of insertions per replicate for each genotype

[i] Percentage of total variants that were insertions

[j] Average total number of complex mutations per replicate for each genotype

[k] Percentage of total variants that were complex mutations

[l] Sum of each type of mutation

Table 2 (cont'd). Mutation spectra								
	***rnr1D57N* (7)[a]**		***rnr1D57N msh2Δ* (4)**		***rnr1D57N msh3Δ* (4)**		***rnr1D57N msh6Δ* (4)**	
Variants	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**
CG>AT	14.00	13.48	12.00	4.58	14.25	7.77	31.00	25.84
CG>GC	4.71	1.40	0.25	0.05	6.25	1.72	1.75	1.36
CG>TA	25.43	9.55	41.5	20.23	20.75	18.49	59.25	52.16
TA>AT	2.57	1.19	0.50	0.30	3.00	6.35	3.75	0.98
TA>CG	3.00	1.03	4.50	1.66	3.25	1.38	9.00	2.27
TA>GC	3.14	4.09	1.00	0.32	3.25	1.16	4.75	2.09
SNV total[d]	**52.86**	**30.74**	**59.75**	**27.14**	**50.75**	**36.87**	**109.50**	**85.79**
% SNV[e]	**75.67**	**37.04**	**70.71**	**28.33**	**60.06**	**47.71**	**86.39**	**86.56**
A/T-1	6.57	12.99	13.50	49.23	11.00	10.17	8.25	3.71
G/G-1	3.14	35.34	4.50	2.81	11.75	11.87	2.50	0.56
>1 bp del	2.14	1.30	0	0	5.25	15.11	0.25	4.08
Deletion total[f]	**11.86**	**49.62**	**18.00**	**52.04**	**28.00**	**37.15**	**11.00**	**8.35**
% Deletion[g]	**16.97**	**59.80**	**21.30**	**54.32**	**33.14**	**48.08**	**8.68**	**8.43**
A/T+1	2.86	1.65	5.25	16.19	3.25	2.54	5.50	4.85
G/G+1	0.57	0.37	0.25	0.03	1.75	0.15	0.25	0.04
>1 bp insert	0.57	0.32	1.25	0.39	0.50	0.53	0	0
Insertion total[h]	**4.00**	**2.34**	**6.75**	**16.61**	**5.50**	**3.22**	**5.75**	**4.89**
% Insertion[i]	**5.73**	**2.82**	**7.99**	**17.34**	**6.51**	**4.17**	**4.54**	**4.93**
MNV	0.57	0.14	0	0	0.25	0.04	0.50	0.85
Replacement	0.57	0.15	0	0	0	0	0	0
Complex total[j]	**1.14**	**0.28**	**0**	**0**	**0.25**	**0.04**	**0.50**	**0.09**
% Complex[k]	**1.64**	**0.35**	**0**	**0**	**0.30**	**0.05**	**0.40**	**0.09**
Sum total[l]	**69.86**	**82.99**	**84.50**	**95.79**	**84.50**	**77.28**	**126.75**	**99.12**
Counts/Freq.	**0.84**		**0.88**		**1.09**		**1.28**	

Table 2 (cont'd). Mutation spectra								
	***rnr1Y285F* (8)[a]**		***rnr1Y285F msh2Δ* (4)**		***rnr1Y285F msh3Δ* (4)**		***rnr1Y285F msh6Δ* (4)**	
Variants	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**
CG>AT	20.25	40.21	39.75	33.14	47.25	50.42	6.00	4.78
CG>GC	5.00	6.52	0.25	0.10	0	0	1.50	7.48
CG>TA	39.00	20.33	45.00	30.34	48.00	53.54	22.00	4.03
TA>AT	3.00	2.42	0.25	0.04	0.50	0.11	0	0
TA>CG	1.50	0.34	1.25	0.35	2.50	1.13	0.50	0.08
TA>GC	3.00	1.04	0.75	0.26	0.50	0.44	0.50	0.16
SNV total[d]	**71.75**	**70.86**	**87.25**	**64.23**	**98.75**	**105.64**	**30.50**	**16.53**
% SNV[e]	**85.67**	**91.83**	**77.56**	**61.75**	**94.95**	**98.26**	**57.28**	**17.61**
A/T-1	4.63	1.30	13.00	28.56	3.75	1.28	4.50	2.45
G/G-1	4.38	3.79	7.75	9.26	0	0	15.00	74.24
>1 bp del	1.25	0.26	0.50	0.13	0	0	3.00	0.57
Deletion total[f]	**10.25**	**5.35**	**21.25**	**37.95**	**3.75**	**1.28**	**22.50**	**77.27**
% Deletion[g]	**12.24**	**6.93**	**18.89**	**36.49**	**3.61**	**1.19**	**42.25**	**82.36**
A/T+1	1.50	0.78	3.75	1.77	1.50	0.59	0.25	0.03
G/G+1	0.13	0.11	0.25	0.06	0	0	0	0
>1 bp insert	0.13	0.08	0	0	0	0	0	0
Insertion total[h]	**1.75**	**0.96**	**4.00**	**1.83**	**1.50**	**0.59**	**0.25**	**0.03**
% Insertion[i]	**2.09**	**1.25**	**3.56**	**1.76**	**1.44**	**0.55**	**0.47**	**0.03**
MNV	0	0	0	0	0	0	0	0
Replacement	0	0	0	0	0	0	0	0
Complex total[j]	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
% Complex[k]	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Sum total[l]	**83.75**	**77.16**	**112.50**	**104.01**	**104.00**	**107.51**	**53.25**	**93.82**
Counts/Freq	**1.09**		**1.08**		**0.97**		**0.57**	

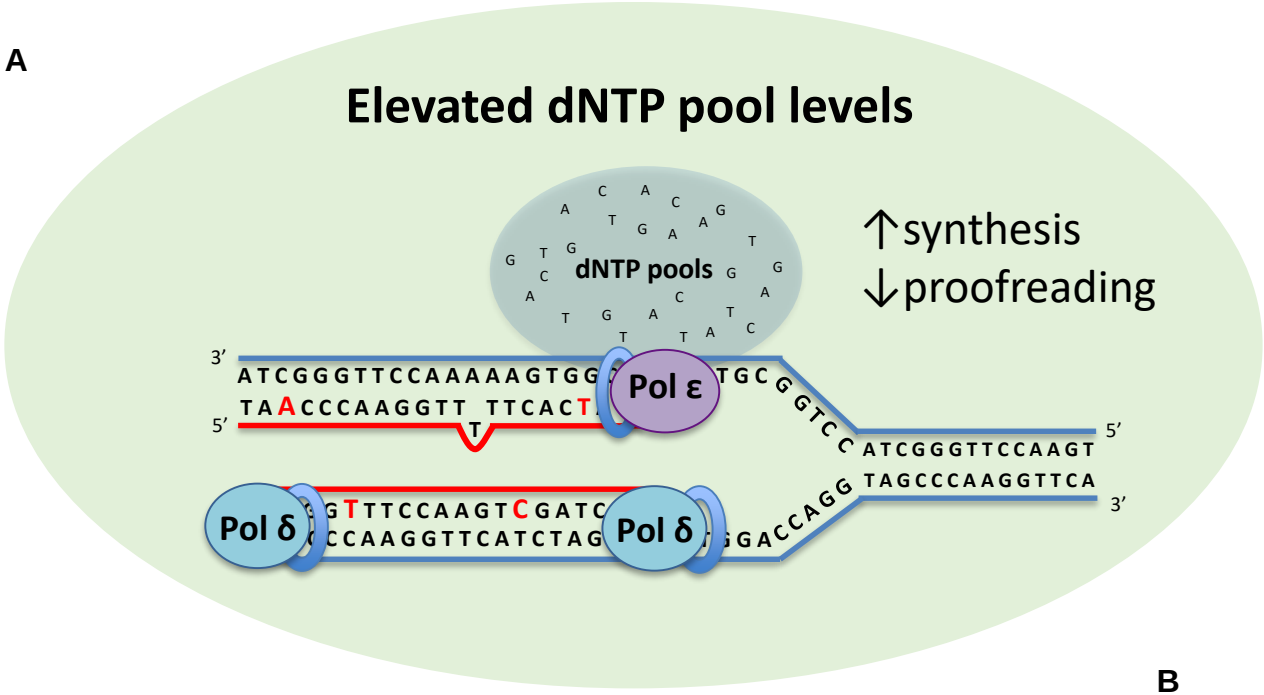
Table 2 (cont'd). Mutation spectra						
	***rnr1Y285F pGAL-RNR1* (6)[a]**		***rnr1Y285F pGAL-RNR1 msh2Δ* (6)**		***rnr1Y285F pGAL-RNR1 msh6Δ* (4)**	
Variants	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**
CG>AT	35.00	21.51	34.83	30.05	41.00	38.98
CG>GC	11.33	3.38	0.17	0.15	0.50	0.13
CG>TA	47.50	20.64	44.67	30.85	34,50	40.57
TA>AT	7.33	2.08	1.17	0.22	1.50	2.88
TA>CG	2.50	6.26	1.00	0.19	3.25	3.06
TA>GC	5.17	7.27	1.50	0.32	0.50	0.16
SNV total[d]	**108.83**	**61.13**	**83.33**	**61.79**	**81.25**	**85.78**
% SNV[e]	**80.22**	**81.63**	**72.15**	**61.81**	**92.86**	**81.87**
A/T-1	10.83	3.77	15.00	22.92	4.00	18.02
G/G-1	6.50	2.70	11.17	11.07	1.25	0.30
>1 bp del	4.00	1.19	0.50	0.09	0	0
Deletion total[f]	**21.33**	**7.66**	**26.67**	**34.07**	**5.25**	**18.31**
% Deletion[g]	**15.73**	**10.22**	**23.09**	**34.09**	**6.00**	**17.48**
A/T+1	3.33	4.77	4.50	3.83	1.00	0.69
G/G+1	0	0	0	0	0	0
>1 bp insert	1.00	0.75	0.83	0.25	0	0
Insertion total[h]	**4.33**	**5.53**	**5.33**	**4.07**	**1.00**	**0.69**
% Insertion[i]	**3.19**	**7.38**	**4.62**	**4.08**	**1.14**	**0.66**
MNV	0.83	0.50	0.17	0.02	0	0
Replacement	0.33	0.07	0	0	0	0
Complex total[j]	**1.17**	**0.57**	**0.17**	**0.02**	**0**	**0**
% Complex[k]	**0.86**	**0.76**	**0.14**	**0.02**	**0**	**0**
Sum total[l]	**135.67**	**74.88**	**115.50**	**99.95**	**87.50**	**83.80**
Counts/Freq.	**1.81**		**1.16**		**1.04**	

Table 2 (cont'd). Mutation spectra								
	***rnr1Y285A* (5)[a]**		***rnr1Y285A msh2Δ* (7)**		***rnr1Y285A msh3Δ* (4)**		***rnr1Y285A msh6Δ* (5)**	
Variants	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**
CG>AT	27.40	38.48	36.29	44.68	3.25	1.40	38.60	1 67.69
CG>GC	1.20	0.32	0.14	0.04	0	0	0	0
CG>TA	34.00	13.68	36.43	25.48	20.25	3.23	33.40	25.62
TA>AT	1.40	3.38	0.43	0.24	0	0	1.20	0.78
TA>CG	1.00	0.27	0.86	0.26	1.25	0.18	4.00	2.90
TA>GC	2.00	2.45	0.14	0.03	0	0	4.60	3.25
SNV total[d]	**67.00**	**58.57**	**74.29**	**70.72**	**24.75**	**4.82**	**81.80**	**99.24**
% SNV[e]	**79.95**	**60.41**	**76.36**	**62.3**	**58.58**	**4.71**	**88.72**	**93.08**
A/T-1	4.20	9.38	9.71	23.59	4.50	0.97	8.80	7.06
G/G-1	10.20	27.98	9.57	16.61	13.00	96.45	0.80	0.18
>1 bp del	0.80	0.27	0	0	0	0	0	0
Deletion total[f]	**15.20**	**37.63**	**19.29**	**40.20**	**17.50**	**97.42**	**9.60**	**7.24**
% Deletion[g]	**18.14**	**38.81**	**19.82**	**35.43**	**41.42**	**95.29**	**10.41**	**6.79**
A/T+1	1.40	0.73	3.00	2.13	0	0	0.80	0.15
G/G+1	0	0	0.14	0.05	0	0	0	0
>1 bp insert	0.20	0.03	0.57	0.36	0	0	0	0
Insertion total[h]	**1.60**	**0.75**	**3.71**	**2.54**	**0**	**0**	**0.80**	**0.15**
% Insertion[i]	**1.91**	**0.78**	**3.82**	**2.24**	**0**	**0**	**0.87**	**0.14**
MNV	0	0	0	0	0	0	0	0
Replacement	0	0	0	0	0	0	0	0
Complex total[j]	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
% Complex[k]	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Sum total[l]	**83.80**	**96.95**	**97.29**	**113.46**	**42.25**	**102.24**	**92.20**	**106.62**
Counts/Freq	**0.86**		**0.86**		**0.41**		**0.87**	

Table 2 (cont'd). Mutation spectra								
	***rnr1Y285A pGAL-RNR1* (6)[a]**		***rnr1Y285A pGAL-RNR1 msh2Δ* (4)**		***rnr1Y285A pGAL-RNR1 msh3Δ* (4)**		***rnr1Y285A pGAL-RNR1 msh6Δ* (4)**	
Variants	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**	**Counts[b]**	**Freq.[c]**
CG>AT	26.33	31.68	25.75	30.55	3.25	3.61	21.50	41.39
CG>GC	1.17	0.23	1.25	0.18	0	0	0.50	0.11
CG>TA	33.17	8.73	41.00	18.80	18.00	3.57	17.75	7.64
TA>AT	1.83	2.26	1.00	0.25	0.50	0.19	2.00	2.01
TA>CG	2.00	0.31	3.50	0.93	0.75	0.12	0.75	0.12
TA>GC	1.50	0.75	3.00	3.21	0	0	2.00	0.66
SNV total[d]	**66.00**	**43.96**	**75.50**	**53.91**	**22.50**	**7.49**	**44.50**	**51.93**
% SNV[e]	**75.29**	**44.59**	**71.89**	**53.22**	**56.96**	**7.45**	**76.72**	**49.20**
A/T-1	7.00	3.40	15.25	29.66	3.50	1.39	1.75	1.98
G/G-1	12.00	50.42	9.75	9.03	13.50	91.61	10.25	50.09
>1 bp del	0.33	0.20	0.075	0.19	0	0	0.50	0.75
Deletion total[f]	**19.33**	**54.02**	**25.75**	**38.88**	**17.00**	**93.00**	**12.50**	**52.81**
% Deletion[g]	**22.05**	**54.79**	**24.18**	**38.38**	**43.04**	**92.55**	**21.55**	**50.03**
A/T+1	1.833	0.45	4.25	6.00	0	0	1.00	0.81
G/G+1	0	0	0	0	0	0	0	0
>1 bp insert	0.17	0.05	1.00	2.50	0	0	0	0
Insertion total[h]	**2.00**	**0.50**	**5.25**	**8.50**	**0**	**0**	**1.00**	**0.81**
% Insertion[i]	**2.28**	**0.511**	**4.93**	**8.40**	**0**	**0**	**1.72**	**0.77**
MNV	0.17	0.05	0	0	0	0	0	0
Replacement	0.17	0.05	0	0	0	0	0	0
Complex total[j]	**0.33**	**0.10**	**0**	**0**	**0**	**0**	**0**	**0**
% Complex[k]	**0.38**	**0.10**	**0**	**0**	**0**	**0**	**0**	**0**
Sum total[l]	**87.67**	**98.58**	**106.50**	**101.28**	**39.50**	**100.48**	**58.00**	**105.55**
	0.89		**1.05**		**0.39**		**0.55**	

Figure 1.

A


Elevated dNTP pool levels
dNTP pools
↑synthesis
↓proofreading
3'
5'
Pol ε
Pol δ
Pol δ
ATCGGGTTCCAAGT
TAGCCCAAGGTTCA

B

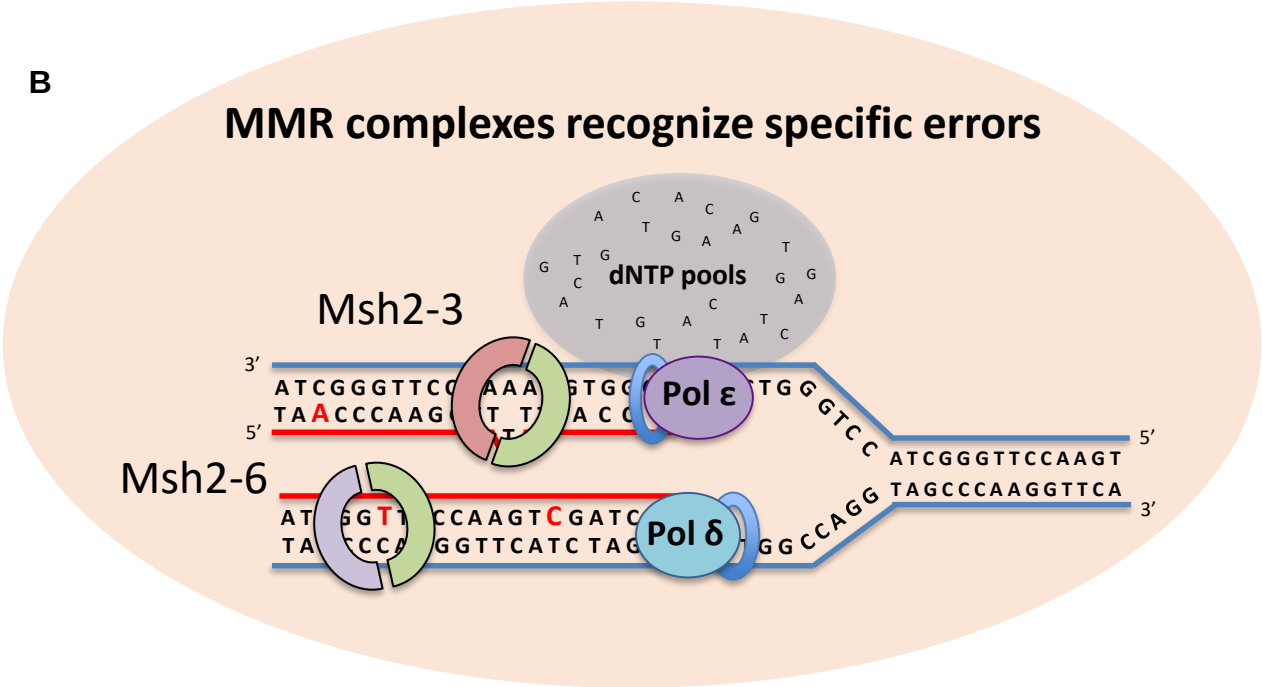

MMR complexes recognize specific errors
dNTP pools
Msh2-3
Msh2-6
3'
5'
Pol ε
Pol δ
ATCGGGTTCCAAGT
TAGCCCAAGGTTCA

Figure 2.

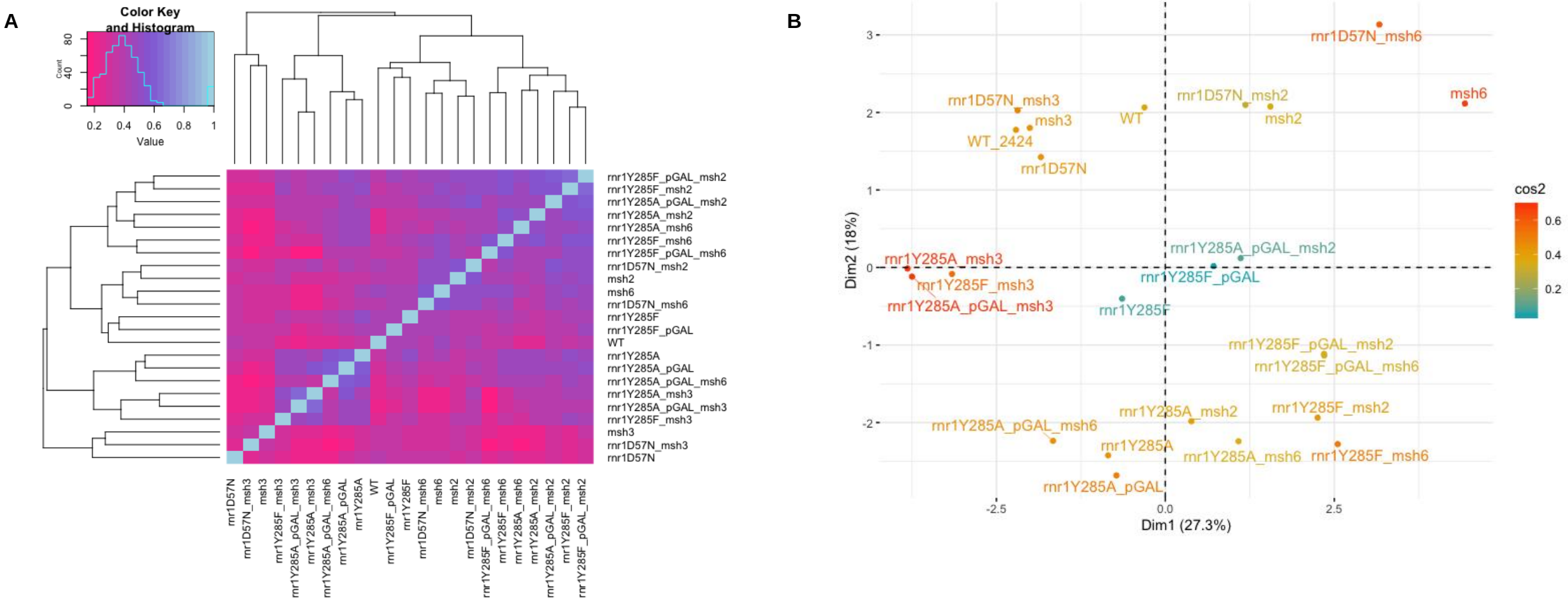

A
Color Key and Histogram
Count
0
40
80
0.2
0.4
0.6
0.8
1
Value
rnr1Y285F_pGAL_msh2
rnr1Y285F_msh2
rnr1Y285A_pGAL_msh2
rnr1Y285A_msh2
rnr1Y285A_msh6
rnr1Y285F_msh6
rnr1Y285F_pGAL_msh6
rnr1D57N_msh2
msh2
msh6
rnr1D57N_msh6
rnr1Y285F
rnr1Y285F_pGAL
WT
rnr1Y285A
rnr1Y285A_pGAL
rnr1Y285A_pGAL_msh6
rnr1Y285A_msh3
rnr1Y285A_pGAL_msh3
rnr1Y285F_msh3
msh3
rnr1D57N_msh3
rnr1D57N
B
rnr1D57N_msh6
rnr1D57N_msh3
msh3
WT
rnr1D57N_msh2
msh2
msh6
WT_2424
rnr1D57N
rnr1Y285A_msh3
rnr1Y285A_pGAL_msh2
rnr1Y285F_msh3
rnr1Y285F_pGAL
rnr1Y285A_pGAL_msh3
rnr1Y285F
rnr1Y285F_pGAL_msh2
rnr1Y285F_pGAL_msh6
rnr1Y285A_msh2
rnr1Y285F_msh2
rnr1Y285A_pGAL_msh6
rnr1Y285A
rnr1Y285A_msh6
rnr1Y285F_msh6
rnr1Y285A_pGAL
Dim2 (18%)
Dim1 (27.3%)
-2.5
0.0
2.5
3
2
1
0
-1
-2
cos2
0.6
0.4
0.2

Figure 3.

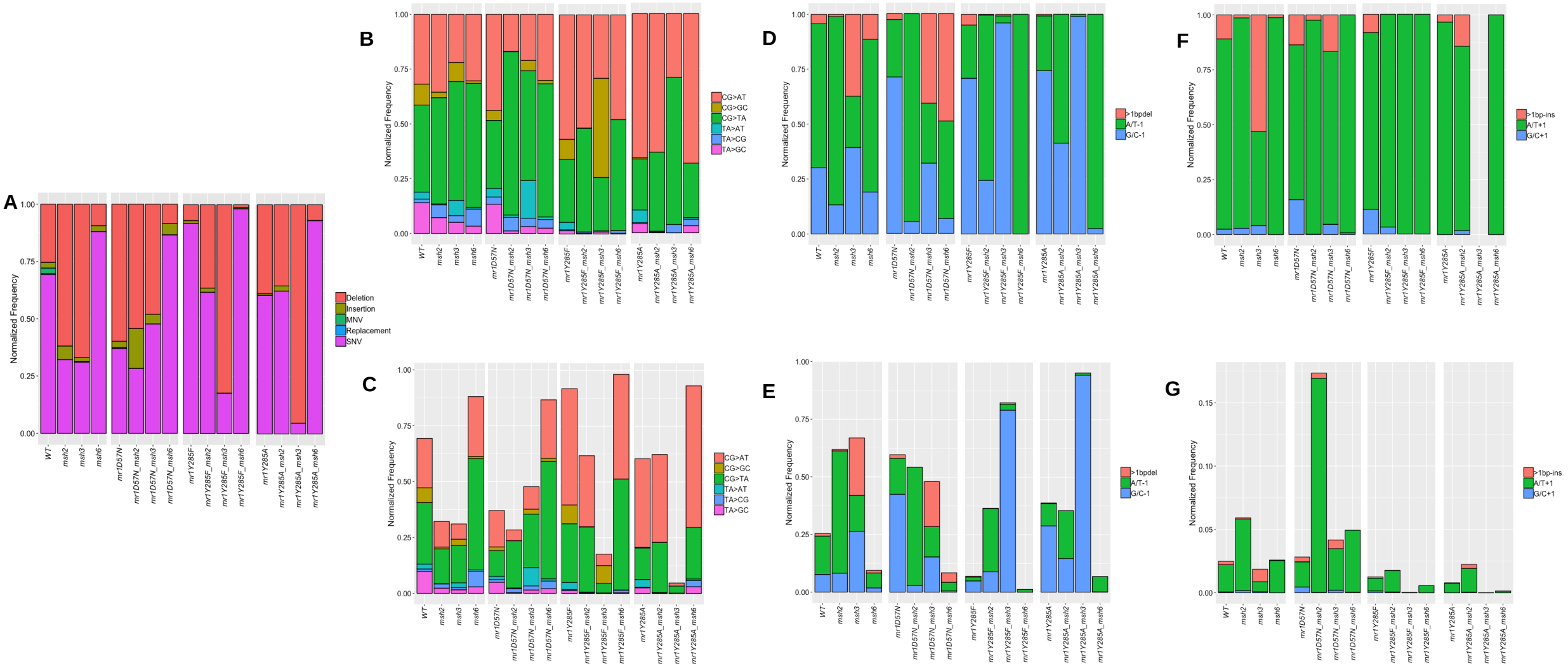

A
B
C
D
E
F
G
Normalized Frequency
Deletion
Insertion
MNV
Replacement
SNV
CG>AT
CG>GC
CG>TA
TA>AT
TA>CG
TA>GC
>1bpdel
A/T-1
G/C-1
>1bp-ins
A/T+1
G/C+1
WT
msh2
msh3
msh6
mr1D57N
mr1D57N_msh2
mr1D57N_msh3
mr1D57N_msh6
mr1Y285F
mr1Y285F_msh2
mr1Y285F_msh3
mr1Y285F_msh6
mr1Y285A
mr1Y285A_msh2
mr1Y285A_msh3
mr1Y285A_msh6

Figure 4.

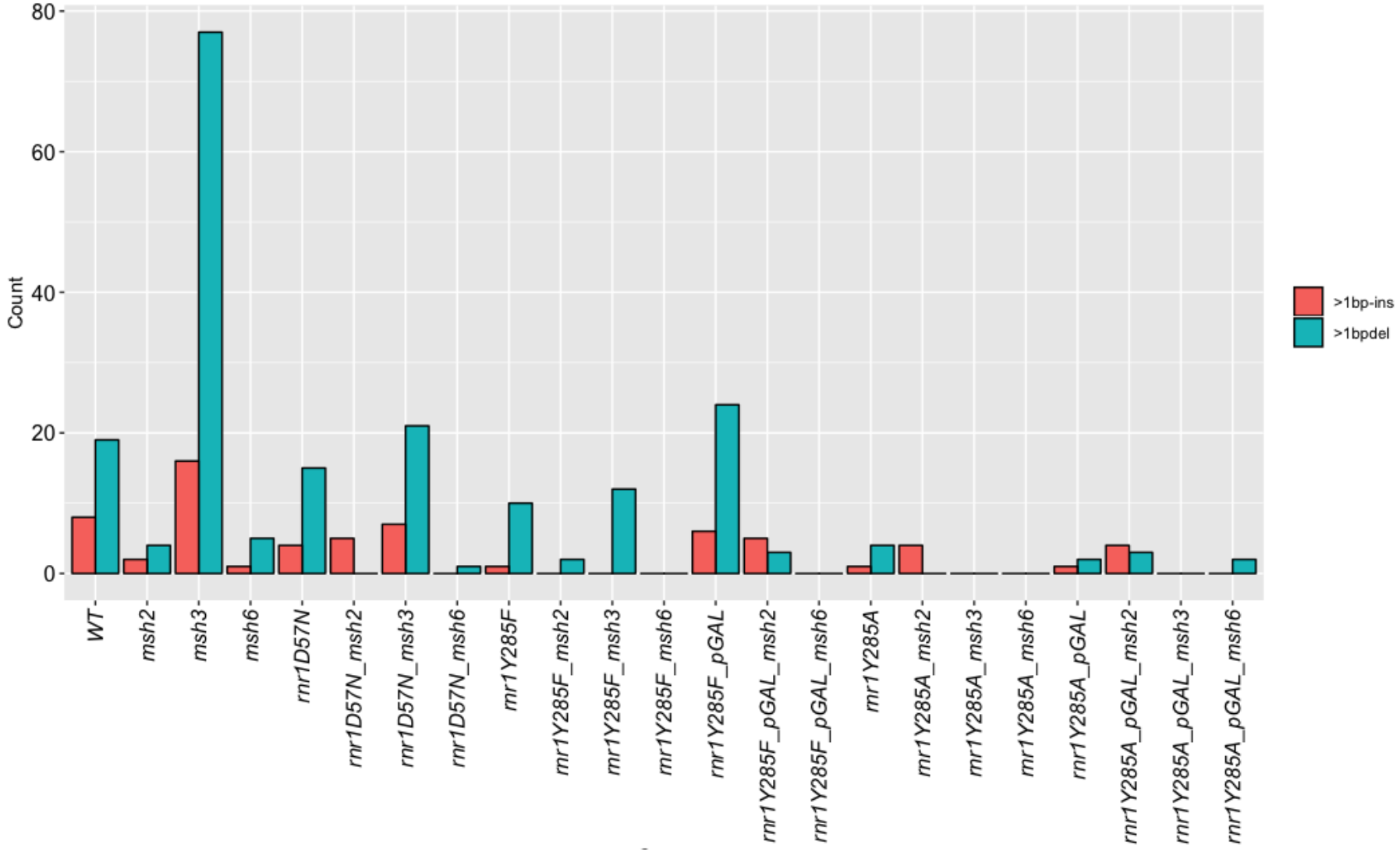

Count
80
60
40
20
0
WT
msh2
msh3
msh6
rnr1D57N
rnr1D57N_msh2
rnr1D57N_msh3
rnr1D57N_msh6
rnr1Y285F
rnr1Y285F_msh2
rnr1Y285F_msh3
rnr1Y285F_msh6
rnr1Y285F_pGAL
rnr1Y285F_pGAL_msh2
rnr1Y285F_pGAL_msh6
rnr1Y285A
rnr1Y285A_msh2
rnr1Y285A_msh3
rnr1Y285A_msh6
rnr1Y285A_pGAL
rnr1Y285A_pGAL_msh2
rnr1Y285A_pGAL_msh3
rnr1Y285A_pGAL_msh6
>1bp-ins
>1bpdel

Figure 5.

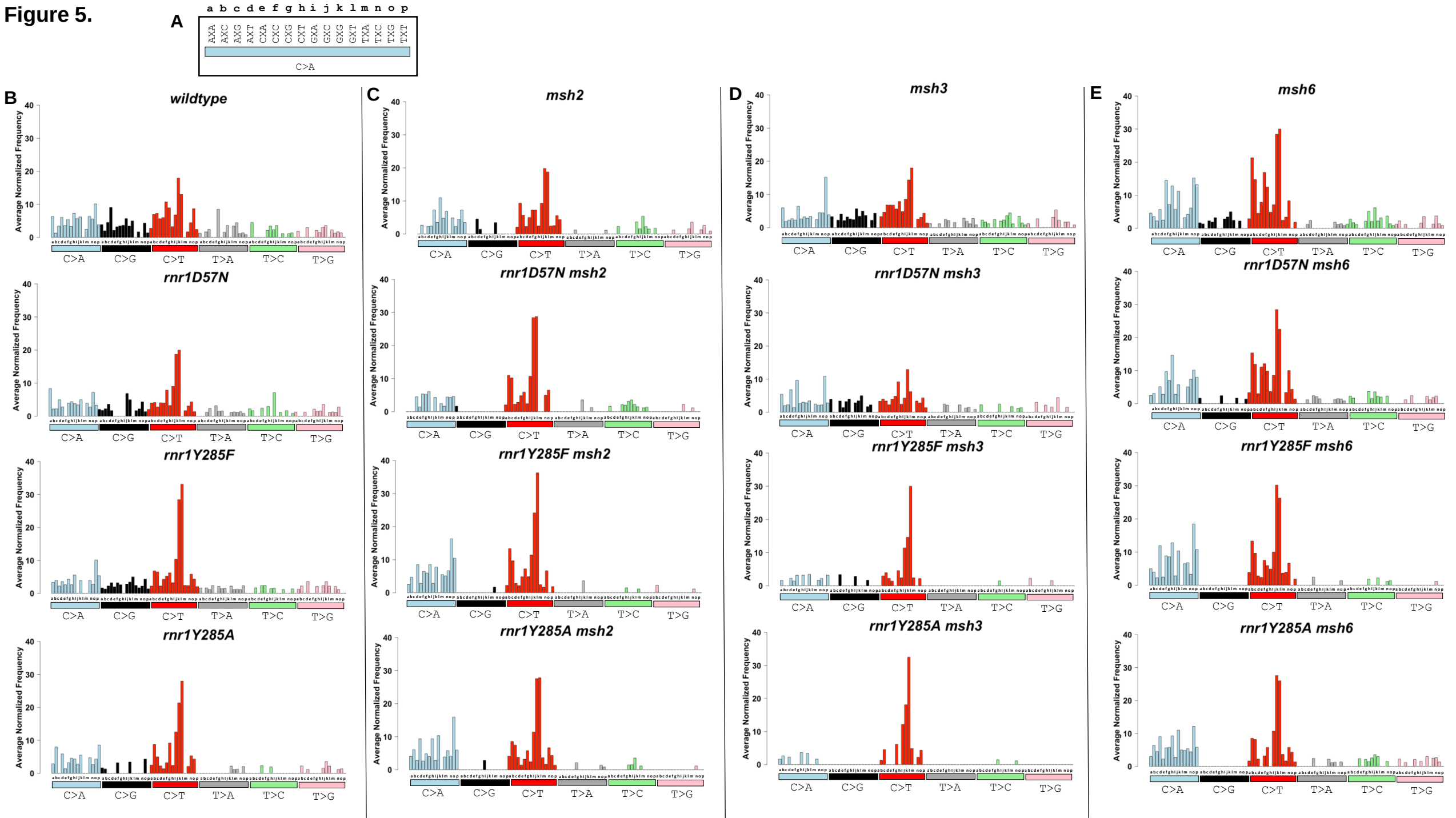

A
a b c d e f g h i j k l m n o p
AXA AXC AXG AXT CXA CXC CXG CXT GXA GXC GXG GXT TXA TXC TXG TXT
C>A
B
wildtype
rnr1D57N
rnr1Y285F
rnr1Y285A
C
msh2
rnr1D57N msh2
rnr1Y285F msh2
rnr1Y285A msh2
D
msh3
rnr1D57N msh3
rnr1Y285F msh3
rnr1Y285A msh3
E
msh6
rnr1D57N msh6
rnr1Y285F msh6
rnr1Y285A msh6
Average Normalized Frequency
0
10
20
30
40
C>A
C>G
C>T
T>A
T>C
T>G

Figure 6.

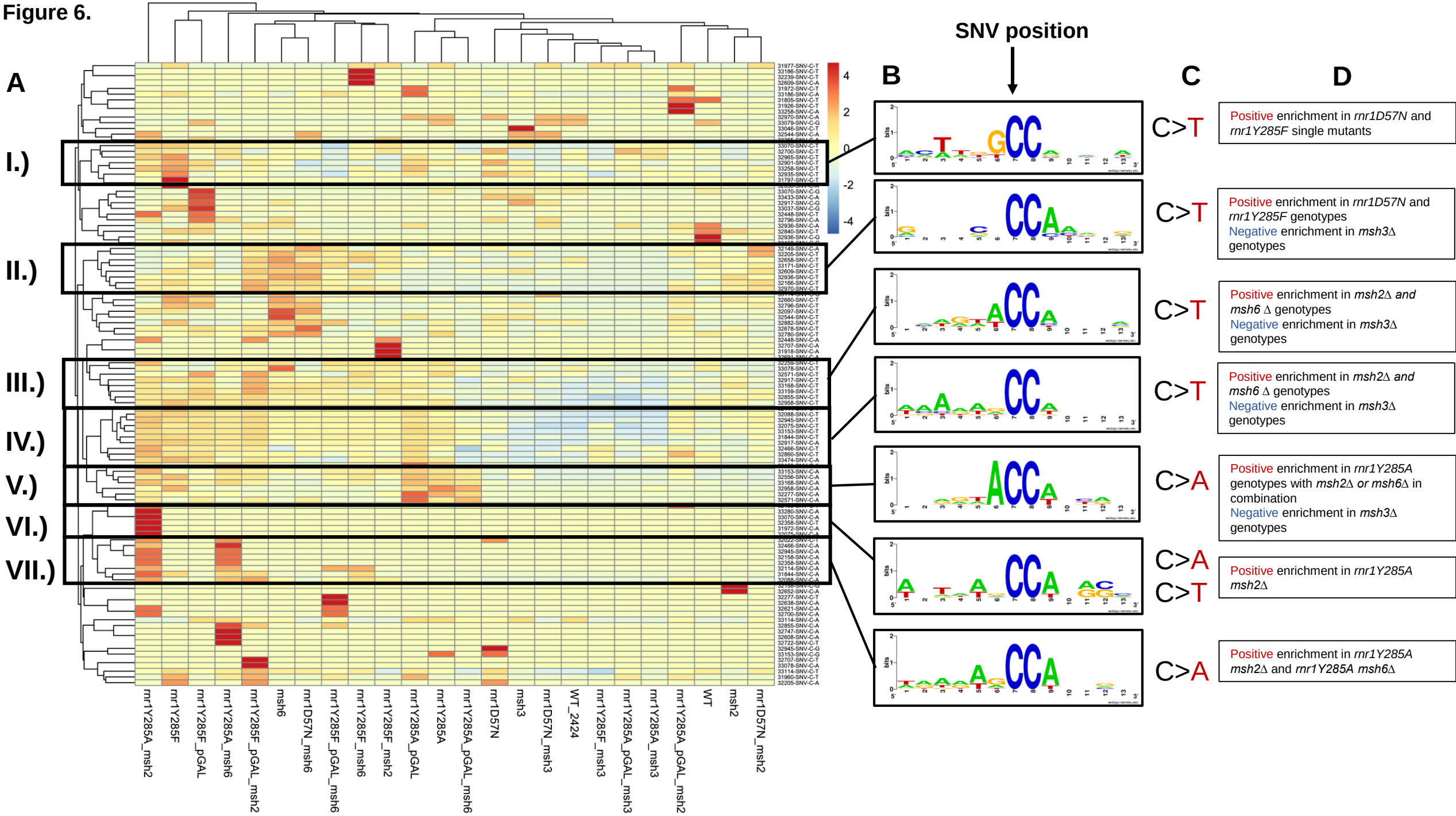

A
I.)
II.)
III.)
IV.)
V.)
VI.)
VII.)
4
2
0
-2
-4
rnr1Y285A_msh2
rnr1Y285F
rnr1Y285F_pGAL
rnr1Y285A_msh6
rnr1Y285F_pGAL_msh2
msh6
rnr1D57N_msh6
rnr1Y285F_pGAL_msh6
rnr1Y285F_msh6
rnr1Y285F_msh2
rnr1Y285A_pGAL
rnr1Y285A
rnr1Y285A_pGAL_msh6
rnr1D57N
msh3
rnr1D57N_msh3
WT_2424
rnr1Y285F_msh3
rnr1Y285A_pGAL_msh3
rnr1Y285A_msh3
rnr1Y285A_pGAL_msh2
WT
msh2
rnr1D57N_msh2
SNV position
B
C
D
C>T
Positive enrichment in rnr1D57N and rnr1Y285F single mutants
C>T
Positive enrichment in rnr1D57N and rnr1Y285F genotypes
Negative enrichment in msh3Δ genotypes
C>T
Positive enrichment in msh2Δ and msh6 Δ genotypes
Negative enrichment in msh3Δ genotypes
C>T
Positive enrichment in msh2Δ and msh6 Δ genotypes
Negative enrichment in msh3Δ genotypes
C>A
Positive enrichment in rnr1Y285A genotypes with msh2Δ or msh6Δ in combination
Negative enrichment in msh3Δ genotypes
C>A
C>T
Positive enrichment in rnr1Y285A msh2Δ
C>A
Positive enrichment in rnr1Y285A msh2Δ and rnr1Y285A msh6Δ

Figure 7.

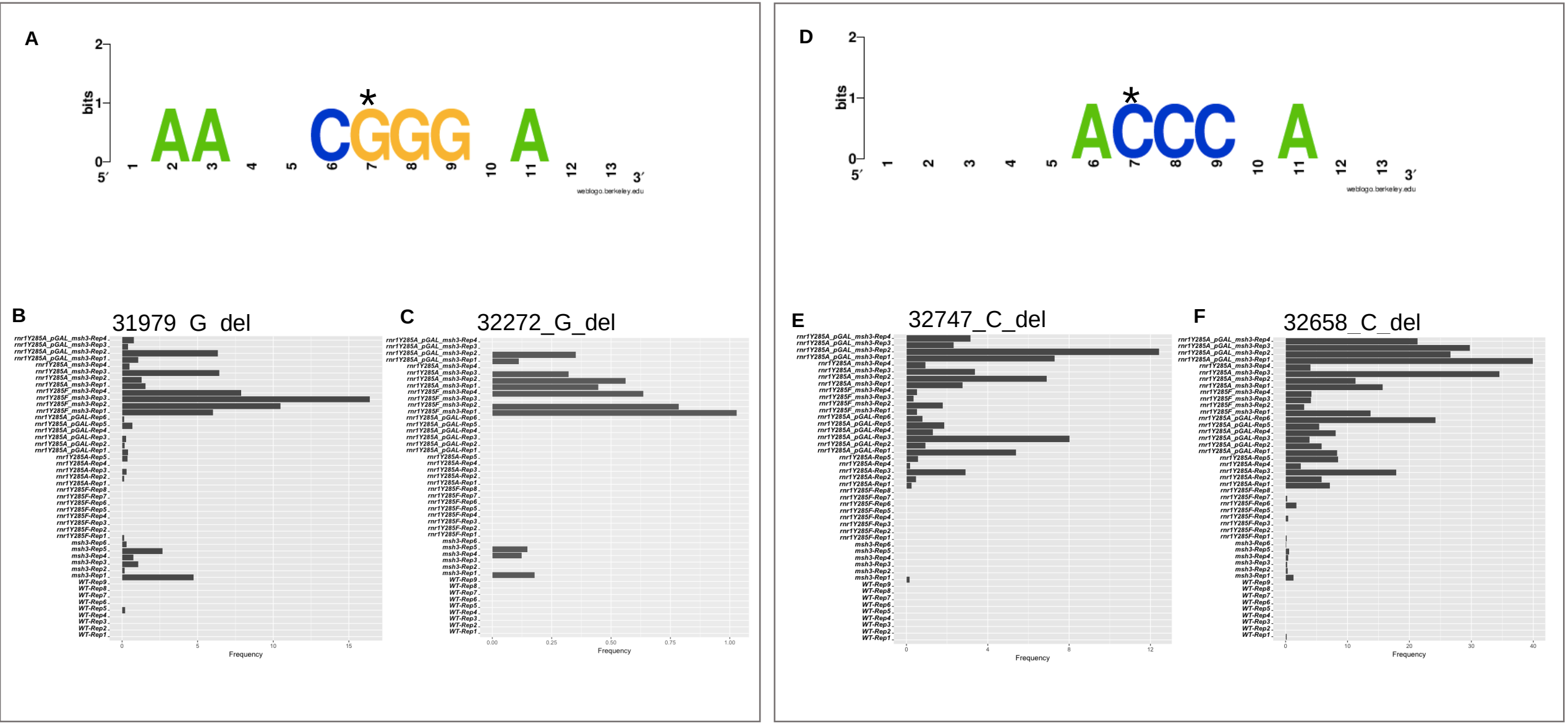
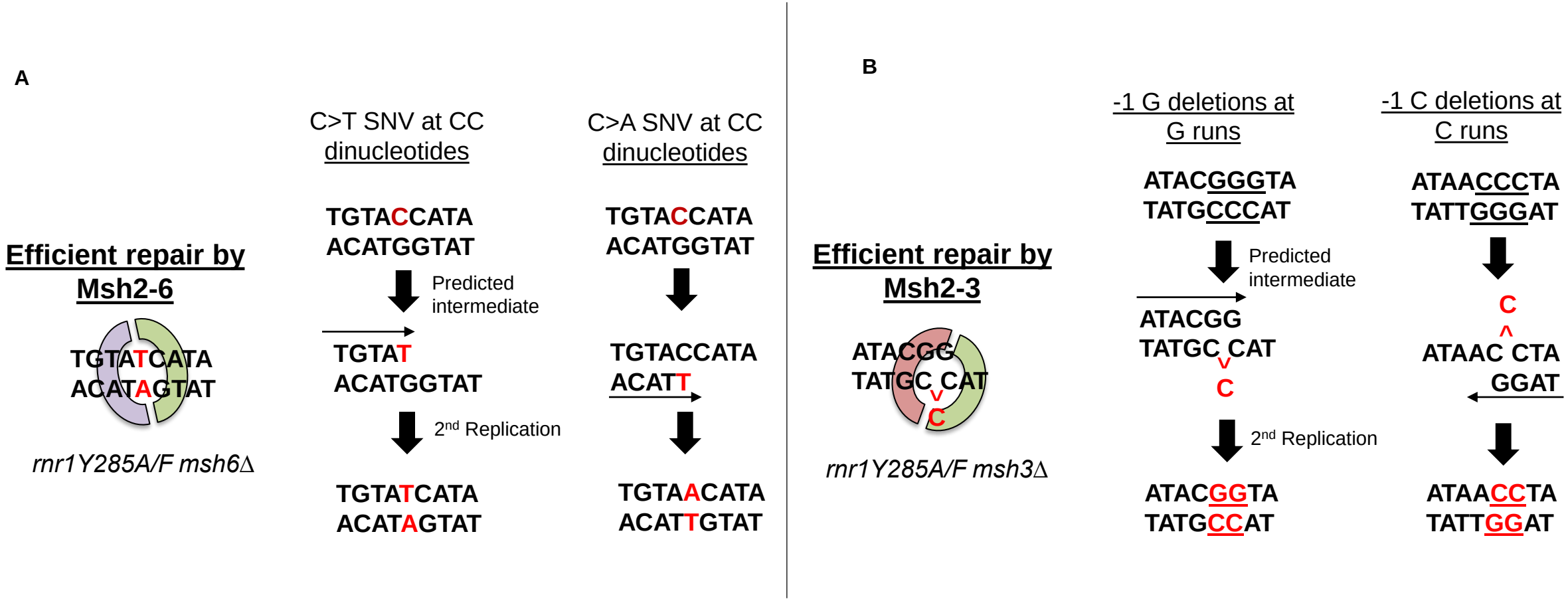

A
bits
AA CGGG A
*
5′
3′
weblogo.berkeley.edu
B
31979_G_del
C
32272_G_del
D
ACCC A
*
E
32747_C_del
F
32658_C_del
Frequency

Figure 8.


A
Efficient repair by Msh2-6
TGTATCATA
ACATAGTAT
rnr1Y285A/F msh6Δ
C>T SNV at CC dinucleotides
TGTACCATA
ACATGGTAT
Predicted intermediate
TGTAT
ACATGGTAT
2nd Replication
TGTATCATA
ACATAGTAT
C>A SNV at CC dinucleotides
TGTACCATA
ACATGGTAT
TGTACCATA
ACATT
TGTAACATA
ACATTGTAT
B
Efficient repair by Msh2-3
ATACGG
TATGC CAT
C
rnr1Y285A/F msh3Δ
-1 G deletions at G runs
ATACGGGTA
TATGCCCAT
Predicted intermediate
ATACGG
TATGC CAT
C
2nd Replication
ATACGGTA
TATGCCAT
-1 C deletions at C runs
ATAACCCTA
TATTGGGAT
C
ATAAC CTA
GGAT
ATAACCTA
TATTGGAT