A modified fluctuation assay reveals a natural mutator phenotype that drives mutation spectrum variation within *Saccharomyces cerevisiae*

Pengyao Jiang[1], Anja R. Ollodart[1,2], Vidha Sudhesh[1], Alan J. Herr[3], Maitreya J. Dunham[1], Kelley Harris[1,4]

1 Department of Genome Sciences, University of Washington, Seattle, WA
2 Molecular and Cellular Biology Program, University of Washington, Seattle, WA
3 Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA
4 Department of Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA

Abstract

Mutations are the source of genetic variation and a prerequisite for evolution. Despite their fundamental importance, however, their rarity makes them expensive and difficult to detect, which has limited our ability to measure the extent to which mutational processes vary within and between species. Here, we use the 1011 *Saccharomyces cerevisiae* collection to measure variation of mutation rates and spectra among strains isolated from a variety of natural and human-related environments. The mutation spectra of variants segregating in different *S. cerevisiae* populations exhibit differences in the relative numbers of specific transition and transversion types, a pattern reminiscent of previously observed mutation spectrum differences between populations of humans, great apes, and mice. Such natural variation is thought to reveal historical differences in the activity of particular mutational processes, but is also potentially complicated by other forces such as admixture, genetic drift, and selection. In order to directly test how much of the observed mutation spectrum variation is caused by heritable differences between extant strains of *S. cerevisiae*, we developed an experimental pipeline to assay *de novo* mutation rates and spectra of individual strains, using the reporter gene *CAN1*. We found a 10-fold range of mutation rate variation among 16 haploid strains surveyed. While many strains exhibit similar mutation spectra, two related strains from the panel's "Mosaic beer" clade, known as AEQ and AAR, share a distinctive mutation spectrum enrichment for C>A mutations. This C>A enrichment found through our experimental pipeline mirrors an enrichment of C>A mutations in rare variants segregating throughout the genomes of AEQ and AAR as well as additional Mosaic beer strains. We deduce that a major axis of *S. cerevisiae* mutation spectrum variation is likely driven by one or more naturally occurring mutator alleles whose action is measurable in a controlled laboratory environment.

Introduction

Mutations are a double-edged sword. At the molecular level, they usually arise as a spontaneous consequence of DNA replication errors or damage and are the ultimate cause of genetic diseases (Nei, 1983; Crow, 1997; Antonarakis and Beckmann, 2006; Sebat *et al.*, 2007; Iossifov *et al.*, 2014). All organisms have evolved complex mechanisms for keeping mutation

rates low and safeguarding their genetic information as it is passed from generation to generation (Beckman and Loeb, 1993; Eisen and Hanawalt, 1999); in multicellular organisms, these mechanisms also safeguard somatic tissues from mutations that can cause cancer and age-related decline (Alexandrov *et al.*, 2013; Loeb, 2016; Risques and Kennedy, 2018). A low mutation rate is essential for long-term population survival, and the larger and more complex a genome is, the lower the mutation rate must be to prevent deleterious mutations from arising faster than natural selection can eliminate them (Eigen, 1971; Drake, 1991; Sung, Ackerman, *et al.*, 2012; Acosta *et al.*, 2015). Over long time scales, however, mutations also serve as the raw material for evolution. Although beneficial mutations are rare occurrences, they are essential for the acquisition of novel phenotypes and adaptations (Gompel *et al.*, 2005; McGregor *et al.*, 2007).

A large body of theory has been written to describe how natural selection might act on the mutation rate to balance these beneficial and deleterious effects (Sturtevant, 1937; Kimura, 1967; Leigh, 1970; Johnson, 1999; André and Godelle, 2006; Sung, Ackerman, *et al.*, 2012). One prediction is that organisms living in more changeable environments might evolve higher mutation rates than organisms living in more stable environments, assuming that the environment determines whether a higher rate of beneficial mutations is likely to counterbalance a higher rate of deleterious mutations. This prediction has been borne out in laboratory evolution experiments, where mutator phenotypes sometimes emerge in populations that are forced to tolerate challenging conditions (Tenaillon *et al.*, 2016; Good *et al.*, 2017) and mutator strains are often observed to take over chemostat populations by producing beneficial mutations at a higher rate than competing non-mutator strains (Chao and Cox, 1983). However, it is less clear how much mutation rate variation exists within and between natural populations, and if such variation exists, whether it is maintained by natural selection. The “drift barrier hypothesis” predicts that mutator alleles will usually be deleterious because they produce more damaging mutations than beneficial ones, but that mutator alleles with relatively small effects may persist in populations because they are not deleterious enough to be efficiently eliminated (Lynch *et al.*, 2016).

Although next-generation sequencing has rapidly increased our ability to measure the genetic variation that currently exists within populations, the extent of mutation rate variation is still more difficult and expensive to measure. One of the original methods for measuring mutation rates is the Luria-Delbrück fluctuation assay (Luria and Delbrück, 1943; Lang and Murray, 2008; Gou, Bloom and Kruglyak, 2019), in which a population of microorganisms is allowed to grow clonally for a controlled length of time, then challenged with a form of artificial selection that kills most cells except for those that have happened to acquire specific resistance mutations. The mutation rate can then be calculated from the number of colonies that manage to grow after this artificial selection is imposed.

Though fluctuation assays are an elegant and efficient way for measuring the mutation rates of specific reporter genes, the results are potentially sensitive to the reporter gene being used and where it is located within the genome (Lang and Murray, 2008, 2011); in addition, they are not applicable to multicellular organisms. These drawbacks have motivated the development of

newer methods that take advantage of high-throughput sequencing, such as mutation accumulation (MA) assays in which a laboratory population is serially bottlenecked for many generations, eliminating most effects of natural selection and allowing mutations to be directly counted by sequencing at the end of the experiment. MA studies have been used to estimate mutation rates in a wide variety of organisms (Lynch *et al.*, 2008; Keightley *et al.*, 2009; Zhu *et al.*, 2014; Farlow *et al.*, 2015; Sharp *et al.*, 2018; Wang *et al.*, 2019). However, it is labor-intensive to maintain MA lineages for enough generations to measure a low mutation rate accurately, which has limited the feasibility of measuring variability of mutation rates within species.

An alternative source of information about mutational processes is genetic variation among related individuals who share common ancestors. Polymorphic sites are easier and cheaper to discover than new mutations, since they are present at a higher density within the genome and often shared among several individuals. Mining polymorphisms for information about mutation rates can be difficult since their abundance is affected by genetic drift and natural selection (Scally and Durbin, 2012; Ségurel, Wyman and Przeworski, 2014; Zhu, Sherlock and Petrov, 2017), but despite these limitations, they have provided surprisingly strong evidence for the existence of historical changes to the mutation spectrum, meaning the tendency of mutations to occur most often in certain nucleotide contexts (Hwang and Green, 2004). In humans, for example, Europeans and South Asians have a significantly higher proportion of TCC>TTC mutations than other human groups (Harris, 2015; Harris and Pritchard, 2017), a pattern that is difficult to explain without a recent population-specific increase in the rate of this type of mutation. This pattern might have been caused by either a genetic mutator or an environmental mutagen, but is not explicable by the action of selection or drift or any other process that modulates the retention or loss of genetic variation.

Polymorphism data has revealed that each human population and great ape species appears to have a distinctive triplet mutation spectrum, which implies that genetic and/or environmental mutators likely emerge relatively often and act within localized populations to increase mutation rates in specific sequence contexts (Harris and Pritchard, 2017; Goldberg and Harris, 2019). However, identifying these hypothetical mutators is a challenging proposition, not least because some population-specific signatures such as the human TCC>TTC enrichment appear to be relics of mutators that are no longer active. A recent study of *de novo* mutations in diverse human families found some evidence of mutation rate variation between human populations (Kessler *et al.*, 2020), but argued that most of this variation was driven by the environment rather than genetics. Given that humans from different populations tend to be born and raised in different environments, it is extremely challenging to determine the degree to which genetics and/or the environment are responsible for variation of the rates and spectrum of *de novo* mutations accumulating within human populations today.

More is known about the genetic architecture of mutagenesis in model organisms, including the single-celled organism *Saccharomyces cerevisiae,* where it is tractable to disentangle genetic mutator effects from environmental ones by accumulating mutations on different genetic

backgrounds in controlled laboratory environments (Huang *et al.*, 2003; Herr *et al.*, 2011; Lang, Parsons and Gammie, 2013; Serero *et al.*, 2014; Stirling *et al.*, 2014). Many *S. cerevisiae* mutator alleles have been discovered using genetic screens, which involve creating libraries of artificial mutants in the lab and determining which ones have high mutation rates (Stirling *et al.*, 2014). Mutation rates can be elevated by up to a thousand-fold in lines where DNA proofreading and repair capabilities are artificially knocked out (Herr *et al.*, 2011; Lang, Parsons and Gammie, 2013; Serero *et al.*, 2014), and quantitative trait loci with more modest effects have been found to underlie a five-fold range of mutation rate variation among a few natural *S. cerevisiae* strains (Gou, Bloom and Kruglyak, 2019). A more complex mutator phenotype has been observed as a result of epistasis between two incompatible alleles found as natural variation in the mismatch repair genes *MLH1* and *PMS1*, although the natural isolates in which these alleles are found appear to have acquired compensatory variants that suppress this mutator phenotype (Argueso *et al.*, 2003; Heck *et al.*, 2006; Bui *et al.*, 2017; Raghavan *et al.*, 2018).

Mild environmental stressors, such as high salt and ethanol, can also alter the mutation rate of *S. cerevisiae* laboratory strains (Liu and Zhang, 2019; Voordeckers *et al.*, 2020). The same environmental perturbations can cause detectable changes to the *S. cerevisiae* mutation spectrum. The mutation spectrum has also been observed to depend on whether *S. cerevisiae* is replicating in a haploid or diploid state (Sharp *et al.*, 2018). In addition environmental mutagens, more complex ploidy, and genetic mutation rate modifiers could all conceivably affect the mutation spectrum of natural variation as it accumulates. However, no study to our knowledge has looked at whether any mutational signatures measured in the laboratory are capable of explaining natural mutation spectrum variation observed in polymorphism data from a model species.

Recently, comprehensive sampling efforts have produced a collection of 1011 natural isolates of *S. cerevisiae* (Peter *et al.*, 2018). This is a uniquely powerful system containing abundant natural variation that accumulated within natural environments during the recent and ancient evolution of *S. cerevisiae*, and the panel is also amenable for experimental accumulation of mutations over laboratory growth. Many genetic polymorphisms differentiate these strains, and these are relics of mutations that accumulated over many generations on divergent genetic backgrounds adapted to diverse environmental conditions, ranging from forests to beverage fermentation pipelines. Both environmental mutagens and genetic mutators may have created differences among the mutation spectra of these 1011 strains, but only genetically determined mutation spectrum differences should have the potential to be reproduced in the spectra of mutations accumulated in a controlled lab environment.

We hypothesized that yeast strains with outlying spectra of natural polymorphisms are more likely to have distinct *de novo* mutation spectra than strains whose polymorphisms have indistinguishable mutation spectra. The same hypothesis underlies previous inferences of *de novo* mutation spectrum variation from polymorphism data (Harris, 2015; Harris and Pritchard, 2017; Dumont, 2019; Goldberg and Harris, 2019), but was not directly testable in any previously

analyzed species. Mutation accumulation experiments have shown that ascomycete and basidiomycete yeast have distinct mutation spectra despite having similar overall mutation rates (Long *et al.*, 2016), but such comparisons have not been performed on more closely related yeast strains. To enable such direct testing for the first time at higher throughput in *S. cerevisiae,* we describe a new Luria-Delbrück-based assay that efficiently measures the spectra of *de novo* mutations in haploid strains using pooled amplicon sequencing. We then use this assay to identify strains with reproducibly measurable mutator phenotypes that explain the spectrum biases of these strains' polymorphisms. Some proportion of natural mutation spectrum variation might not be reproducible in the lab if it is driven by environmental mutagens, bioinformatic artifacts, or extinct genetic mutators, but our assay has the potential to identify which gradients of mutation spectrum variance are driven by extant genotypic differences.

Results

The mutation spectrum of natural variation in *S. cerevisiae*

To measure the mutation spectrum of genetic variation present in the 1011 *S. cerevisiae* natural isolates (Peter *et al.*, 2018), we polarized single nucleotide polymorphisms using the outgroup *S. paradoxus* (Yue *et al.*, 2017), then classified them into two transition types and four transversion types based on their ancestral and derived alleles. Closely related strains were excluded to avoid overrepresentation of certain groups (Materials and Methods). We calculated the proportion of each mutation type among the derived alleles present in each individual strain, utilizing all derived variants present below 50% frequency. In order to minimize bias from ancestral allele misidentification, we excluded strains with extensive, pre-documented introgression from *S. paradoxus* (Peter *et al.*, 2018). Principal component analysis (PCA) on these individual mutation spectra reveals that strains from the same population tend to have more similar mutation spectra than more distantly related strains (Figure 1A, Supplementary Table S1). Some of this structure disappears when SNPs are subsampled to eliminate double-counting of variants that are shared among multiple strains (Supplementary Figure S1), but several clades appear as consistent outliers in both analyses, including the African beer and European wine strains. The compact architecture of the yeast genome makes it infeasible to exclude coding regions and conserved regions, but a PCA constructed using only synonymous protein-coding variants recapitulates similar PC structures inferred using all polymorphisms passing quality filters (Supplementary Figure S2).

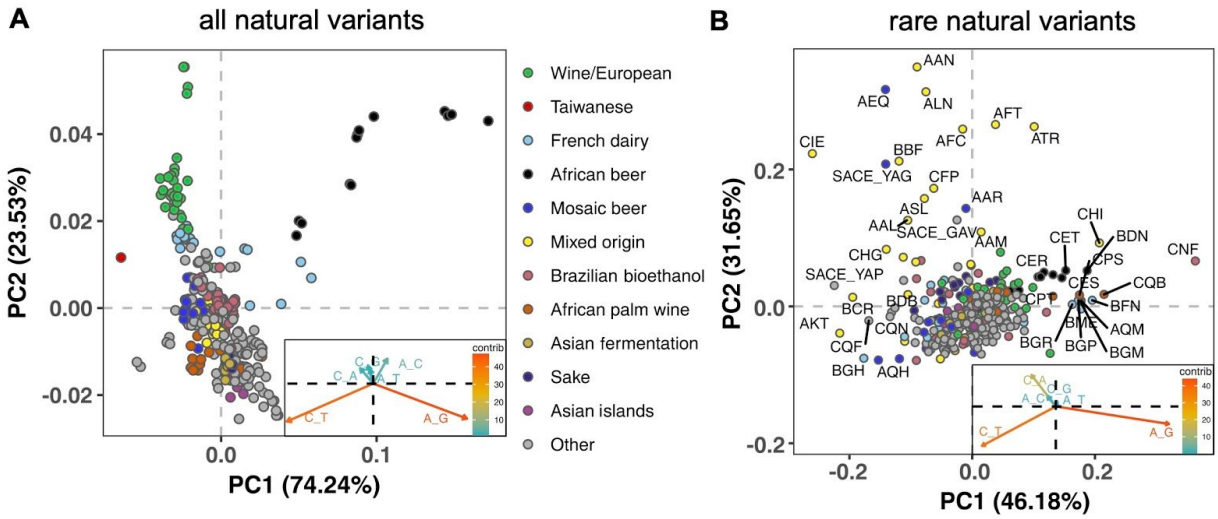

Figure 1. Mutation spectra of natural isolates of *S. cerevisiae*

Principal component analysis of segregating mutation spectrum variation from a subset of the 1011 yeast strains. **A**. Mutation spectrum PCA of all natural variants under 50% derived allele frequency. Each strain's mutation spectrum histogram is projected as a single point, colored to indicate its population of origin (Peter *et al.*, 2018). The inset summarizes the loadings of the first and second principal component vectors. **B**. Mutation spectrum PCA of rare variants (derived allele count 2-4). Singleton variants are excluded to minimize the impact of sequencing error. Strains appearing more than 1.8 standard deviations from the origin along both PC1 and PC2 are labeled with their strain names.

Figure 1A shows that the Taiwanese and African beer populations are outliers along PC1. As seen from the principal component loadings, these two groups mainly differ from the rest in the relative proportions of the two transition types (A>G and C>T): Taiwanese strains are enriched for C>T mutations while African beer strains are enriched for A>G mutations. In contrast, PC2 separates the majority of other populations, such as human-associated strains isolated from wine, dairy, and bioethanol production, along a gradient of varying transition/transversion ratio.

Although strains from the same population tend to cluster together, this trend is less pronounced in the 1011 *S. cerevisiae* genomes than in previously reported mutation spectrum PCAs of humans, great apes, and mice (Harris and Pritchard, 2017; Dumont, 2019; Goldberg and Harris, 2019). That being said, one methodological difference from these previous studies is that we only partition the yeast mutation spectra into six basic types (A>C, A>G, etc.) rather than the 96 trinucleotide-based types used in analyses of vertebrate mutation spectra, a concession to the small size of the yeast genome. We found that the trinucleotide mutation spectra of yeast exhibit similar PCA structure (Supplementary Figure S3), but that the sparsity of yeast triplet spectra appears to limit their utility.

Figure 1B shows a PCA of rare variant mutation spectra from the same collection of strains used in 1A. We define rare variants as those with derived allele counts of 2, 3, and 4 and exclude singletons to minimize the impact of sequencing error. These spectra are noisier than the spectra computed from variants up to 50% frequency, but are potentially more likely to

reflect recently active mutational processes. While these noisier spectra exhibit less clustering by population than those in Figure 1A, a subset of strains from several groups appear as outliers. For example, a few Mixed origin and Mosaic beer strains are outliers along a C>A mutation gradient, and African beer and French dairy strains separate out along an A>G mutation gradient. For completeness, we also examined mutation spectra of singleton variants alone (Materials and Methods), which are the youngest mutations among polymorphisms (Supplementary Figure S4). It resembles the PCA of non-singleton rare variants, except that C>A mutation variation explains a larger variation and becomes the PC1 axis.

Several previous studies have found a puzzling discrepancy between the spectra of *de novo* mutations and polymorphisms in *S. cerevisiae*: polymorphisms have a transition-to-transversion (ts/tv) ratio around 3, compared to only 1 for *de novo* mutations (Agier and Fischer, 2012; Zhu, Sherlock and Petrov, 2017). Our analysis of the 1011 strain collection replicates this finding (Supplementary Figure S5). We also replicate the prior finding that singletons and higher frequency variants have nearly identical ts/tv ratios, but that singletons inferred to be young based on their presence on long shared haplotypes have a lower ts/tv ratio somewhat closer to that of new mutations (Supplementary Figure S5).

A scalable experimental pipeline for measuring mutation rates and spectra

In order to test whether any of the mutation spectrum differences evident from natural variation in different *S. cerevisiae* strains are driven by extant genetic mechanisms that increase the rates of specific mutation types, we set out to measure several strains' *de novo* mutation spectra and rates experimentally. To this end, we developed an experimental pipeline using the reporter gene *CAN1*. Traditional reporter gene fluctuation assays only estimate the overall rate of mutations, but we introduced an extra step that utilizes Illumina sequencing of pooled amplicons derived from *CAN1* mutants to estimate each strain's mutation spectrum as well.

The gene *CAN1* encodes a transport protein that imports arginine and arginine analogs into yeast cells from the surrounding growth media. This means that strains with a functional *CAN1* transporter are sensitive to poisoning by the arginine analog canavanine, while a single loss-of-function mutation can render such cells able to survive on canavanine media (Whelan, Gocke and Manney, 1979). Poisoning a culture with canavanine is thus a very efficient method to select for cells with point mutations in *CAN1*. A limitation of this method is that it only works on genomes that contain exactly one functional copy of *CAN1*, since canavanine resistance is recessive. This means that it cannot be used to measure mutation rates in diploid or polyploid strains directly, which unfortunately include the Taiwanese strains and African beer strains that are PCA outliers in Figure 1A. However, 133 of the 1011 strains are haploid, leaving many strains of interest that are amenable to the assay, including several outliers in the rare variant PCA (Figure 1B).

A schematic overview of our experimental setup is shown in Figure 2. First, we estimated mutation rates using established fluctuation assay methodology (Lang and Murray, 2008; Gou,

Bloom and Kruglyak, 2019), which involves plating multiple independent cultures from each strain being investigated. We then picked a single colony from each plated culture and grew it to saturation in canavanine-containing media. We then selected mutants observed to grow in culture to similar saturation density and pooled them in equal proportions to give each mutant a roughly equal frequency in the pool. Individual pools of mutants from each strain were then subjected to PCR amplification of *CAN1* followed by Illumina sequencing. Individual mutants were called from the sequencing pools using a customized pipeline (Materials and Methods). Mutations collected from different pools of the same strain were combined to calculate the strain's mutation spectrum. We aimed to collect roughly 300 mutants per strain, enough to detect mutation spectrum differences of the magnitude estimated from polymorphism data in several of the 1011 genomes populations.

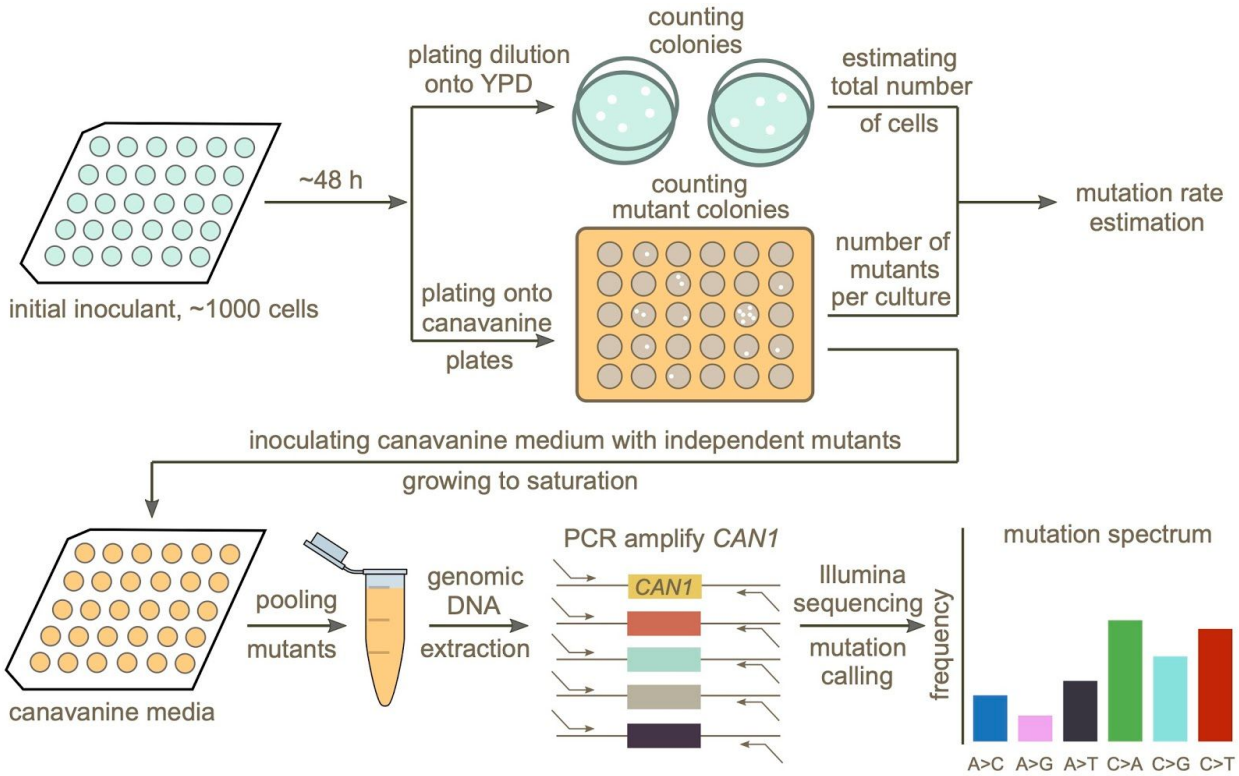

Figure 2. Schematic overview of the experimental pipeline

Overview of the experimental pipeline used to estimate the mutation rate and spectrum for each strain using the reporter gene *CAN1*. First, mutation rates were estimated using fluctuation assays. Independent mutants were then pooled and sequenced to estimate the mutation spectrum of each strain.

Pooling mutants across canavanine media cultures before sequencing allowed us to efficiently estimate mutation spectra at scale, yielding measurements of many individual mutations per library prep. However, pooling too many mutants during this step could have the potential to compromise the pipeline's accuracy by putting the frequency of each mutation too close to the expected frequency of Illumina sequencing errors. To test for this failure mode, we

Sanger-sequenced 38 independent mutants generated using the lab strain LCTL1 (SEY6211-MATα). After pooling these 38 mutants, we performed two replicate library preps and Illumina-sequenced both using our standard procedure. Sanger sequencing identified 37 mutants with single nucleotide mutations plus one containing two adjacent mutations (Supplementary Table S3). We expect each of these mutations, which should be present at a frequency of about 1/38 in the pooled culture, to be easily distinguishable from Illumina sequencing errors that occur at a rate of less than 1% per base.

To identify *bona fide* mutations from each Illumina sequencing pool, we developed a pipeline designed to call mutations present at or above an expected frequency that is inversely proportional to the number of mutants being pooled. In order to minimize false positive mutation calls introduced by sequencing errors, we excluded low coverage regions located at the ends of the amplicons (Materials and Methods). We also identified multinucleotide mutations (MNMs) based on the co-occurrence of variation on the same reads (Averof *et al.*, 2000; Schrider, Hourmozdi and Hahn, 2011), separating these complex mutations from single base substitutions and small indels. When we tested this Illumina sequencing pipeline on the same mutant pool that we previously Sanger sequenced, we detected 37 of the 38 mutations identified by Sanger sequencing, missing only one mutation that occurred at the end of the amplicon located outside our pipeline's callable region. A second Illumina sequencing replicate measured only 36 of these mutations, missing one additional true mutation. Neither Illumina replicate produced any false positives, verifying that the pipeline is accurate enough to permit pooling of up to 40 *CAN1* mutants before each library prep.

Mutation rate variation among haploid natural isolates

We used our pipeline to measure mutation rates and spectra in 16 haploid strains from a wide variety of environments (Supplementary Table S2). Wherever possible, we selected two euploid strains per clade without any copy number variation at the scale of whole chromosome arms. We also selected two lab strains, LCTL1 and LCTL2, to use as controls, since their mutation rates were previously measured. The mutation rate of LCTL1 was measured using a genome-wide mutation accumulation assay by Sharp et al. (2018), while the mutation rate of LCTL2 (GIL 104, a derivative of W303) was measured by Lang et al. (2008) using a *CAN1* fluctuation assay. Two additional strains from the 1011 collection, AAA and ACS, were selected because their mutation rates had been previously measured in another study (Gou, Bloom and Kruglyak, 2019).

We observed a ten-fold range of mutation rate variation in *CAN1* among the strains we surveyed: from 2.1×10^{-7} to 2.1×10^{-6} canavanine resistance mutations per gene per cell division (Figure 3). This range of variation is larger than the five-fold range of mutation rate variation found among six *S. cerevisiae* strains in a recent study (Gou, Bloom and Kruglyak, 2019). All estimates from different replicates of the same strain were generally consistent with each other, though three strains (ACS, LCTR2, and AAR) showed close to a 2-fold difference between our highest and lowest mutation rate measurements. This is within the margin of error observed in

previously published fluctuation assays performed at large scale (Gou, Bloom and Kruglyak, 2019). Among the three strains (LCTL2, AAA, and ACS) with previously published mutation rate measurements, our results fall within a 1.5-fold range of those estimates, with no particular trend of upward or downward bias.

We noticed that AAR and AEQ, the two strains with the highest mutation rates, formed larger colonies during the fixed duration of the experimental growth period compared to many of the other strains tested. This suggests that AAR and AEQ have either unusually high growth rates or unusually large cell sizes. This motivated us to test for correlation between the mutation rates we measured and strain-specific growth rates reported in the literature (Peter *et al.*, 2018), but overall, we found no significant correlation between these attributes (R^2<0.001; p=0.95) (Supplementary Figure S6).

We observed that strains from the Mosaic beer, Sake, African palm wine, and Asian fermentation clades exhibited higher mutation rates than have been previously reported for any natural *S. cerevisiae* strains. The two strains with the highest mutation rates, roughly 10-fold higher than that of the control strain LCTL1, were AAR and AEQ, both from the Mosaic beer clade. While this is milder than some mutator phenotypes that have been artificially generated in the lab, to our knowledge, no comparably high mutation rate has been previously reported in a natural isolate of *S. cerevisiae*, with the exception of the spore derivatives of the incompatible *cMLH1-kPMS1* diploid natural isolate (Raghavan *et al.*, 2018).

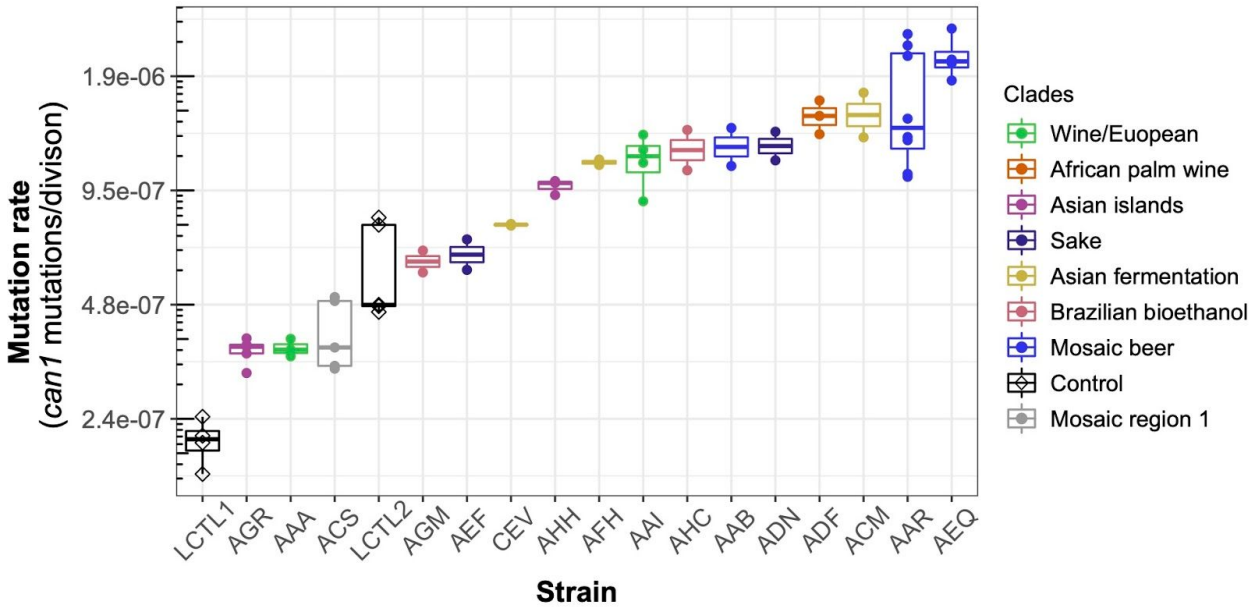

Figure 3. Haploid natural isolates exhibit a 10-fold range of mutation rate variation

Mutation rate variation measured among haploid natural isolates using our *CAN1* reporter gene Luria-Delbrück fluctuation assays. Strains are shown ordered by their mean mutation rates. Mutation rates for each strain were

estimated using at least two replicates, each estimate represented here by a dot. A standard boxplot spans the interquartile confidence interval of possible mutation rates for each strain.

A natural mutator phenotype with a distinctive mutation spectrum

We identified a total of 5571 *CAN1* mutations across all strains, including 4561 point mutations, 837 indels (Supplementary Table S4), and 173 multinucleotide mutations (MNMs) (Supplementary Table S5-S6). 90% of the observed indels are single base-pair indels (754 out of 837), and for simplicity we included only single base-pair indels along with point mutations when reporting each strain's mutation spectrum.

Two of the 4561 point mutations occurred at strain-specific non-reference sites. The remaining 4559 mutations consisted of repeated observations of only 727 unique mutations at 476 positions in *CAN1*. Given that each mutation was observed an average of 6.2 times, it is likely that our dataset contains every possible mutation that causes *CAN1* to lose functionality. We observed 2676 missense mutations, 1866 nonsense mutations and only 17 synonymous mutations. These synonymous mutations made up less than 0.37% of the total point mutations observed; since these are unlikely to have caused *CAN1* to lose functionality, they are likely sequencing errors or hitchhikers that occurred in cells containing other inactivating mutations in *CAN1*. This low synonymous mutation rate further demonstrates the accuracy of our pipeline.

To our knowledge, the largest previous *CAN1* fluctuation assay in *S. cerevisiae* observed point mutations at just 102 distinct positions (Lang and Murray 2008). We observed 100 of these mutant sites in addition to 376 additional mutant sites not previously known to abrogate *CAN1* function. Among the two mutations observed by Lang and Murray that are missing from our dataset, one is located near the end of the *CAN1* amplicon in a region we exclude due to insufficient sequencing coverage in most strains. The other site is the location of a mutation changing the anticodon "CTA" to "TTA," which is synonymous and thus not likely to have affected *CAN1* function.

MNMs are complex mutation events that create multiple nearby substitutions or indels at once, likely as a result of error-prone lesion bypass (Averof *et al.*, 2000; Schrider, Hourmozdi and Hahn, 2011; Stone *et al.*, 2012; Harris and Nielsen, 2014), and we were able to distinguish them from independent sets of point mutations by looking for the presence of multiple mutations on the same Illumina reads (Materials and Methods). We estimate that 3.1% of all mutations are MNMs, similar to the 2.6% reported in a single strain background (Lang and Murray, 2008). Most of our strains have similar ratios of MNMs to single base-pair mutations, except for a few outliers (Figure 4A). For example, AAB has disproportionately more MNMs (Figure 4A and Supplementary Figure S7) while AAR and AEQ have lower ratios of MNMs to single base-pair mutations.

We performed hypergeometric tests to determine whether the mutation spectra we measured from the two control lab strains LCTL1 and LCTL2 were distinct from those measured from

haploid natural isolates and from the spectrum measured from the same LCTL2 strain by Lang et al. (2008) (Materials and Methods). We found the spectra of point mutations we measured from the lab strains LCTL1 and LCTL2 to be statistically indistinguishable from the spectra Lang, et al. (2008) obtained using Sanger sequencing of canavanine-resistant mutants (p = 0.82 for LCTL1 and p = 0.087 for LCTL2). With indels included, our LCTL1 spectrum appears significantly different from that of Lang et al. (2008) (p = 0.0012, Bonferroni corrected p-value: 0.042), but the LCTL2 spectra remain indistinguishable with indels included (p = 0.0054, Bonferroni corrected p-value: 0.189).

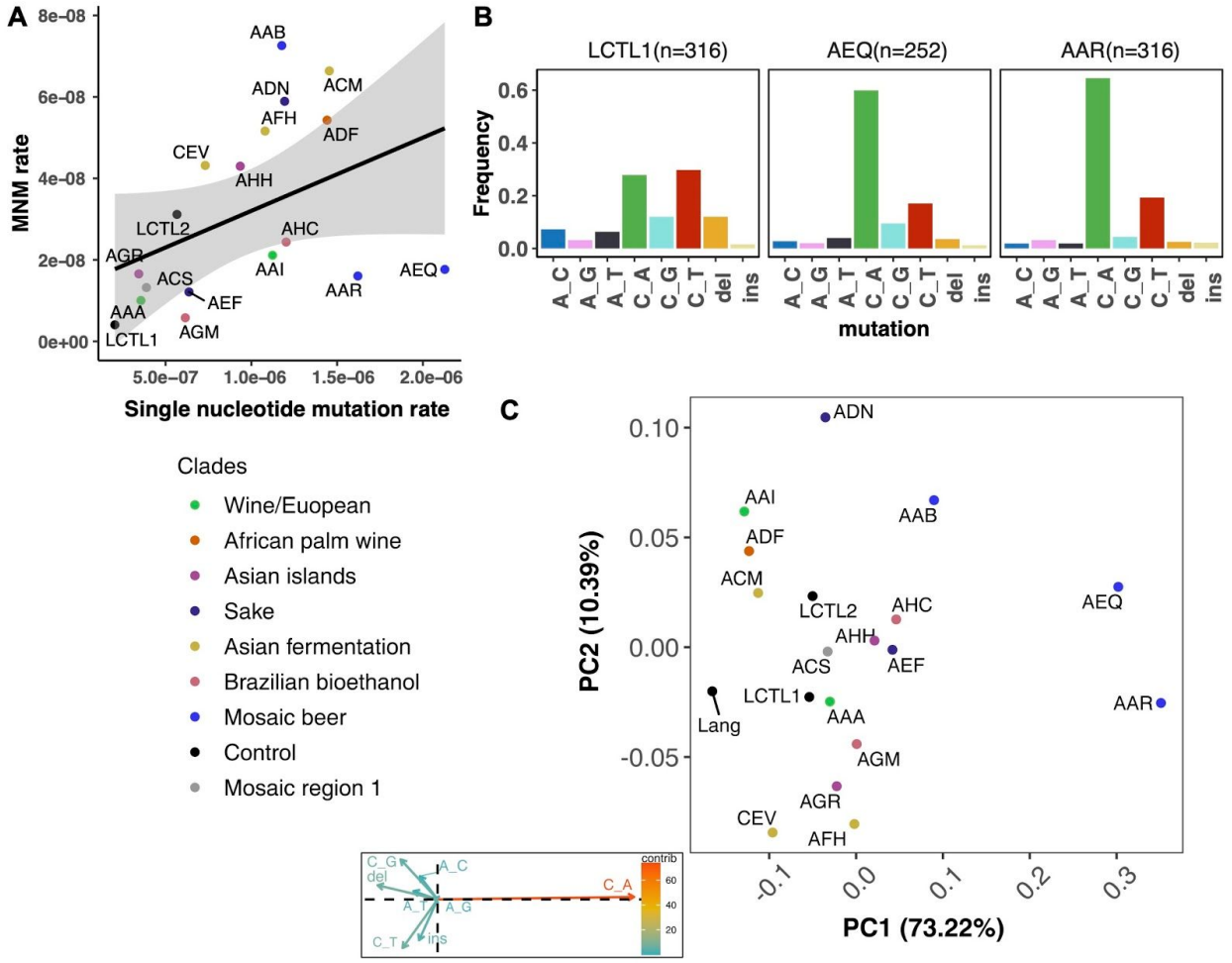

Figure 4. *de novo* mutation rate and spectra in natural isolates

A. Single nucleotide mutation rates plotted against MNM rates across strains. These rates were calculated by multiplying the mean mutation rate estimated using *CAN1* by the proportion of mutations in each strain measured to be either single-nucleotide mutations or MNMs. Here, single nucleotide mutations include both single base pair substitutions and indels. **B**. Mutation spectra in AEQ and AAR show significant enrichment of C>A mutations compared to the control lab strain LCTL1. Only single base-pair indels were used to generate these counts. **C**. A PCA of the same strains' *de novo* mutation spectra compared to the mutation spectrum reported in (Lang and Murray, 2008).

The lab strain LCTL1 appears to have a mutation spectrum that is representative of most natural isolates (Figure 4C, Supplementary Figure S8-S9). Using Bonferroni corrected *p*-values to determine significance, we found the strains AAA, ACS, AGM, AHH, AHC, AEF, ADF, ACM, AGR, AAI, and AAB to have mutation spectra that are statistically indistinguishable from that of LCTL1. We found that CEV and AFH are distinguished only by their high proportions of insertions. ADN showed significant but subtle divergence from LCTL1 in the spectrum of single nucleotide variants (Supplementary Figure S8-S9). In contrast, we measured strikingly divergent mutation spectra from AEQ ($p < 1e-4$) and AAR ($p < 1e-4$), the two strains with 10-fold higher mutation rates than LCTL1. Both strains appear highly enriched for C>A mutations compared to LCTL1 (Figure 4B). The strain AAR's mutation rate estimates appear somewhat bimodal (Figure 3), but C>A mutations are consistently enriched in replicate pools with both lower and higher estimated mutation rates. The main spectrum difference between the two mutation rate modes appears to be a small difference in the C>G mutation proportion (Supplementary Figure S10).

Concordance of the C>A mutator phenotype between inherited variation and *de novo* mutation spectra

In both AEQ and AAR, the proportion of C>A mutations was measured to be elevated nearly 3-fold above the proportion of C>A mutations in LCTL1 and similar strains (Figure 4B). Remarkably, this C>A enrichment appears sufficient to explain the placement of AEQ and AAR as rare variant mutation spectrum outliers that we previously saw in Figure 1B, which was computed from polymorphisms sampled genome-wide, not just within *CAN1*. Both strains have rare variant spectra that are displaced from the population norm along a principal component vector pointing in the direction of increased C>A enrichment. These strains' high mutation rate, C>A-heavy *de novo* mutation spectrum, and concordant C>A-heavy rare variant spectrum all point to the conclusion that these Mosaic beer strains display a naturally-occurring genetically encoded mutator phenotype.

To assess whether the C>A enrichment phenotype observed in AAR and AEQ is likely shared with any of the 1011 strains for which we lack mutation spectrum measurements, we ranked all 1011 strains (excluding close relatives) according to the C>A enrichment of their rare variants (global allele count less than 5). We found that SACE_YAG and BRM, the two strains closest to AAR and AEQ in the global, neighbor-joining phylogeny, clustered with AAR and AEQ in having high C>A fractions within the top 5% in the dataset. The rest of the top-ranking 5% of the strains exhibit some phylogenetic clustering, but no others fall within the Mosaic beer clade (Figure 5C). Instead, they are somewhat dispersed across two large, diverse clades known as "Mosaic region 3" and the "Mixed origin" clade. We also used a bootstrapping method to find strains with enriched C>A fractions, using an empirical *p*-value threshold of 0.05. Many of the same strains are outliers in both tests, including the four Mosaic beer strains (Supplementary Figure S11). The phylogenetic clustering of C>A rare variant enrichment suggests that multiple clades may be genetically predisposed toward accumulating relatively higher rates of this mutation type.

Based on the pattern of C>A enrichment observed in the rare polymorphism data, we hypothesized that a shared mutator allele was responsible for the pattern of C>A enrichment present in the four-strain clade consisting of AAR, AEQ, BRM, and SACE_YAG (Figure 5A,B). In these strains, rarer variants are notably C>A-enriched, but higher frequency variants exhibit weaker enrichment that declines toward the C>A fraction more typical of other strains. In contrast, the four strains most closely related to AAR, AEQ, BRM, and SACE_YAG in the phylogeny exhibit a consistently lower C>A fraction that does not vary with allele frequency (Figure 5B). The concordant enrichment of C>A mutations in rare polymorphisms and *de novo* mutations from the same strains suggests that this C>A enrichment is genetically determined and is not specific to the *CAN1* locus but has affected the entire genome during the recent history of this clade.

Three of the four C>A-enriched mosaic beer strains, AAR, AEQ, and SACE_YAG, are all haploid derivatives of the diploid *Saccharomyces cerevisiae var diastaticus* strain CBS 1782, which was isolated in 1952 from super-attenuated beer (Andrews and Gilliland, 1952). AEQ and AAR differ at roughly 14,000 variant sites (the median pairwise genetic distance in the 1011 strains is 64,000) and SACE_YAG differs from AEQ and AAR at about 11,000 sites each, due to the high level of heterozygosity in the parental diploid strain. The fourth strain with an elevated C>A mutation fraction, BRM, is derived from an independent source: it was isolated in 1988 from a cassava flour factory in Brazil (Laluce *et al.*, 1988). BRM differs at 14,000-17,000 sites from the above mentioned three strains.

Although definitively identifying the genetic variants responsible for C>A enrichment in AEQ and AAR is beyond the scope of this work, we scanned for nonsense and missense mutations in a list of 158 candidate genes. To formulate this list, we combined genes known to play roles in DNA replication and repair with genes that were previously identified to harbor mutator alleles through genetic screens (Supplementary Table S7) (Boiteux and Jinks-Robertson, 2013; Stirling *et al.*, 2014). No candidate premature stop codons were found to be both present in AAR and AEQ and rare (MAF<0.05) in the total population. However, we identified 40 sites with at least one rare non-synonymous allele (MAF<0.05) shared between both AEQ and AAR and absent from the other haploid strains that we experimentally found to have normal mutation spectra (Supplementary Table S8).

One of these missense variants falls within *OGG1*, a gene encoding a glycosylase key to the oxidative stress response that specifically excises 8-oxo-G. *OGG1* null mutants are known to experience high C>A mutation rates (Shockley *et al.*, 2013) and a ten-fold overall increase in mutation rate compared to standard lab strains (Ni, Marsischky and Kolodner, 1999). However, the natural variant we identified in *OGG1* is present not only in the four Mosaic beer strains enriched for C>A rare variants, but is also present in two other strains, AQH and AAQ, whose rare variants are not C>A-enriched. A close examination of mutation spectrum from rare variants revealed that AQH may have a C>A enrichment phenotype that is masked by a high rate of C>T mutations (Supplementary Figure S12). No evidence of C>A enrichment was found in AAQ. This could imply that either the candidate variant in *OGG1* is masked by one or more

additional epistatic variants in the other two strains or that it is not responsible for the observed mutator phenotype. Further study will be required to determine the genetic architecture of the observed mutator phenotype, which might be caused by variation at multiple loci.

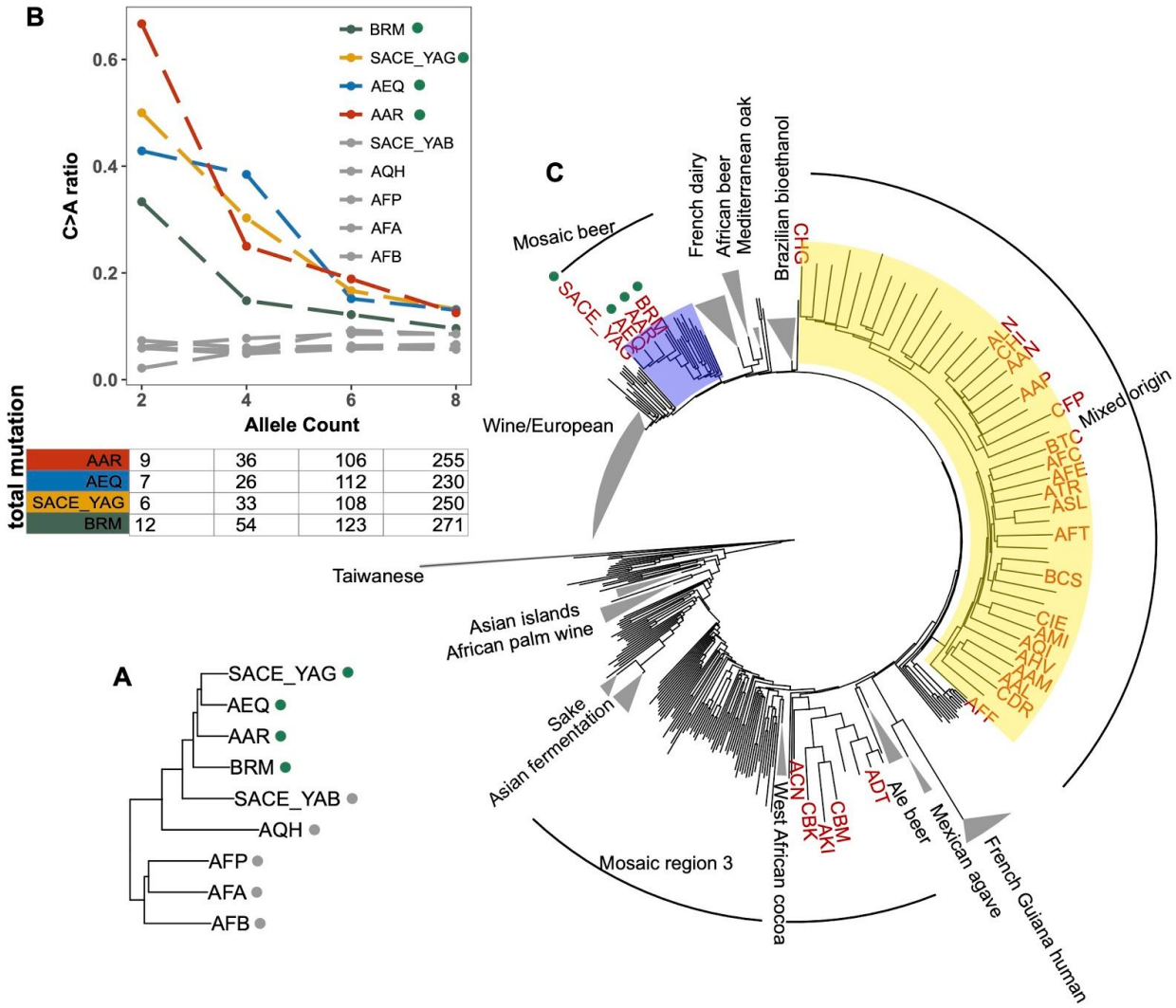

Figure 5. Enrichment of C>A mutations in rare natural variants

A. Phylogeny of AEQ, AAR, and closely related Mosaic beer strains **B.** Top panel: The C>A ratio as a function of minor allele count in Mosaic beer strains that are closely related to AEQ and AAR. C>A ratios in polymorphisms are calculated across allele count (AC) bins with cutoffs of 2, 4, 6, and 8. When computing allele counts, closely-related strains are excluded, and each strain is represented as a diploid in genotype. Bottom panel: total number of variants in each AC bin. **C.** Phylogeny of the 1011 collection with strains in the top 5% of C>A fraction shown in red.

Discussion

To our knowledge, this is one of the first studies to systematically explore the rates and spectra of mutational processes among *S. cerevisiae* natural isolates. Using polymorphism data, we find that the mutation spectrum of *S. cerevisiae* exhibits rich, multidimensional variation. By measuring *de novo* mutation spectra in a representative set of haploid natural strains, we identify a naturally occurring genetically encoded mutator phenotype that appears to be driving at least one dimension of the mutation spectrum variation present in the polymorphism data.

Although the mutation spectrum variation we report here is reminiscent of observed mutation spectrum variation among humans, great apes, and laboratory mouse strains, we note that mutation spectrum divergence among subpopulations of *S. cerevisiae* appears noisier than the separation previously observed among populations and species of vertebrates. While it is possible to infer a human genome's continental group of origin using its mutation spectrum alone, the same is not true of an *S. cerevisiae* genome. Several factors, which are not mutually exclusive, might underlie this difference. One is the presence of pervasive gene flow between *S. cerevisiae* clades (Liti *et al.*, 2009; Schacherer *et al.*, 2009; Peter *et al.*, 2018). Another factor is the small size of the *S. cerevisiae* genome; each strain has two orders of magnitude fewer derived alleles than most vertebrate genomes have. This data sparsity renders individual mutation spectrum measurements relatively noisy and limits our ability to detect how flanking base pairs affect the mutation rate of each site in the genome.

While gene flow and data sparsity might be responsible for the relatively modest magnitude of mutation spectrum divergence between most strains of *S. cerevisiae*, it is also possible that DNA replication and repair are intrinsically more uniform in *S. cerevisiae* than in vertebrates, perhaps because of the greater efficiency of selection against weakly deleterious mutator alleles in a unicellular organism that exists at large effective population sizes and often reproduces asexually. Asexual reproduction should theoretically increase the efficiency of selection against mutator alleles because deleterious variants created by the mutator cannot recombine onto other genetic backgrounds; on the other hand, it can also limit the efficiency of selection against individual deleterious mutations by permamently tethering them to particular genetic backgrounds. Further measurements of mutation spectrum variation within other species will be needed to determine whether the stability observed here is indeed characteristic of unicellular eukaryotes. If mutation spectra tend to be stable within species that have low mutation rates and strong selection against mutation rate modifiers, we might expect to see even less mutation spectrum variation among populations of ciliates like *Paramecium* and *Tetrahymena*, whose mutation rates are substantially lower than that of *S. cerevisiae* (Sung, Tucker, *et al.*, 2012; Long *et al.*, 2016).

Many questions about mutation spectrum variation with *S. cerevisiae* and other species remain unresolved and present important avenues for future work. One obvious unknown is the identity of the gene or genes responsible for the mutator phenotype detected in AEQ and AAR. It is also

unclear whether rare polymorphisms in the Mixed origin and Mosaic region 3 clades are enriched for C>A mutations due to the same genetic mechanisms active in AEQ and AAR. Other genes might underlie the mutation spectrum differences observed among other strains, though our analyses suggest that some mutation spectrum gradients that dominate the common variation PCA are unlikely to be explained by extant mutators. One such gradient is the A>G enrichment in the African beer yeast clade, which is less pronounced in our rare variant PCA (Figure 1B) compared to our PCA extracted from variation of all frequencies (Figure 1A). This is somewhat reminiscent of the frequency distribution of the TCC>TTC mutation "pulse" that distinguishes Europeans and South Asians from other human populations, and may suggest that the African beer A>G enrichment was caused by an extinct mutator allele or a mutagen found in a past environment.

Natural selection might contribute to the mutation spectrum variation within *S. cerevisiae* if certain mutation types are more often beneficial than others and if such asymmetries vary between populations. However, we note that most of the gradient structure observed in our analyses can be reproduced with synonymous mutations alone, meaning that selection is unlikely to explain much of the natural yeast mutation spectrum variation we observe.

The C>A mutations enriched in AEQ, AAR, and their relatives might be a signature of oxidative stress damage; such mutations are a known signature of the repair of 8-oxoguanine lesions, which is consistent with a causal role for these strains' missense substitution in the oxidative stress response gene *OGG1*. At the same time, C>A mutations do not comprise all of the 10-fold excess of mutations measured in AEQ and AAR compared to standard lab strains. Compared to LCTL1, the ratio of C>A mutations to C>T mutations is elevated 3.61-fold in AEQ and 3.33-fold in AAR, which is less than the 10-fold overall elevation of the mutation rate in these strains. These mutation data imply that all mutation types have higher rates in AEQ and AAR compared to other *S. cerevisiae* strains, not just C>A. Further work will be required to test this claim and verify whether the C>A enrichment and broad-spectrum mutation rate increase are driven by the same biochemical mechanism.

A potential limitation of our assay is that we measure mutation spectra using only missense or nonsense mutations that disrupt functionality of *CAN1*, which might accumulate differently than mutations of the same types in other regions of the genome. That being said, the 2676 and 1866 missense and nonsense mutations in our dataset contain numerous instances of all six mutation types that comprise our summary mutation spectrum. Moreover, mutation spectra ascertained from *CAN1* are consistently similar to the spectra measured in MA experiments, with just a slight enrichment of mutations at GC sites (Supplementary Figure S13).

In summary, the results presented in this paper provide the most direct evidence to date that eukaryotic mutation spectra are variable within species (Harris, 2015; Harris and Pritchard, 2017; Dumont, 2019; Goldberg and Harris, 2019). It has been proposed that the best explanation for such mutation spectrum heterogeneity is the frequent emergence of nearly neutral mutator alleles that turn over rapidly as a consequence of weak purifying selection on

the mutation rate. Our *de novo* mutation spectrum measurements provide the first experimental verification of this claim, showing that at least one mutational signature whose activity varies among natural yeast strains is likely caused by an extant mutator allele.

Although our results show that the mutation spectrum bias shared by certain Mosaic beer yeast is genetically encoded, it is worth noting that this C>A gradient is not the principal axis of mutation spectrum variation in the 1011 yeast genomes that we computed from all variants (Figure 1A). It remains to be seen how many other mutator or antimutator alleles might exist within this strain collection and to what extent they can explain the mutation spectrum variation observed among strains from different environments. A broader question still is whether the forces that created *S. cerevisiae*'s mutation spectrum variation are similar to the forces that shaped the distinctive mutation spectra of different human populations and great ape species. If we can identify the genes that underlie natural yeast mutator phenotypes such as the one described in this study, it will likely be more straightforward to test these genes for mutator activity in humans and other species than to discover mutator alleles via any kind of agnostic genome scan.

Acknowledgements

We would like to thank all members of the Harris and Dunham labs for helpful comments and discussions. We thank Nathaniel Sharp and Greg Lang for sharing strains. We also thank Joseph Schacherer for sharing the 1011 strain collection with the Dunham lab. P.J was supported by a Burroughs Wellcome Fund Career Award at the Scientific Interface awarded to K.H. K.H. acknowledges additional support from a Searle Scholarship, a Sloan Research Fellowship, a Pew Biomedical Scholarship, and National Institute of General Medical Sciences Grant 1R35GM133428-01. A.J.H was supported by the National Institute for General Medical Sciences (NIH/NIGMS R01GM118854). A.R.O. was supported by the National Human Genome Research Institute of the NIH under award T32 HG00035. The research of M.J.D. was supported by NIH/NIGMS award P41 GM103533 and a Faculty Scholar grant from the Howard Hughes Medical Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Burroughs Wellcome Fund, the Kinship Foundation, the Sloan Foundation, the Pew Charitable Trust, HHMI, the NIH or NIGMS.

Materials and Methods

Variant filtering and mutation PCA analysis

We filtered the original variants from the 1011 *S. cerevisiae* collection (Peter *et al.*, 2018) by including biallelic SNPs with less than 20% missing genotypes. We restricted to regions in the genomes where reads can be uniquely mapped ("mregions_100_annot_2011.bed" from (Jubin *et al.*, 2014)) and excluded repeat-masked regions. Closely related strains (pairwise genetic distance less than 8000) are excluded from the 1011 dataset. Singletons were excluded when counting individual mutations in Figure 1 to minimize the impact of sequencing errors. Strains

with extensive introgression from *S. paradoxus* (clade 2, 9, 10 from (Peter *et al.*, 2018)) were excluded in order to minimize bias from errors in the inference of ancestral and derived alleles. Ancestral states of mutations were inferred using five *S. paradoxus* sequences (Yue *et al.*, 2017), aligned to the *S. cerevisiae* reference genome R64-1-1 using lastz v1.04.00 (Harris, 2007). Only sites that are fixed in four out of five strains were inferred to be the ancestral alleles, and other sites were ignored. When computing the mutation spectra of strains from variants for Figure 1A, each individual strain was assumed to be diploid, with homozygous derived alleles counted with twice the weight as heterozygous derived alleles. When counting rare variants, homozygous derived alleles were given the same weights as heterozygous derived alleles.

To further minimize confounding of the mutation spectrum by ancestral allele misidentification, only variants with derived allele frequency less than 0.5 were used. Variants that passed all filtering criteria were used to compute a normalized mutation spectrum histogram for each individual strain. When performing PCAs, no more than 30 strains from each population were randomly sampled to minimize bias from uneven sampling. The same strains were used to generate PCA plots in Supplementary Figures S1-S4 (Supplementary Table S1), except that strains with fewer than 8 singletons or rare variants were further excluded when generating Supplementary Figure S4 and Figure 1B. Our definition of singletons varied as a function of ploidy (Supplementary Figure S4-S5): In haploids and homozygous diploids (as defined in (Peter *et al.*, 2018)), a singleton will be fixed in the strain where it occurs (represented as homozygous), but in other types of strains, a singleton is required to be heterozygous. In all cases, a singleton is a variant present in only a single strain.

Fluctuation assays and sequencing

We performed fluctuation assays according to an established protocol (Lang, 2018) with the following modifications: 4μl of overnight inoculant was diluted in 40ml SC-Arginine+2% Glucose media. 50μl of the diluted cultures were distributed in 96-well round-bottom plates (Costar 3788) for each strain. Plates were sealed with Breathe-Easy sealing membrane (Sigma Z380059). SC-Arginine-Serine+Canavanine (60 mg/liter L-canavanine) Omni plates (Nunc OmniTray 242811) were used and dried for 2-4 days in a 30℃ incubator before using. Depending on the strains, 50μl of culture were diluted one- to four-fold when plating on the Omni plates, either to reduce the background or to avoid growth of too many mutant colonies. After plating, the plates were dried and then incubated at 30℃ for 48 hours. Independent mutants from separate cultures were inoculated into 200μl SC-Arginine-Serine+60mg/Liter Canavanine+2% Glucose media, and then grown to saturation over ~43 hours at 30℃ with shaking. Optical densities (ODs) were measured after incubation, and only mutants that reached similar saturation ODs were pooled (150ul each) to achieve equal proportions. Genomic DNA from each pool was extracted using the Hoffman Winston protocol (Hoffman and Winston, 1987). *CAN1* was then PCR amplified using published primers (Lang and Murray, 2008) with 15 cycles. Two independent 25μl PCR reactions were then pooled and cleaned up with a Zymo Clean & Concentrator Kit (D4004). Sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit with customized indices. Sequencing runs with 75 or 150 bp paired-end

reads were performed using an Illumina NextSeq 550 sequencer (Raw reads uploaded to SRA are pending).

Calculation of mutation rates

The rSalvador package (Zheng, 2017) was used to estimate the number of mutation events (*m*) in each fluctuation assay using maximum likelihood under the Lea-Coulson model (Luria and Delbrück, 1943; Lea and Coulson, 1949; Ma, G. Vh. Sandri and Sarkar, 1992). The total number of cells (*Nt*) was measured by counting colonies seeded with dilutions of cells on YPD plates with dilutions ranging from 1:10,000 to 1:40,000. The rate of loss-of-function mutations per *CAN1* gene per cell division was estimated to be *m/Nt*.

Mutation calling

Sequencing reads were first mapped using bowtie v2.2.3 (Langmead and Salzberg, 2012) to the Scer3 S288C reference *CAN1* PCR fragment sequence using primers designed by Lang et al (2008). Mutation coordinates were therefore called relative to the start of the *CAN1* amplicon. Adapters were trimmed using the program trim_galore v0.6.6 (Krueger, no date) and paired-end reads were merged using pear v0.9.11 (Zhang *et al.*, 2014). The command `fastq_quality_filter -q 20 -p 94` was used to remove low quality reads before running bowtie. A MAPQ cutoff of 40 was used for SNPs and a cutoff of 20 was used for indels. Pysamstats v1.1.2 (Miles, no date) was used to compute the frequencies of all possible alleles at each base pair. Sites with read depth less than 200 or with less than 40% coverage of the amplicon were excluded. After the first round of mapping, sites that were fixed in each strain were called and compared to the SNPs in the 1011 collection to confirm strain identity. We then performed a second round of read mapping using the same pipeline except that each strain's reads were mapped to a strain-specific *CAN1* reference sequence.

For each sequencing pool, we let *N* be the number of mutants that were pooled prior to sequencing. Non-reference alleles with frequencies between $0.65 \times 1/N$ and 0.95 were included as evidence of mutations, discarding alleles below this frequency range as likely to be sequencing errors and alleles above this frequency range as likely to be strain-specific SNPs. Adjacent indels were merged if their frequencies differed by less than 10%). MNMs were identified in each pool by first flagging pairs of mutations occurring at similar frequencies (plus or minus 9%) within 10bp of one another and then verifying the coexistence of the two mutations on at least 70% of the paired-end reads where at least one of the two mutations appears. Complex MNMs containing three or more variants were identified by merging MNMs that share a SNP in common. To obtain single nucleotide mutation counts and indel counts, mutations that are part of MNMs were first excluded from each pool. The coordinates of each mutation were converted back from *CAN1*-specific coordinates to genomic positions. Point mutations were further annotated using VEP (McLaren *et al.*, 2016) to further categorize into missense, nonsense or synonymous mutation types.

Allele frequencies were used to estimate the multiplicity of each mutant as follows: First, the mean and standard deviation of all mutant allele frequencies were calculated from each pool. Each allele frequency more than two standard deviations above the mean was then translated into a mutation count by dividing it by the mean allele frequency and then rounding to the nearest integer. Mutations with frequencies less than two standard deviations above the mean are assumed to be mutations with count 1.

Statistically quantifying mutation spectrum differentiation

To compare the mutation spectra between strains, mutations were first classified as one of the 6 general classes of base-substitutions (A>C, A>G, A>T, C>A, C>G, C>T) or as single base-pair insertions or deletions. We then compared the mutation spectra of the two control strains LCTL1 and LCTL2 to all other haploid isolates as well as one spectrum published by Lang et al (2008) (a total of 35 tests) using a pairwise hypergeometric test (Adams and Skopek, 1987), a custom python script (Tracy *et al.*, 2020). In the first round of this test, the paired mutation counts were arranged in a 2 x 8 contingency table. To test the null hypothesis that the two mutation spectra are the same, the hypergeometric probability of the observed table was calculated and compared to the hypergeometric probabilities of 10,000 random tables with the same row and column totals. The number of random tables with a higher hypergeometric probability than the observed provides an estimate of the *p*-value. We used the conservative Bonferroni correction to compute the significance cutoff (0.05/35=0.001429). A second set of Bonferroni-corrected *p*-values was calculated after excluding indels to form a 2 x 6 contingency table. These *p*-values were used to determine how many of the significant mutation spectrum differences were driven by the indel category (Supplementary Figure S8, S9).

References

Acosta, S. *et al.* (2015) 'DNA Repair Is Associated with Information Content in Bacteria, Archaea, and DNA Viruses', *The Journal of heredity*, 106(5), pp. 644–659.

Adams, W. T. and Skopek, T. R. (1987) 'Statistical test for the comparison of samples from mutational spectra', *Journal of molecular biology*, 194(3), pp. 391–396.

Agier, N. and Fischer, G. (2012) 'The mutational profile of the yeast genome is shaped by replication', *Molecular biology and evolution*, 29(3), pp. 905–913.

Alexandrov, L. B. *et al.* (2013) 'Deciphering signatures of mutational processes operative in human cancer', *Cell reports*, 3(1), pp. 246–259.

André, J.-B. and Godelle, B. (2006) 'The evolution of mutation rate in finite asexual populations', *Genetics*, 172(1), pp. 611–626.

Andrews, B. J. and Gilliland, R. B. (1952) 'SUPER-ATTENUATION OF BEER: A STUDY OF THREE ORGANISMS CAPABLE OF CAUSING ABNORMAL ATTENUATIONS', *Journal of the*

Institute of Brewing. Institute of Brewing , 58(3), pp. 189–196.

Antonarakis, S. E. and Beckmann, J. S. (2006) 'Mendelian disorders deserve more attention', *Nature reviews. Genetics*, 7(4), pp. 277–282.

Argueso, J. L. *et al.* (2003) 'Systematic mutagenesis of the Saccharomyces cerevisiae MLH1 gene reveals distinct roles for Mlh1p in meiotic crossing over and in vegetative and meiotic mismatch repair', *Molecular and cellular biology*, 23(3), pp. 873–886.

Averof, M. *et al.* (2000) 'Evidence for a high frequency of simultaneous double-nucleotide substitutions', *Science*, 287(5456), pp. 1283–1286.

Beckman, R. A. and Loeb, L. A. (1993) 'Multi-stage proofreading in DNA replication', *Quarterly reviews of biophysics*, 26(3), pp. 225–331.

Boiteux, S. and Jinks-Robertson, S. (2013) 'DNA repair mechanisms and the bypass of DNA damage in Saccharomyces cerevisiae', *Genetics*, 193(4), pp. 1025–1064.

Bui, D. T. *et al.* (2017) 'Mismatch Repair Incompatibilities in Diverse Yeast Populations', *Genetics*, 205(4), pp. 1459–1471.

Chao, L. and Cox, E. C. (1983) 'Competition between high and low mutating strains of Escherichia coli', *Evolution; international journal of organic evolution*, 37(1), pp. 125–134.

Crow, J. F. (1997) 'The high spontaneous mutation rate: is it a health risk?', *Proceedings of the National Academy of Sciences of the United States of America*, 94(16), pp. 8380–8386.

Drake, J. W. (1991) 'A constant rate of spontaneous mutation in DNA-based microbes', *Proceedings of the National Academy of Sciences of the United States of America*, 88(16), pp. 7160–7164.

Dumont, B. L. (2019) 'Significant Strain Variation in the Mutation Spectra of Inbred Laboratory Mice', *Molecular biology and evolution*, 36(5), pp. 865–874.

Eigen, M. (1971) 'Selforganization of matter and the evolution of biological macromolecules', *Die Naturwissenschaften*, 58(10), pp. 465–523.

Eisen, J. A. and Hanawalt, P. C. (1999) 'A phylogenomic study of DNA repair genes, proteins, and processes', *Mutation research*, 435(3), pp. 171–213.

Farlow, A. *et al.* (2015) 'The Spontaneous Mutation Rate in the Fission Yeast Schizosaccharomyces pombe', *Genetics*, 201(2), pp. 737–744.

Goldberg, M. E. and Harris, K. (2019) 'Great ape mutation spectra vary across the phylogeny and the genome due to distinct mutational processes that evolve at different rates', *bioRxiv*. doi: 10.1101/805598.

Gompel, N. *et al.* (2005) 'Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila', *Nature*, 433(7025), pp. 481–487.

Good, B. H. *et al.* (2017) 'The dynamics of molecular evolution over 60,000 generations',

Nature, 551(7678), pp. 45–50.

Gou, L., Bloom, J. S. and Kruglyak, L. (2019) 'The Genetic Basis of Mutation Rate Variation in Yeast', *Genetics*, 211(2), pp. 731–740.

Harris, K. (2015) 'Evidence for recent, population-specific evolution of the human mutation rate', *Proceedings of the National Academy of Sciences of the United States of America*, 112(11), pp. 3439–3444.

Harris, K. and Nielsen, R. (2014) 'Error-prone polymerase activity causes multinucleotide mutations in humans', *Genome research*, 24(9), pp. 1445–1454.

Harris, K. and Pritchard, J. K. (2017) 'Rapid evolution of the human mutation spectrum', *eLife*, 6. doi: 10.7554/eLife.24284.

Harris, R. S. (2007) *Improved pairwise alignment of genomic dna*. phd. Pennsylvania State University. Available at: https://dl.acm.org/doi/book/10.5555/1414852.

Heck, J. A. *et al.* (2006) 'Negative epistasis between natural variants of the Saccharomyces cerevisiae MLH1 and PMS1 genes results in a defect in mismatch repair', *Proceedings of the National Academy of Sciences of the United States of America*, 103(9), pp. 3256–3261.

Herr, A. J. *et al.* (2011) 'Mutator suppression and escape from replication error-induced extinction in yeast', *PLoS genetics*, 7(10), p. e1002282.

Hoffman, C. S. and Winston, F. (1987) 'A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformaion of Escherichia coli', *Gene*, 57(2), pp. 267–272.

Huang, M.-E. *et al.* (2003) 'A genomewide screen in Saccharomyces cerevisiae for genes that suppress the accumulation of mutations', *Proceedings of the National Academy of Sciences of the United States of America*, 100(20), pp. 11529–11534.

Hwang, D. G. and Green, P. (2004) 'Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 101(39), pp. 13994–14001.

Iossifov, I. *et al.* (2014) 'The contribution of de novo coding mutations to autism spectrum disorder', *Nature*, 515(7526), pp. 216–221.

Johnson, T. (1999) 'Beneficial mutations, hitchhiking and the evolution of mutation rates in sexual populations', *Genetics*, 151(4), pp. 1621–1631.

Jubin, C. *et al.* (2014) 'Sequence profiling of the Saccharomyces cerevisiae genome permits deconvolution of unique and multialigned reads for variant detection', *G3* , 4(4), pp. 707–715.

Keightley, P. D. *et al.* (2009) 'Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines', *Genome research*, 19(7), pp. 1195–1201.

Kessler, M. D. *et al.* (2020) 'De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population', *Proceedings of the National Academy*

of Sciences of the United States of America, 117(5), pp. 2560–2569.

Kimura, M. (1967) ‘On the evolutionary adjustment of spontaneous mutation rates*’, *Genetics research*, 9(1), pp. 23–34.

Krueger, F. (no date) *TrimGalore*. Github. Available at: https://github.com/FelixKrueger/TrimGalore (Accessed: 28 May 2020).

Laluce, C. *et al.* (1988) ‘New amylolytic yeast strains for starch and dextrin fermentation’, *Applied and environmental microbiology*, 54(10), pp. 2447–2451.

Lang, G. I. (2018) ‘Measuring Mutation Rates Using the Luria-Delbrück Fluctuation Assay’, in Muzi-Falconi, M. and Brown, G. W. (eds) *Genome Instability: Methods and Protocols*. New York, NY: Springer New York, pp. 21–31.

Lang, G. I. and Murray, A. W. (2008) ‘Estimating the Per-Base-Pair Mutation Rate in the Yeast Saccharomyces cerevisiae’, *Genetics*, 178(1), pp. 67–82.

Lang, G. I. and Murray, A. W. (2011) ‘Mutation rates across budding yeast chromosome VI are correlated with replication timing’, *Genome biology and evolution*, 3, pp. 799–811.

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* , 3(9), pp. 1453–1465.

Langmead, B. and Salzberg, S. L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nature methods*, 9(4), pp. 357–359.

Lea, D. E. and Coulson, C. A. (1949) ‘The distribution of the numbers of mutants in bacterial populations’, *Journal of genetics*, 49(3), pp. 264–285.

Leigh, E. G. (1970) ‘Natural Selection and Mutability’, *The American naturalist*, 104(937), pp. 301–305.

Liti, G. *et al.* (2009) ‘Population genomics of domestic and wild yeasts’, *Nature*, 458(7236), pp. 337–341.

Liu, H. and Zhang, J. (2019) ‘Yeast Spontaneous Mutation Rate and Spectrum Vary with Environment’, *Current biology: CB*, 29(10), pp. 1584–1591.e3.

Loeb, L. A. (2016) ‘Human Cancers Express a Mutator Phenotype: Hypothesis, Origin, and Consequences’, *Cancer research*, 76(8), pp. 2057–2059.

Long, H. *et al.* (2016) ‘Similar Mutation Rates but Highly Diverse Mutation Spectra in Ascomycete and Basidiomycete Yeasts’, *Genome biology and evolution*, 8(12), pp. 3815–3821.

Luria, S. E. and Delbrück, M. (1943) ‘Mutations of Bacteria from Virus Sensitivity to Virus Resistance’, *Genetics*, 28(6), pp. 491–511.

Lynch, M. *et al.* (2008) ‘A genome-wide view of the spectrum of spontaneous mutations in yeast’, *Proceedings of the National Academy of Sciences of the United States of America*,

105(27), pp. 9272–9277.

Lynch, M. *et al.* (2016) ‘Genetic drift, selection and the evolution of the mutation rate’, *Nature reviews. Genetics*, 17(11), pp. 704–714.

Ma, W. T., G. Vh. Sandri and Sarkar, S. (1992) ‘Analysis of the Luria-Delbrück Distribution Using Discrete Convolution Powers’, *Journal of applied probability*, 29(2), pp. 255–267.

McGregor, A. P. *et al.* (2007) ‘Morphological evolution through multiple cis-regulatory mutations at a single gene’, *Nature*, 448(7153), pp. 587–590.

McLaren, W. *et al.* (2016) ‘The Ensembl Variant Effect Predictor’, *Genome biology*, 17(1), p. 122.

Miles, A. (no date) *pysamstats*. Github. Available at: https://github.com/alimanfoo/pysamstats (Accessed: 28 May 2020).

Nei, M. (1983) ‘Genetic polymorphism and the role of mutation in evolution’, *Evolution of genes and proteins*, 71, pp. 165–190.

Ni, T. T., Marsischky, G. T. and Kolodner, R. D. (1999) ‘MSH2 and MSH6 are required for removal of adenine misincorporated opposite 8-oxo-guanine in S. cerevisiae’, *Molecular cell*, 4(3), pp. 439–444.

Peter, J. *et al.* (2018) ‘Genome evolution across 1,011 Saccharomyces cerevisiae isolates’, *Nature*, 556(7701), pp. 339–344.

Raghavan, V. *et al.* (2018) ‘Incompatibilities in Mismatch Repair Genes MLH1-PMS1 Contribute to a Wide Range of Mutation Rates in Human Isolates of Baker’s Yeast’, *Genetics*, 210(4), pp. 1253–1266.

Risques, R. A. and Kennedy, S. R. (2018) ‘Aging and the rise of somatic cancer-associated mutations in normal tissues’, *PLoS genetics*, 14(1), p. e1007108.

Scally, A. and Durbin, R. (2012) ‘Revising the human mutation rate: implications for understanding human evolution’, *Nature reviews. Genetics*, 13(10), pp. 745–753.

Schacherer, J. *et al.* (2009) ‘Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae’, *Nature*, 458(7236), pp. 342–345.

Schrider, D. R., Hourmozdi, J. N. and Hahn, M. W. (2011) ‘Pervasive multinucleotide mutational events in eukaryotes’, *Current biology: CB*, 21(12), pp. 1051–1054.

Sebat, J. *et al.* (2007) ‘Strong association of de novo copy number mutations with autism’, *Science*, 316(5823), pp. 445–449.

Ségurel, L., Wyman, M. J. and Przeworski, M. (2014) ‘Determinants of mutation rate variation in the human germline’, *Annual review of genomics and human genetics*, 15, pp. 47–70.

Serero, A. *et al.* (2014) ‘Mutational landscape of yeast mutator strains’, *Proceedings of the*

National Academy of Sciences of the United States of America, 111(5), pp. 1897–1902.

Sharp, N. P. *et al.* (2018) 'The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast', *Proceedings of the National Academy of Sciences of the United States of America*, 115(22), pp. E5046–E5055.

Shockley, A. H. *et al.* (2013) 'Oxidative damage and mutagenesis in Saccharomyces cerevisiae: genetic studies of pathways affecting replication fidelity of 8-oxoguanine', *Genetics*, 195(2), pp. 359–367.

Stirling, P. C. *et al.* (2014) 'Genome destabilizing mutator alleles drive specific mutational trajectories in Saccharomyces cerevisiae', *Genetics*, 196(2), pp. 403–412.

Stone, J. E. *et al.* (2012) 'DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in Saccharomyces cerevisiae', *Environmental and molecular mutagenesis*, 53(9), pp. 777–786.

Sturtevant, A. H. (1937) 'Essays on Evolution. I. On the Effects of Selection on Mutation Rate', *The Quarterly review of biology*, 12(4), pp. 464–467.

Sung, W., Ackerman, M. S., *et al.* (2012) 'Drift-barrier hypothesis and mutation-rate evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 109(45), pp. 18488–18492.

Sung, W., Tucker, A. E., *et al.* (2012) 'Extraordinary genome stability in the ciliate Paramecium tetraurelia', *Proceedings of the National Academy of Sciences of the United States of America*, 109(47), pp. 19339–19344.

Tenaillon, O. *et al.* (2016) 'Tempo and mode of genome evolution in a 50,000-generation experiment', *Nature*, 536(7615), pp. 165–170.

Tracy, M. A. *et al.* (2020) 'Spontaneous Polyploids and Antimutators Compete During the Evolution of Saccharomyces cerevisiae Mutator Cells', *Genetics*, 215(4), pp. 959–974.

Voordeckers, K. *et al.* (2020) 'Ethanol exposure increases mutation rate through error-prone polymerases', *Nature communications*, 11(1), p. 3664.

Wang, L. *et al.* (2019) 'The architecture of intra-organism mutation rate variation in plants', *PLoS biology*, 17(4), p. e3000191.

Whelan, W. L., Gocke, E. and Manney, T. R. (1979) 'The CAN1 locus of Saccharomyces cerevisiae: fine-structure analysis and forward mutation rates', *Genetics*, 91(1), pp. 35–51.

Yue, J.-X. *et al.* (2017) 'Contrasting evolutionary genome dynamics between domesticated and wild yeasts', *Nature genetics*, 49(6), pp. 913–924.

Zhang, J. *et al.* (2014) 'PEAR: a fast and accurate Illumina Paired-End reAd mergeR', *Bioinformatics* , 30(5), pp. 614–620.

Zheng, Q. (2017) 'rSalvador: An R Package for the Fluctuation Experiment', *G3* , 7(12), pp.

3849–3856.

Zhu, Y. O. *et al.* (2014) ‘Precise estimates of mutation rate and spectrum in yeast’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(22), pp. E2310–8.

Zhu, Y. O., Sherlock, G. and Petrov, D. A. (2017) ‘Extremely Rare Polymorphisms in Saccharomyces cerevisiae Allow Inference of the Mutational Spectrum’, *PLoS genetics*, 13(1), p. e1006455.