


# Is the Mutation Rate Lower in Genomic Regions of Stronger Selective Constraints?

Haoxuan Liu<sup>1,2</sup> and Jianzhi Zhang <sup>1,\*</sup>

<sup>1</sup>Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

<sup>2</sup>Evolutionary and Organismal Biology Research Center, School of Medicine, Zhejiang University, Hangzhou 310000, China

\*Corresponding author: E-mail: jianzhi@umich.edu.

Associate editor: Weiwei Zhai

## Abstract

**A study of the plant *Arabidopsis thaliana* detected lower mutation rates in genomic regions where mutations are more likely to be deleterious, challenging the principle that mutagenesis is blind to its consequence. To examine the generality of this finding, we analyze large mutational data from baker's yeast and humans. The yeast data do not exhibit this trend, whereas the human data show an opposite trend that disappears upon the control of potential confounders. We find that the *Arabidopsis* study identified substantially more mutations than reported in the original data-generating studies and expected from *Arabidopsis*' mutation rate. These extra mutations are enriched in polynucleotide tracts and have relatively low sequencing qualities so are likely sequencing errors. Furthermore, the polynucleotide "mutations" can produce the purported mutational trend in *Arabidopsis*. Together, our results do not support lower mutagenesis of genomic regions of stronger selective constraints in the plant, fungal, and animal models examined.**

**Key words:** *Arabidopsis*, yeast, human, natural selection, mutation.

A central tenet of evolutionary biology is that mutations occur randomly with respect to their consequences (Luria and Delbruck 1943; Lederberg and Lederberg 1952). This tenet, however, has been repeatedly challenged in the last decade, in a large part due to the availability of large genomic sequence data that allow testing its validity across the genome. For example, by analyzing synonymous polymorphisms in *Escherichia coli*, Martincorena *et al.* reported that genes subject to stronger purifying selection or with higher expressions mutate less often, and proposed that this mutational trend reflects adaptive risk management (Martincorena *et al.* 2012). However, synonymous polymorphisms may be nonneutral (Lind *et al.* 2010; Sharon *et al.* 2018; Shen *et al.* 2022), distorting the estimation of mutation rates. Indeed, a reanalysis based on mutations observed in a mutation accumulation (MA) experiment in the near absence of selection invalidated the polymorphism-based result (Chen and Zhang 2013). More importantly, it was pointed out that selection for modifiers that lower the mutation rate of a gene because of the deleterious effects of mutations in the gene is extremely weak; consequently, selective optimization of gene-specific mutation rates is theoretically untenable (Chen and Zhang 2013). Nevertheless, Xie *et al.* reported that human genes expressed relatively strongly in the testis mutate less often than those expressed relatively weakly, proposing that testis gene expression is regulated for the purpose of optimizing gene-specific germline mutation rates (Xia *et al.* 2020). A subsequent scrutiny, however,

identified several flaws in the analysis and found the original observation unsupported (Liu and Zhang 2020).

More recently, based on exceptionally large data of *de novo* mutations in the model plant *Arabidopsis thaliana*, Monroe *et al.* reported that the mutation rate is 58% lower inside than immediately outside genes and is 37% lower in essential than nonessential genes (Monroe *et al.* 2022). They also observed a positive correlation between the mutation rate of a gene and the nonsynonymous to synonymous substitution rate ratio ( $d_N/d_S$ ) of the gene across the genome. Because  $d_N/d_S$  is commonly regarded as a measure of the protein function-related selective constraint of a gene (lower the  $d_N/d_S$ , higher the constraint), the above finding suggests lower mutation rates for more strongly constrained genes. Monroe *et al.* detected several genomic and epigenomic features that are correlated with the local mutation rate, so proposed that the mutation rate is modulated via these features. Because a genomic or epigenomic feature could be associated with a sizable fraction of the genome, the proposed mechanism allows a modifier to affect the mutation rate of a large number of nucleotide sites so could in principle arise through natural selection (Monroe *et al.* 2022). For example, the authors estimated that a modifier that lowers the mutation rate of all coding sequences of a third of essential genes by 30% could be selectively fixed. Alternatively, the proposed mechanism could have arisen as a byproduct of some other biological processes (Zhang 2022). Regardless, if the reported mutational trend in *Arabidopsis* holds true in diverse evolutionary

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Open Access

lineages, many evolutionary phenomena would require re-interpretation (Zhang 2022). In this study, we investigate the generality of the *Arabidopsis*-based finding. We show that the *Arabidopsis* result is found in neither baker's yeast nor humans. To understand the source of the *Arabidopsis* finding, we examine the mutational data analyzed by Monroe and colleagues. We show that the authors identified substantially more mutations than expected and that many of the mutations called are dubious, contributing to the unusual mutational trend observed.

### The *Arabidopsis* Mutational Trend Is not Found in Yeast or Humans

To examine the generality of the *Arabidopsis* finding, we turned to other species with the largest data of *de novo* mutations—baker's yeast (*Saccharomyces cerevisiae*) and humans (*Homo sapiens*). We focused on the correlation between the mutation rate of a gene and its  $d_N/d_S$  because of its direct relevance to the mutagenesis and evolution of genes. In yeast, we acquired mutational data from three MA studies (Zhu *et al.* 2014; Sharp *et al.* 2018; Liu and Zhang 2019), including 427 MA lines and 3,296 single nucleotide variations (SNVs). The  $d_N/d_S$  ratios were computed by comparing orthologous gene sequences between *S. cerevisiae* and its sister species *S. paradoxus* (Goncalves *et al.* 2011). The human mutational data came from two studies with a total of 217,247 SNVs identified from the genome sequences of 3,450 parents-offspring trios (Jonsson *et al.* 2017; An *et al.* 2018), while the  $d_N/d_S$  ratios were estimated from human and chimpanzee orthologous gene sequences.

Following Monroe *et al.*, we correlated  $d_N/d_S$  with the mutation rate across all genes in yeast and humans, respectively. No significant linear or rank correlation was found in yeast (table 1). Because this result could be due to a lack of statistical power owing to the relatively small number of mutations in the data, we grouped all genes into 50 equal-size bins according to their  $d_N/d_S$  and then computed the mutation rate for each bin. We found the mutation rate to be negatively correlated with  $d_N/d_S$  across the 50 bins, with marginal statistical significance (Spearman's rank correlation =  $-0.26$ ,  $P = 0.065$ ). Hence, if anything, the yeast data signal a trend that is opposite to that in *Arabidopsis*.

In humans, we observed a significant negative (rank) correlation between mutation rate and  $d_N/d_S$  (table 1), contrasting the *Arabidopsis* finding. Selection against mutagenesis should be stronger at genes with relatively high fractions of deleterious mutations than at those with relatively low fractions of deleterious mutations (Kimura 1967), so the observation in humans is not due to selection against mutagenesis. Because mutations have a higher average probability of fixation in less constrained than more constrained genes and because mutations of a highly mutagenic sequence tend to lower the local mutation rate, it has been predicted that the mutation rate should become lower in less constrained than more constrained genes (Oman *et al.* 2022). Our observation is consistent with this

prediction. Notwithstanding, there are several factors known to influence or otherwise be correlated with the mutation rate (or  $d_N/d_S$ ). If these factors are also correlated with  $d_N/d_S$  (or mutation rate), a spurious relationship may result between  $d_N/d_S$  and mutation rate. We thus computed partial correlations between  $d_N/d_S$  and mutation rate by individually or jointly controlling the following six factors—gene length (Lipman *et al.* 2002), DNA curvature (Duan *et al.* 2018), nucleosome occupancy (Li and Luscombe 2020), expression level (Park *et al.* 2012), GC content (Kiktev *et al.* 2018), and replication timing (Stamatoyannopoulos *et al.* 2009). In yeast, all partial correlations remain non-significant (table 1). In humans, the negative correlation between  $d_N/d_S$  and mutation rate becomes non-significant when the six factors are jointly controlled (table 1). We also performed the same analysis in *A. thaliana* using the mutational data from Monroe *et al.*, finding that the positive correlation between  $d_N/d_S$  and mutation rate exists with or without controlling the potential confounders (supplementary table S1, Supplementary Material online).

We ran a multiple linear regression to simultaneously evaluate the potential influences of  $d_N/d_S$  and the above six factors on the mutation rate of a gene (table 1). The results show that the mutation rate is significantly dependent on replication timing and gene expression level in yeast and nucleosome occupancy and replication timing in humans, respectively. However, in neither species is the mutation rate significantly dependent on  $d_N/d_S$ . Therefore, the reported trend in *Arabidopsis* of lower mutation rates of genes with lower  $d_N/d_S$  ratios is absent in yeast and humans.

### Monroe *et al.* Reported Much Higher Mutation Rates Than Did Previous *Arabidopsis* Studies

To investigate why the mutational trend reported by Monroe *et al.* is not replicated in yeast and humans, we looked into the mutational data analyzed by Monroe *et al.*, which comprised three separate datasets: Dataset 1 was derived from a published MA experiment (Weng *et al.* 2019), Dataset 2 was based on MA specifically performed for the study (Monroe *et al.* 2022), and Dataset 3 was a published somatic mutation dataset (Wang *et al.* 2019) (table 2).

The original authors of Dataset 1 reported an SNV mutation rate of  $6.95 \times 10^{-9}$  per site per generation (Weng *et al.* 2019), similar to various previous estimates (Ossowski *et al.* 2010; Yang *et al.* 2015). They identified 2,209 mutations that included SNVs and insertions/deletions (indels) (Weng *et al.* 2019), but Monroe *et al.* reported 3.9 times that number by reanalyzing the published sequencing reads (table 2). The original authors screened for germline (homozygous) mutations, while Monroe *et al.* screened for both germline and somatic (heterozygous) mutations. However, the detectability of somatic mutations, each of which should occur in only one seedling, is expected to be minute, because 40 seedlings were pooled from each sample and sequenced to a

**Table 1.** Relationships among Mutation Rate,  $d_N/d_S$ , and Six Other Factors in Yeast and Humans.

Species	Partial correlation between $d_N/d_S$ and mutation rate across genes			Multiple linear regression <sup>a</sup>		
	Controlled variables	Rank correlation (P-value)	Linear correlation (P-value)	Independent variables	Coefficient	P-value
Yeast	None	-0.0063 (0.68)	-0.0149 (0.33)	$d_N/d_S$	$-5.63 \times 10^{-5}$	0.43
	Expression level	-0.0145 (0.34)	-0.0114 (0.46)	Expression level	$3.85 \times 10^{-10}$	0.03
	Gene length	-0.0202 (0.19)	-0.0149 (0.33)	Gene length	$5.26 \times 10^{-9}$	0.51
	Nucleosome occupancy	-0.0137 (0.37)	-0.0150 (0.33)	Nucleosome occupancy	$-2.55 \times 10^{-7}$	0.67
	Replication timing	-0.0082 (0.59)	-0.0171 (0.26)	Replication timing	$-8.59 \times 10^{-5}$	$2.6 \times 10^{-3}$
	GC content	-0.0230 (0.13)	-0.0138 (0.37)	GC content	$4.24 \times 10^{-4}$	0.44
	DNA curvature	-0.0134 (0.38)	-0.0169 (0.27)	DNA curvature	$5.64 \times 10^{-5}$	0.41
	All of the above	-0.0167 (0.28)	-0.0122 (0.43)			
Humans	None	-0.0633 ( $4.6 \times 10^{-13}$ )	-0.0126 (0.15)	$d_N/d_S$	$9.51 \times 10^{-7}$	0.62
	Expression level	-0.0596 ( $9.2 \times 10^{-12}$ )	-0.0125 (0.15)	Expression level	$-3.53 \times 10^{-9}$	0.38
	Gene length	-0.0432 ( $7.8 \times 10^{-7}$ )	-0.0135 (0.12)	Gene length	$9.41 \times 10^{-12}$	0.14
	Nucleosome occupancy	-0.0625 ( $8.7 \times 10^{-13}$ )	0.0021 (0.81)	Nucleosome occupancy	$8.27 \times 10^{-7}$	$3.9 \times 10^{-6}$
	Replication timing	-0.0635 ( $3.9 \times 10^{-13}$ )	-0.0113 (0.20)	Replication timing	$-6.57 \times 10^{-5}$	$3.8 \times 10^{-3}$
	GC content	-0.0675 ( $1.1 \times 10^{-14}$ )	0.00042 (0.96)	GC content	$3.14 \times 10^{-5}$	0.18
	DNA curvature	-0.0620 ( $1.3 \times 10^{-12}$ )	-0.0036 (0.68)	DNA curvature	$-1.12 \times 10^{-5}$	0.24
	All of the above	-0.0095 (0.28)	0.0044 (0.62)			

<sup>a</sup>Mutation rate is the dependent variable in the multiple linear regression.

**Table 2.** Differences in the Number of Mutations Identified Between Monroe *et al.*'s Study and Former Studies.

Dataset	Sample size	Mutation no. in Monroe <i>et al.</i>	Mutation no. in original studies	Sequencing depth	References
1. Training dataset	107 MA lines $\times$ 25 generations	8,574	2,209	36 $\times$	Weng <i>et al.</i> (2019)
2. New dataset	400 MA lines $\times$ 8.3 generations	359,133	NA	$\sim$ 30 $\times$	Monroe <i>et al.</i> (2022)
3. Somatic dataset	64 somatic samples from two plants	773,141	17	52.3 $\times$	Wang <i>et al.</i> (2019)

depth of 36 $\times$  but Monroe *et al.* required at least three reads to call a mutation. Consequently, the 3.9-fold difference in data size is unlikely to be mainly caused by the inclusion of somatic mutations.

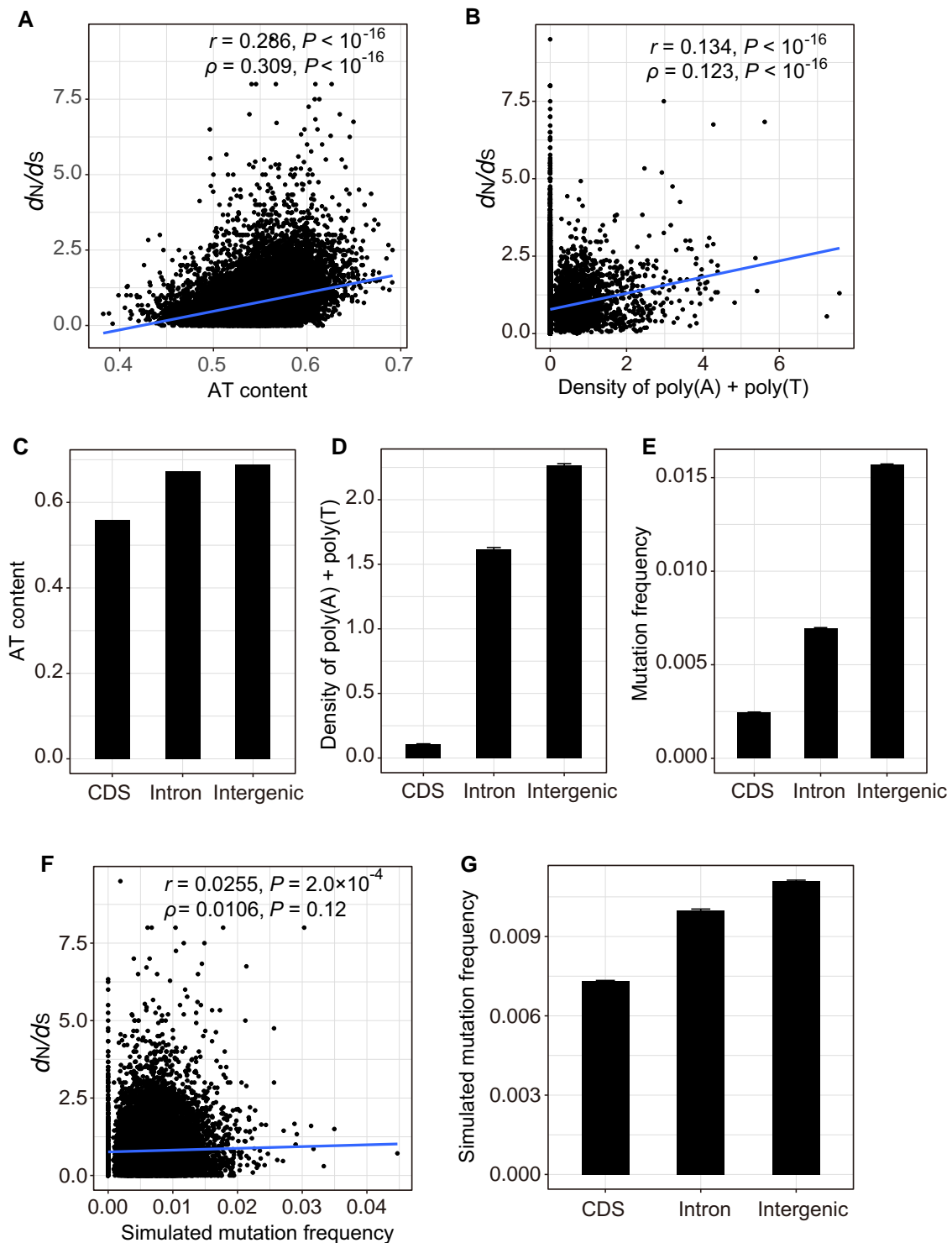
The expected sample size of Dataset 2 is 1.24 times that of Dataset 1 (Dataset 1: 107 lines  $\times$  25 generations/line = 2,675 generations; Dataset 2: 400 lines  $\times$  8.3 generations/line = 3,320 generations). Despite similar sequencing depths for the two datasets, the number of mutations reported by Monroe *et al.* for Dataset 2 is 41.9 times that reported by the same authors for Dataset 1 (table 2).

Dataset 3 came from 64 somatic samples taken from two *A. thaliana* plants (Wang *et al.* 2019). The original authors identified 17 mutations, but Monroe *et al.* reported 773,141 mutations by reanalyzing the published sequencing reads, a 45,479-fold difference (table 2). The mutation rate estimated by the original authors was approximately  $4.35 \times 10^{-9}$  per nucleotide per generation (Wang *et al.* 2019). By contrast, the corresponding rate from Monroe *et al.*'s reanalysis becomes  $2.0 \times 10^{-4}$  per nucleotide per generation, orders of magnitude higher than the mutation rate of any species ever known (Lynch *et al.* 2016).

### Monroe *et al.*'s Mutational Data Include Many Potential Sequencing Errors at Polynucleotide Tracts

To find out the causes of the massive differences in the number of mutations called between Monroe *et al.*'s study and the previous studies, we analyzed the mutations in Dataset 1. Monroe *et al.* reported 8,574 mutations in this

dataset, while the original authors reported 2,209 mutations (Weng *et al.* 2019). We found that 6,326 of the 8,574 mutations were not called in the original study. By reviewing the variant calling file (Weng *et al.* 2019), we found that, 57% of the 8,574 mutations appeared in more than one of the 107 MA lines, which is highly improbable for *de novo* mutations. One of the common errors in Illumina sequencing is caused by polynucleotides (Heydari *et al.* 2019). Indeed, we found that 51% of the 6,326 extra mutations reported by Monroe *et al.* are located within 20 nucleotides from a poly(A) or poly(T) tract (referred to as polynucleotides-associated mutations), while this fraction is 15% for the 2,209 originally reported mutations ( $P < 10^{-5}$ , chi-squared test). Of the 8,574 mutations reported by Monroe *et al.*, those that appeared in more than one line are more likely to be polynucleotides-associated than those that appeared in only one line (49% versus 33%, chi-square test,  $P < 10^{-5}$ ). In addition, most of the extra mutations are based on reads with multiple mismatches and low sequencing qualities (e.g., see supplementary fig. S1A, Supplementary Material online for some mutations found in five samples and supplementary fig. S1B, Supplementary Material online for some mutations found in 12 samples), while the mutations reported in the original study do not suffer from these problems (supplementary fig. S1C, Supplementary Material online). For instance, the median quality score is 67% lower for the extra mutations than the originally identified



**Fig. 1.** Relationships among AT content, density of polynucleotides, and  $d_N/d_S$  in *Arabidopsis thaliana*. (A) Pearson's correlation ( $r$ ) and Spearman's correlation ( $\rho$ ) between the coding region AT content and  $d_N/d_S$  across genes. Each dot represents a gene. The blue line is the linear regression. (B) Correlations between the number of poly(A) + poly(T) tracts per 1000 nucleotides (i.e., density) in the coding sequence (CDS) of a gene and its  $d_N/d_S$ . Each dot represents a gene. The blue line is the linear regression. (C) AT content in coding, intron, and intergenic regions, respectively. Errors are too small to present. (D) Mean density of poly(A) + poly(T) tracts in coding, intron, and intergenic regions, respectively. (E) Number of mutations per site in coding, intron, and intergenic regions, respectively, calculated using the sum of the three datasets in [Monroe et al. \(2022\)](#). (F) Correlations between the no. of mutations per site generated by simulation and  $d_N/d_S$  across genes. Each dot represents a gene. In the simulation, 70% of mutations are randomly distributed at non-polynucleotide sites across the genome while the remaining mutations are randomly distributed among poly(A) and poly(T) tracts. A similar Pearson's correlation is observed upon log transformations of the data. (G) Numbers of simulated mutations per site in coding, intron, and intergenic regions, respectively. In (D), (E), and (G), error bars represent 95% confidence intervals predicted by Poisson distributions.

mutations. This comparison suggests that, for Dataset 1, many extra mutations identified by Monroe *et al.* are unreliable and are likely sequencing errors at polynucleotide tracts. Because Monroe *et al.*'s mutation rate estimates from Datasets 2 and 3 would be even greater than that from Dataset 1, it is expected that the mutations they identified from Datasets 2 and 3 suffer from the same problem if not additional problems.

### False-Positive Mutations at Polynucleotides Can Create the Mutational Trend Observed by Monroe *et al.*

It has been reported that the  $d_N/d_S$  ratio of a gene is correlated with the AT content of the gene, because of the correlation of the gene expression level with both the AT content and  $d_N/d_S$  (Park *et al.* 2013; Zhang and Yang 2015). Indeed, we found that  $d_N/d_S$  is significantly positively correlated with the AT content in *A. thaliana* (fig. 1A) and that genes with higher densities of poly(A) + poly(T) have higher  $d_N/d_S$  ratios (fig. 1B). Therefore, if poly(A) and poly(T) cause over-detection of mutations, a spurious positive correlation could arise between  $d_N/d_S$  and mutation rate.

Monroe *et al.* (2022) also reported that the mutation rate of introns and that of intergenic regions exceed the mutation rate of coding regions in *Arabidopsis*. Interestingly, in *A. thaliana*, the AT content and the density of poly(A) + poly(T) are both the lowest in coding sequences, higher in introns, and highest in intergenic sequences (fig. 1C and D), which closely resembles the observation from Monroe *et al.*'s mutational data (fig. 1E), suggesting that Monroe *et al.*'s observation of mutation rate differences among the three groups of genomic regions could also be artifacts of false-positive detections of mutations around polynucleotides.

To test if over-detection of mutations around polynucleotides is sufficient to generate the mutational patterns reported by Monroe *et al.*, we simulated the same number of mutations as the total number of mutations reported by Monroe *et al.* in the three datasets they analyzed (1,140,848). Because 30% of these mutations are associated with polynucleotides, we randomly distributed 70% of the mutations at non-polynucleotide sites across the genome and the remaining mutations at polynucleotides. From the simulated data, we found a positive correlation between the frequency of mutations and  $d_N/d_S$  across genes (fig. 1F), as well as a variation in mutation frequency among coding regions, introns, and intergenic regions (fig. 1G) that is qualitatively similar to Monroe *et al.*'s results (fig. 1E). These trends disappear when all mutations are randomly distributed across the genome (supplementary fig. S2, Supplementary Material online). These findings suggest that false detections of mutations around polynucleotides are sufficient to produce the purported mutational trend of *Arabidopsis*.

## Conclusion

In summary, we showed that the trend of lower mutation rates in selectively more constrained genes that was

recently reported in *Arabidopsis* is present in neither yeast nor humans. Additionally, no mutation rate difference was found between genic and intergenic regions in prior yeast and fruit fly MA studies (Sharp and Agrawal 2016; Melde *et al.* 2022). We discovered that Monroe *et al.* (2022) identified orders of magnitude more mutations than reported previously from the same data and expected from *A. thaliana*'s known mutation rate. Many of the extra mutations reported by Monroe *et al.* appear to be sequencing errors associated with polynucleotides and these errors have the potential to create the purported unusual mutational trend. Together, our findings suggest that mutation rate is not lower in evolutionarily more constrained genomic regions in any of the plant, fungal, or animal models examined so far. While there is ample theoretical and empirical evidence that the genome-wide mutation rate is subject to natural selection (Kimura 1967; Sniegowski *et al.* 2000; Baer *et al.* 2007; Lynch 2011; Lynch *et al.* 2016; Liu and Zhang 2021), selection optimizing mutation rates of local genomic regions may simply be too weak (Chen and Zhang 2013) to have any effect in any species. While our analysis focused on SNV mutations, it is worth noting that yeast indel mutations occur less frequently in genic than intergenic regions, probably because, relative to intergenic regions, genic regions contain a lower density of repetitive sequences that are prone to indel mutations (Melde *et al.* 2022).

For various reasons, a genomic or epigenomic feature may be correlated with the selective constraint of a genomic region. For instance, because highly expressed genes tend to be under strong selective constraints (Zhang and Yang 2015), a genomic/epigenomic feature of high expression may be found more frequently in genes under stronger selective constraints. If this feature affects the mutation rate directly or indirectly, we might observe a correlation between the local mutation rate and selective constraint. However, given the theoretical understanding of the weakness of selection acting on local mutation rates, one should seriously consider the possibility that such potential correlations are not results of selection on local mutation rates but byproducts of some other processes (Zhang 2022). Our observation of a negative correlation between  $d_N/d_S$  and mutation rate across human genes is a testament to this possibility.

## Materials and Methods

*De novo* mutations were retrieved from three yeast studies (Zhu *et al.* 2014; Sharp *et al.* 2018; Liu and Zhang 2019), two human studies (Jonsson *et al.* 2017; An *et al.* 2018), and three *Arabidopsis* studies (Wang *et al.* 2019; Weng *et al.* 2019; Monroe *et al.* 2022). The  $d_N/d_S$  ratio for each gene in yeast (between *S. cerevisiae* and *S. paradoxus*) and humans (between *Homo sapiens* and *Pan troglodytes*) were obtained from a previous study (Goncalves *et al.* 2011) and Ensembl (<https://www.ensembl.org/Help/View?id=135>), respectively. Gene expression levels in yeast were retrieved from a previous study (Chou *et al.* 2017), and those

in human testis were retrieved from the GTEx project (The GTEx Consortium 2013). The  $d_N/d_S$  ratios (between *A. thaliana* and *A. lyrata*) and expression levels of *A. thaliana* genes were previously published (Monroe *et al.* 2022). Nucleosome occupancy information was acquired from the database NucMap (Zhao *et al.* 2019). Data on DNA replication timing was obtained from previous studies (Muller and Nieduszynski 2012; Concia *et al.* 2018; Pratto *et al.* 2021). DNA curvature was calculated following a previous study (Duan *et al.* 2018).

The aligned read files for the 107 *Arabidopsis* MA lines (Weng *et al.* 2019) were downloaded from NCBI Short Read Archive under the accession number SRP133100 and were visualized by IGV (Robinson *et al.* 2017). A polynucleotide tract is a run of the same seven or more nucleotides. *A. thaliana* has an AT-rich genome that comprises 165,937 instances of poly(A) or poly(T), but only 928 instances of poly(G) or poly(C). The number and location of polynucleotides were determined using a custom Perl script.

In table 1 and supplementary table S1, Supplementary Material online the mutation rate of a gene refers to the mutation rate in the coding region except for humans where the mutation rate of the entire genic region is computed because there were too few mutations in coding regions. Correspondingly, gene length refers to the coding sequence length except in the case of humans, where the total length of exons and introns is considered. Rank correlation, linear correlation, and multiple linear regression were performed in R.

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

We thank three anonymous reviewers for valuable comments. This work was supported by National Institutes of Health research grant R35GM139484 to J.Z.

## Data Availability

All data used in this study are publicly available as described in Materials and Methods.

## References

- The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* **45**:580–585.
- An JY, Lin K, Zhu L, Werling DM, Dong S, Brand H, Wang HZ, Zhao X, Schwartz GB, Collins RL, *et al.* 2018. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**:eaat6576.
- Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet.* **8**:619–631.
- Chen X, Zhang J. 2013. No gene-specific optimization of mutation rate in *Escherichia coli*. *Mol Biol Evol.* **30**:1559–1562.
- Chou HJ, Donnard E, Gustafsson HT, Garber M, Rando OJ. 2017. Transcriptome-wide analysis of roles for tRNA modifications in translational regulation. *Mol Cell.* **68**:978–992.e4.
- Concia L, Brooks AM, Wheeler E, Zynada GJ, Wear EE, LeBlanc C, Song J, Lee TJ, Pascuzzi PE, Martienssen RA, *et al.* 2018. Genome-wide analysis of the *Arabidopsis* replication timing program. *Plant Physiol.* **176**:2166–2185.
- Duan C, Huan Q, Chen X, Wu S, Carey LB, He X, Qian W. 2018. Reduced intrinsic DNA curvature leads to increased mutation rate. *Genome Biol.* **19**:132.
- Goncalves P, Valerio E, Correia C, de Almeida JM, Sampaio JP. 2011. Evidence for divergent evolution of growth temperature preference in sympatric *Saccharomyces* species. *PLoS One* **6**:e20739.
- Heydari M, Miclotte G, Van de Peer Y, Fostier J. 2019. Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. *BMC Bioinf.* **20**:298.
- Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, *et al.* 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**:519–522.
- Kiktev DA, Sheng Z, Lobachev KS, Petes TD. 2018. GC Content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A.* **115**:E7109–E7118.
- Kimura M. 1967. On the evolutionary adjustment of spontaneous mutation rates. *Genet Res.* **9**:23–34.
- Lederberg J, Lederberg EM. 1952. Replica plating and indirect selection of bacterial mutants. *J Bacteriol.* **63**:399–406.
- Li C, Luscombe NM. 2020. Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. *Nat Commun.* **11**:1363.
- Lind PA, Berg OG, Andersson DI. 2010. Mutational robustness of ribosomal protein genes. *Science* **330**:825–827.
- Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. *BMC Evol Biol.* **2**:20.
- Liu H, Zhang J. 2019. Yeast spontaneous mutation rate and spectrum vary with environment. *Curr Biol.* **29**:1584–1591.e3.
- Liu H, Zhang J. 2020. Higher germline mutagenesis of genes with stronger testis expressions refutes the transcriptional scanning hypothesis. *Mol Biol Evol.* **37**:3225–3231.
- Liu H, Zhang J. 2021. The rate and molecular spectrum of mutation are selectively maintained in yeast. *Nat Commun.* **12**:4044.
- Luria SE, Delbruck M. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**:491–511.
- Lynch M. 2011. The lower bound to the evolution of mutation rates. *Genome Biol Evol.* **3**:1107–1118.
- Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet.* **17**:704–714.
- Martincorena I, Seshasayee AS, Luscombe NM. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* **485**:95–98.
- Melde RH, Bao K, Sharp NP. 2022. Recent insights into the evolution of mutation rates in yeast. *Curr Opin Genet Dev.* **76**:101953.
- Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, Klein M, Hildebrandt J, Neumann M, Kliebenstein D, *et al.* 2022. Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* **602**:101–105.
- Muller CA, Nieduszynski CA. 2012. Conservation of replication timing reveals global and local regulation of replication origin activity. *Genome Res.* **22**:1953–1962.
- Oman M, Alam A, Ness RW. 2022. How sequence context-dependent mutability drives mutation rate variation in the genome. *Genome Biol Evol.* **14**:evac032.
- Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**:92–94.

- Park C, Chen X, Yang JR, Zhang J. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. **110**:678–683.
- Park C, Qian W, Zhang J. 2012. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep*. **13**:1123–1129.
- Pratto F, Brick K, Cheng G, Lam KG, Cloutier JM, Dahiya D, Wellard SR, Jordan PW, Camerini-Otero RD. 2021. Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell* **184**:4251–4267.e20.
- Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A, Mesirov JP. 2017. Variant review with the integrative genomics viewer. *Cancer Res*. **77**:e31–e34.
- Sharon E, Chen SA, Khosla NM, Smith JD, Pritchard JK, Fraser HB. 2018. Functional genetic variants revealed by massively parallel precise genome editing. *Cell* **175**:544–557.e16.
- Sharp NP, Agrawal AF. 2016. Low genetic quality alters key dimensions of the mutational spectrum. *PLoS Biol*. **14**:e1002419.
- Sharp NP, Sandell L, James CG, Otto SP. 2018. The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. *Proc Natl Acad Sci U S A*. **115**:E5046–E5055.
- Shen X, Song S, Li C, Zhang J. 2022. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature* **606**:725–731.
- Sniegowski PD, Gerrish PJ, Johnson T, Shaver A. 2000. The evolution of mutation rates: separating causes from consequences. *Bioessays* **22**:1057–1066.
- Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet*. **41**:393–395.
- Wang L, Ji Y, Hu Y, Hu H, Jia X, Jiang M, Zhang X, Zhao L, Zhang Y, Jia Y, et al. 2019. The architecture of intra-organism mutation rate variation in plants. *PLOS Biol*. **17**:e3000191.
- Weng ML, Becker C, Hildebrandt J, Neumann M, Rutter MT, Shaw RG, Weigel D, Fenster CB. 2019. Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics* **211**:703–714.
- Xia B, Yan Y, Baron M, Wagner F, Barkley D, Chiodin M, Kim SY, Keefe DL, Alukal JP, Boeke JD, et al. 2020. Widespread transcriptional scanning in the testis modulates gene evolution rates. *Cell* **180**:248–262.e21.
- Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen JQ, Hurst LD, Tian D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**:463–467.
- Zhang J. 2022. Important genomic regions mutate less often than do other regions. *Nature* **602**:38–39.
- Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. *Nat Rev Genet*. **16**:409–420.
- Zhao Y, Wang J, Liang F, Liu Y, Wang Q, Zhang H, Jiang M, Zhang Z, Zhao W, Bao Y, et al. 2019. Nucmap: a database of genome-wide nucleosome positioning map across species. *Nucleic Acids Res*. **47**:D163–D169.
- Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci U S A*. **111**:E2310–E2318.