12-8-2023

Investigating the impact of transcription on mutation rates

Sarah Patterson
Mississippi State University, spatterson1655@gmail.com

Investigating the impact of transcription on mutation rates

By

Sarah Patterson

Approved by:

Jean-Francois Gout (Major Professor)
Amy Dapper
Donna M. Gordon
Andy Perkins (Committee Member/Graduate Coordinator)
Rick Travis (Dean, College of Arts & Sciences)

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Computational Biology
in the Computational Biology Program

Mississippi State, Mississippi

December 2023

Name: Sarah Patterson

Date of Degree: December 8, 2023

Institution: Mississippi State University

Major Field: Computational Biology

Major Professor: Jean-Francois Gout

Title of Study: Investigating the impact of transcription on mutation rates

Pages in Study: 34

Candidate for Degree of Master of Science

tRNA genes are highly transcribed and perform one of the most fundamental cellular functions. Although a universal pattern observed across all three domains of life is that highly transcribed genes tend to evolve slowly, tRNA genes have been shown previously to evolve rapidly. This rapid sequence evolution could result from relaxed selection, increased mutation rate, or a combination of both. Here, we use mutation-accumulation line sequencing data to show that tRNA genes accumulate more mutations than other gene types. Our results indicate that this elevated mutation rate is a consequence of both elevated transcription-associated mutagenesis and a lack of transcription-coupled repair in tRNA genes. We also identify the gene *MSH2* as being involved in transcription-coupled repair.

DEDICATION

Dedicated to all the friends, family, and cats who supported me through my time as a student. I would never have made it here without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INVESTIGATING THE IMPACT OF TRANSCRIPTION ON MUTATION RATES

Introduction

Mutations are at the very core of evolution. They provide the raw material for genetic diversity within a species and can allow individuals to adapt to new or changing environments. Mutations primarily occur when DNA is replicated due to misincorporations by the DNA polymerases (Clausen et al., 2013; Nick McElhinny et al., 2010) or when chemical damage occurs on one or both strands of the DNA (Marnett, 2000). However, DNA damage can also occur outside of replication and many parameters affect the rate at which such damage occurs. The production of mutations, regardless of means, is referred to as mutagenesis. One particular type of mutagenesis, transcription-associated mutagenesis (TAM), occurs during the process of transcription (Jinks-Robertson & Bhagwat, 2014). During transcription, double-stranded DNA molecules are unwound into single-stranded DNA (ssDNA). This ssDNA is vulnerable to mutagens in the cell and may accumulate chemical damage leading to an increased mutation rate (Ramiro et al., 2003). Therefore, it is expected that the likelihood of sustaining mutations increases with transcription level. Highly expressed genes are typically evolving under stronger purifying selection (Duret & Mouchiroud, 2000; Hastings, 1996; Pál et al., 2001), so it is expected that mutations in these genes tend to be more deleterious and selection should reduce mutation rate in these genes if possible (Xiong et al., 2017).

In this context, tRNA genes present a paradox, in that they are both highly transcribed (Lucas et al., 2023; Nagai et al., 2021) and evolve rapidly. This rapid evolution is likely caused by relaxed selection, higher mutation rate, or a combination of both. A previous study (Thornlow et al., 2018) has come to the conclusion that tRNA genes experience a substantial increase in mutations through increased TAM but undergo purifying selection. This study, however, only looked at long-term patterns of evolution between species, thus failing to consider cellular mechanisms, such as transcription-coupled repair (TCR), that may contribute to the rapid evolution of tRNA genes compared to other gene types.

Transcription-coupled repair (TCR) is a cellular process of mutation rate reduction present in eukaryotes. It allows for repair of damaged sites in the DNA during the process of transcription (Spivak, 2016). TCR relies on the RNA polymerase to detect damage in the template strand and signal for repair at the damaged site (Spivak, 2016). Despite the importance of TCR to transcriptional fidelity, we still don't have a full understanding of this mechanism, including the list of genes which are involved with the process, either directly or indirectly. While some genes are known to be involved in TCR (e.g., CSB (Bradsher et al., 2002), XAB2 (Nakatsu et al., 2000), and UVR-A, B, C, and D (Mellon & Champe, 1996)), there is still a level of uncertainty as to whether or not certain genes are involved in the process. One such gene is *MSH2*, which has previously been demonstrated (van Oosten et al., 2005) to be involved in the global genomic repair (GGR) process, but its potential role in TCR is still unclear (Sweder et al., 1996). Further study of the genes involved in TCR would aid in elucidating the specifics of how the mechanism functions as well as give insight as to why the mechanism does not repair every mutation it comes across.

As TCR is a highly effective, but not perfect, mechanism, it is difficult to discern whether TCR fully compensates (and even possibly offsets) the impact of TAM, possibly leading to reduced mutation rate in highly transcribed genes, or if TAM overwhelms TCR, leading to higher mutation rates in highly transcribed genes. Answering this question has proven to be difficult, as different studies have reached opposite conclusions (X. Chen & Zhang, 2014; Zhu et al., 2014a, 2014b), but the current consensus is that the effect of TAM is almost entirely offset by TCR, leaving only a weak positive correlation between expression level and mutation rate in protein-coding genes (Zhu et al., 2014a). However, not all genes benefit from TCR. Eukaryotic tRNA genes are of particular importance due to their lack of a TCR mechanism despite being transcribed at record high levels. Indeed, tRNA genes are transcribed by RNA polymerase III (RNAPIII) while TCR has been demonstrated to be associated with RNAPII (the polymerase which transcribes protein-coding genes) (Adebali et al., 2017; Hu et al., 2016; Sancar, 2016) and is suspected to be also active with rRNA-transcribing RNAPI (Bradsher et al., 2002; Daniel et al., 2018). Due to this lack of a TCR mechanism in tRNA genes combined with their extreme transcription levels, we predict that tRNA genes should have a higher mutation rate than protein-coding genes in eukaryotes.

A major challenge in testing this hypothesis is to obtain reliable estimates of mutation rates for both types of genes. Indeed, most studies of mutation rates rely on sequencing end-products of mutation accumulation experiments and typically result in the detection of a few hundred to a few thousand mutations at most (Liu & Zhang, 2019, 2021; Sharp et al., 2018; Zhu et al., 2014a). This issue is also true of experiments that rely on trio sequencing, which is the gold-standard method for measuring mutation rates in humans (the 1000 Genomes Project, 2011). Because tRNA genes make up a small proportion of the genome, the chance of detecting

mutations in these genes is fairly low, reducing the statistical power of any comparison of mutation rate between tRNA genes and protein-coding genes. For example, with the 84 mutations found in a trio sequencing experiment by the 1000 Genomes Project (the 1000 Genomes Project, 2011), only 0.001% of those mutations (0.001 mutations) are expected to be found in tRNA genes with a uniform mutation rate across the entire genome (tRNA genes make up only 51,783 of 3,298,430,730 base pairs (~0.001%) in the human genome). As such, we would have no statistical power to detect even a 10-100X increase in mutation rate. While the fraction of the genome occupied by tRNA genes increases in species with more compact genomes, it remains a small fraction of the total genome size. For example, in *S. cerevisiae*, tRNA-coding genes only make up 23,778 of the 12,157,105 base pairs, or ~0.2% of the genome. However, this may be just enough to allow for detection of an increased mutation rate, provided there are several hundred mutations to consider from mutation accumulation (MA) line studies.

To investigate the cause of rapid evolution in tRNA genes, we gathered data from several previously published MA line studies to compare the mutation rate and spectrum of tRNA genes to that of the rest of the genome. Analysis of the data gained from these MA line studies provides insight into the roles of both TAM and TCR in the rapid accumulation of mutations in tRNA genes as compared to other highly transcribed genes, furthering our current understanding of the accelerated evolution of tRNA genes. Additionally, we provide evidence for the involvement of *MSH2* in TCR in eukaryotes as well as evidence indicating the presence of strand-specific mutations and mutation repair biases.

Results

tRNA Genes in S. cerevisiae Have an Elevated Mutation Rate

To investigate the impact of transcription-coupled repair (TCR) on mutation rates, we first analyzed the mutation rate of tRNA genes in the yeast *S. cerevisiae*. tRNA genes make the bulk of the genes transcribed by the RNA Polymerase III (RNAPIII), the only nuclear RNA polymerase with no known TCR mediation activity (Yang et al., 2019). tRNA genes represent about 0.2% of the total genome size of *S. cerevisiae*. While this is only a small fraction of the genome, it is still an order of magnitude more than the fraction of the genome occupied by tRNA genes in other model eukaryotes with large mutation accumulation datasets available (for example, tRNA genes make up only 0.016% of the entire genome in Drosophila melanogaster). Based on these numbers, and assuming a uniform mutation rate along the genome, we would expect to observe, on average, one mutation in tRNA genes for every 500 mutations reported in mutation accumulation line experiments in *S. cerevisiae*. With a few thousand mutations reported across several mutation accumulation (MA) line studies, large differences in mutation rates between tRNA genes and the rest of the genome might be detectable in *S. cerevisiae*.

We re-analyzed data from three independent MA line studies (Liu & Zhang, 2019; Sharp et al., 2018; Zhu et al., 2014a) and used the frequency of mutations observed outside of tRNA genes to compute the expected number of mutations inside tRNA genes, assuming a constant mutation rate across the entire genome. In all three datasets, we found that mutations inside tRNA genes are >5-times more frequent than in the rest of the genome. This excess of mutations inside tRNA genes is highly significant (Figure 1A, $p = 1.15 \times 10^{-10}$) in all three datasets, revealing that tRNA genes experience a higher mutation rate than the rest of the genome in the

yeast *S. cerevisiae*. This heightened mutation rate in tRNA genes is likely due to a combined effect of transcription-associated mutagenesis (TAM) and lack of TCR.

However, tRNA genes are known to have a higher GC content than the rest of the genome (average GC content 52% for tRNA genes vs. 38% for the rest of the genome in *S. cerevisiae*). Additionally, guanine:cytosine (GC) pairs have been found to mutate at a higher frequency than adenine:thymine (AT) pairs (Zhu et al., 2014a). Therefore, we investigated the possibility that the elevated mutation rate seen in tRNA genes could be explained by a higher GC content than the rest of the genome. To correct for this potential bias, we computed the mutation rate at GC and AT pairs independently and applied these rates to GC and AT pairs inside tRNA genes. This correction for GC content made very little difference to the comparison of mutation rates between tRNA genes and the rest of genomes (Appendix, Table 2), indicating that the elevated mutation rate of tRNA genes is not explained by their biased nucleotide composition.

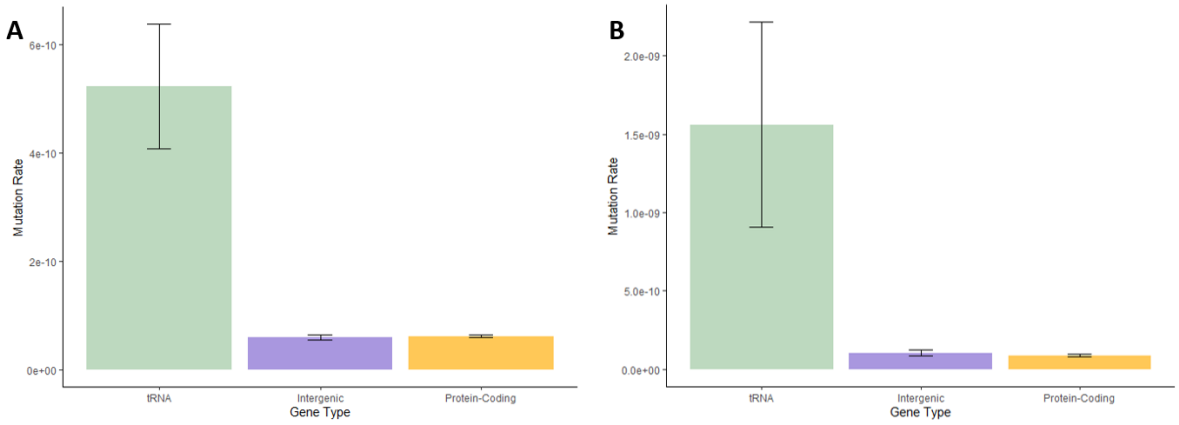

Figure 1 Mutation rates across *S. cerevisiae* and *S. pombe* by gene type.

A) Mutation rates compared by gene type in *S. cerevisiae*. tRNA genes show a ~7.4-fold increase in mutation rate when compared to protein-coding genes and intergenic regions. B) Mutation rates compared by gene type in *S. pombe*. tRNA genes show a ~17.7-fold increase in mutation rate when compared to protein-coding genes and intergenic regions.

Other Eukaryotic Species Also Show Elevated Mutation Rate for tRNA Genes

Several other eukaryotic species were also considered to determine whether or not the elevated mutation rate seen in *S. cerevisiae* was a trend seen across eukaryotes. Of the species considered, only *S. pombe* had sufficient data from previous MA line experiments for analysis. It is important to note that, while *S. cerevisiae* and *S. pombe* belong to the same phylum, their divergence time is estimated to be about half a billion years (Kumar et al., 2022) and their genome structures are highly divergent as illustrated by the many introns present in *S. pombe* (Wood et al., 2002). Therefore, any pattern present in both yeast species would have a good chance of being present in other eukaryotes as well.

In *S. pombe*, after correction for GC content, we found that mutations inside tRNA genes are >14-times more frequent than in the rest of the genome (Figure 1B, $p = 7.45 \times 10^{-10}$). This excess of mutations in tRNA genes combined with that found in tRNA genes in *S. cerevisiae* indicates that the elevated mutation rate seen in tRNA genes in these species is not species-specific, but rather is a trend seen across eukaryotes. To further support this, we analyzed several other eukaryotic species with mutation accumulation lines derived estimates of mutation rate (see Appendix, Table 3). However, the relatively small number of mutations reported in each one of these species, combined with the extremely small fraction of their genome occupied by tRNA genes implies that the expected number of mutations observed in tRNA genes would still be typically less than one, even with a mutation rate 5-times higher for tRNA genes. These small numbers prevent us from having any statistical support for each species individually. However, we note that the mutation rate is higher in tRNA genes for all the species considered (Appendix, Table 3), suggesting that the trend observed in yeast extends to multicellular eukaryotes.

Introducing a Control

tRNA genes show a heightened mutation rate compared to the rest of the genome. However, as tRNA genes are, on average, much more highly expressed than the rest of the genome (Lucas et al., 2023; Nagai et al., 2021), it is unclear whether this difference is due more so to their lack of transcription-coupled repair (TCR) or excess of transcription-associated mutagenesis (TAM). If the effects of TCR were to be removed, it would be possible to tell how much of the difference in tRNA gene mutation rate is contributed by each. We accomplished this by analyzing a dataset from Liu & Zhang (Liu & Zhang, 2021) in which the gene *MSH2* was knocked out in *S. cerevisiae* before performing mutation accumulation lines followed by genome sequencing. *MSH2* is believed to be involved in TCR, although this is still debated (Sweder et al., 1996; van Oosten et al., 2005). When analyzing this dataset, we found that tRNA genes had a mutation rate only ~2 times higher than that of the rest of genome (Figure 2, $p = 3.44 \times 10^{-9}$), compared to the ~5 times increase previously observed in wild-type *S. cerevisiae*. This result strongly suggests that *MSH2* is indeed involved in TCR and that lack of TCR is a major reason why tRNA genes have an elevated mutation rate.

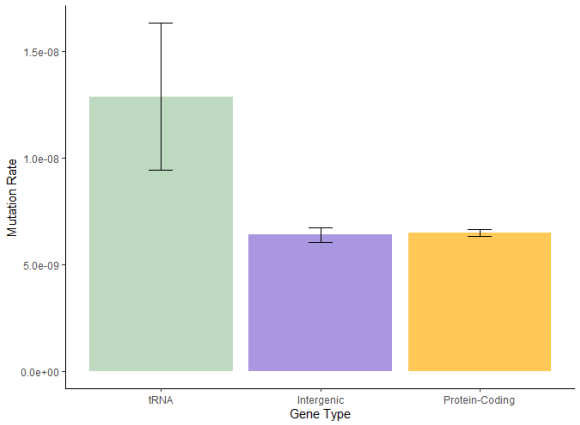

Figure 2 Mutation rates across *S. cerevisiae* with *msh2* knockout by gene type.

tRNA genes show a ~2-fold increase in mutation rate when compared to protein-coding genes and intergenic regions.

Effect of Expression Level on Mutation Rate

Removing the action of TCR might help us quantify the impact of TAM on the mutational input. Indeed, previous studies that aimed at detecting the effect of TAM on mutation rate relied on comparisons of mutation rates for bins of genes of increasing expression level. However, this strategy failed to reveal any large increase in mutation rate associated with elevated expression level and the impact of transcription on mutation has remained difficult to quantify (X. Chen & Zhang, 2014; Zhang et al., 2018; Zhu et al., 2014b). The most likely explanation for this difficulty is that the increased DNA damage caused by TAM in highly expressed genes is almost entirely compensated by the action of TCR.

With TCR activity mostly abolished, the *msh2*-KO MA line dataset represents a unique opportunity to directly measure the impact of transcription on DNA damage. To determine the relationship between expression level and mutation rate in the *msh2*-knockout datasets, we split both protein-coding genes and tRNA genes by expression level as determined by a previous study (Pelechano et al., 2010). We found that the mutation rate increases with increased

expression level in protein-coding genes (Figure 3, $p = 7.59 \times 10^{-5}$), revealing the impact of TAM which was previously hidden by TCR. It is notable that there is no similar difference between highly and lowly expressed tRNA genes. However, it is exceptionally difficult to get reliable estimates of tRNA expression (Lucas et al., 2023; Nagai et al., 2021). Additionally, tRNA genes have extremely high expression levels compared to protein-coding genes (Lucas et al., 2023; Nagai et al., 2021). It is possible that the impact of TAM does not increase linearly with expression level, so that the impact of TAM might reach an asymptote at extremely high levels of expression.

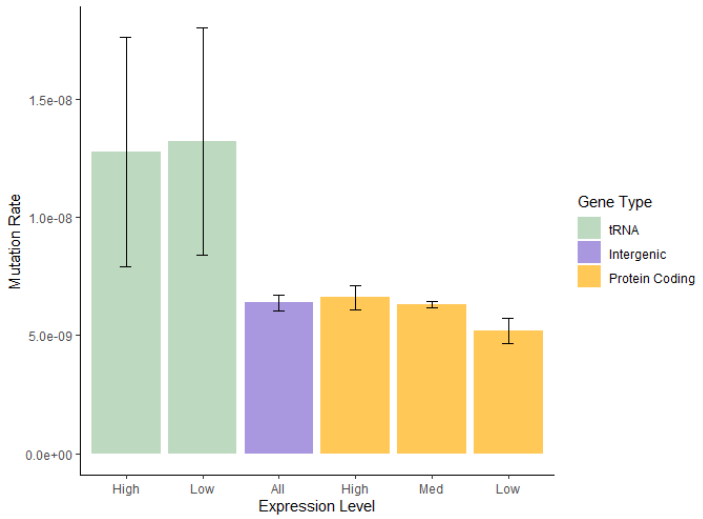

Figure 3 Mutation rate across *S. cerevisiae* with *msh2* knockout by gene type and gene expression level.

tRNA genes, regardless of expression level, show a higher mutation rate than protein-coding genes and intergenic regions. There is no significant difference in mutation rate between highly and lowly expressed tRNA genes. Protein-coding genes show a positive correlation between expression level and mutation rate.

Mutation spectra of prokaryotes

While the *msh2*-knockout dataset is highly informative about the effects of TAM, it is entirely possible that the activity of TCR was not completely abolished by the *msh2* knock out. There is already a natural environment in which there is no gene-specific effect of TCR: bacteria. While tRNA genes in prokaryotes are still much more highly expressed than protein-coding genes (Lucas et al., 2023; Nagai et al., 2021), there is only one RNA polymerase in prokaryotes, and therefore no gene-specific TCR mechanism like that of eukaryotes.

To investigate the impact of TAM on mutation rates in prokaryotes, we analyzed the mutation rate of tRNA genes in the bacterium *E. coli*. We re-analyzed data from three independent MA line studies (Foster et al., 2015, 2018; Zhang et al., 2018) and computed the mutation rates of tRNA genes and the rest of the genome, as we did previously in yeast. In all three datasets, we found mutations in tRNA genes to be >1.5-times more frequent than in the rest of the genome. The excess of mutations in tRNA genes was highly significant (Figure 4, $p = 3.04 \times 10^{-12}$) in all three datasets, indicating that, much like eukaryotes, tRNA genes experience a higher mutation rate than the rest of the genome in *E. coli*. However, because no gene-specific TCR mechanism exists in *E. coli*, this difference has to be explained by another mechanism, most likely the elevated amount of TAM caused by the high expression levels of tRNA genes. We split *E. coli* tRNA genes into two bins according to their expression level and found that the elevated mutation rate was specific to the highly expressed ones, with lowly expressed tRNA genes having a mutation rate barely above that of protein-coding genes (Figure 5, $p = 2.97 \times 10^{-12}$). Therefore, it appears that the elevated mutation rate of tRNA genes in *E. coli* is caused exclusively by their high expression level.

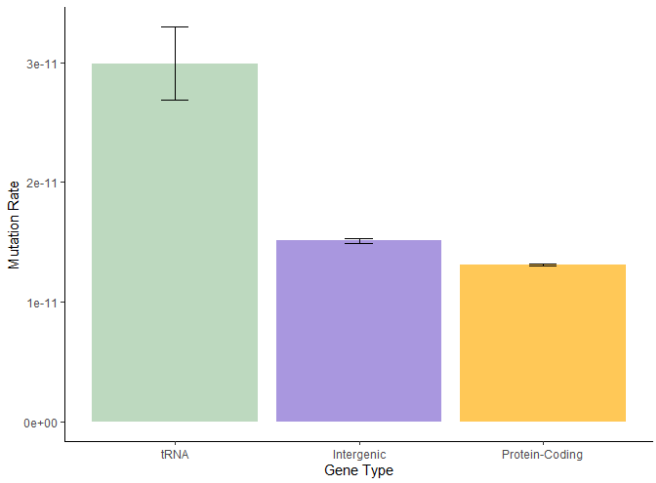

Figure 4 Mutation rates across *E. coli* by gene type.

tRNA genes show a ~2.3-fold increase in mutation rate when compared to protein-coding genes and intergenic regions.

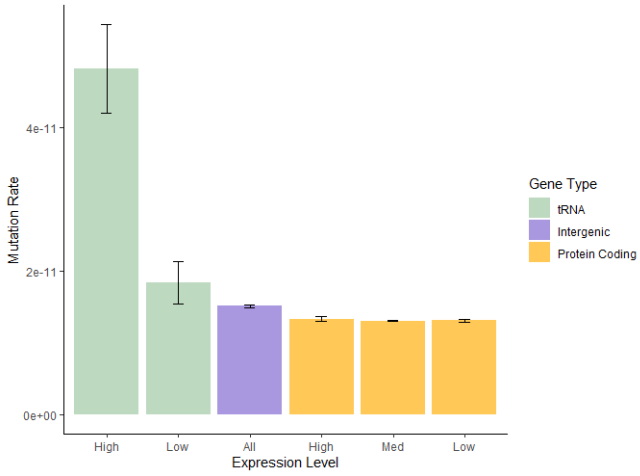

Figure 5 Mutation rate across *E. coli* by gene type and gene expression level.

tRNA genes, regardless of expression level, show a higher mutation rate than protein-coding genes and intergenic regions. Additionally, highly expressed tRNA genes have a significantly higher mutation rate than lowly expressed tRNA genes. Protein-coding genes show no correlation between expression level and mutation rate.

Biases in Type of Repair Targeted by TCR

To determine whether TCR is more effective at repairing certain mutations than others, we ascertained the full mutation spectrum of *S. cerevisiae* in tRNA genes compared to that of protein-coding genes. Mutation rates were computed for each of the 12 possible base-substitutions and were polarized relative to the "coding" strand (i.e., the non-template strand for tRNA genes). Four types of base-substitutions were especially common in tRNA genes: G-to-T, C-to-A, C-to-T, and A-to-G (Figure 6). We found that the two most common types of point mutations in tRNA genes are the complementary G-to-T and C-to-A base-substitutions (Figure 6). This points to a mutational process that can be started and repaired on both strands with the same probability. However, the next two most common types of mutations appear to be strand-specific. Indeed, C-to-T mutations are more frequent than their complementary G-to-A, and A-to-G is also more frequent than its complementary T-to-C. This pattern suggests the presence of strand-specific damage or repair in tRNA genes.

In the *S. cerevisiae msh2* knockout dataset, all mutations showed an elevation of mutation rate compared to the non-knockout datasets regardless of gene type (Appendix, Figure 8). It is particularly noteworthy that, while tRNA mutation rates seem to go up significantly, mutation rates in protein-coding genes appear to increase at a much higher rate. The increase in tRNA mutation rates can be explained by the involvement of *MSH2* in global genomic repair (GGR) while the greater increase in protein-coding mutation rates can be explained both by involvement of *MSH2* in GGR and by involvement of *MSH2* in TCR.

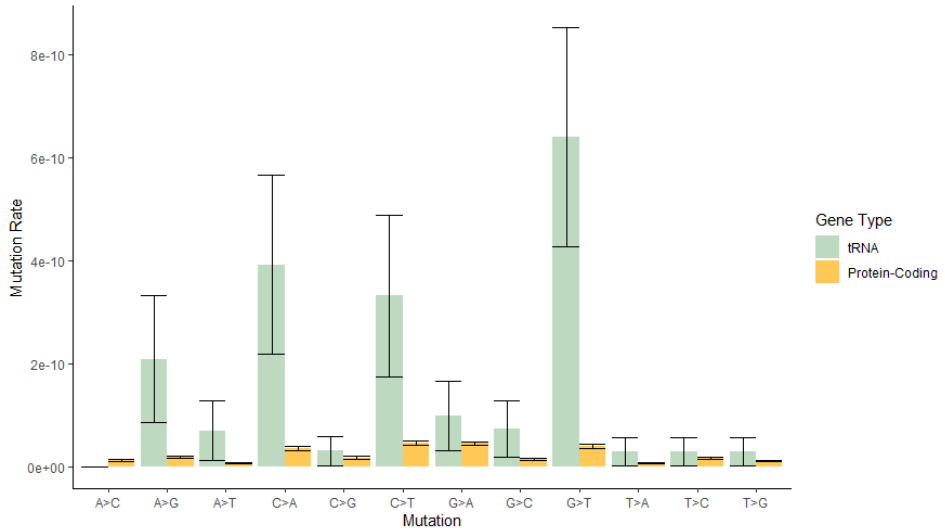

Figure 6 Mutation spectrum of *S. cerevisiae* by gene type.

tRNA genes overall show a heightened mutation rate for most mutation types in *S. cerevisiae* as compared to protein-coding genes. tRNA genes show particularly elevated mutation rates for A-to-G, C-to-A, C-to-T, and G-to-T mutations.

In *E. coli*, every mutation occurred at a significantly higher rate in tRNA genes than in protein-coding genes, with the exception of G-to-A. Contrary to *S. cerevisiae*, where only a subset of the 12 possible base-substitutions have an elevated mutation rate, the fact that almost all types of base-substitutions increase in the same proportion in *E. coli* tRNA genes suggests that the process responsible for this increase is not biased towards either strand.

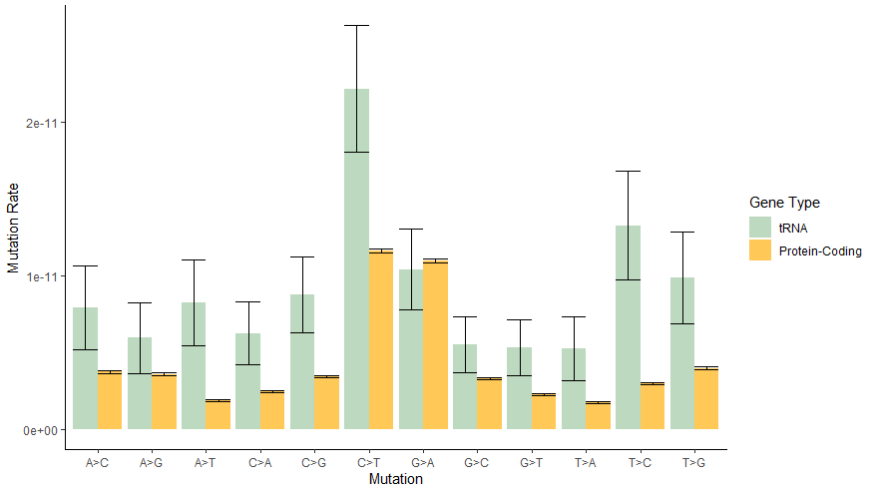

Figure 7 Mutation spectrum of *E. coli* by gene type.

For all but one mutation type (G-to-A), tRNA genes show a heightened mutation rate as compared to protein-coding genes.

Discussion

In this study, we have used publicly available mutation accumulation (MA) lines sequencing data to directly compare mutation rate of tRNA genes to that protein-coding genes. Our results show that tRNA genes have mutation rates about one magnitude higher than the rest of the genome in yeast. A previous study, relying on divergence and polymorphism data in humans and mice concluded that tRNA genes experience elevated mutation rates despite being under strong purifying selection (Thornlow et al., 2018). While Thornlow et al. attributed this elevated mutation rate to higher levels of transcription-associated mutagenesis (TAM), our analysis shows that lack of transcription-coupled repair is the main reason for elevated mutation rate of tRNA genes. Indeed, when *MSH2* is knocked-out, the difference in mutation rate between tRNA and protein-coding genes drops down to only about 2-fold. We note that this observation, combined with the obvious positive correlation between mutation rate and expression level in protein-coding genes upon *MSH2* inactivation, confirms the crucial role played by *MSH2* in

TCR. Previously, *MSH2* was known to be involved in global genomic repair (GGR) but its role in TCR was debated (Sweder et al., 1996; van Oosten et al., 2005).

Because both tRNA and protein-coding genes experience the same amount of repair by TCR in the absence of *MSH2* (i.e., no repair), the difference in mutation rate in this context allows us to isolate the contribution of TAM alone. This result is further supported by our analysis of bacterial MA lines, where there is no difference in TCR between tRNA and protein-coding genes, and the difference in mutation is also about 2-fold. Therefore, we conclude that the elevated mutation rate of tRNA genes is explained by a combination of their extreme transcription levels and lack of transcription-coupled repair.

The types of base-substitutions observed in tRNA genes are also an indicator of the processes responsible for the elevated mutation rate. The two strands of transcribed genes are not equally as likely to be damaged, as the non-transcribed strand (NTS) is more frequently exposed to the cellular environment in a single-stranded state than the transcribed strand (TS) (Ramiro et al., 2003). For this reason, the NTS is more likely to be damaged during transcription and accumulate mutations over time (Polak & Arndt, 2008; Saini et al., 2017). This may, over time, lead to patterns of strand-specific damage, as DNA replication may mispair the damaged base with the incorrect base, leading to a mutation even upon repair of the damaged base (Clausen et al., 2013; Nick McElhinny et al., 2010). In our study, the three most prevalent mutation types all showed a pattern of strand-specific damage, which we attempt to explain here.

The two types of base-substitutions with the highest elevation in tRNA genes are G-to-T and C-to-A. The most likely culprit in producing G-to-T and C-to-A mutations in the DNA is the creation of 8-oxoguanine through oxidative damage to guanine (Bohr et al., 2002). 8-oxoguanine often mispairs with adenine (Zahn et al., 2011), resulting in C-to-A mutations on the strand

opposite to that of the damage (and G-to-T mutation on the damaged strand). Our observation that the two complementary mutations (G-to-T and C-to-A) occur at similarly high rates in tRNA genes is in agreement with the lack of a strand-specific repair mechanism taking place in these genes.

Surprisingly, we found that the remaining two types of base-substitutions with elevated rates in tRNA genes reveal a strand-specific pattern. Indeed, while C-to-T mutation rate is strongly elevated in tRNA genes, the complementary G-to-A mutation is not elevated, contrary to what is expected for mutational mechanisms that act with the same strength on both strands. The same is true of A-to-G, where the complementary T-to-C mutations show no significant elevation in tRNA genes. We interpret the elevated C-to-T mutation rate (and lack of elevated G-to-A) as additional evidence for strand specific cytidine deamination, confirming previous research showing higher rate of deamination on the NTS strand (Bhagwat et al., 2016; Sohail et al., 2003).

Here, we would like to propose an additional mechanism by which such strand-specific patterns could emerge in regions such as tRNA genes where no strand-specific repair mechanism is operating. It is well established that uridines, the main product of cytidine deamination, are removed from DNA by uracil-DNA glycosylases (Krokan et al., 2001). This leaves an abasic site which can eventually be repaired. Presence of an abasic site on the template strand can lead to stalling of the transcribing RNA polymerase and ultimately abandonment of the transcriptional process if the abasic site is not repaired (Y. H. Chen & Bogenhagen, 1993). As most of the tRNA genes considered in this study are essential genes, it is entirely possible that the presence of abasic sites in the template strand would impose a significant fitness cost by preventing transcription of the damage-containing tRNA genes. Therefore, purifying selection would be

responsible for the lack of C-to-T mutations on the template strand (which would show as G-to-A in our graph).

We observe a similar pattern in the strand-specific bias in A-to-G mutations over T-to-C mutations on the non-transcribed strand of tRNA genes. This observation is in direct contrast with previous publications (Zhu et al., 2014a, 2014b), which determined TCR to be the deciding factor in this mutational asymmetry. As there is no TCR mechanism in tRNA genes, the observed mutational asymmetry must be resultant of some other mechanism. It is entirely possible that a similar mechanism to that behind the observed skew between C-to-T and G-to-A mutations is responsible for this bias. Indeed, when adenine is deaminated, hypoxanthine is produced, causing a mispairing of the resultant deoxyinosine, most commonly with cytidine (Case-Green & Southern, 1994). Hypoxanthine is then removed by a DNA-deoxyinosine glycosylase (Mi et al., 2012), creating a highly deleterious abasic site.

Finally, we note that the elevated mutation rate of tRNA genes is a puzzling paradox in the face of their importance and evolution under strong purifying selection (Thornlow et al., 2018). Indeed, natural selection is expected to drive mutation rate to the lowest possible level, which is determined by the relative strengths of selection and drift (Lynch, 2011). In this context, it is surprising that selection did not favor the emergence of TCR for tRNA genes especially since it would most likely only require the addition of one or a few subunits to the RNA polymerase III. One possible solution to this paradox is that TCR comes at a cost of reduced processivity, increasing the amount of time needed to transcribed genes. Because tRNA genes are transcribed at such high levels, the cost of reduced transcription rate might overwhelm the benefits of a lower mutation rate.

Conclusions

We have determined that tRNA genes do indeed experience a much higher rate of mutation than protein-coding genes. However, in eukaryotes, this difference is contributed to by both transcription-coupled repair (TCR) and transcription-associated mutagenesis (TAM). To determine the level to which each of these mechanisms is responsible for the difference in mutation rate between tRNA and protein-coding genes, we determined the difference in mutation rate between the two gene groups in prokaryotes as well. As prokaryotes do not have an RNA polymerase-specific TCR mechanism, we were able to identify the contribution of TAM alone on the heightened mutation rate seen in tRNA genes and determined that TAM on its own does cause a slightly higher mutation rate in tRNA genes than in protein-coding genes.

However, as the difference in mutation rate observed in prokaryotes was less than half that seen in eukaryotes, we determined that the majority of the difference in mutation rate between tRNA and protein-coding genes in eukaryotes comes from the TCR mechanism being available to protein-coding genes but not to tRNA genes.

Materials and Methods

MA Line Materials

The MA line studies used in this experiment are listed by species (and, in the case of *E. coli*, strain) in Appendix Table 1. Genome-wide mutation rates for each dataset within a species or strain were highly similar. As such, for purposes of analysis, all mutation data for each species or strain was combined to create one dataset for each species or strain. All datasets without ample data were cut from the final analysis. For each dataset, an appropriate reference genome

was used to separate genomic regions into tRNA genes, protein-coding genes, and intergenic regions.

For most studies considered, the reference genome used was that which was listed in the original publication. For those studies that did not list a reference genome, the NCBI reference genome with the closest percent match was used. Reference genomes used are listed in Appendix Table 1.

Mutation Rate Estimation and Comparison

Mutation rate of protein-coding genes, as determined by the ratio of number of mutations to number of base positions covered across all generations (Equation 1), was applied to tRNA genes to provide an expected mutation rate.

$$Mutation\ Rate = \frac{\#\ Mutations\ in\ Gene\ Type}{\#\ Bases\ Covered\ in\ Gene\ Type \bullet \#\ Generations} \quad (1)$$

Expected tRNA mutation rate was calculated both with and without consideration of GC content. To correct for GC content, the expected number of mutations for each base in tRNA genes was calculated individually using the mutation rate of the same base in the rest of the genome (Equation 2). This was done to avoid skewing of tRNA mutation rate due to higher GC content.

$$\#\ Expected\ Mutations = \\ \#\ Bases\ Covered\ in\ tRNA\ Genes\ \bullet Base\ Mutation\ Rate\ in\ Rest\ of\ Genome \quad (2)$$

The actual mutation rate of tRNA genes was calculated using MA line mutation data (Equation 1). Actual and expected mutation rates were compared to determine the rate of

mutation in tRNA genes compared to the rest of the genome. Mutation rate for each of the twelve mutations was generated by finding the frequency of each mutation type among all twelve mutation types in tRNA and protein-coding genes in *S. cerevisiae* and *E. coli*. Mutations were represented as they would appear on the coding strand.

Splitting by Expression Level

In *S. cerevisiae*, genes were split by expression level using transcript per million (TPM) values from a previous study (Pelechano et al., 2010). TPM values were log transformed to obtain a normal distribution and split into expression bins based on a histogram model (Appendix, Figure 9). For protein-coding genes, the lowest 10% by TPM value were considered low expression and the highest 10% were considered high expression, with the remaining 80% being considered medium expression.

In *E. coli*, genes were split by expression level by calculating TPM values from previous transcriptomic studies (Larson et al., 2014; Pobre & Arraiano, 2015) using Kallisto (Bray et al., 2016). TPM values were split into expression bins based on histogram models. tRNA genes with multiple copies that could not be confidently assigned as high or low expression for all copies were excluded. Due to a low number of tRNA genes, only genes with low or high expression were kept and genes with mid-level expression were cut from analysis (Appendix, Figure 10). For protein-coding genes, the lowest 10% by TPM value were considered low expression and the highest 10% were considered high expression, with the remaining 80% being considered medium expression.

Program Information

R studio version 4.2.1 was used for all analyses with exception of transcriptomic data analysis to find expression level, which used Kallisto version 0.46. A list of R packages used can be found in Appendix Table 4.

Data Availability

All datasets used for analysis can be found in their respective papers. Sources for both data and reference genome are listed in Appendix Table 1. The code used for analysis can be found at https://github.com/jfgout/RNAPIII-mutation-rate.

REFERENCES

Adebali, O., Sancar, A., & Selby, C. P. (2017). Mfd translocase is necessary and sufficient for transcription-coupled repair in *Escherichia coli*. *The Journal of Biological Chemistry*, *292*(45), 18386–18391. https://doi.org/10.1074/jbc.C117.818807

Assaf, Z. J., Tilk, S., Park, J., Siegal, M. L., & Petrov, D. A. (2017). Deep sequencing of natural and experimental populations of *Drosophila melanogaster* reveals biases in the spectrum of new mutations. *Genome Research*, *27*(12), 1988–2000. https://doi.org/10.1101/gr.219956.116

Bhagwat, A. S., Hao, W., Townes, J. P., Lee, H., Tang, H., & Foster, P. L. (2016). Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, *113*(8), 2176–2181. https://doi.org/10.1073/pnas.1522325113

Bohr, V. A., Stevnsner, T., & de Souza-Pinto, N. C. (2002). Mitochondrial DNA repair of oxidative damage in mammalian cells. *Gene*, *286*(1), 127–134. https://doi.org/10.1016/S0378-1119(01)00813-7

Bradsher, J., Auriol, J., de Santis, L. P., Iben, S., Vonesch, J.-L., Grummt, I., & Egly, J.-M. (2002). Csb is a component of RNA pol I transcription. *Molecular Cell*, *10*(4), 819–829. https://doi.org/10.1016/S1097-2765(02)00678-0

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Case-Green, S. C., & Southern, E. M. (1994). Studies on the base pairing properties of deoxyinosine by solid phase hybridisation to oligonucleotides. *Nucleic Acids Research*, *22*(2), 131–136. https://doi.org/10.1093/nar/22.2.131

Chen, X., & Zhang, J. (2014). Yeast mutation accumulation experiment supports elevated mutation rates at highly transcribed sites. *Proceedings of the National Academy of Sciences*, *111*(39). https://doi.org/10.1073/pnas.1412284111

Chen, Y. H., & Bogenhagen, D. F. (1993). Effects of DNA lesions on transcription elongation by T7 RNA polymerase. *Journal of Biological Chemistry*, *268*(8), 5849–5855. https://doi.org/10.1016/S0021-9258(18)53397-4

Clausen, A. R., Zhang, S., Burgers, P. M., Lee, M. Y., & Kunkel, T. A. (2013). Ribonucleotide incorporation, proofreading and bypass by human DNA polymerase δ. *DNA Repair*, *12*(2), 121–127. https://doi.org/10.1016/j.dnarep.2012.11.006

Daniel, L., Cerutti, E., Donnio, L.-M., Nonnekens, J., Carrat, C., Zahova, S., Mari, P.-O., & Giglia-Mari, G. (2018). Mechanistic insights in transcription-coupled nucleotide excision repair of ribosomal DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(29), E6770–E6779. https://doi.org/10.1073/pnas.1716581115

Denver, D. R., Dolan, P. C., Wilhelm, L. J., Sung, W., Lucas-Lledó, J. I., Howe, D. K., Lewis, S. C., Okamoto, K., Thomas, W. K., Lynch, M., & Baer, C. F. (2009). A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proceedings of the National Academy of Sciences*, *106*(38), 16310–16314. https://doi.org/10.1073/pnas.0904895106

Denver, D. R., Wilhelm, L. J., Howe, D. K., Gafner, K., Dolan, P. C., & Baer, C. F. (2012). Variation in base-substitution mutation in experimental and natural lineages of *Caenorhabditis* nematodes. *Genome Biology and Evolution*, *4*(4), 513–522. https://doi.org/10.1093/gbe/evs028

Dillon, M. M., Sung, W., Sebra, R., Lynch, M., & Cooper, V. S. (2017). Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in *Vibrio cholerae* and *Vibrio fischeri*. *Molecular Biology and Evolution*, *34*(1), 93–109. https://doi.org/10.1093/molbev/msw224

Duret, L., & Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution*, *17*(1), 68–070. https://doi.org/10.1093/oxfordjournals.molbev.a026239

Foster, P. L., Lee, H., Popodi, E., Townes, J. P., & Tang, H. (2015). Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. *Proceedings of the National Academy of Sciences*, *112*(44). https://doi.org/10.1073/pnas.1512136112

Foster, P. L., Niccum, B. A., Popodi, E., Townes, J. P., Lee, H., MohammedIsmail, W., & Tang, H. (2018). Determinants of base-pair substitution patterns revealed by whole-genome sequencing of DNA mismatch repair defective *Escherichia coli*. *Genetics*, *209*(4), 1029–1042. https://doi.org/10.1534/genetics.118.301237

Hastings, K. E. M. (1996). Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. *Journal of Molecular Evolution*, *42*(6), 631–640. https://doi.org/10.1007/BF02338796

Hu, J., Lieb, J. D., Sancar, A., & Adar, S. (2016). Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(41), 11507–11512. https://doi.org/10.1073/pnas.1614430113

Jinks-Robertson, S., & Bhagwat, A. S. (2014). Transcription-associated mutagenesis. *Annual Review of Genetics*, *48*(1), 341–359. https://doi.org/10.1146/annurev-genet-120213-092015

Krokan, H. E., Otterlei, M., Nilsen, H., Kavli, B., Skorpen, F., Andersen, S., Skjelbred, C., Akbari, M., Aas, P. A., & Slupphaug, G. (2001). Properties and functions of human uracil-DNA glycosylase from the UNG gene. In *Progress in Nucleic Acid Research and Molecular Biology* (Vol. 68, pp. 365–386). Academic Press. https://doi.org/10.1016/S0079-6603(01)68112-1

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). TimeTree 5: An expanded resource for species divergence times. *Molecular Biology and Evolution*, *39*(8), msac174. https://doi.org/10.1093/molbev/msac174

Larson, M. H., Mooney, R. A., Peters, J. M., Windgassen, T., Nayak, D., Gross, C. A., Block, S. M., Greenleaf, W. J., Landick, R., & Weissman, J. S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics *in vivo*. *Science (New York, N.Y.)*, *344*(6187), 1042–1047. https://doi.org/10.1126/science.1251871

Liu, H., & Zhang, J. (2019). Yeast spontaneous mutation rate and spectrum are environment-dependent. *Current Biology : CB*, *29*(10), 1584-1591.e3. https://doi.org/10.1016/j.cub.2019.03.054

Liu, H., & Zhang, J. (2021). The rate and molecular spectrum of mutation are selectively maintained in yeast. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-24364-6

Long, H., Behringer, M. G., Williams, E., Te, R., & Lynch, M. (2016). Similar mutation rates but highly diverse mutation spectra in ascomycete and basidiomycete yeasts. *Genome Biology and Evolution*, *8*(12), 3815–3821. https://doi.org/10.1093/gbe/evw286

Lucas, M. C., Pryszcz, L. P., Medina, R., Milenkovic, I., Camacho, N., Marchand, V., Motorin, Y., Ribas de Pouplana, L., & Novoa, E. M. (2023). Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. *Nature Biotechnology*, 1–15. https://doi.org/10.1038/s41587-023-01743-6

Lynch, M. (2011). The lower bound to the evolution of mutation rates. *Genome Biology and Evolution*, *3*, 1107–1118. https://doi.org/10.1093/gbe/evr066

Marnett, L. J. (2000). Oxyradicals and DNA damage. *Carcinogenesis*, *21*(3), 361–370. https://doi.org/10.1093/carcin/21.3.361

Mellon, I., & Champe, G. N. (1996). Products of DNA mismatch repair genes mutS and mutL are required for transcription-coupled nucleotide-excision repair of the lactose operon in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(3), 1292–1297. https://doi.org/10.1073/pnas.93.3.1292

Mi, R., Alford-Zappala, M., Kow, Y. W., Cunningham, R. P., & Cao, W. (2012). Human endonuclease V as a repair enzyme for DNA deamination. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *735*(1), 12–18. https://doi.org/10.1016/j.mrfmmm.2012.05.003

Nagai, A., Mori, K., Shiomi, Y., & Yoshihisa, T. (2021). OTTER, a new method for quantifying absolute amounts of tRNAs. *RNA*, *27*(5), 628–640. https://doi.org/10.1261/rna.076489.120

Nakatsu, Y., Asahina, H., Citterio, E., Rademakers, S., Vermeulen, W., Kamiuchi, S., Yeo, J.-P., Khaw, M.-C., Saijo, M., Kodo, N., Matsuda, T., Hoeijmakers, J. H. J., & Tanaka, K. (2000). XAB2, a novel tetratricopeptide repeat protein involved in transcription-coupled DNA repair and transcription. *Journal of Biological Chemistry*, *275*(45), 34931–34937. https://doi.org/10.1074/jbc.M004936200

Nick McElhinny, S. A., Watts, B. E., Kumar, D., Watt, D. L., Lundström, E.-B., Burgers, P. M. J., Johansson, E., Chabes, A., & Kunkel, T. A. (2010). Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases. *Proceedings of the National Academy of Sciences*, *107*(11), 4949–4954. https://doi.org/10.1073/pnas.0914857107

Pál, C., Papp, B., & Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, *158*(2), 927–931. https://doi.org/10.1093/genetics/158.2.927

Pelechano, V., Chávez, S., & Pérez-Ortín, J. E. (2010). A complete set of nascent transcription rates for yeast genes. *PLoS ONE*, *5*(11), e15442. https://doi.org/10.1371/journal.pone.0015442

Pobre, V., & Arraiano, C. M. (2015). Next generation sequencing analysis reveals that the ribonucleases RNase II, RNase R and PNPase affect bacterial motility and biofilm formation in <i>E. coli<\i>. *BMC Genomics*, *16*(1), 72. https://doi.org/10.1186/s12864-015-1237-6

Polak, P., & Arndt, P. F. (2008). Transcription induces strand-specific mutations at the 5′ end of human genes. *Genome Research*, *18*(8), 1216–1223. https://doi.org/10.1101/gr.076570.108

Ramiro, A. R., Stavropoulos, P., Jankovic, M., & Nussenzweig, M. C. (2003). Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. *Nature Immunology*, *4*(5), Article 5. https://doi.org/10.1038/ni920

Saini, N., Roberts, S. A., Sterling, J. F., Malc, E. P., Mieczkowski, P. A., & Gordenin, D. A. (2017). APOBEC3B cytidine deaminase targets the non-transcribed strand of tRNA genes in yeast. *DNA Repair*, *53*, 4–14. https://doi.org/10.1016/j.dnarep.2017.03.003

Sancar, A. (2016). Mechanisms of DNA repair by photolyase and excision nuclease (Nobel Lecture). *Angewandte Chemie (International Ed. in English)*, *55*(30), 8502–8527. https://doi.org/10.1002/anie.201601524

Schrider, D. R., Houle, D., Lynch, M., & Hahn, M. W. (2013). Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*, *194*(4), 937–954. https://doi.org/10.1534/genetics.113.151670

Sharp, N. P., Sandell, L., James, C. G., & Otto, S. P. (2018). The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. *Proceedings of the National Academy of Sciences*, *115*(22). https://doi.org/10.1073/pnas.1801040115

Sohail, A., Klapacz, J., Samaranayake, M., Ullah, A., & Bhagwat, A. S. (2003). Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. *Nucleic Acids Research*, *31*(12), 2990–2994. https://doi.org/10.1093/nar/gkg464

Spivak, G. (2016). Transcription-coupled repair: An update. *Archives of Toxicology*, *90*(11), 2583–2594. https://doi.org/10.1007/s00204-016-1820-x

Sung, W., Ackerman, M. S., Gout, J.-F., Miller, S. F., Williams, E., Foster, P. L., & Lynch, M. (2015). Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments. *Molecular Biology and Evolution*, *32*(7), 1672–1683. https://doi.org/10.1093/molbev/msv055

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G., & Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences*, *109*(45), 18488–18492. https://doi.org/10.1073/pnas.1216223109

Sweder, K. S., Verhage, R. A., Crowley, D. J., Crouse, G. F., Brouwer, J., & Hanawalt, P. C. (1996). Mismatch repair mutants in yeast are not defective in transcription-coupled DNA repair of UV-induced DNA damage. *Genetics*, *143*(3), 1127–1135.

the 1000 Genomes Project. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, *43*(7), 712–714. https://doi.org/10.1038/ng.862

Thornlow, B. P., Hough, J., Roger, J. M., Gong, H., Lowe, T. M., & Corbett-Detig, R. B. (2018). Transfer RNA genes experience exceptionally elevated mutation rates. *Proceedings of the National Academy of Sciences*, *115*(36), 8996–9001. https://doi.org/10.1073/pnas.1801240115

van Oosten, M., Stout, G. J., Backendorf, C., Rebel, H., de Wind, N., Darroudi, F., van Kranen, H. J., de Gruijl, F. R., & Mullenders, L. H. (2005). Mismatch repair protein Msh2 contributes to UVB-induced cell cycle arrest in epidermal and cultured mouse keratinocytes. *DNA Repair*, *4*(1), 81–89. https://doi.org/10.1016/j.dnarep.2004.08.008

Weng, M.-L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M. T., Shaw, R. G., Weigel, D., & Fenster, C. B. (2019). Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics*, *211*(2), 703–714. https://doi.org/10.1534/genetics.118.301721

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., … Nurse, P. (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature*, *415*(6874), Article 6874. https://doi.org/10.1038/nature724

Xiong, K., McEntee, J. P., Porfirio, D. J., & Masel, J. (2017). Drift barriers to quality control when genes are expressed at different levels. *Genetics*, *205*(1), 397–407. https://doi.org/10.1534/genetics.116.192567

Yang, Y., Hu, J., Selby, C. P., Li, W., Yimit, A., Jiang, Y., & Sancar, A. (2019). Single-nucleotide resolution analysis of nucleotide excision repair of ribosomal DNA in humans and mice. *The Journal of Biological Chemistry*, *294*(1), 210–217. https://doi.org/10.1074/jbc.RA118.006121

Zahn, K. E., Wallace, S. S., & Doublié, S. (2011). DNA polymerases provide a canon of strategies for translesion synthesis past oxidatively generated lesions. *Current Opinion in Structural Biology*, *21*(3), 358–369. https://doi.org/10.1016/j.sbi.2011.03.008

Zhang, X., Zhang, X., Zhang, X., Liao, Y., Song, L., Zhang, Q., Li, P., Tian, J., Shao, Y., AI-Dherasi, A. M., Li, Y., Liu, R., Chen, T., Deng, X., Zhang, Y., Lv, D., Zhao, J., Chen, J., & Li, Z. (2018). Spatial vulnerabilities of the *Escherichia coli* genome to spontaneous mutations revealed with improved duplex sequencing. *Genetics*, *210*(2), 547–558. https://doi.org/10.1534/genetics.118.301345

Zhu, Y. O., Siegal, M. L., Hall, D. W., & Petrov, D. A. (2014a). Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences*, *111*(22). https://doi.org/10.1073/pnas.1323011111

Zhu, Y. O., Siegal, M. L., Hall, D. W., & Petrov, D. A. (2014b). Reply to Chen and Zhang: On interpreting genome-wide trends from yeast mutation accumulation data. *Proceedings of the National Academy of Sciences*, *111*(39). https://doi.org/10.1073/pnas.1413861111

APPENDIX A

APPENDIX: INVESTIGATING THE IMPACT OF TRANSCRIPTION ON MUTATION RATES

Supplementary Tables

Table 1 Reference genomes used by species and study.

Species	Study	Reference Genome
S. cerevisiae	Zhu et al., 2014	GSM4008796
	Sharp et al., 2018	
	Liu & Zhang, 2019	
	Liu & Zhang, 2021	
E. coli, ATCC 8739 Strain	Zhang et al., 2018	CP000946.1
E. coli, MG1655 Strain	Foster et al., 2015	NC_000913.2, NC_000913.3, U00096.2, & U00096.3
	Foster et al., 2018	
A. thaliana	Weng et al., 2019	GCA_000001735.1
B. subtilis	Sung et al., 2015	CM000488.1
C. briggsae	Denver et al., 2012	GCF_000004555.2
C. elegans	Denver et al., 2009	GCF_000002985.6
D. melanogaster	Schrider et al., 2013	GCF_000001215.4
	Assaf et al., 2017	
M. florum	Sung et al., 2012	GCF_000008305.1
R. toruloides	Long et al., 2016	GCA_000320785.2
V. cholerae	Dillon et al., 2017	GCF_001683415.1
V. fischeri	Dillon et al., 2017	GCF_000011805.1

The NC_000913.2 & U00096.2 and the NC_000913.3 & U00096.3 genomes are, for the purposes of this experiment, the same. Each was used interchangeably throughout the project

Table 2 GC content contribution to expected number of mutations in tRNA genes.

Species	Study	Number of Mutations Observed	Number of Mutations Expected With GC Correction	Number of Mutations Expected Without GC Correction
S. cerevisiae	Sharp et al., 2018	34	4.4538	3.8523
	Liu & Zhang, 2019	19	3.1396	2.7522
	Zhu et al., 2014	16	1.9231	1.6896
	Liu & Zhang, 2021	43	25.1155	21.1051
E. coli, ATCC 8739 Strain	Zhang et al., 2018	338	180.2567	169.5089
E. coli, MG1655 Strain	Foster et al., 2015	49	16.0741	15.3481
	Foster et al., 2018	33	21.7208	23.5471

GC content does not seem to have a significant effect on the number of mutations expected to be seen in tRNA genes.

Table 3 tRNA mutations observed in eukaryotic species not included in final results.

Species	Number of Mutations Observed	Number of Mutations Expected with GC Correction	Number of Mutations Expected Without GC Correction
A. thaliana	3	0.1379	0.1385
B. subtilis	1	0.5624	0.5521
C. briggsae	1	0.1426	0.1117
C. elegans	2	0.1693	0.1436
D. melanogaster	3	0.7292	0.6278
M. florum	2	2.9079	1.5596
R. toruloides	3	0.1379	0.1385
V. cholerae	3	0.9940	0.9641
V. fischeri	8	5.6907	5.5494

While there are not enough mutations in these datasets to obtain statistically significant data, it is worth noting that all species, with the exception of *M. florum*, show a significantly higher number of mutations in tRNA genes than expected.

Table 4 R packages used for analysis of data & figure creation.

Package Name	Version Number
BioStrings	2.64.0
BSgenome	1.64.0
doParallel	1.0.17
dplyr	1.0.9
foreach	1.5.1
genomation	1.28.0
GenomicRanges	1.48.0
ggpattern	0.4.3-3
ggplot2	3.3.6
rBLAST	0.99.2
readxl	1.4.0
reshape2	1.4.4
stringr	1.4.0
tidyr	1.2.0

Supplementary Figures

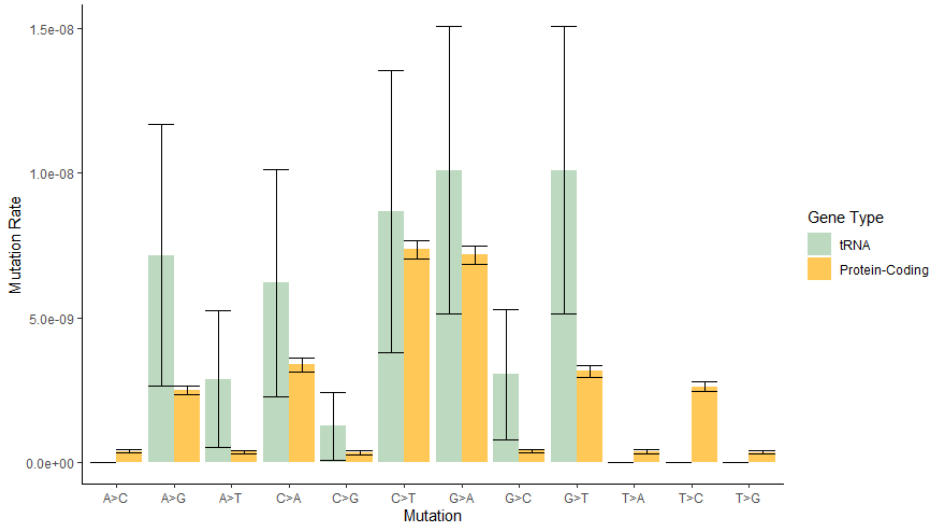

Figure 8 Mutation spectrum of *S. cerevisiae msh2*-knockout by gene type.

Mutation rates between tRNA and protein-coding genes are much closer than when *MSH2* is functional. tRNA mutation rates are still higher than those of protein-coding genes.

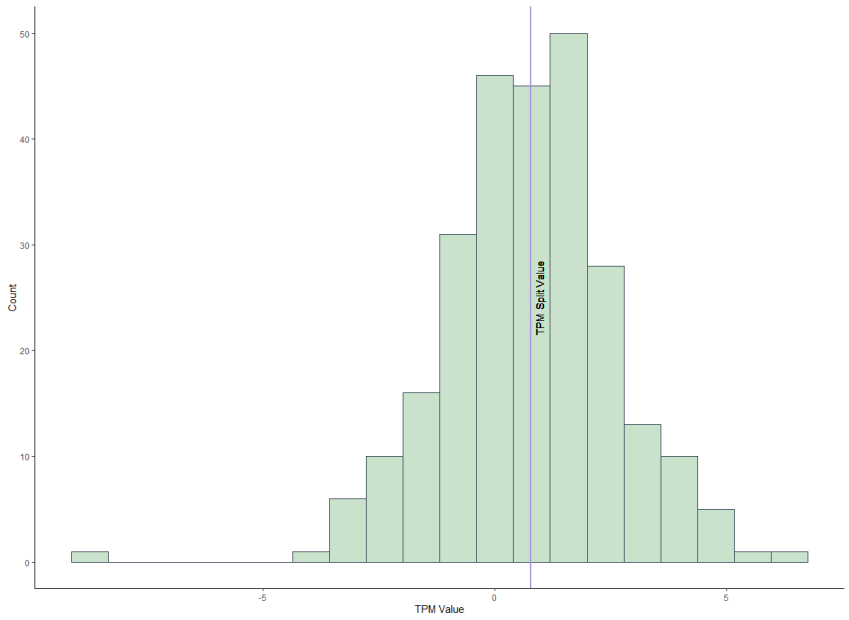

Figure 9 Distribution of tRNA gene expression level in *S. cerevisiae*

Distribution of tRNA genes in *S. cerevisiae* with marking indicating cutoff for high and low log transformed transcript per million (TPM) values. TPM cutoff value was chosen based on the data available. Split value was 0.7760.

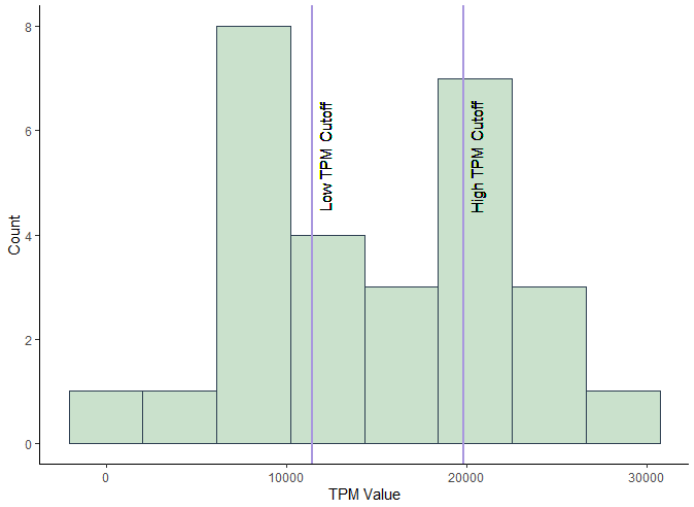

Figure 10 Distribution of tRNA gene expression level in *E. coli*

Distribution of tRNA genes in *E. coli* with markings indicating cutoff for high and low transcript per million (TPM) values. TPM cutoff values were chosen based on the data available. Low TPM cutoff was 11,427.6 and high TPM cutoff was 19,858.2.