Multiplexing mutation rate assessment: determining pathogenicity of Msh2 variants in Saccharomyces cerevisiae

Anja R. Ollodart,1,2 Chiann-Ling C. Yeh,2 Aaron W. Miller,2† Brian H. Shirts,2 Adam S. Gordon,3 and Maitreya J. Dunham4,2,*

1Molecular Cellular Biology Program, University of Washington, Seattle, WA 98195, USA,
2Genome Sciences Department, University of Washington, Seattle, WA 98195, USA,
3Department of Laboratory Medicine, University of Washington, Seattle, WA 98195, USA, and
4Department of Pharmacology, Northwestern University, Chicago, IL 60208, USA

Abstract

Despite the fundamental importance of mutation rate as a driving force in evolution and disease risk, common methods to assay mutation rate are time-consuming and tedious. Established methods such as fluctuation tests and mutation accumulation experiments are low-throughput and often require significant optimization to ensure accuracy. We established a new method to determine the mutation rate of many strains simultaneously by tracking mutation events in a chemostat continuous culture device and applying deep sequencing to link mutations to alleles of a DNA-repair gene. We applied this method to assay the mutation rate of hundreds of Saccharomyces cerevisiae strains carrying mutations in the gene encoding Msh2, a DNA repair enzyme in the mismatch repair pathway. Loss-of-function mutations in MSH2 are associated with hereditary nonpolyposis colorectal cancer, an inherited disorder that increases risk for many different cancers. However, the vast majority of MSH2 variants found in human populations have insufficient evidence to be classified as either pathogenic or benign. We first benchmarked our method against Luria–Delbrück fluctuation tests using a collection of published MSH2 missense variants. Our pooled screen successfully identified previously characterized nonfunctional alleles as high mutators. We then created an additional 185 human missense variants in the yeast ortholog, including both characterized and uncharacterized alleles curated from ClinVar and other clinical testing data. In a set of alleles of known pathogenicity, our assay recapitulated ClinVar’s classification; we then estimated pathogenicity for 157 variants classified as uncertain or conflicting reports of significance. This method is capable of studying the mutation rate of many microbial species and can be applied to problems ranging from the generation of high-fidelity polymerases to measuring the frequency of antibiotic resistance emergence.

Keywords: mutation rate; chemostat; MSH2; yeast

Introduction

Mutation rate is the timer for many different error-prone processes: how many cycles of PCR before the polymerase makes a mistake, how long before the bacterial infection becomes resistant to existing medications, or how quickly DNA damage will result in the uncontrolled growth of a cancerous tumor. An example of the latter is germ line variants in mismatch repair (MMR) pathway genes. These have strong implications for human health and are responsible for the cancer risk syndrome known as hereditary nonpolyposis colorectal cancer (HNPCC) (Lynch et al. 2015; Peltoniäki 2016). In patients carrying pathogenic alleles, increased surveillance can detect cancers early, improving treatment outcomes (Gupta et al. 2019). There are a large number of human MMR variants, and many are classified as variants of uncertain significance (VUS) (Starita et al. 2017). Functional data obtained from model organisms can be used to assess potential pathogenicity of these variants (Richards et al. 2015; Gordon et al. 2019). As such, urgently needed is a new way to screen variants to determine whether they increase mutation rate, which may contribute to their pathogenicity.

Existing methods for measuring mutation rate are tedious and not scalable for the challenge of functionally testing hundreds or thousands of VUS. Methods to study mutation rate all have their advantages and disadvantages (Foster 2006). In microbial systems, Luria–Delbrück fluctuation tests, mutation accumulation lines, and Novick–Szilard chemostat mutation accumulation experiments are the most widely used (Luria and Delbrück 1943; Lynch et al. 2008). These methods currently require initiating populations with single clones, which necessarily limits the ability to multiplex experiments. For mutants in MMR, there are some additional specific methods such as measuring sensitivity to 6-thioguanine or N-methyl-N’-nitro-N-nitrosoguanidine, which correlates with functionality of subunits of the MMR complex including Msh2 and Mlh1 (Drost et al. 2013; Houlleberghs et al. 2017;
Bouvet et al. 2019; Jia et al. 2020). Genome editing methods such as CRISPR-Cas9 provide a convenient way to introduce variants into human cells, where signatures of MMR deficiency can then be tracked (Rath et al. 2019). Another alternative is cell-free systems, which allow for using the human protein in an assay that checks for ability to repair DNA (Drost et al. 2010, 2012, 2019). While this provides an easy way to see all mutations caused by errors in replication, each variant must be expressed and purified separately, and this strategy is not amenable to being pooled. Also, while these systems work well for MMR complex proteins, they do not generalize to mutation rate variability caused by dysfunction in other protein complexes. Computational strategies theoretically could scale to all possible sites in all proteins of interest and have been demonstrated to be good predictors of destabilizing variants (Nielsen et al. 2017; Ablidgaard et al. 2019; Stein et al. 2019). However, they still require further validation using lower throughput methods. To address some of these problems, we wanted to generate a new experimental protocol to do multiplexed, direct assessment of mutation rate that was amenable to any molecular pathways for which mutations can be a read out in an easy to culture and genetically tractable organism, *Saccharomyces cerevisiae*.

In addition to ease of use, yeast is a good model system to study effects on MMR because much of the sequence and function of the pathway is conserved between humans and yeast (Boiteux and Jinks-Robertson 2013). Indeed, many discoveries about MMR originate in studies of *S. cerevisiae*, as the MMR complex is more highly conserved with its human orthologs than that of Escherichia coli (Strand et al. 1993, 1995). In addition to the general biology of the MMR complex, yeast has been used to determine the mutation rate of MMR alleles using traditional fluctuation assays, mutation accumulation lines, qualitative patch assays, and fluorescence-based assays (Drotschmann et al. 1999; Gammie et al. 2007; Demogines et al. 2008; Lang and Murray 2008; Martinez and Kolodner 2010; Lang et al. 2013; Shor et al. 2019). These assays all test the full functionality of Msh2; however, none of them allow for pooled assessment of many alleles simultaneously. Although medium-throughput assays exist that take advantage of automated liquid-handling systems (Gou et al. 2019), these still require considerable effort if studying the effect on mutation rate of many different alleles and require maintaining clonal populations.

In an effort to overcome scaling problems inherent to other methods for measuring mutation rate, we have developed a chemostat-based assay that utilizes pools of variants to assay hundreds of alleles simultaneously. The chemostat is a continuous culture device that maintains a constant population size over time by diluting out an actively growing culture at the same rate at which it is growing—defined as steady state. Chemostat conditions have a limiting nutrient, the metabolite which is at a level such that the microbe’s growth rate is slowed but not stopped. This combination of slower growth rate and stable population size makes this device ideal for studies determining the rate of biological processes. Novick and Szilard pioneered the use of chemostats to determine mutation rate over 70 years ago, and the approach has been used to study microbial mutation rates in many studies since (Novick and Szilard 1950, 1951; Fox 1955, p. 1; Kubitschek and Bendikeyt 1964; Paquin and Adams 1983). The continuous dilution in the chemostat means an increase in the frequency of a neutral mutation is a result of *de novo* mutation, as opposed to other methods, where such an increase could be explained by both *de novo* mutation and exponential growth. If the neutral mutation is also selectable, such as with some types of drug resistance (*e.g.*, canavanine resistance in yeast), the mutation rate can be calculated by simply tracking the frequency of resistance over time. Combining this very old technique with next-generation sequencing opens up the possibility for high-throughput study of pools of allelic variants. While studies using traditional fluctuation assays have many replicates of mutation frequency—the number of mutants in a population—they tend to only provide one or two estimates mutation rate—the rate of increase of mutation frequency. We wanted a method in which we could provide many mutation rate calculations at once. We applied this new method to study mutation rate differences caused by missense SNP variants in MSH2, a gene that is associated with HNPCC. Msh2 is a part of the MMR complex, which in combination with Msh3, Msh6, Mlh1, and Pms2—named Pms1 in *S. cerevisiae*—binds and fixes small mismatches and indels (Boiteux and Jinks-Robertson 2013). Msh2 is an integral part of the recognition complex (Edelbrock et al. 2013). We completed a proof of principle with previously published variants of MSH2 and found that the pooled assay recapitulated the results of traditional Luria–Delbruck fluctuation tests, qualitative patch assays, and yeast two-hybrid assays (Gammie et al. 2007). We then assayed an additional 185 MSH2 missense variants curated from ClinVar, a public repository of clinical variant interpretations derived from diagnostic genetic testing. To do so, we recreated these variants in the homologous sites of yeast MSH2, barcoded them along with control WT clones, and measured their mutation rates in a pooled format. Of the 28 variants of known pathogenicity, 100% recapitulated the functional consequence implied by previous clinical interpretation. We then examined 157 VUS from ClinVar and identified 50 variants with significantly different mutation rates from WT as measured by our assay. In addition to ClinVar classifications, data were also compared to tumor sequencing in cancer patients (Shirts et al. 2018); of the 25 VUS for which tumor data were available, 64% had clinical findings that were consistent with our functional data. In total, our data set represents ~6000 individual mutation rate calculations. These data taken together show that our new multiplexed mutation rate assay is an accurate and scalable assay to study the mutation rate of many strains in a pooled format.

**Materials and methods**

**Growth in the chemostat**

For individually inoculated chemostats, 1 ml of overnight culture was inoculated into 230–245 ml of glucose-limited media (calcium chloride 0.1 g/L, sodium chloride 0.1 g/L, magnesium sulfate heptahydrate 0.5 g/L, potassium phosphate monohydrate 1 g/L, ammonium sulfate 5 g/L, glucose 0.8 g/L). Pools were thawed from the freezer and inoculated straight into the chemostat. Cultures were kept at 30°C and allowed 2 days to grow to saturation and the pumps were turned on to a rate of 40 ml/hour or ~five replacement volumes per day. Chemostats were allowed to reach steady state as determined by optical density and later confirmed by stabilization of CFU counts. Samples were then collected starting at ~15 generations. Sampling the nonselective population involved spinning down ~2 × 10⁶ cells, as well as plating ~200 cells onto SC-histidine for accurate counts. For each chemostat, to determine the number of canavanine-resistant mutants, sufficient culture to reach an estimated countable number of colonies (~200) was plated onto SC-arginine-serine-histidine + 60 mg/L canavanine to select for loss-of-function (LOF) mutations in CAN1. For pools, ~6 × 10⁸ cells were
plated, in addition to those used for counts, onto 15 cm SC-histidine + 60 mg/ml canavanine plate and then allowed to
grow at 30°C for 3 days at which point they were scraped for
downstream analysis.

Generation of sequencing libraries
For both the nonselective and mutant population, cells were vor-
texed vigorously with acid washed beads in resuspension buffer
for 3 minutes and then put through the Mini-prep Wizard kit.
They were then concentrated using the PCR cleanup Wizard kit.
For unbarcoded pools, MSH2 was amplified from the plasmid
vector using 15 rounds of PCR to prevent overamplification. Then
Nextera sequencing libraries were generated using the Nextera-
XT kit. The average library size was 500bp and sequenced using a
Nextera 500 with paired end 150 bp reads at a depth of ~30,000
reads over the length of MSH2. The run was conducted according
to the manufacturer’s specifications.

For barcoded pools, PCR amplification of the amplicon con-
taining the barcode was done using 15–22 cycles of PCR using
custom Nextera Primers, listed in Supplementary Table S3. A
comparison of PCR replicate barcode values is in Supplementary
Figure S5. Sufficient amplification was determined by qPCR. The
amplicon was purified using dual sized Sepharose bead cleanup
to isolate the 250bp ampiclon. Samples were then pooled at
equal molar ratios and run on the NextSeq 550 using paired end
reads of both 75 or 150 cycles using custom read and index pri-
mers (Supplementary Table S3) at a read depth of ~1.7 million
reads for the mutant population and 0.5 million reads for the
nonselective population. Number of reads roughly corresponded
with the number of colonies collected for the mutant pool and 100x
coverage of the known number of barcodes for the nonse-
lective population.

Data analysis of direct sequencing of unbarcoded plasmid
Reads were demultiplexed using bc12, allowing for no mis-
matches in the index read. Reads were then processed first by
Trim Galore (Krueger 2019) to remove adaptors, then reads were
collapsed using PEAR (Zhang et al. 2014) then aligned to the yeast
MSH2 sequence using Bowtie2 (Langmead and Salzberg 2012), a
SAM file was generated using Samtools (Li et al. 2009), and then
the make up at each base pair was generated using Pysamstats
(Miles 2019). Data were manipulated in Excel, and then data
points were graphed in R (Supplementary Files S2 and S3) using
ggplot (Wickham 2009).

Data analysis pipeline for barcoded pools
Reads were shortened to the barcode length using a custom py-
thon script, fed into PEAR to combine forward and reverse reads,
then fed into Enrich (Fowler et al. 2011) to count barcodes. These
counts were fed into a custom R script (Supplementary File S2)
which manipulated data and plotted using ggplot2. All data re-
quired to run the custom R script are in Supplementary File S3, in
tabs labeled with the variable names.

Competition of msh2Δ and WT
GFP driven by the TEF2 promoter was introduced to the msh2Δ
and his3Δ strain by mating. Competitions were set up by individ-
ually inoculating 20 ml glucose limited chemostats with 1 ml of
saturated culture of each stated strain. The strains were
allowed to grow up for 2 days before the pumps were turned on.
After reaching steady state after ~10 generations, cultures were
mixed half and half, and GFP percentage was monitored twice a
day via flow cytometry. Fitness effects were calculated by taking
the slope of the ln of the GFP-tagged to non-GFP-tagged strains
over time.

Making of unbarcoded pools
Plasmid DNA was extracted from E. coli strains sent from Alison
Gammie (Gammie et al. 2007), pooled, and used to transform
YMD4328: FY4 msh2Δ and his3Δ to ~20x coverage.

Mapping human variants on the yeast Msh2 protein
Variants in human MSH2 found in ClinVar were mapped to the
yeast MSH2 sequence.

Putative functional alleles uncovered in humans were adapted
for testing in S. cerevisiae using a pairwise protein sequence align-
ment of the two orthologs. The human MSH2 protein was aligned
(NP_000242.1, uniprot: P43246) with the orthologous Msh2 in
S. cerevisiae (strain S288c, uniprot: P25847) using UniProt’s Clustal
Omega webtool with default parameters (https://www.uniprot.
org/align/). The pairwise alignment was validated by comparing
with similar human- to-yeast missense allele adaptations in the
literature (Gammie et al. 2007).

Generating barcoded variants
Variants containing a homolog within the yeast allele were then
ordered as gene products from Twist Biosciences. The gene prod-
ucts were ligated into the pRS413 vector containing the yeast
MSH2 promoter using Gibson and terminator and used to trans-
form DH5α cells at 30X coverage. DNA was extracted using the
Mira-Prep protocol (Pronobis 2016) and then digested with
SacI to linearize. A barcode along with randomized sequence
were inserted into the linearized vector using Gibson assembly
and then used to transform DH5α cells. Transformants were col-
lected such that there was 5X barcode coverage for each allele.
DNA was extracted again using the Mira-prep protocol and
digested with SacI to linearize any unbarcoded alleles and trans-
formed once more to take advantage of E. coli’s inability to
be transformed by linear DNA that does not have homology
overhangs. Colony PCR was done and 0% of clones contained no
barcode and ~6% contained 2–3 barcodes. DNA was once again
extracted with the Mira-Prep protocol and then used to transform
YMD4328: FY4 msh2Δ and his3A flo1A FY4. The flo1A is present to
reduce the prevalence of flocs (Hope et al. 2017). Transformants
were collected such that there was 20X coverage of each barcode.
These were then pooled for future experiments.

Fluctuation assay validation
For the chosen validated alleles, the strain construction was the
same as the barcoded alleles except the variants were kept sepa-
rate and no barcode was inserted. Fluctuation assays were
performed as described in Lang (2018), Lea and Coulson (1949),
and Luria and Delbrück (1943). Briefly, each variant was grown
up to saturation in SC-histidine + 2% glucose and then diluted
1:10,000 in either SC-histidine + 2% glucose for WT-like alleles or
SC-histidine + 0.1% glucose for null-like alleles. The diluted cells
were placed in a 96-well plate—25 μl for null like alleles, 50 μl for
WT-like—covered with Breathe-Easy sealing membrane, and
allowed to incubate at 30°C for 48 hours, without shaking.
Twenty-four wells selected from across the plate were pooled to
determine the average total number of cells, and 68 cultures
were plated onto SC-arginine-serine-histidine + 60 mg/L canava-
nine to select for LOF mutations in CAN1. The four corner wells
were omitted from analysis due to evaporation. Mutation rate
was estimated by using traditional $F_0$ method (Luria and Delbrück 1943) as well as with the rSalvador package (Zheng 2017) that uses the Lea–Coulson model (Lea and Coulson 1949; Ma et al. 1992). This package was implemented as described in Jiang et al. (2021).

**Pac-bio analysis**

Plasmid fragments containing the barcode and variant were isolated from E. coli using the Wizard mini-prep kit, amplified using PCR with Kapa-HiFi, and cleaned up by digesting with DpnI and purifying with Ampure beads. Fragments were prepared for PacBio sequencing using the SMRTbell™ Template Prep Kit 1.0 (Pacific Biosciences) and sent to University of Washington PacBio Sequencing Services for sequencing and Sequel II circular consensus sequence (CCS) analysis (Wenger et al. 2019). BAM files of CCS reads were aligned to the plasmid reference using BWA/0.7.13 mem (Li 2013). Reads that were aligned to the reference sequence were piped to a new BAM file with Samtools/1.9 (Li et al. 2009) and analyzed with cigar strings to validate alignments. Barcodes were then extracted and two barcode-variant maps were generated. One file contains all the barcode-variant reads and the other contains the highest quality read for each unique barcode. Errors found in these files were corrected using a multiple sequence alignment (MUSCLE 3.8.31) (Edgar 2004) of reads sharing the same barcode. Final reads were derived from consensus sequences from these alignments. Ambiguous sequences were fixed by aligning sequences to the highest quality reads using the Needleman-Wunsch algorithm (EMBOSS 6.4.0) (Rice et al. 2000).

**Clinical comparisons**

Clinical comparisons were made using retrospective data gathered from clinical laboratory databases for testing performed as part of standard clinical care between 2014 and 2019. This retrospective analysis was done under University of Washington IRB 00007284.

**Results**

A new method for multiplex mutation rate assessment

The chemostat is a continuous culture device that matches the growth rate of an organism to the dilution rate, stabilizing population size and environmental conditions throughout an experiment. Many ways to study mutation rate take advantage of drugs for which WT organisms are sensitive but an LOF mutation causes resistance, which makes it straightforward to track mutant frequency (Whelan et al. 1979; Boeke et al. 1987). However, to determine rate, one must know the number of generations that have elapsed since the mutational event, which is difficult in batch culture. Luria–Delbrück fluctuation assays and mutation accumulation lines use different tactics to convert mutational frequency into a mutation rate. In continuous culture, since the population size stays stable, an increase in resistance is not due to an increase in a lineage, as long as certain underlying assumptions are met, as described in the following section. In the assay we have developed, outlined in Figure 1, we can track many lineages in a pooled manner to determine all of their mutation rates at once. We do this by keeping track of de novo mutational events on selective media containing a drug, while controlling for any changes in overall population size by monitoring growth on nonselective media.

Next-generation sequencing of the plasmid recovered from both the nonselective and selective media allows us to track the various lineages over time. This assay is amenable to both barcoded and unbarcoded libraries. With unbarcoded libraries, we use shotgun sequencing of the allele isolated from the plasmid, using the mutation within the gene itself as a way to track the variant over the course of the experiment. In barcoded libraries, the barcode and variant are first linked using long read sequencing, after which amplicon sequencing of just the barcode reveals the frequency of each variant at each time point. Our method is amenable to both types of analysis to make it more generalizable. In both cases, the increase in resistance frequency over time for all lineages can be calculated, giving us their mutation rates.

Our first application of this method utilizes S. cerevisiae and focuses on variants of Msh2, a clinically relevant DNA repair enzyme. However, this assay is amenable to the study of any microbial strain that can be cultured inside a chemostat and any molecular pathway where neutral mutations yield a phenotype that can be selected outside the chemostat.

**Conditions required for mutation rate assessment in the chemostat**

To accurately measure mutation rate in the chemostat, certain criteria must be satisfied. First, the readout of mutation rate should be neutral in fitness. Second, resistance should accumulate linearly over the assay period indicating no spontaneous nonneutral mutations arose and reached a detectable frequency. Lastly, there should be no inherent fitness effect due to the variants we seek to characterize. Failure to satisfy these criteria invalidates the assay.

To address the first assumption, we used resistance to canavanine as a read out. LOF mutations in CAN1, which encodes an arginine transporter, prevent uptake of the toxic arginine analog canavanine since the transporter is nonfunctional. In our assay, we use the native CAN1 contained within the genome of our transformed strain to select for mutational events. The use of canavanine to select for LOF mutations within CAN1 has been successfully used to assay mutation rate in yeast previously (Paquin and Adams 1983; Lang and Murray 2008). It has a minimal 1.015% fitness benefit under glucose limitation, the conditions used in this study (Gresham et al. 2008).

To determine the timeframe over which we could observe linear accumulation of resistance, we assayed the mutation rate of msh2Δ strains containing either WT MSH2 (MSH2) or pRS413 (msh2Δ) using the same conditions as future pooled experiments. Previous work has shown complementation by plasmid-borne WT MSH2 and that variants that abrogate activity elevate the mutation rate in yeast (Strand et al. 1993; Drotschmann et al. 1999; Gammie et al. 2007). Chemostats were individually inoculated with MSH2 and msh2Δ strains, and samples were plated every 24 hours to determine which timepoints correspond to the range for linear accumulation of Can<sup>+</sup> mutants (Supplementary Figure S1). We found that between ~12 and 50 generations, resistance to canavanine accumulates at rates consistent with previous literature (Table 1) (Gammie et al. 2007; Lang et al. 2013). The observed lag in linear accumulation can be explained by the approach of the population to steady state, the point at which population growth rate and chemostat dilution rate are balanced [reviewed in (Gresham and Dunham 2014)]. After 50 generations, selection on adaptive mutations is likely the cause of the nonlinear increase (Paquin and Adams 1983; Adams et al. 1985). From this, we determined that all future experimental timepoints must be taken between ~12 and 50 generations to accurately determine the mutation rate. If we are to multiplex this assay, null-like msh2 variants should not have a large fitness effect, otherwise we cannot determine the difference
between a de novo mutation and expansion or contraction of a resistance lineage.

In a head-to-head competition between WT and msh2Δ, we found a 1.097% fitness defect associated with the msh2Δ over the short course of our experiment (Supplementary Figure S1). This means we will likely slightly underestimate mutation rate of high mutators. However, by correcting our mutant frequency by the relative population frequency of each variant, we can mitigate the effects of both the strain manipulation and resistance to canavanine.

**Pools of alleles accumulate mutations at expected rates**

Mutation rate can vary even under very similar conditions, and thus multiple replicate assays must be done to obtain accurate measurements. We created a pooled assay to easily increase the number of replicates for each individual mutation rate assessment. We first established a proof of principle assay with 46 previously published alleles of MSH2 (Gammie et al. 2007) (Figure 2, Supplementary Table S1). We transformed the msh2Δ strain with a pool of the MSH2 alleles contained within plasmids, inoculated aliquots of the pool into four independent 200 ml chemostats, and from each collected six samples on selective and

<table>
<thead>
<tr>
<th>Table 1 Mutation rates of control pure cultures and pools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant genotype</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>MSH2</td>
</tr>
<tr>
<td>msh2Δ</td>
</tr>
<tr>
<td>Unbarcoded pool</td>
</tr>
<tr>
<td>Barcoded pool</td>
</tr>
</tbody>
</table>

* Rates, mutations per generation.

**Figure 1** A schematic outlining the multiplexed mutation rate method. A pool of alleles is inoculated into a chemostat. Samples are plated onto nonselective media or media containing canavanine to select for LOF mutations in CAN1. Plasmid is isolated from the canavanine selected plates as well as from the nonselective pool. The assay can handle both unbarcoded and barcoded plasmids using a shot gun or barcode sequencing, respectively. In both cases, the frequency of the allele on selective canavanine media is divided by its presence in the total pool and tracked over time to generate the mutation rate. With barcoded plasmids, barcoded WT can be used to determine the fold change of variants against an internal control.
nonselective media over 50 generations, as determined from the control experiments described above. Plasmid was recovered from all samples, plasmid-borne $\textit{MSH2}$ alleles were amplified by PCR and subjected to shotgun short read sequencing, and the frequency of each allele from the canavanine-resistant pool was normalized by the frequency in the nonselective pool and converted into colony counts (see Materials and methods).

The average mutation rate of this pool is 13X over wild type (Table 1), which approximates what is expected in a pool containing both LOF and WT-like alleles. In Figure 2, the mutation rates of each allele as measured at $\textit{CAN1}$ are plotted. We compared the results to the phenotype found in previous work using qualitative patch assays, Luria–Delbrück fluctuation tests, and yeast two-hybrid assays between $\textit{Msh2}$ and its subunit partners $\textit{Msh3}$ and $\textit{Msh6}$ (Figure 2, Supplementary Table S1). With the exception of three alleles ($\textit{K873E}$, $\textit{C67Y}$, and $\textit{M707I}$), alleles previously described as WT-like all grouped together, as did the LOF alleles. $\textit{C67Y}$ was classified as LOF due to a lack of subunit interaction and a qualitative patch assay in previous work. This lack of subunit interaction may not be reflected in our mutation rate assay and perhaps explains the lack of correspondence to the qualitative measurement. $\textit{K873E}$ and $\textit{M707I}$ both showed an LOF phenotype measured at $\textit{CAN1}$ but were found to be WT-like when testing for dinucleotide instability. These alleles exist at the edge of the classification between LOF and WT-like in the compared study and could potentially explain why the results are discordant. These data show a grouping of the high mutators and low mutators, indicating that our new method can largely replicate the results of previous efforts to measure allele-specific effects on mutation rate.

Simultaneous measurement of mutation rate among 2000 different lineages

To determine the limit on the number of alleles that can be assayed at once, additional alleles of $\textit{Msh2}$ were curated from the clinical database ClinVar (Landrum et al. 2014). These human $\textit{Msh2}$ variants were mapped onto the yeast $\textit{MSH2}$ gene; only sites with a residue conserved between both orthologs were considered. Twenty-eight alleles with known pathogenicity were included as well as 216 VUS. Alleles in the Walker A ATPase domain and the linker region, which are more highly conserved between humans and yeast, were given precedence (Gammie et al. 2007). Alleles were barcoded a median of five times (Supplementary Figure S7) and the barcode and variant were associated with long-read Pacific Biosciences sequencing. Of the 244 alleles synthesized, 185 variants covered by 1237 barcodes were able to be assayed. In addition, 737 barcodes were associated with WT, giving a robust internal control (Figure 3A). The 59 variants not assayed were due to low barcode coverage of the variant in the pool, low read coverage during sequencing, or because the barcode lineage showed evidence of a fitness altering mutation (Supplementary Figures S2, S3, and S6). The number of variants that can be assayed is dependent on the composition of the pool, with higher mutators being easier to assay than lower mutators, results from this pool of $\textit{Msh2}$ variants indicate that the assay is capable of tracking ~2000 barcodes at once.

**Internal WT barcodes can identify differences in mutation rate among alleles**

The addition of barcodes, while requiring additional work and cost, can provide internal controls to the pooled mutation rate assay, as each genotype can be tracked in multiple independent lineages. The cumulative mutation rate of the barcoded pool containing novel variants of $\textit{Msh2}$ was $3.96 \times 10^{-6}$ CanR/generation, a 26-fold increase over WT (Table 1). The WT barcodes in this assay showed a median mutation rate between $1.09 \times 10^{-6}$ and $3.59 \times 10^{-6}$—approximately 10-fold higher than the expected rate (Figure 3A). Traditional fluctuation assays completed on a strain...
containing the WT plasmid recapitulated individual chemostat assays, indicating that it was not a strain or complementation issue (Supplementary Figure S8). While the source of this global increase in mutation rate is unknown, it can be controlled for by the internal WT barcodes. All variant barcodes in the pool were compared to the internal median WT mutation rate of that pool to calculate fold change over WT for each variant. Inclusion of barcoded WT allows for a robust internal control for mutation rate assessment, which mitigates spurious sources of increased mutation.

### Allele-specific mutation rates recapitulate ClinVar pathogenicity classifications

To determine whether the variants in the barcoded pool recapitulated clinical variant interpretations reported from diagnostic testing, we first looked at variants with known pathogenicity scores in ClinVar. The control variants used have a review status of two or higher, which is a classifier for strong evidence for variant classification in ClinVar (Figure 3B). Variants that shared the same ClinVar classification were found to have similar mutation rates. All five benign variants s "23 out of 23 pathogenic" for correctness and likely pathogenic alleles showed a significant increase in mutation rate compared to WT. Due to the limited number of previously characterized variants in the pool, it is difficult to determine true sensitivity and specificity scores; however, these data lend confidence to our ability to bin variants into pathogenic or benign categories. Based on the results of known variants, we have created four bins: variants that do not differ significantly from WT are potentially benign (1). Those with values ranging between 1.3× and 1.4× over WT are likely intermediate (2). Those which are significantly higher than WT are potentially pathogenic (3). The lowest fold change that showed a significant difference from WT was 1.4×, so we classified those which are above this threshold but do not reach significance as possibly pathogenic (4). Traditional fluctuation assays on variants which had heightened mutation rates but did not reach significance recapitulated the higher mutation rate (Supplementary Table S4), lending credence to this possibility pathogenic classification. We were heartened to see our assay recapitulates previous clinical interpretations and that the use of control alleles allowed for the generation of bins to provide information on the VUS assayed.

### Estimating pathogenicity of VUS

We were able to assay 157 SNP VUS in this assay. Of 157 variants, 50 showed a significant difference in mutation rate in comparison to WT and were classified as potentially pathogenic. The mutation rates of these 50 variants are given in Table 2. A summary of the fold changes of all variants is found in Supplementary Table S1. We found that one variant, K449N, had a significantly lower mutation rate than WT, though this result did not replicate in an individual fluctuation assay on this variant (Supplementary Table S4). The mechanism of this possible lowered mutation rate and whether it is a biologically relevant is unknown and could provide an interesting point of study if later confirmed. The 50 VUS that showed a significant increase in mutation rate ranged from 1.39-fold over WT to 13-fold over WT. No alleles assayed had a full LOF phenotype—which would be characterized by a 30-fold increase in mutation rate. This may reflect the dynamic range of the assay or it may reflect that no true LOF alleles existed in the data set. In an attempt to address this, a validation set of nine alleles was subjected to traditional fluctuation assays (Supplementary Table S4). hMSH2 variant R680P recapitulated the 13-fold increase found in the pooled assay—indicating that the upper range of the values found in the assay are correct. For future experiments, it may be wise to include barcoded deletion strains at a low frequency to directly address the dynamic range of the assay. Taken in total, this data set provides evidence of pathogenicity for an additional 157 VUS of MSH2; 83 will be classified as potentially benign, 7 intermediate, 17 possibly pathogenic, and 50 potentially pathogenic. While additional study would be required before these classifications could inform clinical diagnosis, these data represent a first indication of the effects of these mutations on function and could be used as a line of evidence according to American College of Medical Genetics (ACMG) criteria (Richards et al. 2015).

### Associating variant data with clinical and tumor sequencing phenotypes

To more accurately compare clinical data with the outputs of this screen, the fold change calculations were converted to scores. Table 3 contains information on variants that have clinical findings associated with them. Clinical summaries were gathered from data provided to the University of Washington...
Our data largely recapitulate the results from Jia et al. (2021) and tumors (Thompson et al. 2013; Ranola et al. 2018; Shirts et al. 2018; Li et al. 2020). Personal and family history of colorectal or endometrial cancer provide weak evidence of pathogenicity while personal and family history lacking HNPCC-associated cancers provide weak evidence against pathogenicity (Li et al. 2020). Tumor characteristics of microsatellite instability (MSI-H) and loss of Msh2 and Msh6 on immunohistochemistry staining provide moderate evidence supporting pathogenicity (Thompson et al. 2013; Li et al. 2020). Presence of alternative pathogenic variants in MSH2 or other genes that explain these tumor or other tumor characteristics provides evidence against pathogenicity, while a second somatic pathogenic variant at heterozygous frequency or loss of heterozygosity in tumor provides moderate and strong evidence supporting pathogenicity, respectively (Shirts et al. 2018). Formal strategies for combining each of these types of data with functional data are outside the scope of this effort. Rather we provide relevant clinical summaries with an overall assessment of whether the clinical information is consistent or inconsistent with functional scores. Of the 25 variants identified, 16 (64%) had clinical data that were consistent with functional data, 4 (16%) had clinical data that were inconsistent with functional data, and 5 (20%) had clinical data that were equivocal or fell in the indeterminate functional score range. Some discordance is expected given that functional data are only one component of the ACMG guidelines for clinical variant interpretation; this discordance is consistent with results of past studies seeking to use data with functional data are outside the scope of this effort. Additionally, we provide relevant clinical summaries with an overall assessment of whether the clinical information is consistent or inconsistent with functional scores. Of the 25 variants identified, 16 (64%) had clinical data that were consistent with functional data, 4 (16%) had clinical data that were inconsistent with functional data, and 5 (20%) had clinical data that were equivocal or fell in the indeterminate functional score range. Some discordance is expected given that functional data are only one component of the ACMG guidelines for clinical variant interpretation; this discordance is consistent with results of past studies seeking to use other clinical criteria to classify variants (Thompson et al. 2013; Shirts et al. 2018; Li et al. 2020).

### Discussion

We developed a new method for high-throughput mutation rate assessment that combines a mid-20th century method to determine mutation rate with 21st-century next-generation sequencing. This allows for the pooling and multiplexing of mutation rate assessment that has not been accomplished before. We were able to complete over 6000 individual mutation rate calculations over ~2000 lineages covering ~200 variants of MSH2, a critical component of MMR. Although we included a high frequency of WT sequences, our analysis indicates many of these could be substituted with additional VS to increase throughput at minimal cost to accuracy. The assay is limited by the canavine-resistant subpopulation within a 200 ml chemostat, which is dependent on the mutation rates of the lineages in a pool. One could increase the number of variants to be assayed in a single experiment by increasing the volume of the chemostat, though the logistics of expanding the volume beyond the 2 L size of available commercial fermenters may be difficult. Another possible modification would be to utilize alternative marker loci that generate selectable mutations at higher rates than the WT CAN1 sequence. In addition, the inclusion of barcoded null mutants may provide an additional control to better normalize the results to established mutation rates.

In this work, we estimated the pathogenicity of 157 variants of uncertain or conflicting significance derived from clinical testing. These results provide a key piece of information to testing labs seeking to assign pathogenicity to variants for which little other evidence is available. We have found that our results largely match clinical data obtained from tumors (Shirts et al. 2018). These data in combination with a recent deep mutation scan on MSH2 in a human cell line (Jia et al. 2020) will allow for accurate reclassification of uncertain and conflicting variants in this gene. Our data largely recapitulate the results from Jia et al.

### Table 2 Mutation rates of significantly different alleles

<table>
<thead>
<tr>
<th>Human genotypes</th>
<th>Yeast genotypes</th>
<th>Count</th>
<th>Fold induction CAN</th>
<th>IQRc</th>
<th>Sigd</th>
</tr>
</thead>
<tbody>
<tr>
<td>K449N</td>
<td>K469N</td>
<td>10</td>
<td>0.56</td>
<td>0.28</td>
<td>***</td>
</tr>
<tr>
<td>G761R</td>
<td>G780R</td>
<td>31</td>
<td>1.39</td>
<td>0.96</td>
<td>***</td>
</tr>
<tr>
<td>M672R</td>
<td>M691R</td>
<td>32</td>
<td>1.61</td>
<td>3.37</td>
<td>***</td>
</tr>
<tr>
<td>R711Q</td>
<td>R730Q</td>
<td>25</td>
<td>1.62</td>
<td>2.63</td>
<td>***</td>
</tr>
<tr>
<td>P349A</td>
<td>P361A</td>
<td>61</td>
<td>1.79</td>
<td>2.03</td>
<td>*</td>
</tr>
<tr>
<td>V695L</td>
<td>V714L</td>
<td>12</td>
<td>2.08</td>
<td>1.77</td>
<td>***</td>
</tr>
<tr>
<td>Q681E</td>
<td>Q700E</td>
<td>20</td>
<td>2.09</td>
<td>2.69</td>
<td>***</td>
</tr>
<tr>
<td>R359K</td>
<td>R371K</td>
<td>7</td>
<td>2.25</td>
<td>1.26</td>
<td>***</td>
</tr>
<tr>
<td>M672V</td>
<td>M691V</td>
<td>37</td>
<td>2.27</td>
<td>2.37</td>
<td>**</td>
</tr>
<tr>
<td>H783Y</td>
<td>H802Y</td>
<td>11</td>
<td>2.38</td>
<td>1.82</td>
<td>***</td>
</tr>
<tr>
<td>S676F</td>
<td>S695P</td>
<td>8</td>
<td>2.48</td>
<td>4.55</td>
<td>*</td>
</tr>
<tr>
<td>A609P</td>
<td>A627P</td>
<td>12</td>
<td>2.56</td>
<td>3.36</td>
<td>**</td>
</tr>
<tr>
<td>G761V</td>
<td>G780V</td>
<td>19</td>
<td>2.71</td>
<td>2.82</td>
<td>***</td>
</tr>
<tr>
<td>A700E</td>
<td>A719E</td>
<td>9</td>
<td>2.78</td>
<td>3.79</td>
<td>**</td>
</tr>
<tr>
<td>H783D</td>
<td>H802D</td>
<td>7</td>
<td>2.79</td>
<td>4.53</td>
<td>*</td>
</tr>
<tr>
<td>G683E</td>
<td>G702E</td>
<td>54</td>
<td>2.88</td>
<td>3.62</td>
<td>***</td>
</tr>
<tr>
<td>A689V</td>
<td>A708V</td>
<td>11</td>
<td>3.17</td>
<td>2.58</td>
<td>***</td>
</tr>
<tr>
<td>F696S</td>
<td>F715S</td>
<td>5</td>
<td>3.27</td>
<td>1.66</td>
<td>***</td>
</tr>
<tr>
<td>R524H</td>
<td>R542H</td>
<td>9</td>
<td>3.49</td>
<td>5.85</td>
<td>***</td>
</tr>
<tr>
<td>R524C</td>
<td>R542C</td>
<td>13</td>
<td>3.61</td>
<td>7.10</td>
<td>***</td>
</tr>
<tr>
<td>K675E</td>
<td>K694E</td>
<td>8</td>
<td>3.66</td>
<td>7.49</td>
<td>*</td>
</tr>
<tr>
<td>T724M</td>
<td>T743M</td>
<td>34</td>
<td>3.83</td>
<td>3.51</td>
<td>*</td>
</tr>
<tr>
<td>P622Q</td>
<td>P640Q</td>
<td>5</td>
<td>3.95</td>
<td>2.00</td>
<td>**</td>
</tr>
<tr>
<td>P622A</td>
<td>P640A</td>
<td>9</td>
<td>4.12</td>
<td>2.37</td>
<td>**</td>
</tr>
<tr>
<td>S676L</td>
<td>S695L</td>
<td>5</td>
<td>4.18</td>
<td>1.43</td>
<td>**</td>
</tr>
<tr>
<td>R524L</td>
<td>R542L</td>
<td>12</td>
<td>4.18</td>
<td>2.22</td>
<td>***</td>
</tr>
<tr>
<td>G683W</td>
<td>G702W</td>
<td>18</td>
<td>4.51</td>
<td>5.92</td>
<td>**</td>
</tr>
<tr>
<td>G695R</td>
<td>G713R</td>
<td>14</td>
<td>5.04</td>
<td>8.82</td>
<td>**</td>
</tr>
<tr>
<td>T724R</td>
<td>T743R</td>
<td>18</td>
<td>5.76</td>
<td>6.85</td>
<td>**</td>
</tr>
<tr>
<td>G683W</td>
<td>G702W</td>
<td>18</td>
<td>5.81</td>
<td>6.00</td>
<td>**</td>
</tr>
<tr>
<td>G695R</td>
<td>G713R</td>
<td>14</td>
<td>5.04</td>
<td>8.82</td>
<td>**</td>
</tr>
<tr>
<td>Q680E</td>
<td>Q709E</td>
<td>15</td>
<td>6.30</td>
<td>8.01</td>
<td>**</td>
</tr>
<tr>
<td>A609P</td>
<td>A708P</td>
<td>9</td>
<td>6.27</td>
<td>5.14</td>
<td>**</td>
</tr>
<tr>
<td>P670R</td>
<td>P689R</td>
<td>6</td>
<td>6.31</td>
<td>6.23</td>
<td>**</td>
</tr>
<tr>
<td>G692W</td>
<td>G713W</td>
<td>23</td>
<td>6.33</td>
<td>6.60</td>
<td>*</td>
</tr>
<tr>
<td>G669D</td>
<td>G688D</td>
<td>18</td>
<td>6.63</td>
<td>11.83</td>
<td>*</td>
</tr>
<tr>
<td>G692E</td>
<td>G713E</td>
<td>14</td>
<td>5.04</td>
<td>8.82</td>
<td>**</td>
</tr>
<tr>
<td>Q681E</td>
<td>Q700E</td>
<td>11</td>
<td>6.95</td>
<td>7.11</td>
<td>***</td>
</tr>
<tr>
<td>T677R</td>
<td>T696R</td>
<td>13</td>
<td>7.41</td>
<td>9.49</td>
<td>**</td>
</tr>
<tr>
<td>P349S</td>
<td>P361S</td>
<td>47</td>
<td>7.48</td>
<td>8.01</td>
<td>**</td>
</tr>
<tr>
<td>G358V</td>
<td>G350V</td>
<td>14</td>
<td>7.98</td>
<td>15.14</td>
<td>***</td>
</tr>
<tr>
<td>G669C</td>
<td>G688C</td>
<td>12</td>
<td>8.08</td>
<td>5.35</td>
<td>**</td>
</tr>
<tr>
<td>R621Q</td>
<td>R639Q</td>
<td>4</td>
<td>8.29</td>
<td>2.07</td>
<td>*</td>
</tr>
<tr>
<td>N617D</td>
<td>N690D</td>
<td>13</td>
<td>8.61</td>
<td>6.56</td>
<td>*</td>
</tr>
<tr>
<td>R621P</td>
<td>R639P</td>
<td>13</td>
<td>9.28</td>
<td>10.78</td>
<td>**</td>
</tr>
<tr>
<td>L310R</td>
<td>L305R</td>
<td>14</td>
<td>9.50</td>
<td>13.27</td>
<td>***</td>
</tr>
<tr>
<td>G693Y</td>
<td>G712Y</td>
<td>4</td>
<td>9.70</td>
<td>3.71</td>
<td>***</td>
</tr>
<tr>
<td>R680P</td>
<td>R699P</td>
<td>9</td>
<td>13.00</td>
<td>17.19</td>
<td>**</td>
</tr>
</tbody>
</table>

**a** Count, the number of times a variant was assayed in total. 
**b** For stated genotype, mutation rate (mutations per cell division) was calculated and then compared to the WT mutation rate. Barcode and chemostat replicates are combined to calculate median fold change. 
**c** Interquartile range of all barcode and chemostat replicates. 
**d** Significance is calculated by a Wilcoxon rank-sum test with the Benjamini–Hochberg correction for multiple hypothesis testing. 

- * p < 0.05; ** p < 0.01; *** p < 0.001; **** p < 0.0001.
Table 3 Summary of variants with clinical or tumor data

<table>
<thead>
<tr>
<th>Human genotype</th>
<th>Scorea</th>
<th>95% Clb</th>
<th>Sigc</th>
<th>ClinVard</th>
<th>Clinical information</th>
<th>CSTf</th>
</tr>
</thead>
<tbody>
<tr>
<td>P27L</td>
<td>−1.06</td>
<td>0.87</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H colon cancer with loss of PMS2 by IHC, under age 30 at diagnosis. Also has heterozygous VUS in MLH1.</td>
<td>Y</td>
</tr>
<tr>
<td>K449N</td>
<td>−0.84</td>
<td>0.44</td>
<td>**</td>
<td>VUS</td>
<td>Heterozygous germ line. Ovarian cancer. No family history of cancer.</td>
<td>Y</td>
</tr>
<tr>
<td>A545S</td>
<td>−0.11</td>
<td>0.34</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. No personal history of cancer. Family history early onset colorectal cancer and breast cancer.</td>
<td>Y</td>
</tr>
<tr>
<td>A733T</td>
<td>0.11</td>
<td>0.93</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. Breast cancer. Family history of colon, ovarian, and brain cancers.</td>
<td>Y</td>
</tr>
<tr>
<td>P616R</td>
<td>0.17</td>
<td>0.52</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H endometrial tumor with loss of MSH2 and MSH6 by IHC. Two other clearly pathogenic LOF MSH2 mutations identified in the tumor make this less likely to be pathogenic</td>
<td>Y</td>
</tr>
<tr>
<td>Q374R</td>
<td>0.20</td>
<td>0.42</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H endometrial tumor with loss of MLH1-MSH2, MSH6, and PMS2 by IHC. The tumor had two other clearly pathogenic somatic mutations in MSH2.</td>
<td>Y</td>
</tr>
<tr>
<td>V6551</td>
<td>0.21</td>
<td>0.38</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. Colorectal cancer under age 30. No family history of cancer.</td>
<td>N</td>
</tr>
<tr>
<td>W764R</td>
<td>0.27</td>
<td>0.68</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H colon tumor with loss of MSH2 and MSH6 by IHC. Evidence of LOH for MSH2 variant in tumor.</td>
<td>N</td>
</tr>
<tr>
<td>L599S</td>
<td>0.33</td>
<td>0.39</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. Seen in patient with breast cancer and family history of breast and colorectal cancer.</td>
<td>Y</td>
</tr>
<tr>
<td>V78I</td>
<td>0.36</td>
<td>0.70</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. MSS ovarian tumor with other pathogenic variants with no LOH for MSH2 variant in tumor.</td>
<td>Y</td>
</tr>
<tr>
<td>I770V</td>
<td>0.42</td>
<td>0.52</td>
<td>ns</td>
<td>VUS</td>
<td>Heterozygous germ line. Colorectal cancer diagnosed under age 30. Seen with heterozygous germ line VUS in APC gene.</td>
<td>—</td>
</tr>
<tr>
<td>G761R</td>
<td>0.47</td>
<td>0.38</td>
<td>*</td>
<td>VUS</td>
<td>Suspected germ line variant in prostate tumor. IHC for MSH2, MSH6 equivocal. MSI equivocal. Apparent MSH2 LOH in tumor.</td>
<td>—</td>
</tr>
<tr>
<td>H785R</td>
<td>0.75</td>
<td>0.51</td>
<td>ns</td>
<td>VUS</td>
<td>Homozygous germ line. MSS colorectal cancer. The tumor also had POLE mutation and ultramutator phenotype.</td>
<td>—</td>
</tr>
<tr>
<td>P349A</td>
<td>0.84</td>
<td>0.30</td>
<td>***</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H colorectal cancer diagnosed over age 50. Tumor with loss of MLH1 and PMS2 explained by double somatic MLH1 mutation in tumor.</td>
<td>N</td>
</tr>
<tr>
<td>Q681E</td>
<td>1.06</td>
<td>0.48</td>
<td>**</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H colorectal cancer diagnosed over age 50. Tumor with loss of MLH1 and PMS2 explained by double somatic MLH1 mutation in tumor.</td>
<td>Y</td>
</tr>
<tr>
<td>A609P</td>
<td>1.36</td>
<td>0.73</td>
<td>**</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H colon cancer with loss of MSH2 and MSH6 by IHC. Under age 50 at diagnosis. Tumor had one somatic pathogenic MSH2 mutation seen at heterozygous frequency in the tumor.</td>
<td>Y</td>
</tr>
<tr>
<td>P696S</td>
<td>1.71</td>
<td>0.98</td>
<td>*</td>
<td>VUS</td>
<td>Heterozygous germ line. Personal history of phenochromocytoma, family history of renal and brain cancer.</td>
<td>N</td>
</tr>
<tr>
<td>S676L</td>
<td>2.06</td>
<td>0.70</td>
<td>**</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H colon cancer diagnosed under age 50. Tumor had loss of MSH6 by IHC. Seen with MSH2 p. G827R somatic mutation listed below. This was seen with another heterozygous pathogenic mutation in MSH2 (1760-1 G &gt; A).</td>
<td>Y</td>
</tr>
<tr>
<td>C693R</td>
<td>2.32</td>
<td>0.80</td>
<td>****</td>
<td>LP</td>
<td>Somatic mutation in tumor. MSI-H colon tumor with loss of MSH2 by IHC.</td>
<td>Y</td>
</tr>
<tr>
<td>G692V</td>
<td>2.42</td>
<td>0.58</td>
<td>****</td>
<td>LP</td>
<td>Somatic mutation in tumor. MSI-H neuroendocrine tumor with loss of MSH2 and MSH6 by IHC. Seen in a tumor with a germ line likely pathogenic variant in MSH2 (p.I30R) that had loss of heterozygosity.</td>
<td>—</td>
</tr>
<tr>
<td>G827R</td>
<td>2.58</td>
<td>0.51</td>
<td>****</td>
<td>VUS</td>
<td>Somatic mutation in tumor. MSI-H colon cancer diagnosed under age 50. Tumor had loss of MSH6 by IHC. Seen with the germ line variant p. S676L listed above.</td>
<td>Y</td>
</tr>
<tr>
<td>G692W</td>
<td>2.66</td>
<td>0.61</td>
<td>****</td>
<td>VUS</td>
<td>Heterozygous germ line. MSI-H endometrial tumor with loss of MSH2 and MSH6 by IHC. Cancer diagnosed over age 50. Variant reported by another laboratory to segregate with HNPCC in one family. A variant at the same position (p.G692R, NM_000179.2: c.2074 G to C) is classified as likely pathogenic (class 4) by the InSiGHT consortium.</td>
<td>Y</td>
</tr>
<tr>
<td>G827E</td>
<td>2.80</td>
<td>0.63</td>
<td>****</td>
<td>VUS</td>
<td>Heterozygous germ line. Pancreatic cancer diagnosed over age 80. Loss of MSH2 and MSH6 by IHC.</td>
<td>Y</td>
</tr>
<tr>
<td>N671D</td>
<td>3.11</td>
<td>0.53</td>
<td>****</td>
<td>VUS</td>
<td>Heterozygous germ line. Personal history of colon polyps and family history of colon, uterine, and other cancers.</td>
<td>Y</td>
</tr>
<tr>
<td>L310R</td>
<td>3.25</td>
<td>0.74</td>
<td>****</td>
<td>P</td>
<td>Heterozygous germ line. Seen in a family with multiple MSI-H colon cancers that had loss of MSH2. Co-segregation likelihood ratio for pathogenicity in the family was 44:1. See (Tsai et al., 2019) for complete pedigree information.</td>
<td>Y</td>
</tr>
</tbody>
</table>

a For stated genotype, variant mutation rate (mutants per generation) was compared to the WT mutation rate. The score represents log2(median fold change).

b 95% confidence interval on log2(fold change).

c Significance is calculated by a Wilcoxon rank-sum test with the Benjamini–Hochberg correction for multiple hypothesis testing. * P < 0.05; ** P < 0.01; *** P < 0.001; **** P < 0.0001.

d Initial clinical classification as stated in ClinVar. B, benign; LP, likely pathogenic; P, pathogenic; VUS, variant of uncertain significance.

e Clinical information collected from UW Laboratory Medicine. MSI-H, microsatellite instability-high; IHC, immunohistochemistry.

f CST, consistent, tumor or clinical data is consistent with functional score. Y = Yes, N = No, — = undetermined.
(Supplementary Figure S9). This will provide physicians better guidance on whom to screen for HNPCC and how often.

Our multiplexed assay works for any protein that affects mutation rate. Thus, it can be used to assess mutation rate variation arising from changes in other MMR proteins, as well as in proteins that act in other DNA repair pathways. Assaying variants in the sequence context of the native human cDNA could be possible for genes that complement the orthologous yeast gene knock-out (Vogelsang et al. 2009; Kachroo et al. 2015); however, our initial attempts to recapitulate the complementation of mhl1 and pms1—the S. cerevisiae orthologue of PMS2—with human MLH1 and PMS2, other critical MMR genes associated with HNPCC, were unsuccessful. This, however, does not mean that assessing mutation rate of human alleles of DNA repair enzymes is not possible and in fact could be a very interesting line of study. We further note that our assay has additional limitations in replicating human biology: for example, it is not capable of assessing alleles that perturb splicing, RNA stability, or gene regulation, given that these processes are significantly different in yeast vs human cells.

The current set of alleles could also be tested in different genetic backgrounds, such as in the ubiquitin ligase san1Δ background (Arlow et al. 2013) which could give additional information on the stability of Msh2 variants and could determine the mechanism behind an increased mutation rate for some variants. It could also be done in the background containing deletions or change of function mutations in other proteins in the MMR complex. Both variant library and background are completely mutable in this system.

Our method should be widely applicable and can be used to answer many other questions associated with mutation rate outside of clinical variant interpretation. While this assay is not appropriate for organisms or strains that are not culturable within the chemostat, that still leaves a large set of questions that can be answered. Accurate, multiplexed measurement of mutation rate variation could be used to screen polymerases for increased or decreased fidelity, to screen the yeast deletion collection for knock-outs that increase mutation rate, or to uncover differences in mutation rate among natural variants of yeast. In conclusion, this method is broadly applicable to many different problems in which mutation rate is a factor and can be used to estimate the pathogenicity of clinically relevant DNA repair enzymes.

Data availability
All sequencing data are available at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA662579. Strains and plasmids are available upon request. Supplementary material available at figshare: https://doi.org/10.25386/genetics.14132609. Supplementary File S1 contains figures and tables referenced in this text and Supplementary File S2 contains the custom script used in this work. Supplementary File S3 contains strains and plasmids used in this text, tables from Supplementary File S1 in excel format, as well as data sets required to run the custom script in Supplementary File S2.

Supplementary material is available at https://doi.org/10.25386/genetics.14132609.

Acknowledgments
The authors would like to thank Alison Gammie and Mark Rose for the strains used in this work. Josh Cuperus gave guidance on computational tools used in this project which the authors are grateful for. The authors appreciate Jolie Carlisle’s help with plasmid preps and initial attempts to develop this assay in an alternative system. Pengyao Jiang provided the authors with the scripts implementing rSalvador and helpful comments on the manuscript, for which they are thankful. The authors express their sincere gratitude to Jessica Hartmann for design help on figures. The authors would like to thank Bryce Taylor, Pengyao Jiang, Jeet Patel, Meghan Garrett, Amanda Riley, Kurt Berckmueller, and Patrick Nugent for helpful edits on this manuscript.

Funding
Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health (NIH) under award T32CA080416 (A.W.M.), by the National Human Genome Research Institute of the NIH under award T32 HG00035 (A.R.O. and A.W.M.), National Science Foundation (NSF) Graduate Research Fellowship DGE-1762114 (C.C.Y.), and by the National Institute of General Medical Sciences of the NIH under award P41 GM103533 (M.J.D.). M.J.D. acknowledges prior support as a Rita Allen Foundation Scholar and as a Senior Fellow in the Genetic Networks program at the Canadian Institute for Advanced Research. This material is based in part upon work supported by the NSF under Cooperative Agreement No. DBI-0939454. The research of M.J.D. was supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute.

Conflicts of interest
The authors declared no conflicts of interest.

Literature cited


et al. 2016. The Miraprep: a Protocol that
Richards S, Aziz N, Bale S, Bick D, Das S, ACMG Laboratory Quality
et al. 2016. Update on Lynch syndrome genomics. Fam

Communicating editor: J. Surtees