Targeted next-generation sequencing reveals complex mutation spectra in *rnr1 msh* genetic backgrounds

1. Dept. of Biochemistry, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, New York, 14203
2. University at Buffalo Genomics and Bioinformatics Core, Buffalo, New York 14203

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

#ACS RSG-14-235-01

Abstract

Multiple pathways contribute to maintaining high-fidelity DNA replication, including the regulation of free dNTP levels and both the selectivity and exonuclease domains of the replicative polymerases. Elevated free dNTP levels are a well-established source of mutagenesis due to increased DNA polymerase error and decreased proof-reading, and are likely a hallmark of cancer cells. Replication errors are typically substrates for the mismatch repair (MMR) system, which recognizes misincorporation and insertion/deletion errors and targets them for repair. Here, we developed a high-throughput, targeted deep-sequencing approach to examine the consequences of elevated dNTP pools and DNA polymerase error repair via MMR, both alone and in combination. The combination of altered dNTP pools and compromised MMR together has the potential to alter the mutational landscape. Importantly, mutation signatures of human tumors are used to tailor cancer treatment and inform clinical outcome. We used *Saccharomyces cerevisiae* as a model system to document how mutation spectra are modified by different combinations of elevated dNTP levels and reduced MMR. We sequenced pools of mutated (canavanine-resistant) colonies with 1) altered dNTP pools (*rnr1* alleles), 2) with deletions in MMR recognition factors (*msh* alleles) and 3) combinations of *rnr1* and *msh* alleles. The depth of sequencing allowed us to delineate *CAN1* regions that are systematically susceptible to mutagenesis. We combined variant type and positional information to develop genotype-specific mutation fingerprints. We developed computational methods to quantify the contribution of two different genotypes to the underlying mutation spectra in double mutants, to assess additive, epistatic or synergistic effects. Individually, altered dNTP pools, even very modest changes, and compromised MMR led to distinct mutational profiles. Furthermore, the increased and altered mutation profiles in *rnr1* backgrounds allowed us to identify novel specificity of Msh2-Msh3 for single base deletions in repetitive GC runs, mutations commonly observed in MMR-deficient cancers. Notably, the mutation profiles of double mutants were not a simple combination of the single mutant signatures, indicating a more complex effect on mutagenesis. We propose that establishing mutation spectra from the ground up will provide useful information when interpreting mutational signatures in human tumors.

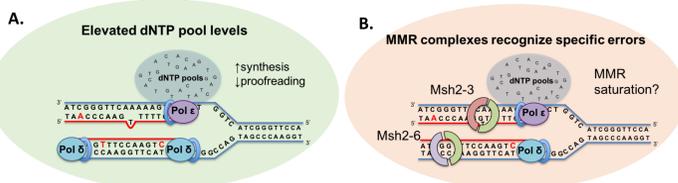## dNTP pool regulation is essential for high fidelity DNA replication

Figure 1. Elevations in dNTP levels lead to mutagenesis targeted to the MMR Pathway. A. The misregulation of free dNTP levels impacts the fidelity of DNA replication. Limiting dNTPs levels lead to fork arrest, possible fork collapse and double-strand DNA breaks. In contrast, elevated nucleotide pools shift the equilibrium of replicative polymerases towards synthesis over proofreading, referred to as next nucleotide effect, leading to increased mutagenesis. **B.** The MMR pathway directs repair of errors in DNA replication. The two MMR recognition complexes, Msh2-Msh6 and Msh2-Msh3, have separate but overlapping specificity for replication errors. Msh2-Msh6 directs repair of mispairs and small (1 base) insertion/deletion loops (in/dels); Msh2-Msh3 preferentially directs repair of larger in/dels. Elevated and altered dNTP levels increased mutagenesis and altered mutational profiles, allowing us to discover novel differential repair specificity of Msh2-Msh3 versus Msh2-Msh6.

Mutation rates increase synergistically when MMR is deleted in the presence of skewed increases in dCTP and dTTP
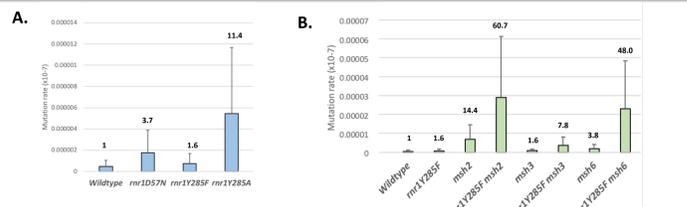

Figure 2. Mutation rates at can1. RNR modulates dNTP levels through allosteric regulation, increasing dNTPs to precise levels and ratios during DNA replication. A mutation in the activity site leads to a balanced 2-fold increase in dNTP levels (*rnr1D57N*). Mutations in the specificity site, *rnr1Y285F* and *rnr1Y285A*, lead to a 3-fold and 20-fold skewed increase in dCTP and dTTP, respectively. Mutation rates were determined using a canavanine resistance assay. Canavanine is a toxic analog of arginine and both are transported into the cell through the arginine permease transporter, encoded by *CAN1*. Inactivating mutations in *CAN1* lead to canavanine resistance. Growth was scored under selective and permissive conditions and mutation rates were calculated using FluCalc (http://flucalc.ase.tufts.edu/). Error bars indicate 95% confidence intervals. **A.** Mutation rates in *rnr1* alleles; *rnr1Y285A* results in the highest increase/imbalance in dNTPs and the highest mutation rate **B.** Mutation rates increase synergistically when *rnr1* alleles are paired with deletions in MMR genes (Fold change over wildtype above each bar). Shown are mutation rates in the *rnr1Y285F* background. Highest mutations rates were observed in the absence of MMR (*msh2Δ*). Loss of Msh2-6 (*msh6Δ*) or Msh2-Msh3 (*msh3Δ*), had differential effects, consistent with distinct substrate specificity.

A novel selection technique using next-generation sequencing to determine mutation spectra
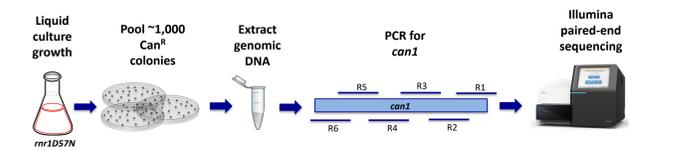

Figure 3. Pooled colony sequencing workflow. Past approaches to study the mutation spectra under elevated dNTP levels utilize Sanger sequencing on individual Can[R] colonies. We have developed a pooled colony selection approach combined with Illumina 2x300 paired-end sequencing technology to increase the number of genetic events analyzed. Strains with altered dNTP pools and/or deletions of MMR initation factors were grown in liquid culture, diluted and plated in the presence of canavanine to propagate individual colonies. ~1,000 Can[R] colonies were pooled and used as template for PCR. All of these colonies should have a least one mutation in the *CAN1* gene; *can1* was sequenced in 6 overlapping regions using Illumina MiSeq technology. Reads were aligned to the SacCer3 reference genome and variants were called using CLC Genomics Workbench. VCF files were analyzed using custom python scripting and R for downstream analysis.

Pooled colony approach increases total genetic events analyzed

	Xu 2008.	Chabes 2014.	Pooled Colony Approach
Average n/strain	19.5	191.5	1000
Strains sequenced	6	4	150
Total genetic events	117	766	150000

Table 1. The number of genetic events sequenced in previous Sanger approaches. The use of next-generation sequencing makes it feasible to sequence significantly more colonies from a greater number of genetic backgrounds, with over 100 times more events analyzed.

Permissive variant filtering removes background mutations
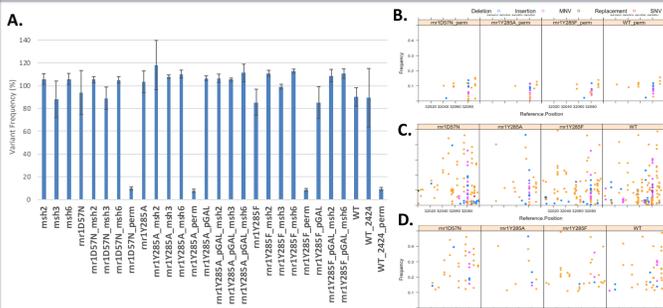

Figure 4. Samples grown under permissive conditions show low variant frequency. In addition to samples selected in the presence of canavanine, we sequenced pools of 1,000 colonies grown under permissive conditions in wildtype, *rnr1D57N*, *rnr1Y285F* and *rnr1Y285A* genotypes. **A.** Selected samples exhibited a total variant frequency of ~100%; permissive samples consistently exhibited variant frequencies of ~8%. We developed a permissive variant filter to remove low frequency variants that represent background mutagenesis in our assay. **B.** Variant frequency for one region of *can1* in permissive samples. **C.** The variants in the same region of *can1* in canavanine resistant samples. **D.** The variants from C after applying the permissive variant filter.

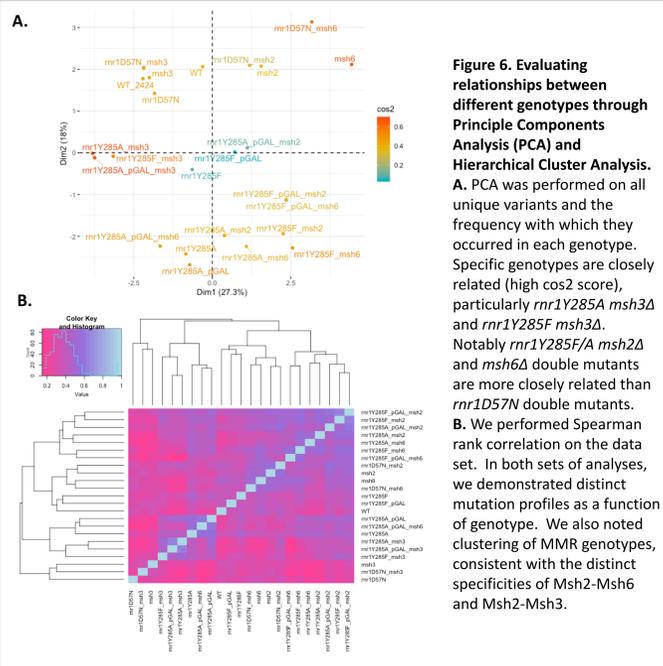## Distinct genotypes are uniquely related

Figure 6. Evaluating relationships between different genotypes through Principle Components Analysis (PCA) and Hierarchical Cluster Analysis. A. PCA was performed on all unique variants and the frequency with which they occurred in each genotype. Specific genotypes are closely related (high cos2 score), particularly *rnr1Y285A msh3Δ* and *rnr1Y285F msh3Δ*. Notably *rnr1Y285F/A msh2Δ* and *msh6Δ* double mutants are more closely related than *rnr1D57N* double mutants. **B.** We performed Spearman rank correlation on the data set. In both sets of analyses, we demonstrated distinct mutation profiles as a function of genotype. We also noted clustering of MMR genotypes, consistent with the distinct specificities of Msh2-Msh6 and Msh2-Msh3.

Elevations in dNTP levels and MMR deletions lead to distinct mutation spectra
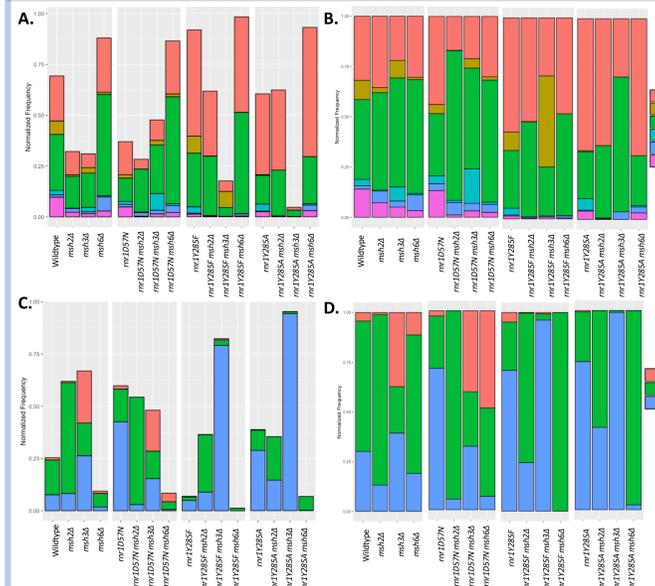

Figure 6. Single nucleotide variant (SNV) and Deletion mutation spectra. A. The normalized mutation frequency of the 6 different types of SNVs out of the total variants in a sample. In some backgrounds (*rnr1Y285F/A msh6Δ*) SNVs represent over 90% of variants **B.** The SNV mutation spectra normalized out of 100% SNVs. **C.** The mutation spectra of deletion frequency out of total variants. Notably *rnr1Y285F/A msh3Δ* are almost completely dominated by G/C deletions, indicating Msh2-Msh3 is uniquely able to correct this type of error; these errors are virtually eliminated in *rnr1Y285A msh6Δ*. **D.** The deletion spectra normalized out of 100% deletions.

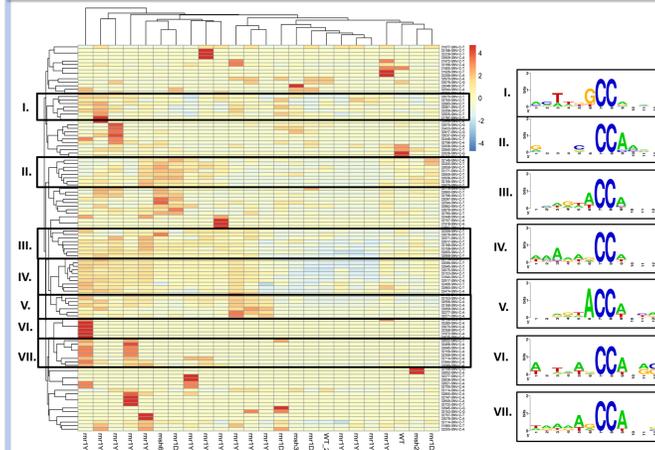## SNVs at CC dinucleotides are enriched in *msh6Δ* genotypes

Figure 7. Analysis at CC dinucleotide reveals differential enrichment of distinct errors Meta-analysis of the data revealed a significant proportion of SNVs occurred in dinucleotide context, specifically CC dinucleotides. We subset the data to include all positions across *CAN1* containing CC dinucleotides to analyze variant frequency across all genotypes. Notably there are distinct enrichment patterns which correspond to unique motifs surrounding the mutated dinucleotide. Of particular interest are III,IV, and V, which show negative enrichment in *msh3Δ* genotypes (blue), but high enrichment in *msh6Δ* genotypes (orange/red).

Single base G/C deletions are differentially enriched in *rnr1Y285A msh3Δ* genotypes
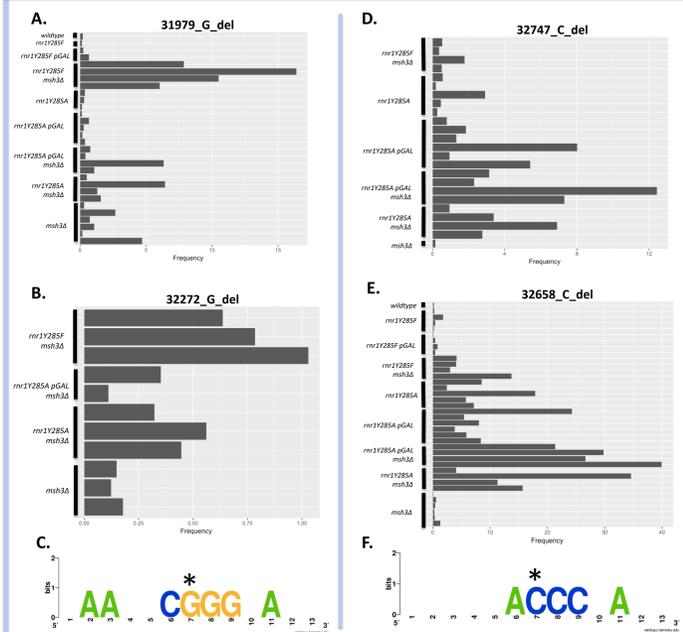

Figure 8. Analysis at CC dinucleotide reveals differential enrichment of distinct errors Meta-analysis also revealed a large proportion of single base deletions occurring in G/C context, particularly in *rnr1Y285F/A* and *msh3Δ* backgrounds. We examined the variants that occurred in multiple biological replicates and at the highest frequency in *rnr1Y285F/A msh3Δ* for epistatic or synergistic effects at these positions across *CAN1*. **A.** The frequency of the single base G deletion at position 31979 was synergistic in *rnr1Y285F msh3Δ* backgrounds. **B.** The variant frequency of G deletion at 32272 was synergistic in *rnr1Y285F/A msh3Δ*. **C.** Motif enrichment of the 12 base window surrounding both deletions. Notably, the synergistic increase in variant frequency appears to be driven by *msh3Δ*. **D.** The variant frequency for the single base C deletion at 32747 was synergistic in *rnr1Y285A msh3Δ*. **E.** The variant frequency for the single base C deletion at 32658 was synergistic in *rnr1Y285A msh3Δ*. **F.** Motif enrichment surrounding both C deletions. This type of error in this sequence context seems to be driven by dCTP and dTTP imbalances and the repair of this error is specific the Msh2-3.

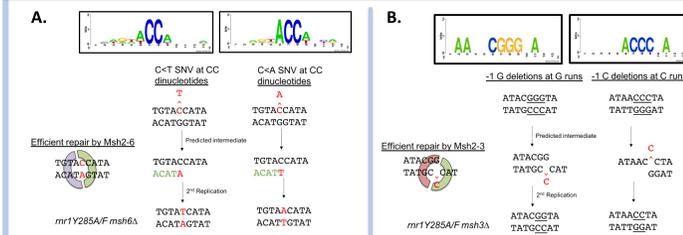## Mechanisms and conclusions

Figure 9. Predicted mechanism of mutagenesis and repair specificity. A. The predicted mechanism of SNV incorporation at two CC dinucleotides enriched in *rnr1Y285A/F msh6Δ*. Both misincorporation events can be explained by incorporation of a nucleotide in excess, dCTP or dTTP, while the correct nucleotide, dGTP, remains most limited. This type of SNV mispair in G/C rich context is efficiently recognized by Msh2-6 and targeted for repair. In the absence of Msh2-6 (*msh6Δ*) and elevated dNTPs these types of errors dominate mutation spectra. **B.** The predicted mechanism of the single base G/C deletions which accumulate in *rnr1Y285A/F msh3Δ* backgrounds. Slippage events on the template strand are explained by an even more limited pool of dGTP when the other three dNTPs are in excess. This single base G/C deletion is efficiently recognized and targeted for repair by Msh2-3, but not Msh2-Msh6. Our novel approach which allowed us to determine mutation spectra with significant depth of sequencing and identify new indicator mutations that are diagnostic of dNTP pool imbalances and/or loss of Msh2-Msh6 or Msh2-Msh3 MMR function.