

RESEARCH ARTICLE

Extremely Rare Polymorphisms in *Saccharomyces cerevisiae* Allow Inference of the Mutational Spectrum

Yuan O. Zhu^{1,2,3}, Gavin Sherlock¹, Dmitri A. Petrov^{2*}

1 Department of Genetics, Stanford University, Stanford, CA, United States of America, **2** Department of Biology, Stanford University, Stanford, CA, United States of America, **3** Genome Institute of Singapore, Singapore

* dpetrov@stanford.edu



OPEN ACCESS

Citation: Zhu YO, Sherlock G, Petrov DA (2017) Extremely Rare Polymorphisms in *Saccharomyces cerevisiae* Allow Inference of the Mutational Spectrum. PLoS Genet 13(1): e1006455. doi:10.1371/journal.pgen.1006455

Editor: Shamil R. Sunyaev, Brigham and Women's Hospital, Harvard Medical School, UNITED STATES

Received: April 13, 2016

Accepted: November 3, 2016

Published: January 3, 2017

Copyright: © 2017 Zhu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are available on NCBI SRA: accession number PRJNA315044

Funding: YOZ was supported by the A*STAR National Science Scholarship PhD. GS was supported by R01 HG003328. DAP was supported by the NIH grants R01GM100366 and R01GM097415. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The characterization of mutational spectra is usually carried out in one of three ways—by direct observation through mutation accumulation (MA) experiments, through parent-offspring sequencing, or by indirect inference from sequence data. Direct observations of spontaneous mutations with MA experiments are limited, given (i) the rarity of spontaneous mutations, (ii) applicability only to laboratory model species with short generation times, and (iii) the possibility that mutational spectra under lab conditions might be different from those observed in nature. Trio sequencing is an elegant solution, but it is not applicable in all organisms. Indirect inference, usually from divergence data, faces no such technical limitations, but rely upon critical assumptions regarding the strength of natural selection that are likely to be violated. Ideally, new mutational events would be directly observed before the biased filter of selection, and without the technical limitations common to lab experiments. One approach is to identify very young mutations from population sequencing data. Here we do so by leveraging two characteristics common to all new mutations—new mutations are necessarily rare in the population, and absent in the genomes of immediate relatives. From 132 clinical yeast strains, we were able to identify 1,425 putatively new mutations and show that they exhibit extremely low signatures of selection, as well as display a mutational spectrum that is similar to that identified by a large scale MA experiment. We verify that population sequencing data are a potential wealth of information for inferring mutational spectra, and should be considered for analysis where MA experiments are infeasible or especially tedious.

Author Summary

The mutational spectrum is central to our understanding of molecular evolution. However, mutational spectra are difficult to study because spontaneous mutations are rare, difficult to observe, and a large number of events is required to detect subtle differences between mutational bias, selection and selection like forces. The possibility of estimating mutational spectra from population polymorphism data, with neither the need for tedious

experiments nor the restrictions and biases of lab conditions, is a crucial step in overcoming such difficulties. We show that with sufficiently broad population sequencing and proper identification of young polymorphisms, it is possible to recapitulate the experimental yeast mutation spectrum. This holds implications for future applications to all species where population sequencing is possible.

Introduction

Knowledge of the mutational spectrum is central to the study of molecular evolution. However, mutational spectra are difficult to characterize because spontaneous mutations are scarce and thus rarely observed in large enough numbers for precise measurements. In addition, mutational spectra vary across species, between individuals, and across genomic segments, placing a demand for methods that can identify a large set of mutational events genome-wide, while remaining applicable to a wide range of species.

One direct approach to the study of spontaneous mutations on a genome-wide scale is through mutation accumulation (MA) experiments. MA experiments allow the accumulation of mutations under minimal selection conditions in a controlled lab environment, usually over many generations [1–4]. If following individual clonal lineages is not feasible, minimal selection conditions are usually achieved in unicellular cultures through repeated extreme bottlenecks, sometimes down to a single individual, such as in *Saccharomyces cerevisiae* [5–13], *Dictyostelium discoideum* [14], *Arabidopsis thaliana* [1], and *Chlamydomonas reinhardtii* [15,16]. It can also be achieved through generations of inbreeding in species such as *Drosophila melanogaster* [17–19], or rhabditid nematods [20]. The final progeny are then sequenced and compared to the starting ancestor to identify *de novo* mutations that occurred within the span of the experiment. The throughput of this process has been greatly aided by recent advances in next generation sequencing, and MA experiments have thus provided significant insights into overall mutation rates, relative frequencies of mutation classes, mutational biases, and repair pathways.

While powerful, MA experiments face certain limitations that cannot be easily rectified. One limitation is technical. Many species cannot be considered for lab studies due to space, life span, ecological, or ethical limitations, if they can be maintained under lab conditions at all. The other limitation is theoretical. Genome stability can be dependent upon environmental factors and life cycle stages [21–23]. For many organisms, including the majority of microbes, such parameters are difficult to characterize. The complex habitats of ‘wild’ populations are thus important but unknown, and therefore cannot be replicated in the lab. In addition, a complex network of genes and pathways regulate DNA repair. Differences in genes involved in DNA fidelity-associated pathways may result in the mutation spectrum varying across sub-populations or even individual strains. As MA experiments usually involve less than a handful of genomic backgrounds that are extremely well adapted to a lab environment, it is possible that they are not representative of the mutational patterns in the species as a whole.

In addition, most MA experiments utilize a relatively small number of lines that are allowed to accumulate relatively large number of mutations for a fairly long period of time. While it is possible to shorten MA experiments, this is often accomplished through the use of mismatch-repair (MMR) impaired strains that accumulate mutations at an artificially fast rate. Such experiments are used to survey large numbers of mutations in a short period of time in a fashion that is specific to the MMR pathway affected. For example, recent work on conditional or complete MMR defect [10, 24–26], nucleotide pool imbalance [27], and replicative polymerase

variants [9,13] has made use of such systems. These experiments are powerful but extremely specific means of probing the DNA replication and repair system, and all mentions of MA experiments in the rest of this paper do not specifically refer to MMR based studies.

In regular MA experiments, where the aim is to study 'natural' mutations spectrum, only 'wild-type' strains are used. For such studies, the MA approach is certainly economical, in that the sequence of a single genome can reveal the presence of a large number of mutations. But the savings come with the cost of two possible sources of bias. First, the MA lines lose fitness as they accumulate mutations and less fit lines might have a very different mutational bias compared to the more fit, naturally occurring lines [28,29]. Second, some MA lines might go extinct—indeed, in most MA experiments they invariably do [7]. The extinct lines are likely to contain some of the most deleterious mutations that will be missed in the final sample of mutations; thus the sequencing of the surviving lines necessarily does not provide a fully unbiased sample of mutations.

An alternative approach to MA experiments relies on the identification of mutations from sequencing of genomes of natural strains. Unlike controlled laboratory experiments, such sequencing can be carried out with most species. Sampling from natural populations further removes many potential biases introduced by lab conditions and experimental set up. Methods that infer mutational spectra from sequence data usually rely upon the assumption that mutations at certain genomic locations are strictly neutral, such as pseudogenes or dead transposable elements [30] that are presumably under no selection pressure, or mutations that lead to a synonymous change in a protein-coding sequence. If this assumption holds, it can be shown that the rate of substitution between species at these sites would directly reflect variation in mutation rates [31–33]. However, it is increasingly apparent that almost no mutations are truly neutral, and even very mild selection or selection-like forces such as biased gene conversion can significantly influence patterns of substitution [34–38]. The overwhelming majority of substitutions observed from sequence data would therefore be survivors of selection and selection like forces, albeit to varying degrees. While extremely informative in their own right, these are necessarily highly biased subsets of the true spectrum of spontaneous mutations.

While divergence data are almost certainly biased by selection, existing polymorphisms within a population need not all be. Segregating alleles can be effectively neutral if they are observed while still under the selection-drift barrier. Because spontaneous mutations necessarily enter the population at a frequency of $1/N$, where N is the number of the chromosomes in the population, identifying a cohort of extremely rare polymorphisms will enrich for very young mutations [39]. Mutational spectra from rare variants through deep population sequencing has already been employed in viral systems such as HIV [40], where the main challenge lies in accurately calling extremely rare variants from a heterogeneous viral population [41–43]. Rare variants have also been applied to characterizing context dependent mutational patterns in 202 human genes [44], although in species where single individual sequencing is accessible and populations are not homogeneous, population structure must be accounted for [45].

One elegant solution would be limiting analysis to *de novo* variants in parent offspring genome comparisons, such as the comparison of family trios in drosophila, butterfly, and humans [46–49]. In many other species, it is not always possible to identify relatedness between individuals ahead of time and selectively sequence parent-offspring genomes. In such instances the relatedness of sampled genomes or genomic regions must be estimated *post hoc*. For a hypothetical organism that reproduces asexually and does not undergo recombination, relatedness between individuals simply involves genomic sequence identity. If two genomes are nearly identical, any variant between them is likely a relatively young mutation that occurred after their last common ancestor. In actual datasets, recombination and/or sexual

reproduction result in genomes with mosaic evolutionary history across genomic segments. To obtain recent mutations from such sequences, regions of identity by descent (IBD) would be more appropriate. However, proper IBD analysis requires haplotype information, which may not always be available, or might be difficult to impute in species such as yeast where ploidy can vary between $1n$ and $4n$ in natural isolates [50].

In the absence of IBD information, on the basis that rare polymorphisms are younger on average, the density of unique SNPs serves as a proxy for IBD information. Genomes with close relatives in the dataset share most of their polymorphisms with at least one other strain and carry few unique mutations, most of which will be young, while genomes with no close relatives share fewer polymorphisms and appear to carry an excessively large number of unique mutations (singletons), most of which will be old. The density of singletons in a genome or genomic region [51], as defined by all polymorphisms present in a sampled population, can serve as a measure of the age of rare variants on that genome.

To test the practicality and accuracy of this technique, we sequenced 141 individual strains of *Saccharomyces cerevisiae* to high genomic coverage and analyzed the mutational spectrum that could be obtained from identified young mutations. By comparing how closely our results matched both theoretical expectations and the mutational spectrum derived from a large-scale MA experiment in yeast, we determined that we could recapitulate the mutation spectrum of a species through broad population sequencing, that is, the sequencing of a large number of individuals.

Results

To sample a set of non-experimental individuals from a relatively diverse population, we sequenced 141 *S. cerevisiae* strains in their natural ploidy states [52]. The majority of these strains were clinical isolates, with around a dozen well-studied commercial and lab strains. Because yeasts are known opportunistic pathogens, this set of strains likely represents the diversity in human-associated yeast populations. SNPs were only called in comparison to the reference sequence of S288C in non-repeat regions after meeting filter requirements (S1 Fig). Excluding one strain where sequencing failed due to contamination, a final set of 423,387 SNPs passed these quality filters (Methods). The site frequency spectrum of the observed population of polymorphisms shows the expected gamma shape of population sequencing datasets, with a small bump around $\text{freq} = 1$ (S2 Fig).

New spontaneous mutations, as a group, should show none of the classical signatures of selection. Three criteria were employed as indicators of our ability to identify very young SNPs: 1) the percentage of nonsynonymous polymorphism (%Pn), 2) the transition transversion (Ts/Tv) ratio, and 3) the GC equilibrium percentage (GCeqm). In divergence data, the ratio of nonsynonymous changes tends to be much lower than the ratio of 0.75 expected in the absence of selection, Ts/Tv values are usually > 2.5 , and the GCeqm (roughly) matches the genomic GC content (which is 38% in yeast). The mutations from a previous large-scale genome-wide MA experiment in yeast yield a %Pn value close to the neutral expectation of 0.75, a Ts/Tv value of 1, and a GCeqm of 32% [12]. We therefore explored our ability to obtain similar values from our polymorphism data.

We first segregated SNPs by their frequencies in the population and summarized all three values for each frequency class. We expected that with decreasing frequency of polymorphisms, the proportion of young SNPs should increase, and the three values should approach those observed in MA experiment (Fig 1 green dotted lines). While the %Pn and Ts/Tv ratios did shift towards MA values, especially in the lowest SNP frequencies, the changes did not reach expected MA values. However, a similar trend was not seen for the value of GCeqm (Fig 1). Indeed, even at the frequency of $1/141$, none came close to matching MA values.

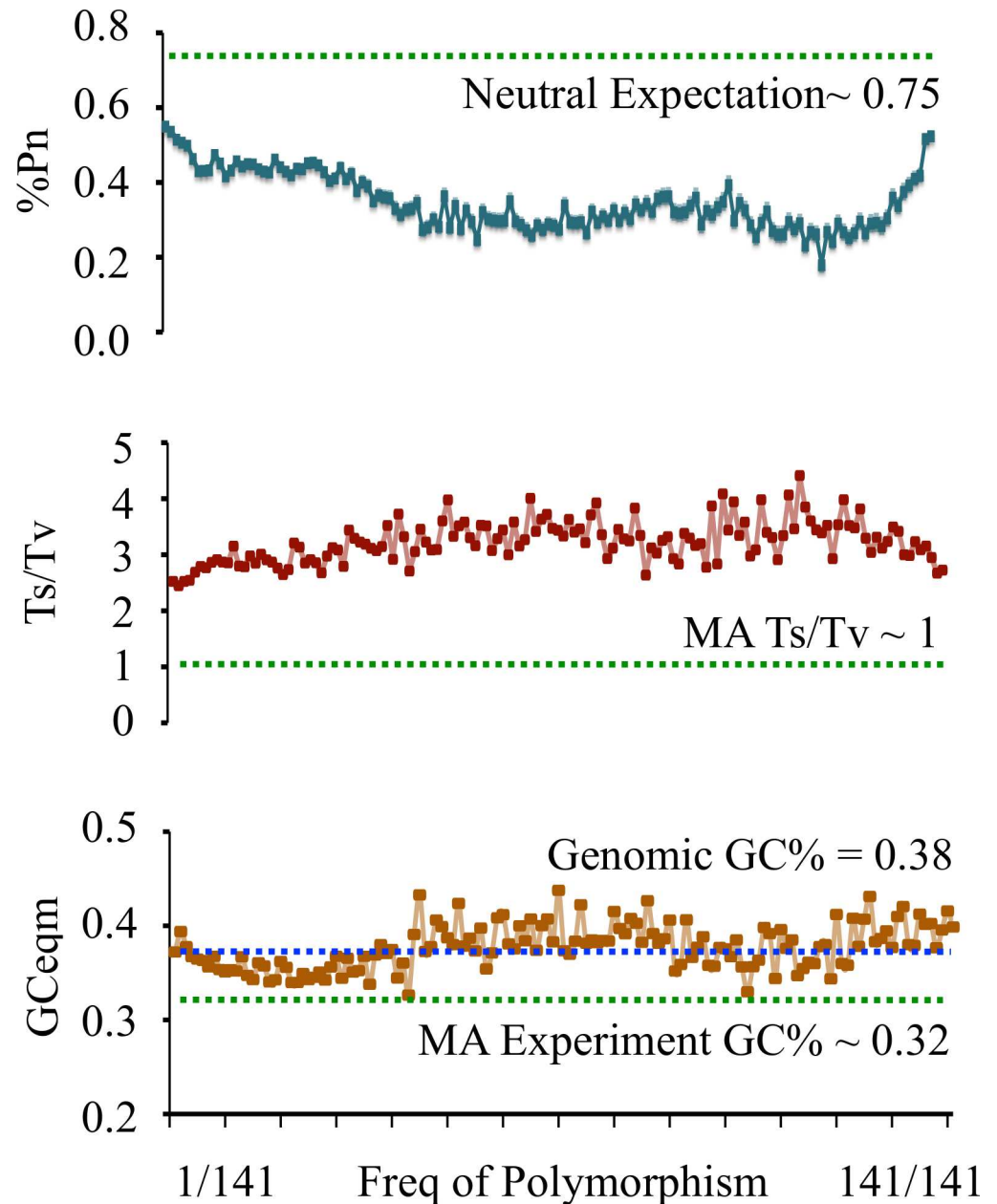


Fig 1. %Pn, Ts/Tv and GCeqm trends across SNP frequency. %Pn and Ts/Tv values show small shifts towards MA/neutral expectations in the lowest SNP frequencies (highlighted in box). X-axis—SNP frequency. Y-axis—%Pn, Ts/Tv, GCeqm.

doi:10.1371/journal.pgen.1006455.g001

Because there is substantial population structure in the sampled strains [52] we tested whether controlling for relatedness between strains could further refine our analysis, this time focusing on just the singletons. We used the density of singletons/kb as a measure of singleton age. For example, if a chromosome carried n singletons, each of the n singletons is given the 'age' of n/length of the chromosome in kb, approximating the time unit it takes for a mutation to occur once per 1 kb since its last common ancestor with the closest sampled relative. Often, chromosomes will carry multiple singletons, and though the singleton mutations must have occurred at different times, it was impossible to accurately identify the order in which these

mutations happened. We chose to be conservative in our age categorization and assign the same age to all singleton mutations on a given chromosome.

We binned SNPs by age into groups of roughly the same sample size, with higher resolution at the youngest ages, ranging from 0.001/kb through 2.25/kb. We then tested whether patterns derived from the younger age groups came closer to the MA experimental values. Plots of the %Pn, Ts/Tv, and GC equilibrium values for each age group showed a clear trend in which the 5 youngest categories (ages <0.005/kb) matched MA values for both Pn/Ps and Ts/Tv ratios (Fig 2). Surprisingly, for GCeqm, the youngest singleton classes suggested an average value of

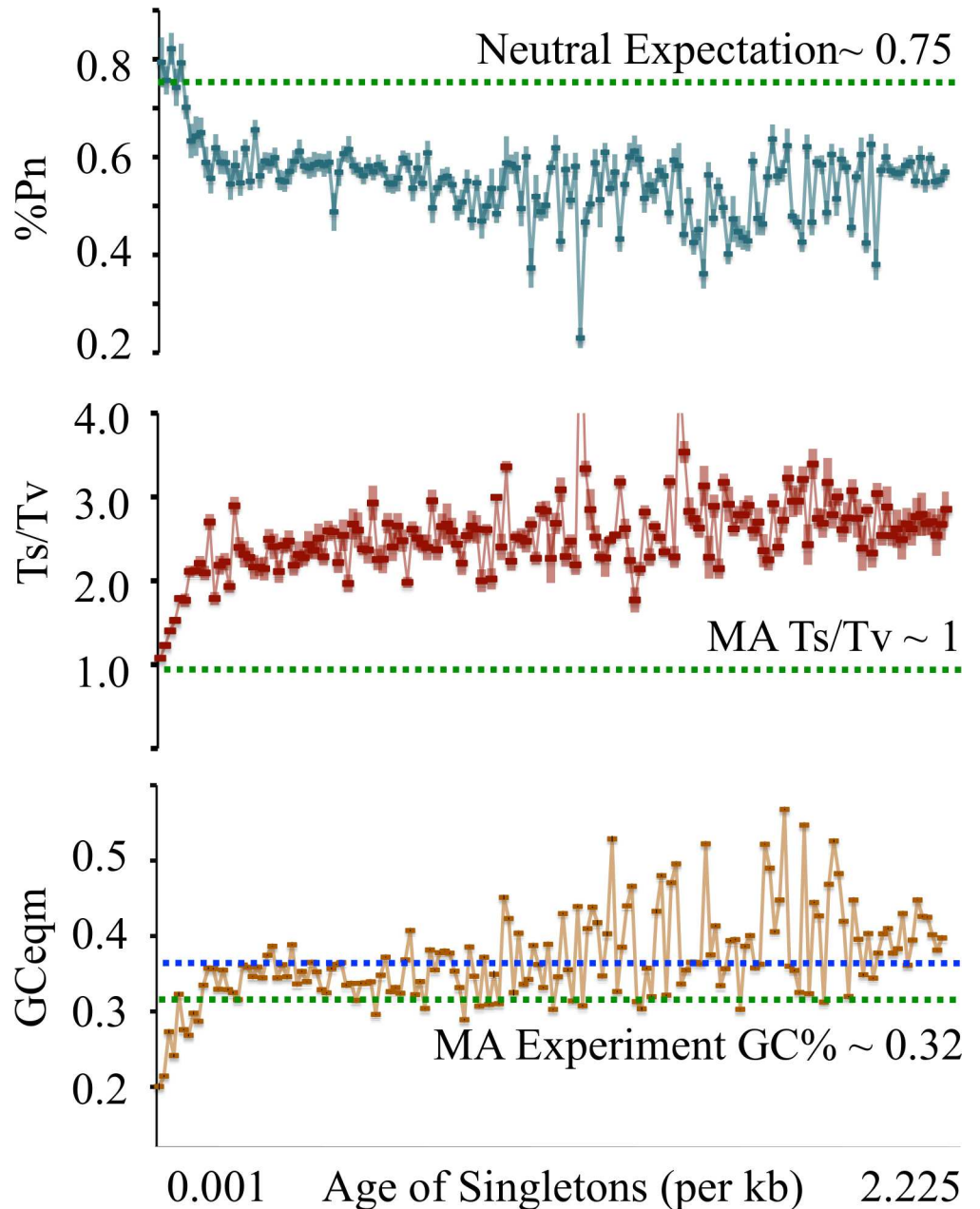


Fig 2. %Pn, Ts/Tv and GCeqm trends across SNP singleton age. %Pn and Ts/Tv values reach MA/neutral expectations in the lowest SNP frequencies (highlighted in box). GCeqm surpasses MA experimental values. X-axis—singletons from youngest to oldest. Y-axis—%Pn, Ts/Tv, GCeqm.

doi:10.1371/journal.pgen.1006455.g002

around 25%, below the 32% derived from the MA experiments (Fig 2). While mutations are indeed AT biased, this value is more extreme than previously reported. To ensure that the youngest singletons as a group were not dominated by low quality SNPs, we noted that coverage depth, genotype qualities, and mapping qualities were not significantly different between young singletons with density <0.005/kb as compared to older singletons, and SNP quality was capped at a minimum of 20 (S3 Fig).

There were 829 singletons of ages <0.005/kb that matched the Pn/Ps and Ts/Tv values from the MA experiment. Coincidentally, this sample size is similar to the 864 SNPs from the MA experiment. Because MA results were based on a single homozygous diploid strain that was exposed to a constant, stable environment, the mutation spectra of a population that is far less homogenous may be different. To determine how the mutation spectra presented by the young singletons differ from old singletons, or from MA data, we calculated the relative mutation rates for all six possible nucleotide changes (Fig 3). Young singleton rates for each nucleotide change were compared to corresponding old singletons and MA rates (Z-test, Bonferroni corrected). There were significant differences in rates between young singletons and old

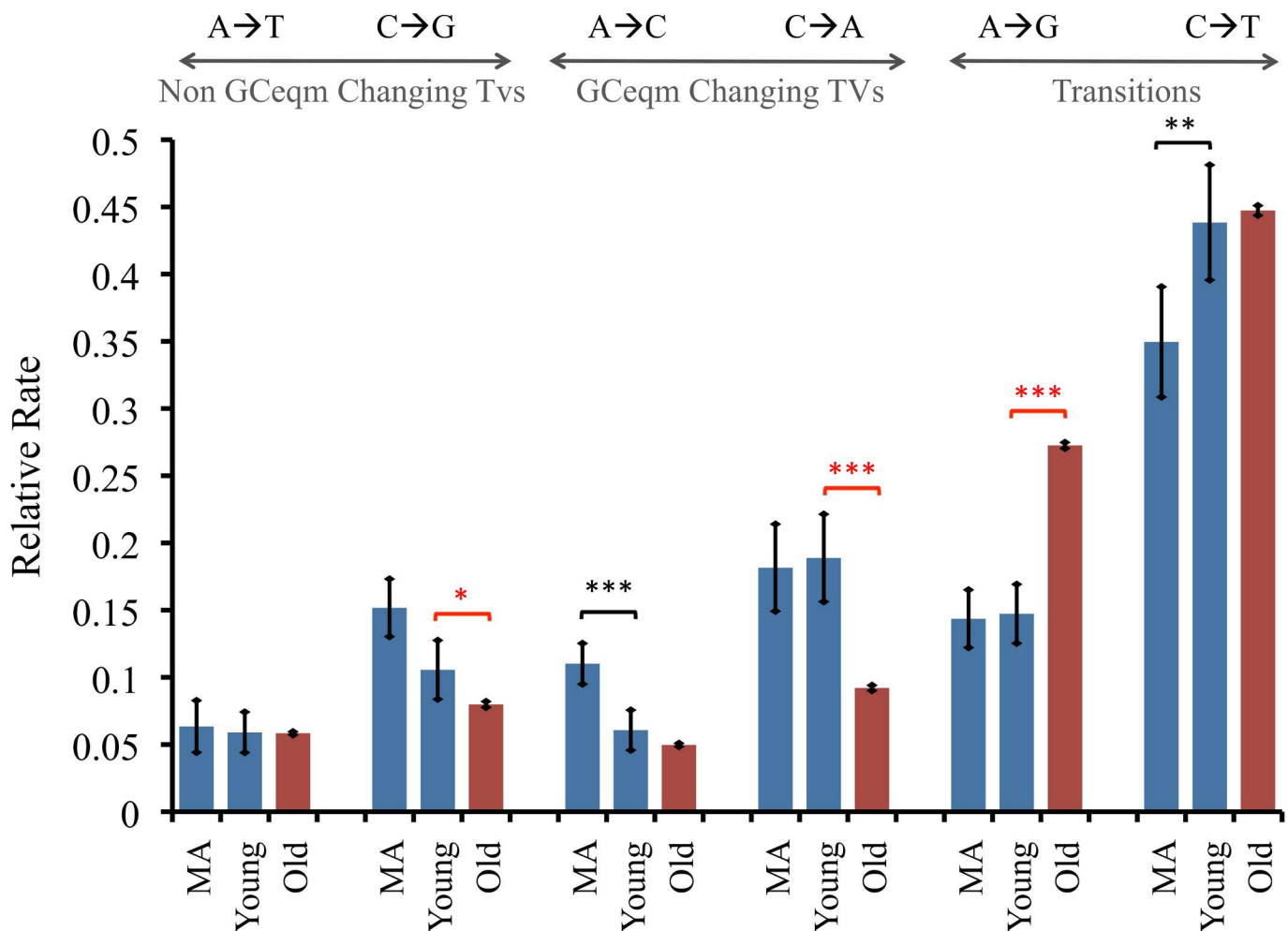


Fig 3. Relative mutation rates of the 6 nucleotide changes for MA experimental, young singletons, old singletons. Significant differences between rates in young singletons as compares to the other two datasets are annotated (Z-test, Bonferroni corrected). Error bars indicate S.E. * indicated $p < 0.05$, ** indicate $p < 0.01$, *** indicate $p < 0.001$.

doi:10.1371/journal.pgen.1006455.g003

singletons, but also between young singletons and the MA mutations. We further pursued the context dependent difference in mutation rates previously found in MA data, and divided singletons into groups based on their neighboring bases. A previous MA experiment showed a potentially elevated mutation rate at the middle nucleotide C in CCG and TCG environments, suggestive of low but detectable levels of methylation [12]. However, this particular bias was not clearly observed in the young singletons. Indeed, the highest rate was observed at ACG sites in the young singletons. Intriguingly, all four *CG sites had higher mutation rates in the old singletons (Fig 4). The biological significance of these results remains to be determined as there is more recent evidence that there is in fact no methylation in *S. cerevisiae* [53]. Our results do suggest, however, that there might be subtle differences between MA estimates and mutational biases in nature. Additional data should be able to resolve this question.

We tested if this classification system could potentially be employed for another mutation class—indels. Indels have been more difficult to study and analyze than SNPs due to their exceedingly rare nature (observed at least an order of magnitude less often than SNPs) and their strong fitness effects (that do not usually allow them to persist in natural populations). In most MA experiments, indels are observed in very low numbers in unique sequences, particularly in coding regions. Broad population sequencing allows larger numbers of such events to be observed, but mapping errors can increase false discovery rate (FDR) around repetitive regions. We filtered and aged indels following the same protocol as SNPs, and utilized the percentage of indels seen within coding regions (which span ~70% of the analyzed portion of the yeast genome) as the main signature for the action of selection. We confirmed that GC content of genomic sequences ± 10 bp of 3,389 high quality singleton indels were not significantly biased, but the incidence of simple tandem repeats (STRs) were more common than expected by chance ± 10 bp of indels, particularly for A/T monomers (Fig 5). This is in spite of the prior masking of 600Mb of known repetitive sequences. The indel singletons also did not occur randomly within the genome, with only 20% found in coding sequences, although this may be partly due to context dependent variation in error rates [13,54–55]. However, the youngest indels of age < 0.002 /kb were clearly less constrained by selection than older indels (Fig 6).

Discussion

Identifying young mutations when they first enter the population, and more critically before they have had a chance to rise above the selection drift barrier, is the equivalent of directly observing spontaneous mutations in a non-lab setting. Young mutations are necessarily rare, and can be captured through extremely broad population sequencing. To minimize noise from older alleles that appear rare simply due to biased sampling, linkage information can be leveraged. So far, mutational spectra deduced through extremely young alleles from broad population sequencing has not been cross-validated by MA experiments.

We sequenced 141 *S. cerevisiae* strains and were able to identify a subset of singletons that appeared to exhibit almost no signatures of selection, indicating their extremely young age. They also described a mutational spectrum similar to that previously detailed in a large-scale MA experiment, concurrently verifying the results from both techniques. However, the neighbor-dependent mutational trends appeared to vary across the datasets. While it is possible that neighbor dependency could vary across strains, this would require more data to clarify.

Applications to other species

In yeast, we used singleton density per Kb of genomic sequence as the arbitrary genomic unit for singleton counts. Any genomic unit or segment can conceivably be used, as long as they

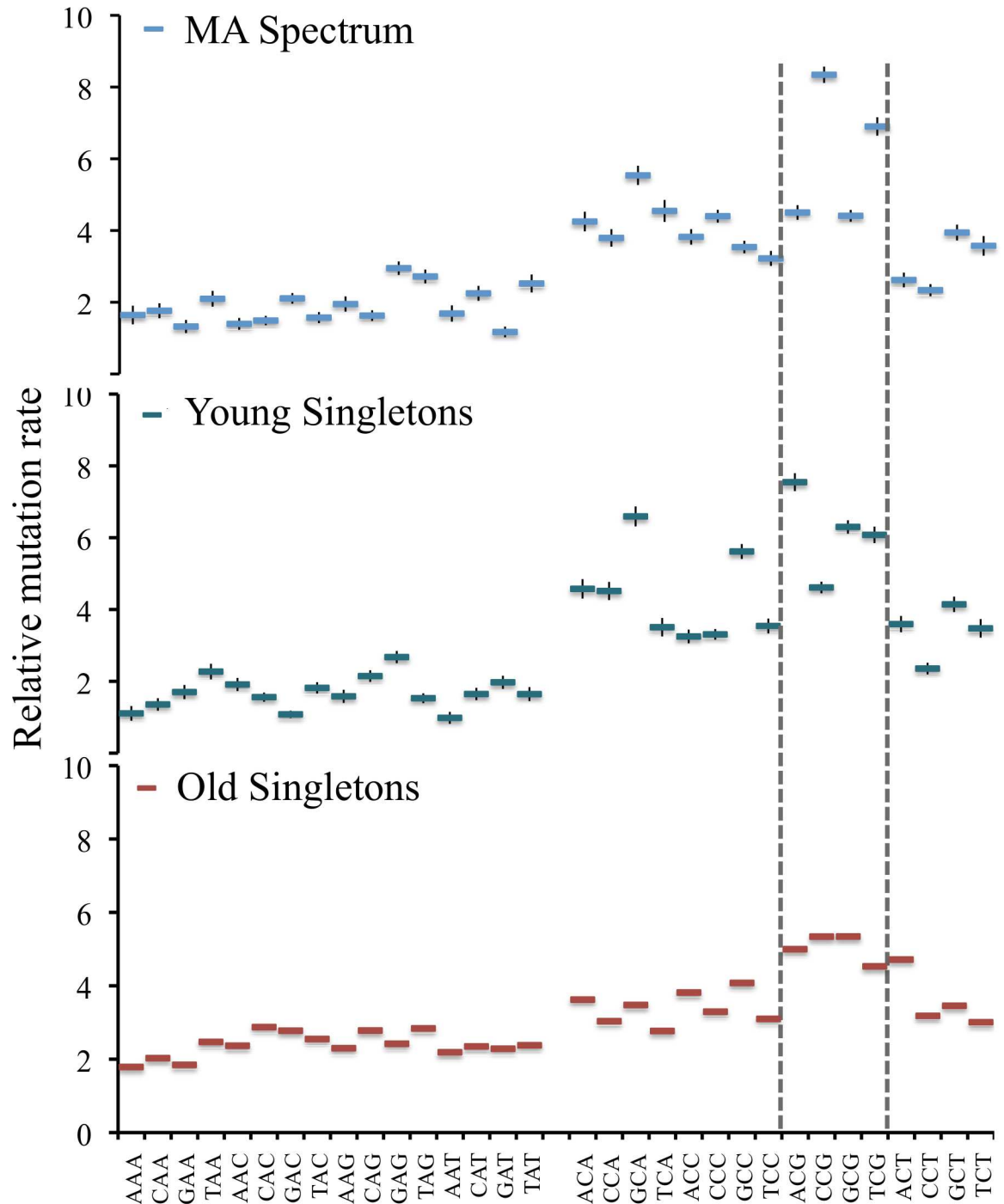


Fig 4. Neighbor dependent mutation rates in MA data, young singletons, and old singletons. X-axis—The 32 neighbor environments sorted. Y-axis—The relative rates in each environment.

doi:10.1371/journal.pgen.1006455.g004

are long enough such that mutations within that region are rare, but not so long that the sample does not contain individuals closely related enough as to be nearly identical across all of it.

It is also important to note that yeast has an atypical life cycle that is neither obligate asexual or sexual. It is thought to reproduce predominantly through clonal means with occasional

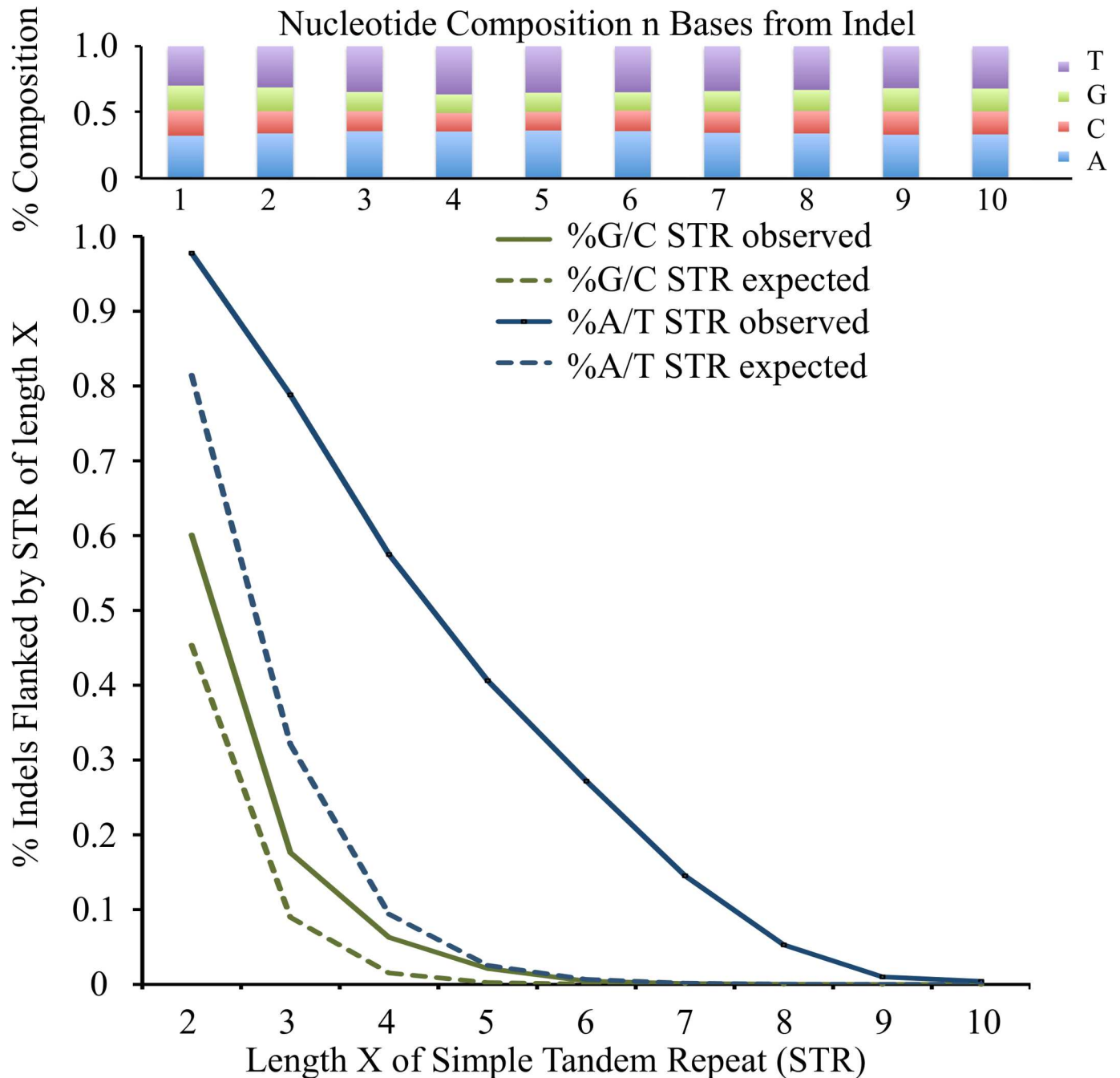


Fig 5. Top panel: base composition of ± 10 bp flanking singleton indel positions. X-axis—the ± 1 -10th base from indel position, 5'→3' oriented. Y-axis—composition of A,C,G,T bases at that position across all indel flanking regions. Bottom panel: % of indels flanked by single base simple tandem repeats (STRs). X-axis—the length of monomer STR. Y-axis—the % of indels flanked by a monomer STR at least that long.

doi:10.1371/journal.pgen.1006455.g005

sexual reproduction (reviewed in [56]). Yeast also has a marked tolerance for large-scale copy number changes (e.g. [57]). It may even be highly tolerant of hybridization with closely related species (e.g. [58–60]), and is known to carry introgression from sister species (e.g. [57]) as well as more distant relatives (e.g. [61]). The impact of such irregular life cycles (found in many fungi/moss species) on segregating sites within a population sample is not clear. A similar

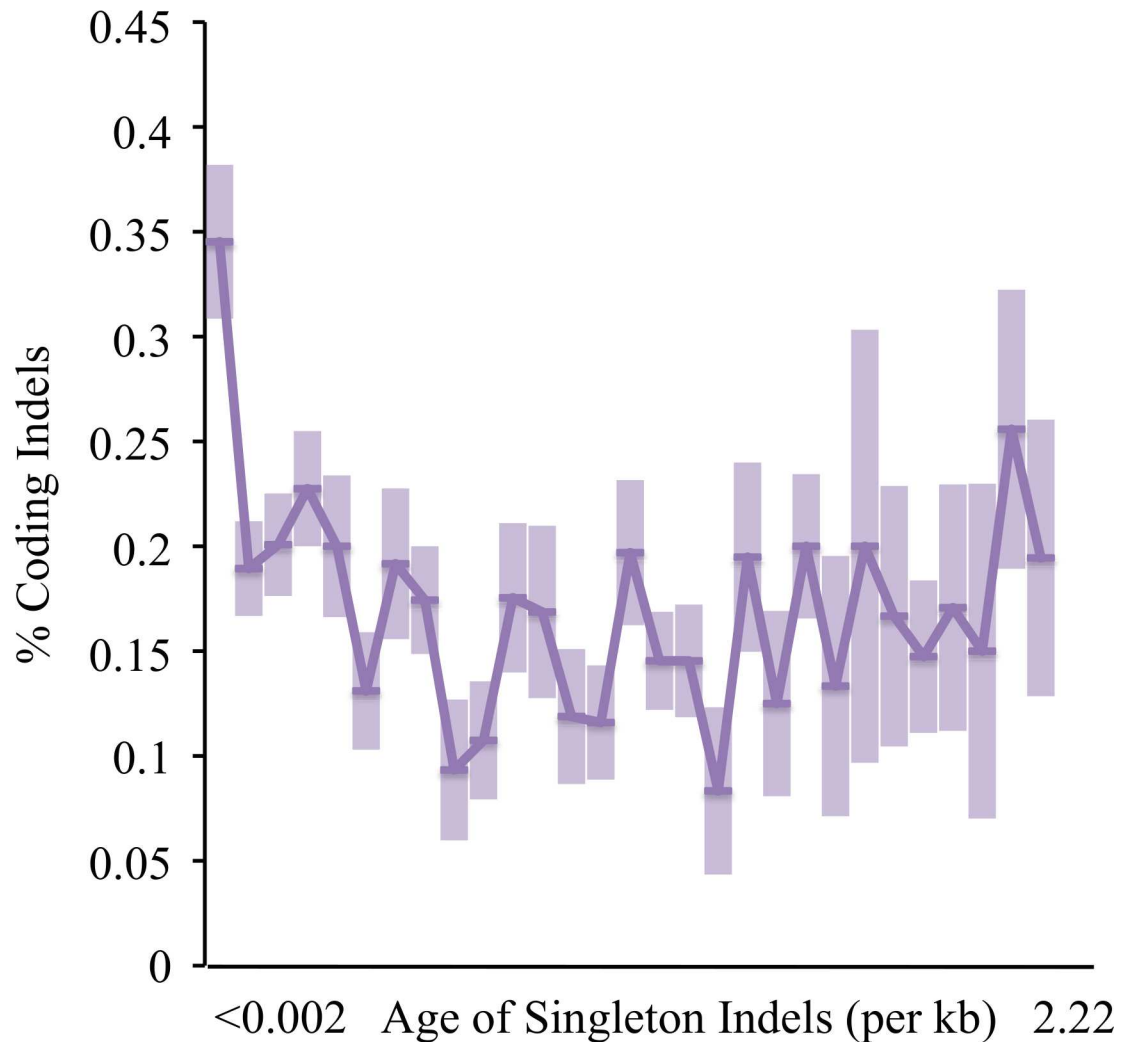


Fig 6. % of indels within coding sequences by age. X-axis—age of indel. Y-axis—percentage of indels found within coding regions.

doi:10.1371/journal.pgen.1006455.g006

study in more species may help to resolve this question. In obligate sexual reproducers, there may still be large variations in mutation rate, recombination rate, or population diversity that can make sampling closely related genomes difficult without prior knowledge. For such species, more care must be taken during sample collection.

Another point to keep in mind is that this method can only identify what selection doesn't immediately remove. There is a practical limit to how closely related individuals from a random population sampling can be, unless the population is extremely inbred, or there is genealogy information. The youngest mutation we can identify is consequently lower bound by how recently the two closest related individuals diverged. If selection is so strong that many mutations have already been removed within that short divergence time, we would be limited to only describing the trends that follow. An extreme example of such scenarios would be lethal mutations, although this was also seen to a certain degree in indels, which are removed by selection at a rate that is 10 times as fast as nonsynonymous mutations [19], and unsurprisingly never reached neutral expectations in our analysis.

Advantages and limitations

The power of this method lies in numbers. Sequencing of just 141 strains was able to give us 829 putative young SNPs, a number nearly matching that from a large diploid yeast MA experiment involving nearly ~311,000 generations under controlled lab conditions [7]. A subset of these young mutations may have accumulated during lab propagation for DNA extraction, but the large numbers suggest that the majority were in fact ‘natural’ mutations. In addition, we identified 168 singleton indels with an age of $<0.002/\text{Kb}$, a class of mutation only very rarely seen in experimental settings. We were able to show that even here, where selection acts strongly and quickly to change the overall signature, some trends can still be observed with a decrease in indel age.

There are multiple benefits unique to this analysis. First, instead of correcting for population structure, an issue common to most population samples, it takes advantage of the varying degrees of relatedness in a sample set to classify singletons into age groups. These continuous age groups, in addition to investigating whether young mutations match neutral expectations, also allow observation of trends across time. Second, unlike methods dealing with divergence data, there are no phasing or haplotype issues. Young singletons are necessarily the derived allele, and they are so rare the effect of linkage is negligible. The yeast strains used were in fact in various states of natural ploidy [52], as can be expected of a natural population sample. However, note that $<1\%$ of sites were suspected of carrying more than 2 alleles, and were not considered in this analysis.

One major limitation of this method is that it doesn’t provide the ability to accurately estimate mutation rate, which is something that naturally follows from an MA experiment. The means of accurately estimating generation time separating such closely related individuals is beyond the scope of this manuscript. A second issue is that the number and identity of singletons will heavily depend upon which and how many strains are sequenced. As more strains are sequenced, some singletons will be lost, while others will be identified. A logical further extension of this approach would be to try to age not just singletons, but doubletons, tripletons etc., based on population frequency and shared haplotype lengths, though it is unclear how much this would modify the overall conclusions.

As broad population sequencing becomes increasingly accessible, the amount of information we can extract from resulting sequence data becomes the limiting factor to their scientific value. Well-described mutational spectra form one area of molecular evolution for which extensive work has been difficult to amass, and which can benefit from this new application.

Materials and Methods

DNA library construction, read mapping, and variant calling protocol was detailed in earlier publication [52]. Briefly, DNA was extracted from liquid cultures using a modified glass bead lysis protocol. 500bp paired-end Illumina sequencing libraries were prepared at The Genome Institute, Washington University School of Medicine, and run on an Illumina HiSeq to an average of 100-fold coverage. Resulting fastq files were mapped to the reference genome with bwa v0.5.9 [62], sorted and indexed with samtools v0.1.18 [63], and assigned strain IDs with picard tools v1.55. Duplicated read pairs were removed and remaining reads locally realigned with GATK v2.1–8 [64]. The UnifiedGenotyper was used to call candidate variants across each sample independently. The resulting VCF files were filtered for variants with $\text{MQ}>40$, $\text{GQ}>20$, $\text{Qual}>20$, coverage depth $>8X$, >2 reads and $>15\%$ of reads supporting alternative variant. Around 600kb of the genome—annotated in the SGD database as simple repeats, centromeric regions, telomeric regions, or LTRs were excluded from analysis due to their susceptibility to mismapping and associated miscalls. Custom scripts were written to parse, identify,

count, and summarize variants for every frequency, age, mutation type, and neighborhood category. Error bars were calculated as sampling errors where possible, or else estimated with 500 bootstraps.

Supporting Information

S1 Fig. Mapping quality across SNPs of varying frequencies.

(TIF)

S2 Fig. Site frequency spectrum of SNPs called from 141 MA strains

(TIF)

S3 Fig. Violin plots of mapping quality (MQ), read coverage (Cov), SNP quality (Qual), and genotype quality (GQ) for young singletons and old singletons.

(TIF)

Acknowledgments

We thank members of the D.A.P. and G.S. laboratories for helpful discussions and comments.

Author Contributions

Conceptualization: DAP.

Data curation: YOZ GS DAP.

Formal analysis: YOZ GS DAP.

Funding acquisition: YOZ GS DAP.

Investigation: YOZ GS DAP.

Methodology: YOZ GS DAP.

Project administration: GS DAP.

Resources: YOZ GS DAP.

Software: YOZ GS DAP.

Supervision: GS DAP.

Validation: YOZ GS DAP.

Visualization: YOZ GS DAP.

Writing – original draft: YOZ GS DAP.

Writing – review & editing: YOZ GS DAP.

References

1. Ossowski S, Schneeberger K, Lucas-Iledó JI, Warthmann N, Clark RM, Shaw RG, et al. The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. *Science* 01 Jan 2010; Vol. 327, Issue 5961, pp.92–94 doi: [10.1126/science.1180677](https://doi.org/10.1126/science.1180677) PMID: [20044577](https://pubmed.ncbi.nlm.nih.gov/20044577/)
2. Denver DR, Wilhelm LJ, Howe DK, Gafner K, Dolan PC, Baser CF. Variation in base-substitution mutation in experimental and natural lineages of *Caenorhabditis nematodes*. *Genom Bio Evol.* 2012; 4(4), 513–22.

3. Rutter MT, Roles A, Conner JK, Shaw RG, Shaw FH, Schneeberger K, et al. Fitness of Arabidopsis Thaliana Mutation Accumulation Lines Whose Spontaneous Mutations Are Known. *Evolution*. 2012 Jul; 66(7):2335–9. doi: [10.1111/j.1558-5646.2012.01583.x](https://doi.org/10.1111/j.1558-5646.2012.01583.x) PMID: [22759306](https://pubmed.ncbi.nlm.nih.gov/22759306/)
4. Long HA, Paixão T, Azevedo RB, Zufall RA. Accumulation of spontaneous mutations in the ciliate Tetrahymena thermophila. *Genetics*. 2013 Oct; 195(2):527–40. doi: [10.1534/genetics.113.153536](https://doi.org/10.1534/genetics.113.153536) PMID: [23934880](https://pubmed.ncbi.nlm.nih.gov/23934880/)
5. Joseph SB, Hall DW. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: more beneficial than expected. *Genetics*. 2004 Dec; 168(4), 1817–25. doi: [10.1534/genetics.104.033761](https://doi.org/10.1534/genetics.104.033761) PMID: [15611159](https://pubmed.ncbi.nlm.nih.gov/15611159/)
6. Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA*. 2008 Jul 8; 105(27):9272–7. doi: [10.1073/pnas.0803466105](https://doi.org/10.1073/pnas.0803466105) PMID: [18583475](https://pubmed.ncbi.nlm.nih.gov/18583475/)
7. Hall DW, Mahmoudizad R, Hurd AW, Joseph SB. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: another thousand cell generations. *Genet Res (Camb)*. 2008 Jun; 90(3):229–41.
8. Nishant KT, Wei W, Mancera E, Argueso JL, Schlattl A, Delhomme N, et al. The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS Genet*. 2010 Sep 9; 6(9): e1001109 doi: [10.1371/journal.pgen.1001109](https://doi.org/10.1371/journal.pgen.1001109) PMID: [20838597](https://pubmed.ncbi.nlm.nih.gov/20838597/)
9. Lujan SA, Clausen AR, Clark AB, MacAlpine HK, MacAlpine DM, Malc EP, et al. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res*. 2014 Nov; 24(11):1751–64 doi: [10.1101/gr.178335.114](https://doi.org/10.1101/gr.178335.114) PMID: [25217194](https://pubmed.ncbi.nlm.nih.gov/25217194/)
10. Serero A, Jubin C, Loeillet S, Legoix-Né, Nicolas AG. Mutational landscape of yeast mutator strains. *Proc Natl Acad Sci U S A*. 2014 Feb 4; 111(5):1897–1902 doi: [10.1073/pnas.1314423111](https://doi.org/10.1073/pnas.1314423111) PMID: [24449905](https://pubmed.ncbi.nlm.nih.gov/24449905/)
11. Stirling PC, Shen Y, Corbett R, Jones SJM, Hieter P. Genome destabilizing mutator alleles drive specific mutational trajectories in *Saccharomyces cerevisiae*. *Genetics*. 2014 Feb; 196(2): 403–412 doi: [10.1534/genetics.113.159806](https://doi.org/10.1534/genetics.113.159806) PMID: [24336748](https://pubmed.ncbi.nlm.nih.gov/24336748/)
12. Zhu YO, Siegal ML, Hall DW, Petrov DA. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci U S A*. 2014 Jun 3; 111(22):E2310–8 doi: [10.1073/pnas.1323011111](https://doi.org/10.1073/pnas.1323011111) PMID: [24847077](https://pubmed.ncbi.nlm.nih.gov/24847077/)
13. Lujan SA, Clark AB, Kunkel TA. Differences in genome-wide repeat sequence instability conferred by proofreading and mismatch repair defects. *Nucleic Acids Res*. 2015 Apr 30; 43(8):4067–74. doi: [10.1093/nar/gkv271](https://doi.org/10.1093/nar/gkv271) PMID: [25824945](https://pubmed.ncbi.nlm.nih.gov/25824945/)
14. Saxer G, Havlak P, Fox SA, Quance MA, Gupta S, Fofanov Y, et al. Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba *Dictyostelium discoideum*. *PLoS One*. 2012; 7(10):e46759. doi: [10.1371/journal.pone.0046759](https://doi.org/10.1371/journal.pone.0046759) PMID: [23056439](https://pubmed.ncbi.nlm.nih.gov/23056439/)
15. Ness RW, Morgan AD, Colegrave N, Keightley PD. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. *Genetics* 2012 Dec; 192(4):1447–54. doi: [10.1534/genetics.112.145078](https://doi.org/10.1534/genetics.112.145078) PMID: [23051642](https://pubmed.ncbi.nlm.nih.gov/23051642/)
16. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A*. 2012 Nov 6; 109(45): 18488–18492 doi: [10.1073/pnas.1216223109](https://doi.org/10.1073/pnas.1216223109) PMID: [23077252](https://pubmed.ncbi.nlm.nih.gov/23077252/)
17. Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Charlesworth B, et al. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* 445, 82–85 (4 January 2007) | doi: [10.1038/nature05388](https://doi.org/10.1038/nature05388) PMID: [17203060](https://pubmed.ncbi.nlm.nih.gov/17203060/)
18. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genom Res*. 2009. 19:1195–1201.
19. Schrider DR, Houle D, Lynch M, Hahn MW. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*. 2013 Aug; 194(4):937–54. doi: [10.1534/genetics.113.151670](https://doi.org/10.1534/genetics.113.151670) PMID: [23733788](https://pubmed.ncbi.nlm.nih.gov/23733788/)
20. Baer CF, Shaw F, Steding C, Baumgartner M, Hawkins A, Houppert A, et al. Comparative evolutionary genetics of spontaneous mutations affecting fitness in rhabditid nematodes. *Proc Natl Acad Sci USA* 2005 Apr 19; 102(16):5785–90. doi: [10.1073/pnas.0406056102](https://doi.org/10.1073/pnas.0406056102) PMID: [15809433](https://pubmed.ncbi.nlm.nih.gov/15809433/)
21. Clancy S. DNA damage & repair: mechanisms for maintaining DNA integrity. *Nature Education*. 2008; 1(1):103
22. Branzei D, Foiani M. Regulation of DNA repair throughout the cell cycle. *Nature Reviews Molecular Cell Biology* 2008; 9:297–308. doi: [10.1038/nrm2351](https://doi.org/10.1038/nrm2351) PMID: [18285803](https://pubmed.ncbi.nlm.nih.gov/18285803/)

23. Huang D, Piening BD, Paulovich AG. The preference for error-free or error-prone postreplication repair in *Saccharomyces cerevisiae* exposed to low-dose methyl methanesulfonate is cell cycle dependent. *Mol Cell Biol*. 2013 Apr; 33(8):1515–27. doi: [10.1128/MCB.01392-12](https://doi.org/10.1128/MCB.01392-12) PMID: [23382077](https://pubmed.ncbi.nlm.nih.gov/23382077/)
24. Zanders S, Ma X, Roychoudhury A, Hernandez RD, Demogines A, Barker B, et al. Detection of heterozygous mutations in the genome of mismatch repair defective diploid yeast using a Bayesian approach. *Genetics*. 2010 Oct; 186(2):493–503. doi: [10.1534/genetics.110.120105](https://doi.org/10.1534/genetics.110.120105) PMID: [20660644](https://pubmed.ncbi.nlm.nih.gov/20660644/)
25. Lang G, Parsons L, Gammie AE. Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. *G3(Bethesda)*. 2013 Sep 4; 3(9):1453–65.
26. Ma X, Rogacheva MV, Nishant KT, Zanders S, Bustamante CD, Alani E. Mutation hotspots in yeast caused by long-range clustering of homopolymeric sequences. *Cell Rep*. 2012 Jan 26; 1(1): 36–42 doi: [10.1016/j.celrep.2011.10.003](https://doi.org/10.1016/j.celrep.2011.10.003) PMID: [22832106](https://pubmed.ncbi.nlm.nih.gov/22832106/)
27. Watt DL, Buckland RJ, Lujan SA, Kunkel TA, Chabes A. Genome-wide analysis of the specificity and mechanisms of replication infidelity driven by imbalanced dNTP pools. *Nucleic Acids Res*. 2016 Feb 29; 44(4): 1669–1680. doi: [10.1093/nar/gkv1298](https://doi.org/10.1093/nar/gkv1298) PMID: [26609135](https://pubmed.ncbi.nlm.nih.gov/26609135/)
28. Trindade S, Perfeito L, and Gordo I. Rate and effects of spontaneous mutations that affect fitness in mutator *Escherichia coli*. *Philos Trans R Soc Lond B Biol Sci*. 2010; 365(1544):1177–1186 doi: [10.1098/rstb.2009.0287](https://doi.org/10.1098/rstb.2009.0287) PMID: [20308092](https://pubmed.ncbi.nlm.nih.gov/20308092/)
29. Heilbron K, Toll-Riera M, Kojadinovic M, MacLean RC. Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics* 2014; 197(3):981–990 doi: [10.1534/genetics.114.163147](https://doi.org/10.1534/genetics.114.163147) PMID: [24814466](https://pubmed.ncbi.nlm.nih.gov/24814466/)
30. Petrov DA, Lozovskaya ER, Hartl DL. High intrinsic rate of DNA loss in *Drosophila*. *Nature* 1996; 384:346–349. doi: [10.1038/384346a0](https://doi.org/10.1038/384346a0) PMID: [8934517](https://pubmed.ncbi.nlm.nih.gov/8934517/)
31. Kimura M. Evolutionary rate at the molecular level. *Nature* 1968; 217:624–626. PMID: [5637732](https://pubmed.ncbi.nlm.nih.gov/5637732/)
32. Kimura M. *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press, Cambridge, U.K. 1983 ISBN 0-521-23109-4.
33. Nei M, Kumar S. *Molecular Evolution and Phylogenetics*. Oxford Univ. Press. 2000. ISBN:0-19-512584-9(hbk); 0-19-513585-7(pbk).
34. Ellegren H, Smith NG, Webster MT. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* 2003; 13(6):562–568. PMID: [14638315](https://pubmed.ncbi.nlm.nih.gov/14638315/)
35. Ochman H. Neutral mutations and neutral substitutions in bacterial genomes. *Mol Biol Evol* 2003; 20(12):2091–6. doi: [10.1093/molbev/msg229](https://doi.org/10.1093/molbev/msg229) PMID: [12949125](https://pubmed.ncbi.nlm.nih.gov/12949125/)
36. Duret L & Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Ann R Genom & Hum Gen* 2009; 10:285–311.
37. Lawrie DS, Petrov DA, Messer PW. Faster than neutral evolution of constrained sequences: the complex interplay of mutational biases and weak selection. *Genom Bio Evol* 2011; 3:383–95.
38. Kousathanas A, Oliver F, Halligan DL, Keightley PD. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol Bio Evol* 2011; 28(3):1183–91.
39. Messer P. Measuring rates and patterns of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* 2009; 182:1219–1232; doi: [10.1534/genetics.109.105692](https://doi.org/10.1534/genetics.109.105692) PMID: [19528323](https://pubmed.ncbi.nlm.nih.gov/19528323/)
40. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Res*, 2007; 17(8):1195–1201 doi: [10.1101/gr.6468307](https://doi.org/10.1101/gr.6468307) PMID: [17600086](https://pubmed.ncbi.nlm.nih.gov/17600086/)
41. Wlim A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res*. 2012 Dec; 40(22):11189–201. doi: [10.1093/nar/gks918](https://doi.org/10.1093/nar/gks918) PMID: [23066108](https://pubmed.ncbi.nlm.nih.gov/23066108/)
42. Chen-Harris H, Borucki MK, Torres C, Slezak TR, Allen JE. Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. *BMC Genomics* 2013; 14:96. doi: [10.1186/1471-2164-14-96](https://doi.org/10.1186/1471-2164-14-96) PMID: [23402258](https://pubmed.ncbi.nlm.nih.gov/23402258/)
43. Isakov O, Borderia AV, Golan D, Hamenahem A, Celniker G, Yoffe L, et al. Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. *Bioinformatics* 2015; 31(13):2141–50 doi: [10.1093/bioinformatics/btv101](https://doi.org/10.1093/bioinformatics/btv101) PMID: [25701575](https://pubmed.ncbi.nlm.nih.gov/25701575/)
44. Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, St Jean PL, et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res* 2013; 23(12):1974–84 doi: [10.1101/gr.154971.113](https://doi.org/10.1101/gr.154971.113) PMID: [23990608](https://pubmed.ncbi.nlm.nih.gov/23990608/)
45. Moore CB, Wallace JR, Wolfe DJ, Frase AT, Pendergrass SA, Weiss KM, et al. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000

- genomes project data. PLoS Genet 2013; 9(12):e1003959 doi: [10.1371/journal.pgen.1003959](https://doi.org/10.1371/journal.pgen.1003959) PMID: [24385916](https://pubmed.ncbi.nlm.nih.gov/24385916/)
46. Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet 2011; 43(7), 712–4. doi: [10.1038/ng.862](https://doi.org/10.1038/ng.862) PMID: [21666693](https://pubmed.ncbi.nlm.nih.gov/21666693/)
 47. Keightley PD, Ness RW, Halligan DL, Haddrill PR. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* fullsib family. Genetics. 2014 Jan; 196(1):313–20. doi: [10.1534/genetics.113.158758](https://doi.org/10.1534/genetics.113.158758) PMID: [24214343](https://pubmed.ncbi.nlm.nih.gov/24214343/)
 48. Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, et al. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. Mol Biol Evol. 2015 Jan; 32(1):239–43. doi: [10.1093/molbev/msu302](https://doi.org/10.1093/molbev/msu302) PMID: [25371432](https://pubmed.ncbi.nlm.nih.gov/25371432/)
 49. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of *de novo* mutations in humans. Nature Genet 2015; 47(7),822–826 doi: [10.1038/ng.3292](https://doi.org/10.1038/ng.3292) PMID: [25985141](https://pubmed.ncbi.nlm.nih.gov/25985141/)
 50. Ezov TK, Boger-Nadjar E, Frenkel Z, Katsperovski I, Kemeny S, Nevo E, et al. Molecular-genetic biodiversity in a natural population of the yeast *Saccharomyces cerevisiae* from “Evolution Canyon”: micro-satellite polymorphism, ploidy and controversial sexual status. Genetics. 2006 Nov; 174(3):1455–1468 doi: [10.1534/genetics.106.062745](https://doi.org/10.1534/genetics.106.062745) PMID: [16980391](https://pubmed.ncbi.nlm.nih.gov/16980391/)
 51. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2,000 years. 2016 bioRxiv
 52. Zhu YO, Sherlock G, and Petrov DA. Whole Genome Analysis of 132 Clinical *Saccharomyces cerevisiae* Strains Reveals Extensive Ploidy Variation G3 (Bethesda). 2016 Aug 9; 6(8):2421–34.
 53. Capuano F, Muelleder M, Kok RM, Blom HJ, Ralser M. Cytosine DNA methylation is found in *Drosophila melanogaster* but absent in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and other yeast species. Analytical Chemistry 2014 86 (8): 3697–3702. doi: [10.1021/ac500447w](https://doi.org/10.1021/ac500447w) PMID: [24640988](https://pubmed.ncbi.nlm.nih.gov/24640988/)
 54. Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, et al. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science. 2009 Jan 16; 323(5912): 401–404. doi: [10.1126/science.1163183](https://doi.org/10.1126/science.1163183) PMID: [19074313](https://pubmed.ncbi.nlm.nih.gov/19074313/)
 55. Tolstorukov MY, Volfovsky N, Stephens RM, Park PJ. Impact of chromatin structure on sequence variability in the human genome. Nat Struct Mol Biol. 2011 Apr 18(4): 510–515. doi: [10.1038/nsmb.2012](https://doi.org/10.1038/nsmb.2012) PMID: [21399641](https://pubmed.ncbi.nlm.nih.gov/21399641/)
 56. Liti G. The fascination and secret wild life of the budding yeast *S. cerevisiae*. eLife 2015; 4:e05835
 57. Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. Genome Res. 2012; 22(5):908–24 doi: [10.1101/gr.130310.111](https://doi.org/10.1101/gr.130310.111) PMID: [22369888](https://pubmed.ncbi.nlm.nih.gov/22369888/)
 58. Martini AV, Martini A. Three newly delimited species of *Saccharomyces sensu stricto*. Antonie Van Leeuwenhoek 1987; 53(2):77–84 PMID: [3662481](https://pubmed.ncbi.nlm.nih.gov/3662481/)
 59. Tamai Y, Momma T, Yoshimoto H, Kaneko Y. Co-existence of two types of chromosome in the bottom fermenting yeast, *Saccharomyces pastorianus*. Yeast 1998; 14(10):923–33 doi: [10.1002/\(SICI\)1097-0061\(199807\)14:10<923::AID-YEA298>3.0.CO;2-I](https://doi.org/10.1002/(SICI)1097-0061(199807)14:10<923::AID-YEA298>3.0.CO;2-I) PMID: [9717238](https://pubmed.ncbi.nlm.nih.gov/9717238/)
 60. Rainieri S, Kodama Y, Kaneko Y, Mikata K, Nakao Y, Ashikari T. Pure and mixed genetic lines of *Saccharomyces bayanus* and *Saccharomyces pastorianus* and their contribution to the lager brewing strain genome. Appl Environ Microbiol 2006; 72(6):3968–74 doi: [10.1128/AEM.02769-05](https://doi.org/10.1128/AEM.02769-05) PMID: [16751504](https://pubmed.ncbi.nlm.nih.gov/16751504/)
 61. Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, et al. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. PNAS 2009; 106(38):16333–8 doi: [10.1073/pnas.0904673106](https://doi.org/10.1073/pnas.0904673106) PMID: [19805302](https://pubmed.ncbi.nlm.nih.gov/19805302/)
 62. Li H & Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25(14), 1754–60. doi: [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324) PMID: [19451168](https://pubmed.ncbi.nlm.nih.gov/19451168/)
 63. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole genome resequencing. Genome Res 2009; 19(6), 1124–1132. doi: [10.1101/gr.088013.108](https://doi.org/10.1101/gr.088013.108) PMID: [19420381](https://pubmed.ncbi.nlm.nih.gov/19420381/)
 64. Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20(9), 1297–1303. doi: [10.1101/gr.107524.110](https://doi.org/10.1101/gr.107524.110) PMID: [20644199](https://pubmed.ncbi.nlm.nih.gov/20644199/)