Title
Transfer RNA genes experience exceptionally elevated mutation rates.

Permalink
https://escholarship.org/uc/item/109529nv

Journal
Proceedings of the National Academy of Sciences of the United States of America, 115(36)

ISSN
1091-6490

Authors
Thornlow, Bryan P
Hough, Josh
Roger, Jacquelyn M
et al.

Publication Date
2018-09-01

DOI
10.1073/pnas.1801240115

Peer reviewed
Transfer RNA genes experience exceptionally elevated mutation rates

Bryan P. Thornlow, Josh Hough, Jacquelyn M. Roger, Henry Gong, Todd M. Lowe, and Russell B. Corbett-Detig

*Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064; and Genomics Institute, University of California, Santa Cruz, CA 95064

Edited by Andrew G. Clark, Cornell University, Ithaca, NY, and approved July 30, 2018 (received for review February 5, 2018)

Transfer RNAs (tRNAs) are a central component for the biological synthesis of proteins, and they are among the most highly conserved and frequently transcribed genes in all living things. Despite their clear significance for fundamental cellular processes, the forces governing tRNA evolution are poorly understood. We present evidence that transcription-associated mutagenesis and strong purifying selection are key determinants of patterns of sequence variation within and surrounding tRNA genes in humans and diverse model organisms. Remarkably, the mutation rate at broadly expressed cytosolic tRNA loci is likely between 7 and 10 times greater than the nuclear genome average. Furthermore, evolutionary analyses provide strong evidence that tRNA genes, but not their flanking sequences, experience strong purifying selection acting against this elevated mutation rate. We also find a strong correlation between tRNA expression levels and the mutation rates in their immediate flanking regions, suggesting a simple method for estimating individual tRNA gene activity. Collectively, this study illuminates the extreme competing forces in tRNA gene evolution and indicates that mutations at tRNA loci contribute disproportionately to mutational load and have unexplored fitness consequences in human populations.

Transfer RNAs (tRNAs) are essential to protein synthesis across all of life. Their primary function is in translation of the genetic code into the corresponding amino acid sequences that make up proteins. Thus, tRNA molecules are critical for virtually all cellular processes, and the genes encoding tRNA molecules have been highly conserved over evolutionary time (1, 2). Mitochondrial tRNAs have been the subject of many studies, as mutations in these genes lead to a large number of maternally inherited genetic diseases (3). However, eukaryotic genomes contain ~10- to 20-fold as many tRNA genes encoded in their nuclear chromosomes, which are required for cytosolic protein translation (2, 4). Despite their importance to the cell, there has been little study of evolutionary conservation or pathogenic mutations in cytosolic tRNA genes (5, 6). tRNAs are required in exceptionally large quantities, and therefore tRNA genes may experience greater levels of transcription than even the most highly transcribed protein-coding genes (7, 8). In turn, this may lead to high levels of transcription-associated mutagenesis (TAM). As the largest, most ubiquitous RNA gene family, cytosolic tRNAs constitute an ideal gene set for studying the interplay between natural selection and elevated mutation rates.

Transcription affects the mutation rates of transcribed genes (9) through the unwinding and separation of cDNA strands (10). During transcription, a nascent RNA strand forms a hybrid DNA–RNA complex with a template DNA strand. While the complementary tract of non template DNA is temporarily isolated, it is chemically reactive and thus accessible by potential mutagens (10). Transcription can lead to the formation of noncanonical DNA structures, which can hinder repair pathways and promote errors by the polymerase (11). The RNA strand can also reamplify to the template DNA strand, prolonging isolation and increasing vulnerability to mutations (12, 13). Furthermore, if transcription and DNA replication occur concomitantly at a particular locus, collisions between RNA polymerase and the DNA replication fork may also damage DNA (9, 11, 14). In human cancer cells, increased transcription and replication induce torsional stress and collisions (11).

Several cellular agents have also been shown to cause damage in highly expressed genes (15). Among the most notable sources of mutation associated with transcription is activation-induced cytidine deaminase (AID) (16). AID accompanies RNA polymerase II and deaminates cytosine nucleotides. To resolve the resulting base-pair mismatch, the opposing guanine is converted to adenine and uracil to thymine, resulting in excess C→T mutations on the non template strand and excess G→A mutations on the template strand (9, 17). AID is a member of the APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like) gene family, many of which are involved in double-stranded break repair in transcription (9). Some members of the APOBEC family act strongly at short genes, suggesting increased activity at tRNA loci (18, 19). For example, APOBEC3B causes 1,000-fold more DNA damage at tRNA loci than at other genomic regions in yeast (19). AID also acts on highly transcribed genes in immune B cells, causing transition mutations and double-stranded breaks (9). Due to the strong association of the APOBEC family with transcription, relative excesses of C→T and G→A mutations are a signature of TAM (9).

To conserve mature tRNA sequence identity in the presence of an elevated mutation rate, tRNA genes should experience strong purifying selection. tRNA transcription requires sequence-specific binding of transcription factors to the internal box A and box B

Significance

While transcription-associated mutagenesis (TAM) has been demonstrated for protein-coding genes, its implications in shaping genome structure at transfer RNA (tRNA) loci in metazoans have not been fully appreciated. We show that cytosolic tRNAs are a striking example of TAM because of their variable rates of transcription, well-defined boundaries, and internal promoter sequences. tRNA loci have a mutation rate approximately 7- to 10-fold greater than the genome-wide average, and these mutations are consistent with signatures of TAM. These observations indicate that tRNA loci are disproportionately large contributors to mutational load in the human genome. Furthermore, the correlations between tRNA locus variation and transcription indicate that prediction of tRNA gene expression based on sequence variation data is possible.
promoter elements (20). Once transcribed, precursor tRNAs must fold properly to undergo maturation, which can be disrupted by sequence-altering mutations. The unique structure of tRNAs dictates processing by RNases, addition of modifications, accurate recognition by aminoacyl tRNA synthetases, incorporation into the translating ribosome, and accurate positioning of the anticodon relative to mRNA codons (21, 22). Because of the need to maintain sequence specificity, DNAs encoding the mature portions of tRNAs are well conserved (21). Therefore, we expect that a large proportion of mutations arising in tRNA genes will be deleterious and will quickly be purged by natural selection.

While most human tRNA genes do not have external promoters (20, 21), tRNA transcripts include leader and trailer sequences extending roughly two to five nucleotides upstream and 5–15 nucleotides downstream of the annotated mature tRNA gene, based on the position of the genomically encoded poly(T) transcription termination sequence. Aside from the termination sequence, these flanking sequences appear to have limited sequence-specific functionality in most cases (23–26). Very early in maturation, all tRNA flanking sequences are removed by RNase P (22–24) and RNase Z (22, 27). Because these flanking genomic sequences are frequently unwound and therefore vulnerable to TAM, we expect that these regions will experience mutation rates similar to those of tRNAs. Whereas tRNA genes should experience purifying selection, the flanking regions should be neutral or under weak selection. Here we investigate the patterns of conservation, divergence, and within-species variation of cytosolic tRNAs in humans and other model organisms to elucidate the forces shaping the evolution of this essential RNA gene family.

Results and Discussion

Flanking Regions of tRNA Genes Are Highly Variable Despite Strong Conservation of Mature tRNA Sequences. To estimate evolutionary conservation, we examined PhyloP, which measures the conservation of each human genomic position across 100 vertebrate species (28), by position within each tRNA locus (Methods). Positive PhyloP scores indicate strong conservation, and negative scores indicate accelerated evolution. To study the effects of evolution on a shorter timescale, we also estimated sequence divergence between human and Macaca mulatta at each tRNA locus. Mature tRNA sequences are highly conserved across all positions, based on both average PhyloP score (Fig. L4 and Dataset S1) (28) and M. mulatta alignment (Fig. 1B). However, the inner 5′ flanking region (20 bases upstream of the tRNA; see Methods) is roughly four times more divergent than the untranscribed reference regions. We also found increased rates of divergence in the inner 3′ flanking region, which is roughly three times more divergent than the reference regions (Fig. 1B). Both the outer 5′ flank (21–40 bases upstream of the tRNA) and the outer 3′ flank (11–40 bases downstream of the tRNA) are also roughly 1.5 times more divergent than the reference regions. For tRNAs that contain introns (2), we find that intronic variation correlates with flanking variation (SI Appendix, Fig. S1). Furthermore, intergenic regions within clusters of active tRNAs show similar patterns in their PhyloP scores (SI Appendix, Fig. S2).

We also studied population-level variation at low-frequency SNPs (minor allele frequency <0.05) for each tRNA locus. Low-frequency SNPs are evolutionarily young and are less affected by selection (29). Consistent with our divergence analyses, we find that low-frequency SNPs are more common across both the tRNA gene sequence and flanking regions than in untranscribed reference regions (Fig. 1C). Although the inner flanking regions are most polymorphic, the mature tRNA sequences have about twice as many low-frequency SNPs as reference regions. Overall, our results are consistent on multiple timescales, indicating that tRNAs and flanking sequences are prone to mutation. Indeed, of the 247 sites in the genome that have the lowest possible PhyloP scores, −20 (28, 30), 14 are 10–15 bases upstream of the start of an active tRNA gene, indicating disproportionate enrichment (hypergeometric test, P < 1.6e-48) and that tRNA flanking regions are among the least conserved in the genome. Nonetheless, mature tRNA gene sequences are strongly conserved by purifying selection, which purges mutations.

Transcription Is Correlated with Variation in tRNA and Flanking Regions. We hypothesized that, if transcription-associated mutagenesis drives variation among tRNA loci, highly active tRNA genes would show the greatest mutation rates. Because tRNA transcript abundance measures are often not attributable to individual loci due to identical gene copies and difficulty sequencing full-length tRNAs, we estimated relative transcriptional activity based on chromatin state data from the Epigenomic Roadmap Project (31). Based on these data, we classified human tRNA genes as “active” if they are located in expressed regions in several cell lines and otherwise as “inactive” (Methods and Fig. 2). In some cases, multiple cell lines correspond to a single tissue or organ, so tissue-specific tRNAs (e.g., the brain-specific arginine tRNA in mouse [6]) are considered active. We find that active tRNA genes are significantly more conserved than inactive tRNA loci (Mann–Whitney U test, P < 8.4e-53), and the flanking regions of active tRNAs are significantly more divergent than the flanking regions of inactive tRNAs (P < 7.9e-41). The peak measure of divergence between human and M. mulatta tRNA genes in the inner 5′ flanking regions is roughly five times greater in active tRNAs than in inactive tRNAs (Fig. 2E and F). Active tRNAs in human populations also have significantly more low-frequency SNPs per site than inactive tRNAs across the entire locus, including the tRNA and flanking regions (P < 3.7e-36) (SI Appendix, Fig. S3). Inactive tRNAs are still significantly more conserved (P < 2.02e-12) and polymorphic (P < 0.007) than the untranscribed reference regions, and their flanks are significantly more divergent than the reference regions (P < 1.36e-16).

That the peak in both divergence and polymorphism in all species is consistently 12–15 nucleotides upstream of the mature tRNA sequence is curious. At the most divergent position, 55% of all tRNA loci differ between human and M. mulatta, and 15% of human tRNA loci have a low-frequency SNP (Fig. 1). Furthermore,
virtually all active tRNA loci differ at this nucleotide between human and M. mulatta, and 25% have a low-frequency SNP at this site (SI Appendix, Fig. S3B). This implies that this region either does not face uniform selective pressures or is not uniformly vulnerable to TAM. While distant flanking sequences can affect tRNA expression in yeast (32), few studies have shown that flanking regions affect expression in higher eukaryotes (33). Transcription initiation is long relative to elongation (34, 35), which may lead to prolonged isolation of the non-template DNA strand at the initiation site and increased vulnerability to TAM. A poised initiation complex might also increase the likelihood of collisions between Pol3 and the replication fork (14). Thus, frequent initiation at highly transcribed tRNA loci may contribute to the nonuniform pattern of variation.

This may also explain the increased variation in the outer 3′ flank relative to the outer 5′ flank, as positioning of downstream transcription termination sites varies among tRNA genes (2, 36), whereas transcription start site positions are more consistent. While most tRNAs do not have clear TATA boxes, the TATA-binding protein (TBP) still binds to the DNA duplex ≈25 nucleotides upstream of the tRNA (37), which coincides with a decrease in variability. Furthermore, while both flanking regions for many other Pol3-transcribed genes are divergent, the 5′ flanking regions are generally more divergent than the 3′ flanking regions, suggesting that the underlying mechanism is not tRNA-specific (Dataset S1). However, additional studies are necessary to support the assertion that this pattern is due to transcription.

Two orthogonal analyses strengthen the observed correlations between gene expression and variation at tRNA loci. First, we find a significant correlation between the TBP intensity peaks (38–40) and conservation of the mature tRNA sequence (Spearman’s r = 0.64, P < 2.2e−16) across all human tRNAs and the opposite relationship in the flanking regions (Spearman’s r = −0.64, P < 2.2e−16) (Fig. 2). TBP ChIP-seq data directly reflect transcriptional activity for each locus, as its occupancy is significantly correlated with and required for transcription (20, 41–45). Second, mature tRNA sequence read counts are strongly correlated with tRNA conservation (Spearman’s r = 0.18, P < 0.001) and flanking region divergence (Spearman’s r = −0.61, P < 2.2e−16) (Fig. 2 and SI Appendix, Fig. S4). These read counts were collected from a single HEK cell line by Zheng et al. (46) using DM-tRNA-seq, a specialized tRNA-sequencing method. These correlations are consistent with the idea that more highly transcribed tRNAs vary more in their flanking regions.

**Patterns of Divergence and Conservation Can Be Leveraged to Predict tRNA Gene Expression.** Regardless of how tRNA expression is measured, we find highly significant correlations between gene expression and tRNA sequence conservation. The consistency of these correlations across methods of measurement and across species indicates that it may be possible to predict relative tRNA with DNA sequence conservation patterns and other correlates of tRNA transcriptional activity (e.g., tRNAscan-SE bit scores). Indeed, active and inactive tRNAs are largely distinguishable using only flank and gene PhyloP data (SI Appendix, Fig. S5). As sequencing technology becomes more accessible, predicting tRNA gene expression levels through analysis of DNA data is enticing. Such a model could make future tRNA gene annotation more detailed and cost-effective.

**Variation Patterns Observed at tRNAs Are Not Observed in Most Other Gene Families.** Applicability of this proposed tool is likely best suited for tRNAs, other Pol3 genes, and unique classes of highly expressed protein-coding genes such as histones. Among the histone protein-coding genes less than 1,000 nucleotides in length, the average PhyloP score per nucleotide across the coding sequence and flanking regions is 3.449 and −2.052, respectively, comparable to tRNA loci (SI Appendix, Fig. S6). In contrast, most genes transcribed by RNA Pol2 do not appear to be good targets (Dataset S1). For example, ribosomal proteins are very highly transcribed (47) and have well-conserved exons, but their introns and flanking regions are not as divergent as tRNA flanking regions (28, 48). tRNAs are likely ideal for studying TAM because they have predictable transcript start and end sites, internal promoters, and high transcription rates.

**Patterns of Low-Frequency SNPs Are Consistent with TAM.** In TAM, repair pathways activated in response to deaminations lead to excess conversions between guanine and adenine and between thymine and cytosine on the coding strand (9, 17). Across all tRNA loci, we found that the most common low-frequency SNPs are C→T and G→A and that these mutations are significantly more common in both tRNA genes and flanking regions than in untranscribed reference regions (Fisher’s exact test, P < 0.05 for all comparisons) (Fig. 3). Removal
The estimated DFE indicates that high proportions of deleterious mutations in tRNAs are under strong selection. (A) Estimated DFE of new deleterious mutations for tRNA genes and inner 3′ flanking regions shown in human, mouse, A. thaliana, and D. melanogaster. Proportions of deleterious mutations are shown for each bin of purifying selection strength, estimated on a scale of NeS. Species are arranged by increasing effective population size. (B–D) Low-frequency SNPs plotted as a function of NeS for mouse (B), A. thaliana (C), and D. melanogaster (D).
may have different functions, and this may affect strength of selection at each locus as well. Indeed, a significantly greater proportion of sites are invariant (Fisher’s exact test, \( P < 7.50e-5 \)) and fewer sites are divergent (\( P < 3.85e-8 \)) in active single-copy human tRNA genes than in active multicopy human tRNA genes. We observe the same patterns in the inner 5' \( (P < 5.87e-5; P < 0.025 \) ) and inner 3' \( (P < 8.90e-5; P < 4.04e-4 \) ) flanks of active tRNA genes, suggesting increased transcription of active multicopy tRNA genes. However, few SNP data are available for multicopy tRNAs compared with single-copy tRNAs, limiting our ability to identify consistent differences among tRNA subgroups.

### tRNA Loci Contribute Disproportionately to Mutational Load

Our discovery of a highly elevated mutation rate at tRNA loci suggests that tRNA genes may contribute disproportionately to mutational load, the reduction in individual fitness due to deleterious mutations (55, 56). To estimate the relative mutation rates at active tRNA loci, we calculated the average ratios of \( \theta' \), the proportion of mutations and SNP density in the flanking regions instead of the tRNAs on the coding strand and \( \theta'' \) in genomic regions annotated as state 5 or 7 in at least 4 of the 127 tissues analyzed are active tRNAs, and we consider the remaining 254 tRNAs to be inactive. To classify mouse tRNAs, we used a 15-state Hidden Markov Model based on Chip data in which states 5 and 7 corresponded to regions near active promoters (50). We considered the 272 tRNAs in genomic regions annotated as state 5 or 7 in at least 3% of tissues as active and the remaining 188 tRNAs as inactive.

### Aligning tRNAs

We aligned all tRNAs across all species using covariance models (60) and assigned coordinates to each position in each tRNA and flank based on the Sprinzl numbering system (61). We averaged the PhyloP, divergence, and low-frequency SNP data for all sites assigned to the same Sprinzl coordinate for their respective tRNA loci. Because some tRNAs have variations in structure (2), this alignment was necessary for position-wise comparisons between tRNAs. We filtered tRNAs with fewer than 50 aligned bases from our analysis. We divided the genome into conserved and divergent regions with a phastCons log odds score greater than 0; ref. 28 was present 4–10 bases upstream or downstream of a tRNA, the tRNA was excluded from our analyses, as these regions might contribute to the secondary structure of mature tRNAs and be subject to anomalous levels of selection. We also excluded nuclear-encoded mitochondrial tRNA genes.

### Parsing Variation Data

We analyzed human variation data from the African superpopulation of 661 humans from phase 3 of the 1000 Genomes Project (62). We acquired D. melanogaster variation data for the Siavonga, Zambia populations from the Drosophila Genome Nexus Database (58, 59). We obtained M. musculus and A. thaliana data from Waterston et al. (63) and the Arabidopsis Genome Initiative (64), respectively. All nonhuman data were aligned and genotypes curated as described in ref. 65.

Within each gene, flank, or reference region, we considered positions with minor allele frequencies between 0 and 0.05 to be low-frequency SNPs. We also determined the frequency each class of mutations (e.g., A→G) within each region of each tRNA locus where the identity of each base is defined according to the coding strand sequence. We found the frequency of divergences and low-frequency SNPs by position across all tRNAs and flanking regions. For conservation studies across multiple species, we used the PhyloP track (28) from the University of California, Santa Cruz (UCSC) Genome Browser (48) and calculated the average score for each position within the tRNAs and flanking regions. No PhyloP data were available for A. thaliana (28). For direct comparisons between the species of interest and an outgroup, we used the Multiz track from the UCSC Table Browser (66) and the Stitch MAfs tool from Galaxy (67) to create sequence alignments. Details are available in SI Appendix.

### Correlating Variation to Cell-Line Read Counts

Zheng et al. (46) used demethylation sequencing to detect tRNAs within HEK293T cells (46, 68). We used Spearman’s rank correlation tests to correlate mature tRNA transcript read counts and tRNA and flanking region conservation. Because Zheng et al. (46) sequenced mature tRNAs, which are often encoded by multiple genes, we excluded identical genes to control for the correlation between gene copy number and overall expression (Fig. 2 C and D and refs. 32 and 46).
Separately, we summed the average PhyloP scores at these loci and corre-
lated the summed scores to total tRNA read counts (SI Appendix, Fig. 54).

Estimating the Distribution of Fitness Effects. We estimated the DFE for each
species using the method of Keightley et al. (69) and the DFE- 
for details.

Estimating the Mutation Rate in Active tRNA Genes. We used the equation
\( \theta = k G^2 \) (defined in ref. 29) to estimate the mutation rate at active tRNA
loci. We calculated the ratios of \( \theta \) in active tRNA flanking regions to \( k \) in the
reference regions for \( k = 1,2,3 \) and bootstrapped by tRNA loci to calculate
95% CIs. See SI Appendix for more details.

ACKNOWLEDGMENTS. We thank Craig Mello for helpful discussions and
input on this project, Brian Lin for the Fig. 1 legend, Andrew Holmes for
assisting with mouse data analysis, and the R.B.C.-D. and T.M.L. laboratories
for suggestions and feedback. This work was supported by National Human
Genome Research Institute/NIH Grant 2R01HG006753-04A1 (to T.M.L.) and
National Institute of General Medical Sciences/NIH Grant R35GM128932-01 (to R.B.C.-D.). B.P.T. was funded by NIH Training Grant T32 HG008345.