# Transfer RNA genes experience exceptionally elevated mutation rates

**Bryan Thornlow[a], Josh Hough[a], Jackie Roger[a], Henry Gong[a], Todd Lowe[a,b,1], and Russell Corbett-Detig[a,b,1]**

[a]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz CA 95064; [b]Genomics Institute, University of California Santa Cruz, Santa Cruz CA 95064

[1]To whom correspondence should be addressed. E-mail: lowe@soe.ucsc.edu or russcd@gmail.com

This manuscript was compiled on December 6, 2017

H.G. performed initial analyses. B.T. performed research and analyzed data, with assistance from J.R.. J.H. performed the DFE analyses. B.T., T.L. and R.C.-D. conceived and designed research. B.T., T.L. and R.C.-D. wrote the paper. J.H., J.R., H.G., T.L. and R.C.-D. edited the paper.

The authors declare no conflict of interest.

**Transfer RNAs (tRNAs) are a central and necessary component for the biological synthesis of new proteins, and they are among the most highly conserved and most frequently transcribed sequences across all of life. Despite their clear significance for fundamental cellular processes, however, the forces governing tRNA evolution are poorly understood. Here, we present evidence that transcription-associated mutagenesis and strong purifying selection are key determinants of patterns of sequence polymorphism and divergence within and surrounding tRNA genes across several diverse model organisms. Remarkably, our results indicate that the mutation rate at broadly expressed tRNA loci is between 8.7 and 13.8 times greater than the genome-wide average. Furthermore, evolutionary analyses provide strong evidence that tRNA loci, but not their flanking sequences, experience strong purifying selection, acting in direct response to this elevated mutation rate. Finally, we also find a highly significant correlation between tRNA expression levels and the mutation rates in their immediate flanking regions, suggesting the possibility of predicting gene expression levels based on relative mutation rates and sequence variation data among tRNA gene loci. Our results provide novel insight into individual tRNA gene evolution, and imply that tRNA loci contribute disproportionately to mutational load in human populations.**

## Significance Statement

tRNAs are essential for the production of all proteins in all tissues across life and are therefore among the most highly transcribed loci in the genome. Our study shows that the frequent transcription of tRNAs results in a highly elevated mutation rate at tRNA loci that is between 8- and 14-fold higher for tRNAs than for the rest of the genome. We also show that the strength of natural selection, which acts to remove sequence-altering mutations, is extremely strong in tRNAs, but is relaxed in introns and regions flanking tRNAs. Finally, our results indicate that mutation rates in non-functional tRNA flanking regions are similarly elevated, and levels of genetic variation correlate strongly with expression. These observations suggest that a predictive model could facilitate future studies of tRNA function.

Transfer RNAs (tRNAs) play an essential role in protein synthesis across all of life. Their

primary function is in the 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). The necessity of tRNAs in large quantities also makes them among the most highly transcribed loci in the genome. Indeed, many tRNA genes may experience greater levels of transcription than even the most highly transcribed protein-coding genes (2, 3). Such high levels of transcription suggest that tRNA genes may experience high levels of transcription-associated mutagenesis (TAM) compared to the rest of the genome, making the tRNA gene family an excellent model system for studying the interplay between natural selection and elevated mutation rates.

tRNA | transcription | mutagenesis | TAM

Transcription affects the mutation rates of transcribed genes (4) through the unwinding and separation of complementary DNA strands (5). In particular, 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 (5). In addition, if transcription and DNA replication occur concomitantly at a particular locus, collisions between RNA Polymerase and the DNA replication fork are possible, which may also result in damage to DNA (6). Several cellular agents have also been shown to cause damage in highly expressed genes (7). Among the most notable sources of mutation associated with high transcription is activation-induced cytidine deaminase (AID) (8). AID accompanies RNA Polymerase II and converts cytosine to uracil, causing a relative excess of cytosine to thymine substitutions in the non-template strand and guanine to adenine substitutions on the template strand (9). More highly transcribed genes are especially vulnerable to mutation by AID, making it a clear diagnostic signature of TAM (4). Because tRNA loci are so often unwound for transcription, these regions are therefore expected to experience elevated mutation rates due to TAM, with deamination via AID being one of the primary mutational mechanisms.

In order to conserve mature tRNA sequence identity in the presence of an elevated mutation rate, it is expected that tRNA genes should experience strong selective pressures. At the gene level, tRNA transcription requires sequence-specific binding of transcription factors to the internal A and B box promoter elements (10). Once transcribed, precursor tRNAs must fold properly to undergo a complex process of maturation, which can be disrupted at any step by sequence-altering mutations. The unique structure of tRNAs dictates processing by RNases, addition of an assortment of modifications, accurate recognition by highly specific aminoacyl tRNA synthetases, incorporation into the translating ribosome, and accurate positioning of the anticodon relative to mRNA codons (11, 12). As a consequence of the need to maintain high sequence-specificity, DNA encoding the mature portions of tRNAs are exceptionally well conserved segments of the genome (11). Therefore, we expect that a large proportion of mutations arising in tRNA genes will be deleterious, and therefore experience strong purifying selection.

While most human tRNA genes do not have external promoters (10, 11), tRNA transcripts generally include leader and trailer sequences, extending roughly 2-5 nucleotides upstream of the annotated mature tRNA gene, and 5-15 nucleotides downstream of the mature tRNA sequence,

based on the position of the genomically encoded poly-T transcription termination sequence. However, these sequences have limited functionality in most cases (13–16). For example, very early in the tRNA maturation process, all tRNAs undergo removal of their 5' leader sequences by RNase P (13, 14) and removal of their 3' trailers by RNase Z (17). Because these flanking sequences are frequently unwound and therefore vulnerable to TAM, we expect that tRNA flanking regions will experience similar mutation rates to tRNA genes. Whereas tRNAs should experience purifying selection, we expect that flanking regions should be nearly neutral or under very weak selection.

Despite these clear predictions for an elevated mutation rate of tRNA loci and strong purifying selective pressure on tRNA genes compared to their flanks, there has been no attempt to quantify the overall impact of mutation and selection on patterns of sequence variability in tRNAs or their flanking regions. Here we investigate the patterns of conservation, divergence and within-species variation of tRNAs in humans and several other model organisms (*Mus musculus*, *Arabidopsis thaliana*, and *Drosophila melanogaster*).

## Results & Discussion

**Flanking regions of tRNA genes are highly variable despite strong conservation of mature tRNA sequences.** To estimate evolutionary conservation, we averaged phyloP data, a measure of the conservation of each human genomic position across 100 vertebrate species (18), by position within each tRNA locus (see Methods). To study the effects of evolution on a shorter time-scale, we also aligned the human and *Macacca mulatta* (Rhesus macaque) genomes and counted the non-gap nucleotide mismatches in the alignments as divergent positions for each tRNA locus. Our analyses indicate that mature tRNA sequences are highly conserved at all positions, based on both average phyloP score (18) (Figure 1A, Supplementary Table 1) and *M. mulatta* alignment (Figure 1B). However, the "inner" flanking regions, defined based on inflection points in the data (see Methods), show significant divergence by the same measures. The inner 5' flanking region, defined as the 20 bases upstream of the tRNA, is the most divergent segment of these regions on average, with roughly four times the rate of divergence between human and *M. mulatta* as the untranscribed reference regions (Figure 1B). We found similarly increased rates of divergence in the inner 3' flanking region, which was roughly three times as divergent between human and *M. mulatta* as the untranscribed reference regions (Figure 1B). Both the outer 5' flank (21-40 bases upstream of the tRNA) and outer 3' flank (11-40 bases downstream of the tRNA) are also roughly 1.5 times as divergent as are the untranscribed reference regions. Furthermore, we find that intergenic regions within clusters of active tRNAs (Figures S1A, S1B) show similar patterns in their phyloP scores, with increased divergence extending hundreds to thousands of bases up and downstream of each tRNA gene. Conversely, we find that the intergenic regions in clusters of inactive tRNA genes do not show this pattern nearly as strongly (Figure S1C).

To focus our evolutionary timescale further and eliminate any bias due to multi-species sequence alignment errors, we also studied relative levels of tRNA variation within a human population by observing the occurrence of low-frequency single nucleotide polymorphisms (SNPs) (minor allele frequency < 0.05%) for each tRNA gene locus. Consistent with our phyloP and *M. mulatta* divergence analyses, we find that low-frequency SNPs are more common across the entire tRNA

locus, including the mature sequence and flanking regions, relative to untranscribed reference regions (Figure 1C). Although the inner 5' and inner 3' flanking regions are the most polymorphic, the mature tRNA sequences have about twice as many low-frequency SNPs per site as untranscribed reference regions. Overall, our results are remarkably consistent on multiple timescales (across vertebrates, between primates, within human populations), indicating that functional tRNA sequences are highly conserved across species but prone to mutations at the individual level, and tRNA flanking regions are both more divergent and more polymorphic than untranscribed, non-genic sequences.

**Transcription contributes to variation in tRNA and flanking regions.** We observed that conservation and divergence patterns varied between tRNA loci, and hypothesized that highly active tRNA genes would show the greatest mutation rates if transcription-associated mutagenesis is a primary driver of variation among tRNA loci. Because tRNA transcript abundance measures are often not attributable to individual loci for multiple reasons, including redundant gene copies, variation in pre-tRNA processing and tRNA degradation rates, and difficulty sequencing full-length tRNAs, we estimated relative transcriptional activity based on chromatin state data from the Epigenomic Roadmap Project ((21); Cozen et al., in preparation). We classified human tRNA genes as "active" if they were in regions of active chromatin and near transcription start sites in at least 3% of the 127 tissues for which genome-wide epigenomic data was available (Figure 2A, see Methods). We considered the remaining tRNAs "inactive". No isotypes are over-represented in any of the epigenomic groups (Cozen, et al., in preparation). We found that active tRNAs were significantly more conserved than inactive tRNAs (Mann-Whitney U, $p < 8.40e-53$), and the flanking regions of active tRNAs were significantly more divergent than the flanking regions of inactive tRNAs ($p < 7.98e-61$). Indeed, the peak measure of divergence in the inner 5' flanking regions is roughly five times greater in active tRNAs than in inactive tRNAs (Figure S2A). Active tRNAs also had significantly more low-frequency polymorphisms per site than inactive tRNAs across the entire locus, including the tRNA and flanking regions ($p < 3.72e-36$). Inactive tRNAs were still significantly more conserved ($p < 2.02e-12$) and polymorphic ($p < 0.007$) than the untranscribed reference regions, and their flanks were significantly more divergent than the reference regions ($p < 1.36e-16$).

That the peak by all three measures is consistently about 12 to 13 nucleotides upstream of the mature tRNA sequence is a curious result. At the most divergent position, roughly 55% of all tRNA loci showed a difference between human and *M. mulatta* (Figure 1B) and roughly 15% of tRNA loci have a low-frequency SNP at this site (Figure 1C). Furthermore, among tRNA loci believed to be active, virtually all loci showed a change at this nucleotide between human and *M. mulatta*, and roughly 25% have a low-frequency SNP at this site (Figure S2A, S2B). This implies that this region either does not face uniform selective pressures or is not uniformly vulnerable to TAM. While it has been suggested that distant flanking sequences may affect tRNA expression levels in yeast (22), few studies to our knowledge have shown that the immediate flanking regions have an effect on expression in humans or other higher eukaryotes (23). Importantly, the duration of transcription initiation is long relative to the process of transcription itself (24, 25), which would presumably 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 (6). Thus, frequent initiation at highly transcribed tRNA loci could contribute to the pattern of variation we observe.

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 is highly variable among tRNA genes (19, 26), whereas transcription start site positions are more consistent. Indeed, we find that the TATA boxes for tRNA-SeC-TCA-1-1, RNase P and U6 RNA are all approximately 25 nucleotides upstream of the start of the gene (27). While most tRNAs do not have clear TATA boxes, the TATA-Binding Protein (TBP) still binds non-specifically to the DNA duplex at this position (28), which seems to coincide with the sudden decrease in variability. Furthermore, we find that, 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 (Supplementary Table 1). However, additional studies will be necessary to conclusively support the assertion that this strong mutation pattern is due entirely or in large part to the process of transcription rather than due to a correlated process.

Two additional and orthogonal analyses strengthen the observed correlations between gene expression and variation at tRNA loci. First, we found a significant correlation between the TATA-Binding Protein (TBP) intensity peaks (29–31) (see Methods) and the average level of divergence in the flanking regions (Spearman's rho = -0.64, $p < 2.2e-16$) (Figure 2D), as well as a correlation between the peaks and the average level of conservation of the mature tRNA sequence (Spearman's rho = 0.64, $p < 2.2e-16$) across all human tRNAs (Figure 2C). The TBP peak data for these transcription factors provide an estimate of the level of transcription for each tRNA, and are consistent with the idea that more highly transcribed tRNAs show higher levels of variability in their transcribed regions.

Second, we found significant correlations between the mature tRNA sequence read counts and tRNA conservation (Spearman's rho = 0.18, $p < 0.001$) and flanking region divergence (Spearman's rho = -0.61, $p < 2.2e-16$) when we exclude mature tRNA sequences encoded for by more than one gene (Figure 2E,F), as well as when we sum the average levels of tRNA conservation (Spearman's rho = 0.12, $p < 0.027$) and flanking region divergence (Spearman's rho = -0.68, $p < 2.2e-16$) for genes encoding identical tRNAs to account for correlations between read count and gene copy-number (22, 32) (Figure S3). These read counts were collected from a single human embryonic kidney cell line by Zheng et al (32) using DM-tRNA-seq, a specialized sequencing method developed for tRNAs which overcomes modifications that impede standard small RNA sequencing methods.

**Patterns of divergence and conservation can be leveraged to develop a predictive model for tRNA gene expression.** Regardless of whether tRNA expression is estimated based on epigenetic chromatin marks across many cell types, TBP transcription factor occupancy across multiple cell lines, or by relative transcript abundance within one cell line, we find highly significant correlations between gene expression and tRNA conservation, flanking region divergence, and tRNA locus polymorphism. The consistency of these correlations indicates that it may be possible to predict tRNA expression based solely on DNA sequence conservation patterns. Genome-wide chromatin-IP and ChIP-seq data are resource-intensive to collect. As sequencing technology is rapidly becoming more affordable and accessible, the prospect of making predictions of tRNA gene expression levels through analysis of multiple alignments and variant sites within populations is enticing. Creating and refining such a model would make future tRNA gene annotation significantly easier and cost-effective.

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. For example, we find that among the shortest histone protein coding genes (cutoff less than 1,000 nucleotides in length), the average phyloP score per nucleotide is 3.4485, indicating a comparable level of conservation to tRNA genes. Consistent with tRNA genes, their immediate 5' flanking regions are also more divergent than are their immediate 3' flanking regions, on average. However, most genes transcribed by RNA Pol2, including protein coding genes, lincRNAs, miRNAs, snoRNAs, and others, generally do not appear to be good targets based on analysis of representatives of each. For example, glyceraldehyde-3-phosphate dehydrogenase and ribosomal proteins are very highly and very broadly transcribed (33). These genes have extremely well conserved exons, but their introns and flanking regions are not nearly as divergent as tRNA flanking regions, based on phyloP data (18, 27). It is possible that high intron and flanking region divergence in protein-coding genes is still indicative of a high transcription rate, but the degree of variation in these genes occupies a much smaller range, and would therefore be more difficult to incorporate into a model. Additionally, microRNAs such as miR-21 and miR-25 are highly conserved and highly abundant (27, 33), but they are processed out of longer pri-miRNA transcripts, and do not show highly divergent flanking regions at fixed upstream positions, based on phyloP data (Supplementary Table 1) (18, 27). That tRNAs are the best examples for studying signatures of TAM can be attributed to the combination of several unique characteristics, including consistently predictable transcript start and end sites, internal promoters, and extremely high transcription rates. Other highly transcribed genes have conserved functional elements in their flanking regions that may obscure the effects of TAM at these loci.

**tRNA flanking regions are among the least conserved sites in the human genome.** Upon scanning the human genome for the least conserved base pairs, we found 247 sites in the genome that had scores of -20, the lowest possible score on the phyloP scale (18, 34). Fifteen of these sites were within 20 base pairs of an active tRNA, based on chromatin-IP data. Of these, 14 sites were found in the inner 5' flanking region of the tRNA, between 10 and 15 base pairs upstream of the first base of the mature tRNA sequence. We found that this set of minimum-phyloP-score sites was enriched for sites within tRNA flanking regions (Hypergeometric test, $p < 1.65e\text{-}48$), indicating that the least conserved sites in the genome are disproportionately found in tRNA flanking regions. This indicates that the flanking regions of some active tRNA genes are among the least conserved regions, and perhaps have among the highest mutation rates, of any in the genome.

**Patterns of low-frequency SNPs indicate transcription-associated mutagenesis (TAM).** Prior studies of TAM in protein coding genes indicate that transcription is a mutagenic process, in that the untranscribed strand becomes more vulnerable to damage, either through collisions between the DNA replication fork and RNA Polymerase, or by other molecules such as deaminases (4, 7, 9). It has been shown that repair pathways activated in response to deaminations lead to excess conversions between guanine and adenine and between thymine and cytosine nucleotides on the coding strand (4, 9). To test this prediction, we analyzed the relative frequencies of all low-frequency SNPs for each region of tRNA loci. Across all tRNA loci, we found that the most common low-frequency SNPs are C→T, G→A, T→C and A→G (transitions), and that these mutations are significantly more common in both tRNA flanking regions and the tRNA gene, relative to untranscribed reference regions (Fisher's exact test, $p < 0.05$ for all comparisons)

(Figure 3). Furthermore, the relative excesses of these SNPs are most pronounced in active tRNA loci (Figure S4A). In contrast, and consistent with observed levels of divergence, these relative changes are barely discernible when considering only inactive tRNA loci (Figure S4B).

It is important to note that, due to the necessity of preserving tRNA secondary structure, we would expect transition mutations (e.g., A-U to G-U base pairs, C-G to U-G base pairs) to be more common than transversions, regardless of the underlying mechanism, as they should impair function less often. However, the strong mutational skew expected of regions affected by TAM is even more pronounced in regions flanking tRNAs. While some pre-tRNAs may have extended secondary structure that could influence the relative SNP frequencies, such pre-tRNA 5' leader sequences tend to be a maximum of five nucleotides long in mammals (unpublished observations).

Prior studies have implicated that CpG sites are significantly more prone to mutations than other nucleotides (35). Therefore, to determine whether TAM was the primary cause of these relative excesses, we repeated our analysis after excluding all CpG sites. We found that CpG sites had no effect on the substitution patterns that we observed in the polymorphism data (Figure S5).

**tRNA flanking region variation in other model organisms is consistent with variation observed in humans.** To test whether the patterns of polymorphism and divergence that we observed in tRNAs and flanking regions also occurred in other species, we repeated our analyses for tRNAs in *Mus musculus*, *Drosophila melanogaster* and *Arabidopsis thaliana*. Consistent with our results from human data, we found similar patterns of sequence conservation of tRNA loci across all species investigated (Figure S6). In particular, mature tRNA sequences were highly conserved and the flanking regions were highly divergent (Figures S6A, S6D). Particularly striking are the similarities in the outgroup comparisons in the inner 5' flank (Figures S6B, S6E, S6G). The 5' flanks were more divergent than the 3' flanks and the most divergent sites were roughly 10-15 bases upstream of the tRNA in all species. These results are consistent with our human data (Figure 1) and suggest the possibility that an underlying molecular mechanism drives these convergent patterns of polymorphism and divergence across species.

We also tested whether the correlation between gene expression and variation was conserved across species (Figure S7). To do this, we utilized chromatin-IP data across nine mouse tissues and classified mouse tRNAs based on their breadth of expression. By this measure, active mouse tRNAs were more strongly conserved than their inactive counterparts (Mann-Whitney *U* test, $p < 1.81e-19$), and their flanks were more divergent ($p < 7.04e-22$) (Figure S7A, S7D), consistent with our results from the human data (Figure 2A,B). Active mouse tRNAs also had more low-frequency SNPs in their flanking regions than did inactive mouse tRNAs ($p < 2.23e-4$) (Figures S7C, S7F). Consistent with the human data, inactive mouse tRNAs were also more conserved ($p < 1.76e-8$) and their flanking regions more divergent ($p < 2.37e-4$) than the untranscribed reference regions (Figure S7D). Such consistency indicates that the mechanism underlying these patterns works similarly in human and mouse.

The patterns of low-frequency SNPs are also consistent across all species. The greatest levels of polymorphism are found in the inner 5' flanking regions for all species studied. The frequency

spectra of the low-frequency SNPs also show excess A→G, G→A, C→T and T→C SNPs on the coding strand in all species analyzed (Figure S8) Additionally, as was observed in humans, active mouse tRNAs show a greater excess of these SNPs (Figure S9A) than do inactive mouse tRNAs (Figure S9B). Consistent with our analysis of human tRNAs, these patterns suggest that deamination of the non-coding strand due to TAM and the DNA repair mechanisms acting in response to deamination are especially common at these loci (4, 9, 36).

In humans, we do not have chromatin-IP data for germline tissues and cannot correct for the fact that only mutations in these tissues are heritable, but we have no evidence that active tRNAs are suppressed in the germline. However, the nine mouse tissues for which we had chromatin-IP data included testes and mouse embryonic stem cells. Virtually all of the tRNAs that are inactive in both stem cells and testes are also inactive in the other tissues. Because only germline mutations are heritable, we expect that only germline expression causes elevated mutation rates at these loci. That virtually all active tRNAs are expressed in the germline and that those not expressed in the germline are inactive is consistent with our findings and suggests that estimates of tRNA expression derived from somatic tissues are sufficient for studying the genetic consequences of exceptional transcription rates.

**Functional tRNA sequences experience strong purifying selection in all species studied.** Our analysis of the distribution of fitness effects (DFE) of deleterious mutations demonstrates that tRNAs evolve under strong purifying selection in all of the species we analyzed. In contrast, regions flanking tRNAs were inferred to be either neutral or subject to weak selection ($N_eS < 10$) (Figure 4). These results are consistent with our estimates of evolutionary conservation in tRNA regions, as well as elevated levels of polymorphism observed in the flanks (Figure 1). Our estimates of the proportions of new mutations falling into each $N_eS$ range of the DFE for tRNAs indicated that there were far fewer nearly neutral mutations ($N_eS < 1$) and substantially more strongly deleterious mutations ($N_eS > 100$) in *D. melanogaster* and *A. thaliana* than in human or mouse populations (Figure 4). Given that estimates of $N_e$ in human (7, 000; 37) and mouse (25, 000 − 120, 000; 38) are substantially lower than in *A. thaliana* (300,000; 39) and especially *D. melanogaster* (> 1,000,000; 40), this difference in the inferred strength of selection may be due to differences in long term $N_e$. That the *A. thaliana* life cycle involves selfing may also contribute to these differences. In addition, the human and mouse genomes contain far more tRNAs (610 and 471, respectively) than *D. melanogaster* (295) (19), and this increased redundancy could also affect the inferred fitness effects across these species, as mutations in high copy-number tRNAs are potentially less deleterious than those affecting unique tRNAs. However, given that there are 700 tRNA genes in the *A. thaliana* genome (19, 41), redundancy alone is unlikely to fully account for the between-species differences in the DFE.

Several tRNAs are known to contain introns (19). We analyzed the introns separately and found that intronic variation correlates with flanking variation in tRNAs; that is, tRNAs with the most variable flanks also had the most variable introns (Figure S10). We considered using introns as selectively neutral regions for estimating DFE, but found that these regions comprised only 619 nucleotides in total, fewer than the total number of human tRNAs (19). As such, this sample size was too small to reliably use in our DFE analysis.

**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 segregating mutational load in humans. To obtain an estimate of the contribution of tRNA loci to this load, we used the ratio of the rate of low-frequency SNPs in tRNA flanking regions to that in untranscribed reference regions (between 8.7 and 13.8) to estimate the tRNA mutation rate relative to the neutral mutation rate in humans (1.45e-8, (42)). Given that there are 25,852 base pairs of tRNA sequence for active tRNAs in the human genome, we estimate that the per generation rate of deleterious mutation arising from tRNAs per diploid genome ($U_{tRNA}$) is 0.01. Using previous estimates of the rate of deleterious amino acid mutation per diploid genome (0.35, (43)), this implies that tRNAs may contribute 2.3% of deleterious mutations as protein coding sequences. Given that tRNAs make up only 0.0009% of the human genome (19), this further implies that mutations in tRNAs contribute to mutational load, the reduction in individual fitness due to segregating deleterious mutations (44, 45), with an effect disproportionate to their total lengths. Although such calculations are clearly approximate, they nevertheless highlight that mutations at tRNA loci are likely an important source of fitness and disease variation in human populations.

## Conclusions

Our findings demonstrate the fundamental importance of tRNA sequences, which are highly conserved despite the continual influx of mutations by TAM at a higher rate than anywhere else in the genome. Our results are consistent across a broad range of taxonomically diverse species, indicating that elevated mutation rates due to TAM and strong purifying selection are widespread across life, and may be a good predictor of relative tRNA gene transcription levels. The conflict between extreme TAM and strong purifying selection at tRNA loci is potentially an unappreciated source of genetic disease, and may have a profound impact on the fitness of human populations.

## Materials and Methods

**Defining tRNA loci and flanking regions.** We used tRNA coordinates from GtRNAdb (19) for the human, *M. musculus*, *D. melanogaster*, and *A. thaliana* genomes. For each species, we defined untranscribed reference regions to use as negative controls. To find these regions in the human genome, we searched 10 kilobases upstream of each tRNA and selected a 200-nucleotide tract. If this tract was within a highly transcribed region of the genome (as determined by genome-wide chromatin-IP data (21)), overlapped a conserved element (defined as a region with a phastCons log odds score greater than 0 (18)), was within 1,000 nucleotides of a known gene (27), or overlapped an untranscribed reference region assigned to another tRNA, we selected a new tract 1,000 bases further upstream, and repeated until we found an acceptable region. For the mouse genome, we checked only known genes, previously assigned untranscribed reference regions, and conserved elements, as analogous genome-wide chromatin-IP data of the caliber used for humans was not readily available for other species. For the *D. melanogaster* and *A. thaliana* genomes, we began our searches only 1,000 bases upstream of each tRNA, and searched for 200-nucleotide tracts that were at least 100 nucleotides away from any annotated genetic element (46, 47). This adjustment was made due to the relatively high functional densities of the genomes of these species.

For each tRNA in all species, we defined the inner 5' flank as the 20 bases immediately upstream of the 5' end of the tRNA on the coding strand, and the outer 5' flank as the 20 bases directly upstream of the inner 5' flank. Likewise, the inner 3' flank refers to the 10 bases directly downstream of the tRNA on the coding strand, and the outer 3' flank refers to the 30 bases downstream of these 10 bases. We made these decisions based on

inflection points in our data, as the flanking regions up to 20 bases upstream and 10 bases downstream of tRNA genes seemed to have less variation. Further, while no studies to our knowledge report the length of tRNA leader sequences in eukaryotes in general, we found that transcription usually ends about 10 bases downstream of mature tRNA sequences (26, 48).

**Classifying tRNAs based on breadth of expression.** The Roadmap Epigenomics Consortium compiled genome-wide epigenomic data across 127 human tissues and cell lines in order to characterize the state of chromatin across the genome (21). Cozen, et al. (in preparation) analyzed the regions surrounding each tRNA in each epigenome sample, and performed a clustering analysis to classify each genomic region according to its most common epigenomic state. They then classified all human tRNAs based on the epigenomic state annotation in the genome. In the corresponding model, regions in state 1 are near transcription start sites, and regions in states 4 and 5 are not near transcription start sites but are nonetheless likely to be transcribed. tRNAs in state 1 in at least 3% of tissues are referred to here as “active tRNAs”, and we consider the remaining tRNAs to be “inactive”.

We followed a similar approach to classify mouse tRNAs. We used data from a 15-state Hidden Markov Model based on chromatin-IP data in which states 5 and 7 corresponded to regions proximal to active promoters (49). tRNAs in genomic regions annotated as state 5 or 7 in at least 3% of tissues were considered to be "active", and all other tRNAs were considered "inactive". These classifications were not conducted in other species due to lack of available data.

**Aligning tRNAs.** We aligned all tRNAs across all species using covariance models (41) and assigned coordinates to each position in each tRNA and flank based on the Sprinzl numbering system (20). Using these alignments, we created files assigning a Sprinzl coordinate to each genomic coordinate within tRNA sequences or flanking regions for each species studied. For example, the first nucleotide at the 5’ end of each tRNA was assigned Sprinzl coordinate 1. To create Figures 1 and 2A and B, 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 insertions, deletions and variations in structure (e.g. Leucine tRNAs often have an extended V-loop (19)), this alignment was necessary for position-wise comparisons between tRNAs. Additionally, some low-scoring tRNAs did not align well using these methods, and Sprinzl coordinates could not be properly assigned. We set a filter such that tRNAs with fewer than 50 aligned bases were excluded.

Some tRNAs are known to have extended leading or trailing sequences that are well conserved across species and potentially contribute to the secondary structure of the tRNAs (16). We determined whether any conserved elements (regions with a phastCons log odds score greater than 0 (18)) were present by using the Vertebrate Multiz Alignment & Conservation track in the UCSC Genome Browser (27) for the regions 4-10 bases up or downstream of each human tRNA. If a conserved element was present within this region, the tRNA was excluded from our analyses, as these flanking regions might contribute to the secondary structure of mature tRNAs, and would therefore be subject to higher levels of selection than the vast majority of tRNA flanking regions.

We also excluded nuclear-encoded mitochondrial tRNAs from our analyses. These tRNAs are transferred from mitochondrial genomes and therefore are not subject to the same evolutionary pressures as the vast majority of tRNAs. Additionally, alignments of these tRNAs across species is dubious, as these transfers likely occurred following speciation, and many of these genes are without true orthologs in other species. Therefore, excluding these genes would better explain the patterns of mutation and selection affecting most tRNA genes.

**Parsing variation data.** We analyzed human variation data from the African superpopulation of humans, consisting of 661 individuals from Kenya, Nigeria, Sierra Leone, The Gambia and Barbados, from Phase 3 of the 1000 Genomes Project (50). For *D. melanogaster*, we acquired variation data for the Siavonga, Zambia

populations from the *Drosophila* Genome Nexus Database (46, 47). *Mus musculus castaneus* raw data were obtained from Waterston, et al (51) and the *A. thaliana* data were obtained from the Arabidopsis Genome Initiative (52). All non-human data were aligned and genotypes curated as described in Corbett-Detig et al (53).

Within each tRNA, flank, or untranscribed reference region, we considered positions with minor allele frequencies greater than 0 but less than 0.05 to be low-frequency single nucleotide polymorphisms (SNPs). This is based on the idea that SNPs with low minor allele frequencies are generally due to new mutations, on which selection is less of a factor (54). Therefore, these are expected to more closely reflect the neutral mutation rate and spectrum. We also determined the frequency of the 12 possible classes of mutations (e.g. A→G, T→A) within each region of each tRNA where the identity of each base is defined according to the coding strand sequence. Using the alignments, we found the frequency of divergences and low-frequency SNPs by position across all tRNAs and flanking regions, and we obtained 95% confidence intervals for each point estimate by non-parametric bootstrapping across tRNA loci.

For conservation studies across multiple species, we used the phyloP track (18) (across 100 vertebrate species for the human data, across 60 vertebrate species for mouse), and across 27 insect species for the *D. melanogaster* data) from the UCSC Genome Browser (27, 34) and calculated the average score for each position within the tRNAs and flanking regions. The phyloP track assigns scores to each nucleotide in the genome based on alignments to other species, where the score represents the -log p-values under a null model of neutral evolution. Positive scores indicate strong conservation, negative scores indicate accelerated evolution, and sites with scores of zero are undergoing change at a rate consistent with neutral genetic drift (18). When plotting this data, we multiplied the average phyloP scores by negative one, such that sites undergoing accelerated change would have high positive scores, and sites that were strongly conserved would have negative scores (Figures 1A and D, 2A and B). We also performed non-parametric bootstrapping across tRNA loci to determine 95% confidence intervals for all positions. No analogous genome-wide phyloP data was available for *A. thaliana* (18).

For direct comparisons between the species of interest and an outgroup, we used the Multiz Alignment & Conservation track from the UCSC Table Browser (34) and the Stitch MAFs tool from Galaxy (55) to create sequence alignments of the regions of interest in the human and mouse genomes. For the human genome, we downloaded the hg19 human reference genome from the UCSC Genome Browser and aligned to the *Macacca mulatta* reference genome (rheMac2) (56), also from the UCSC Genome Browser (27). We also compared the mouse (*Mus musculus*, mm10) and rat (*Rattus norvegicus*, rn6) genomes (34), and the *A. thaliana* (TAIR10) and *A. lyrata* (v.1.0) genomes (57, 58) using the same methods. For *D. melanogaster*, we used an alignment of the dm6 genome to the droYak2 (*D. yakuba*) genome (59). Non-gap nucleotide mismatches in the alignments were classified as divergent sites. To account for the possibility that multiple substitutions occurred at a single site, we applied a Jukes-Cantor correction to the average rate of divergence at each position (60).

**Transcription factor binding.** The ENCODE Project Consortium used ChIP-Seq data to identify binding regions for regulatory factors (29–31), including the TATA-binding protein (TBP) and several Pol3 transcription factors in the human genome (10). These data were taken from the UCSC Genome Browser (27) in the form of peak calls, in which the intensity of a given peak correlates with a greater frequency of transcription factor binding to that region. For each human tRNA, we found the strongest TBP peak in the 50 base pairs immediately upstream of the tRNA, across the GM12878, H1-hESC, HeLa-S3, HepG2 and K562 cell lines. We chose to use TBP instead of Pol3 peaks because, although TBP is not specific to Pol3 genes, we found that this data was a stronger and more reliable indicator of transcriptional activity. We also calculated the average phyloP score across the flanking regions for each tRNA (18), and performed a Spearman's rank correlation test to quantify the relationship between these data. We repeated this test for the maximum peaks for BDP1 and RPC155 as well, but searched for peaks within the mature tRNA sequence instead, as this is where these transcription factors bind (10). Further, we used only HeLa-S3 and K562 cell data for the BDP1

and RPC155 tests, as this was the only data available for these peaks (31). This ChIP-Seq data was available only for the human genome, so other species were excluded from this part of our analysis.

**Correlating variation to cell-line read counts.** Zheng, et al (32) developed a high-throughput demethylation sequencing pipeline in order to efficiently detect tRNAs within human embryonic kidney (HEK293T) cells (32, 61). We performed Spearman's rank correlation tests to determine the relationship between their mature tRNA read counts and tRNA gene and flanking region conservation. Because Zheng, et al (32) sequenced mature tRNA sequences, which are sometimes encoded by multiple genes, and because we had variation data by gene, we needed to account for this discrepancy. For example, tRNA-Val-CAC-1-1 and tRNA-Val-CAC-1-2 are two distinct genes with different degrees of variation, but they encode the same mature tRNA. For tRNAs encoded at only one locus, we evaluated the correlation between the read counts and the levels of variation at that locus. However, for tRNAs encoded at multiple loci, we took two approaches. First, we excluded these genes entirely, to eliminate the need to control for the correlation between gene copy-number and overall expression (Figure 2E, F) (22, 32). In a separate analysis, we summed the average phyloP scores at these loci, and evaluated the correlation between these totals to the tRNA read counts (Figure S3).

**Finding genome-wide minimum phyloP scores.** We used the UCSC Table Browser (34) to determine which sites in the human genome had phyloP scores of -20, the lowest possible phyloP score. These are the least conserved sites in the human genome across the 100 vertebrate species compared in this track (18). Using genomic coordinates of tRNAs from GtRNAdb (19), we determined what proportion of these sites overlapped tRNA genes or flanking regions and performed hypergeometric tests to quantify associations between these data.

**Estimating the distribution of fitness effects.** We estimated the distribution of fitness effects (DFE) for each species by maximum likelihood using the method of Keightley et al (62), implemented in the DFE-α software. The method is based on site frequency spectra (SFS) obtained from within-species SNP data, and assumes a simple model of recent demographic change to correct the SFS at functional sites for possible skews caused by demography. We used a two-epoch model of demographic change and estimated the DFEs for tRNAs, inner 3' and inner 5' flanking regions for each species. Each of these classes of sites was assumed to be subject to mutation, selection and drift, with gamma-distributed DFEs and an initial shape parameter (β) of 0.5. We also estimated the DFE for sites that are likely to be evolving neutrally (outer 5' flanking regions), which were used as the presumably untranscribed reference regions for generating the expected allele frequency distributions. For each class of putatively selected sites, we analyzed folded site frequency spectra, and the fitness effects of new deleterious mutations were estimated on a scale of $N_eS$, where $N_e$ is a measure of the recent effective population size and S is the strength of selection on a new mutation.

**ACKNOWLEDGMENTS.** We would like to thank Craig Mello for helpful discussions and input on this project, Brian Lin for supplying the legend for Figure 1, as well as Andrew Holmes for sending the mouse chromatin-IP data and for suggestions on this project. We also thank the rest of the Corbett and Lowe Labs for their suggestions, which made our manuscript clearer and made our research more efficient. B.T. was funded by a T-32 training grant (T32 HG008345) during this study.

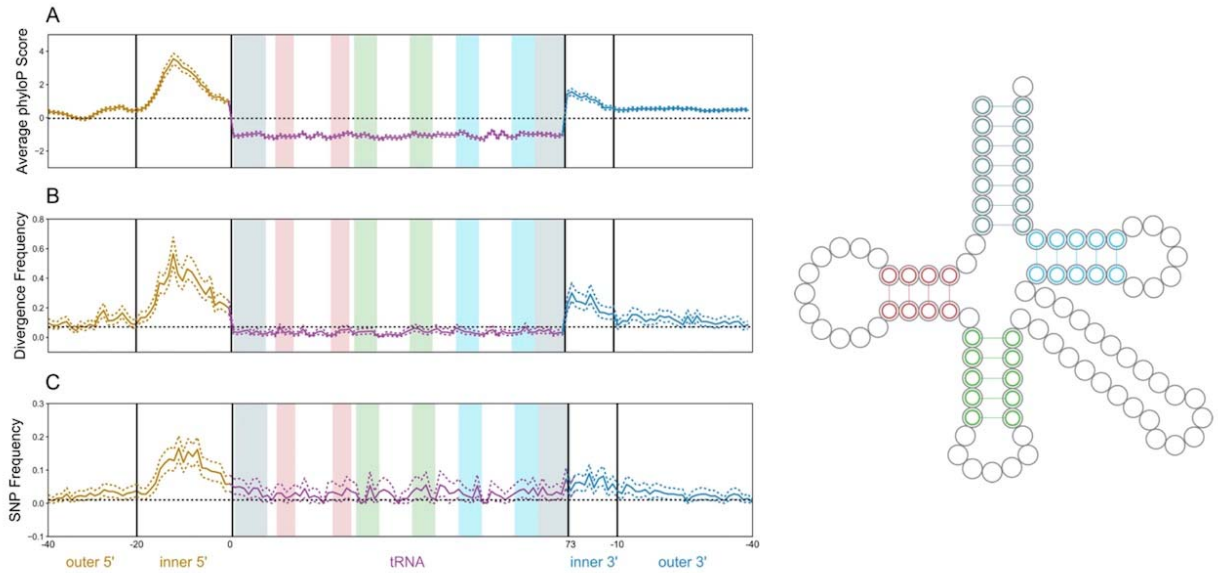


**Fig. 1. Strong pattern of variation in regions flanking human tRNA genes relative to vertebrates, upon comparison to Rhesus macaque, and within the human population. A**: The negative of the average phyloP score (comparing humans to 100 vertebrate species) is plotted for each position within the tRNA and flank, across all human tRNAs. For consistency in plotting, we multiplied the average score at each position by -1, so that more highly divergent regions would have higher, positive scores. **B**: Divergence at non-gap alignments between the hg19 and rheMac2 genomes at each position within tRNAs and their flanking regions. **C**: The frequency at which each position within tRNAs and flanks have a low-frequency SNP (minor allele frequency less than or equal to 0.05) across all human tRNAs. The black dotted line in each plot represents the average value across the untranscribed reference regions used in this study. The acceptor stem (gray), D-stem (red), C-stem (green) and V-stem (blue) are highlighted within the tRNA, both in the plots and in the legend to the right, which shows the secondary structure of the tRNA (19). The black vertical lines separate the inner and outer flanking regions. The 20 bases upstream and 10 bases downstream of each tRNA are considered the inner 5' and inner 3' flanking regions, respectively, as these regions tended to show a marked increase in variation relative to the outer flanking regions (see Methods). The dotted lines surrounding the plots depict 95% confidence intervals, calculated by bootstrapping by tRNA loci.

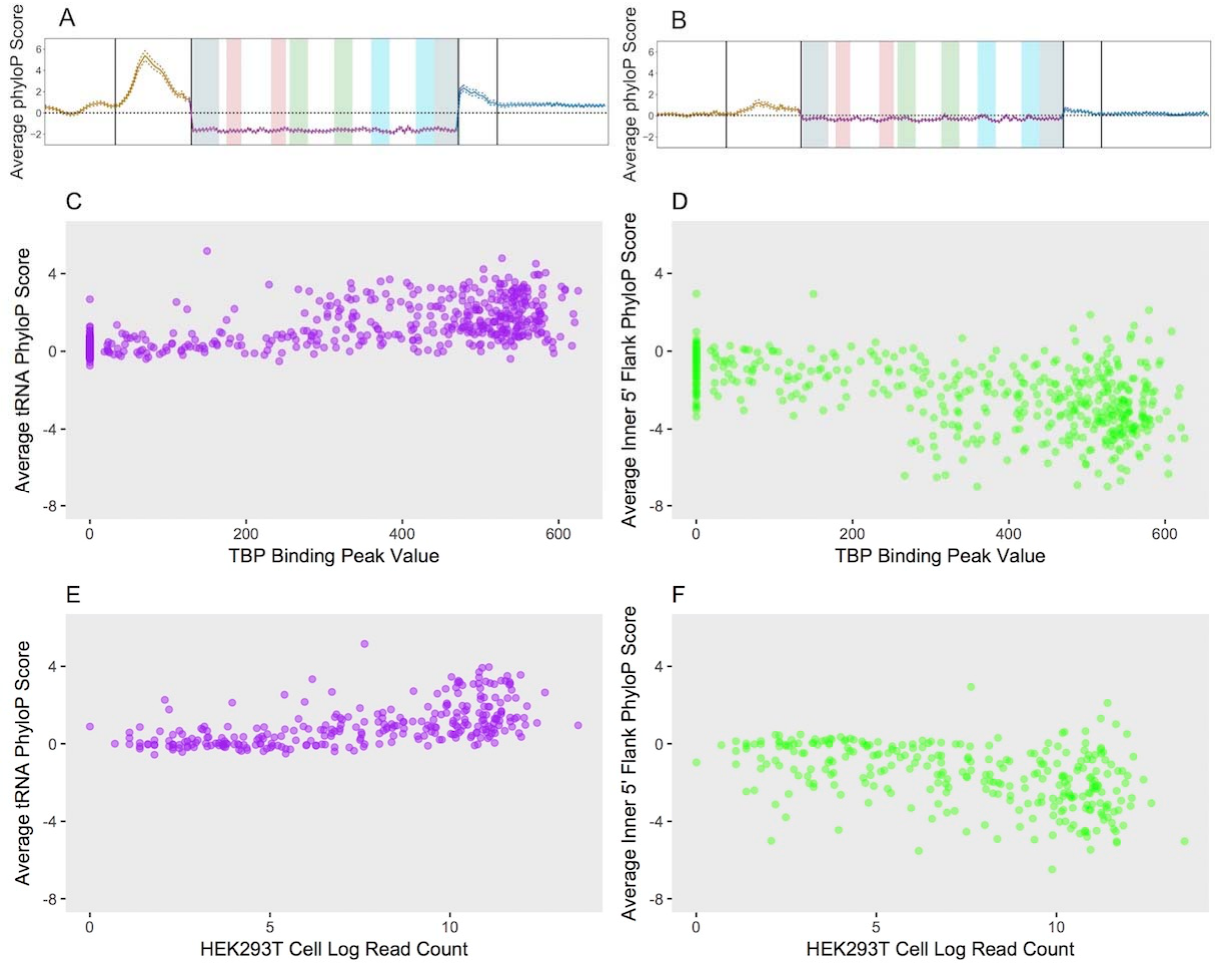


**Fig. 2. Three measures of tRNA expression are significantly correlated to both tRNA conservation and flanking region divergence. A**: The negative of the average phyloP score (comparing humans to 100 vertebrate species) is plotted for each position within the tRNA and flank, across all active human tRNA loci. For consistency with Figure 1A, we used the negative of the phyloP score. **B**: The negative of the average phyloP score is plotted for each position within the tRNA and flank, across all inactive human tRNA loci, following the same format as **A**. **C**: Each tRNA's average phyloP score across its mature sequence is plotted against the value of the TBP peak corresponding to that tRNA. **D**: Each tRNA's average phyloP score across its inner 5' flanking region (20 nucleotides upstream of each tRNA gene) is plotted against the value of the TBP peak corresponding to that tRNA. **E**: The average phyloP score for each tRNA gene encoding a unique mature tRNA sequence is plotted against the log of the HEK293T cell read count for that tRNA (28). **F**: The average phyloP score across the inner 5' flanking region for each tRNA gene encoding a unique mature tRNA sequence is plotted against the log of the HEK293T cell read count for that tRNA (28). Several tRNAs are encoded by multiple sequentially identical genes. Because these would be expected to produce more tRNAs, and therefore have inflated read counts, we excluded these tRNAs from plots E and F. These tRNAs are included in Figure S3.

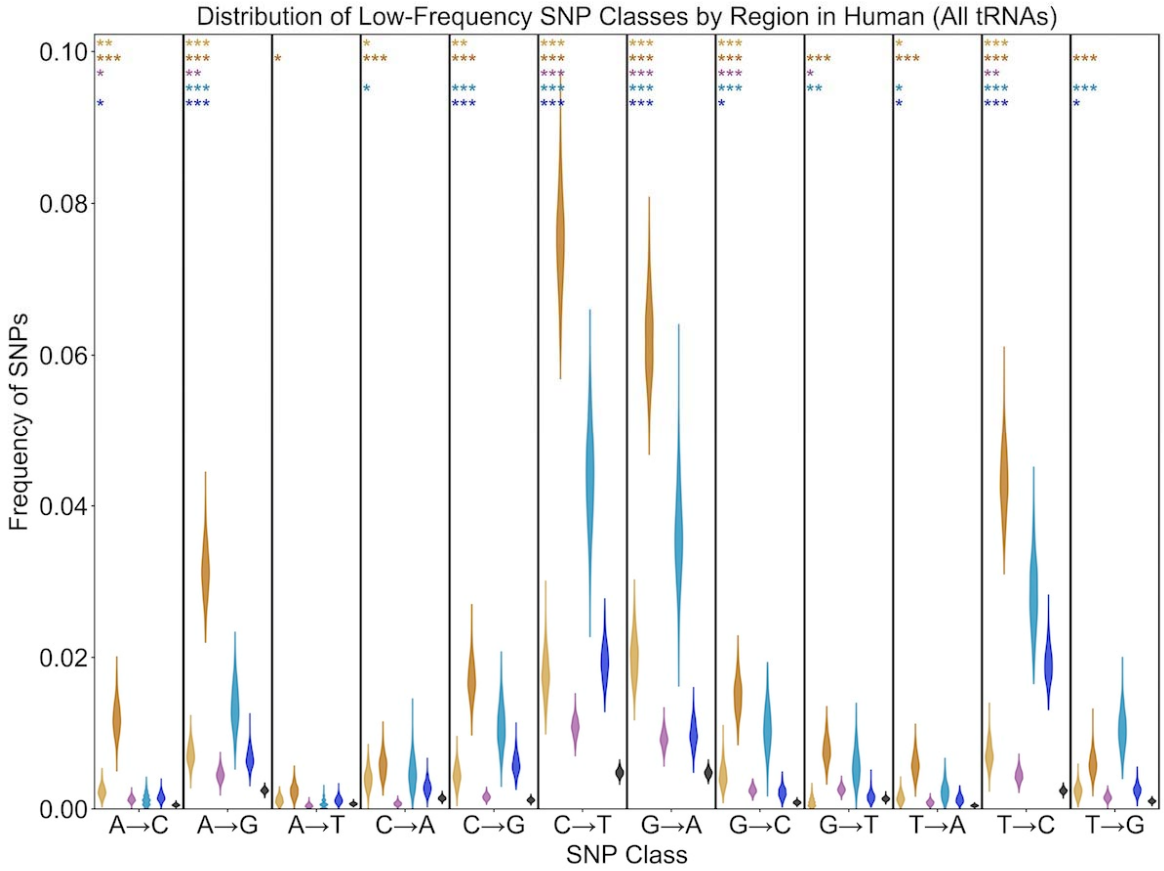


**Fig. 3. SNP classes most common in regions affected by TAM are also most common at tRNA loci.** The distribution of each class of low-frequency polymorphisms, here defined as a SNP with a minor allele frequency less than or equal to 0.05, is shown by region across all human tRNAs. At the top, the significance levels of Fisher's exact tests comparing the SNP distribution within each region of the tRNA and flank (outer 5' flank is yellow, inner 5' flank is orange, tRNA is purple, inner 3' flank is cyan, outer 3' flank is blue) to that of the untranscribed reference region (black) are represented by stars. One star represents a p value $\leq 0.05$, two stars represents a p value $\leq 0.005$, and three stars represents a p value $\leq 0.0005$.

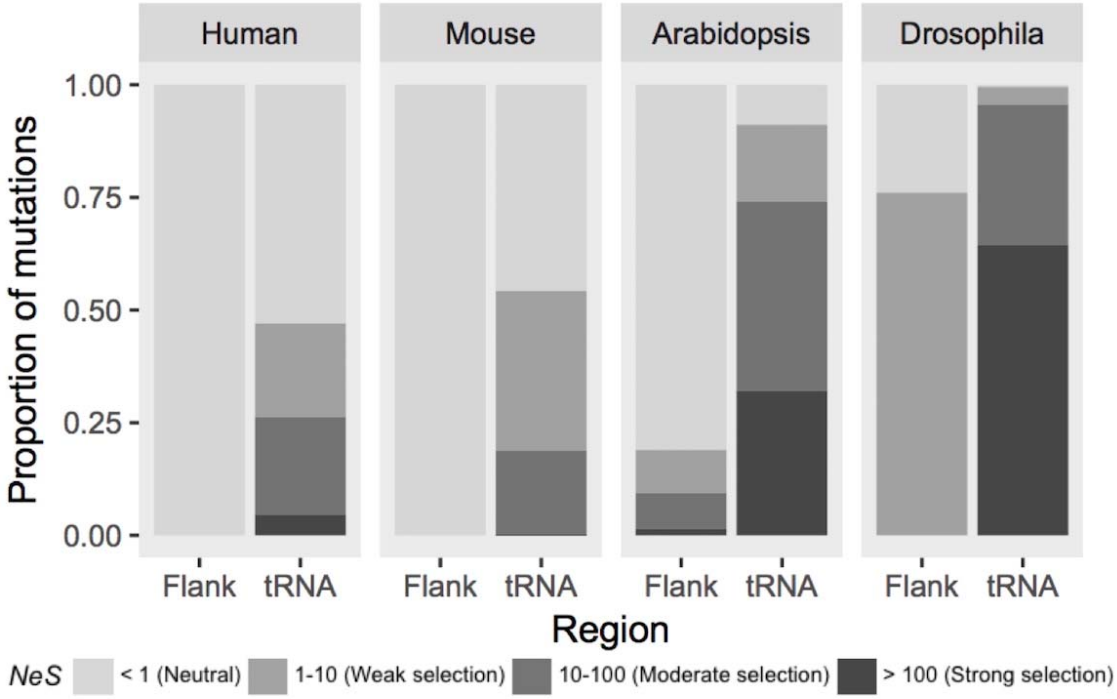


**Fig. 4. Estimated Distribution of Fitness Effects (DFE) indicates that tRNAs show high proportion of deleterious mutations are under strong selection.** Estimated DFE of new deleterious mutations for tRNA genes and inner 3' flanking regions are shown in human, mouse, *A. thaliana* and *D. melanogaster*. The proportions of deleterious mutations are shown for each bin of purifying selection strength, estimated on a scale of $N_eS$, where $N_e$ is a measure of the recent effective population size and S is the strength of selection. The species are arranged by increasing $N_e$.

## References

1. Higgs PG, Jameson D, Jow H, Rattray M (2003) The evolution of trna-leu genes in animal mitochondrial genomes. *Journal of Molecular Evolution* 57(4):435–445.

2. Kirchner S, Ignatova Z (2015) Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nat. Rev. Genet.* 16(2):98–112.

3. Molla-Herman A, Valles AM, Ganem-Elbaz C, Antoniewski C, Huynh JR (2015) tRNA processing defects induce replication stress and Chk2-dependent disruption of piRNA transcription. *EMBO J.* 34(24):3009–3027.

4. Jinks-Robertson S, Bhagwat AS (2014) Transcription-associated mutagenesis. *Annu. Rev. Genet.* 48:341–359.

5. Gnatt AL, Cramer P, Fu J, Bushnell DA, Kornberg RD (2001) Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science* 292(5523):1876– 1882.

6. Helmrich A, Ballarino M, Tora L (2011) Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* 44(6):966– 977.

7. Timakov B, Liu X, Turgut I, Zhang P (2002) Timing and targeting of P-element local transposition in the male germline cells of Drosophila melanogaster. *Genetics* 160(3):1011–1022.

8. Gomez-Gonzalez B, Aguilera A (2007) Activation-induced cytidine deaminase action is strongly stimulated by mutations of the THO complex. *Proc. Natl. Acad. Sci. U.S.A.* 104(20):8409–8414.

9. Green P, Ewing B, Miller W, Thomas PJ, Green ED (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* 33(4):514–517.

10. White RJ (2011) Transcription by RNA polymerase III: more complex than we thought. *Nat. Rev. Genet.* 12(7):459–463.

11. Zhang J, Ferre-D'Amare AR (2016) The tRNA Elbow in Structure, Recognition and Evolution. *Life (Basel)* 6(1).

12. Sun C, et al. (2017) Roles of tRNA-derived fragments in human cancers. *Cancer Lett.* 414:16– 25.

13. Ziehler WA, Day JJ, Fierke CA, Engelke DR (2000) Effects of 5' leader and 3' trailer structures on pre-tRNA processing by nuclear RNase P. *Biochemistry* 39(32):9909–9916.

14. Hopper AK (2013) Transfer RNA post-transcriptional processing, turnover, and subcellular dynamics in the yeast Saccharomyces cerevisiae. *Genetics* 194(1):43–67.

15. Hasler D, et al. (2016) The Lupus Autoantigen La Prevents Mis-channeling of tRNA Fragments into the Human MicroRNA Pathway. *Mol. Cell* 63(1):110–124.

16. Lee YS, , et al. (2009) A novel class of small rnas: trna-derived rna fragments (trfs). *Cell* 23(23):2639–2649.

17. Maraia RJ, Lamichhane TN (2011) 3' processing of eukaryotic precursor tRNAs. *Wiley Interdiscip Rev RNA* 2(3):362–375.

18. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20(1):110–121.

19.Chan PP, Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* 37(Database issue):D93–97.

20. Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* 26(1):148–153.

21.Kundaje A, et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317–330.

22. Bloom-Ackermann Z, et al. (2014) A comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool. *PLoS Genet.* 10(1):e1004084.

23.Doran JL, Bingle WH, Roy KL (1988) Two human genes encoding tRNA(GCCGly). *Gene* 65(2):329–336.

24. Dieci G, Sentenac A (1996) Facilitated recycling pathway for RNA polymerase III. *Cell* 84(2):245–252.

25. Ciesla M, Boguta M (2008) Regulation of RNA polymerase III transcription by Maf1 protein. *Acta Biochim. Pol.* 55(2):215–225.

26. Orioli A, et al. (2011) Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. *Nucleic Acids Res.* 39(13):5499–5512.

27.Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res.* 12(6):996–1006.

28. Juo ZS, et al. (1996) How proteins recognize the TATA box. *J. Mol. Biol.* 261(2):239–254.

29. Dunham I, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.

30. Myers RM, et al. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9(4):e1001046.

31. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* 26(12):1351–1359.

32. Zheng G, et al. (2015) Efficient and quantitative high-throughput tRNA sequencing. *Nat. Methods* 12(9):835–837.

33.Thul PJ, et al. (2017) A subcellular map of the human proteome. *Science* 356(6340).

34. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32(Database issue):D493–496.

609 35. Schmidt S, et al. (2008) Hypermutable non-synonymous sites are under stronger negative selection. *PLoS Genet.* 4(11):e1000281.

610 36. Saini N, et al. (2017) APOBEC3B cytidine deaminase targets the non-transcribed strand of tRNA genes in yeast. *DNA Repair (Amst.)* 53:4–
611 14.

612 37. Tenesa A, et al. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome research* 17(4):520–526.

613 38. Phifer-Rixey M, et al. (2012) Adaptive evolution and effective population size in wild house mice. *Molecular biology and evolution*
614 29(10):2949–2955.

615 39. Cao J, et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* 43(10):956–963.

616 40. Shapiro JA, et al. (2007) Adaptive genic evolution in the *Drosophila* genomes. *Proceedings of the National Academy of Sciences*
617 104(7):2271–2276.

618 41. Lowe TM, Chan PP (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.*
619 44(W1):W54–57.

620 42. Narasimhan VM, et al. (2017) Estimating the human mutation rate from autozygous segments reveals population differences in human
621 mutational processes. *Nat Commun* 8(1):303.

622 43. Keightley PD (2012) Rates and fitness consequences of new mutations in humans. *Genetics* 190(2):295–304.

623 44. Haldane JBS (1937) The effect of variation of fitness. *The American Naturalist* 71(735):337– 349.

624 45. Agrawal AF, Whitlock MC (2012) Mutation load: The fitness of individuals in populations where deleterious alleles are abundant. *Annual*
625 *Review of Ecology, Evolution, and Systematics* 43(1):115–135.

626 46. Lack JB, et al. (2015) The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197
627 from a single ancestral range population. *Genetics* 199(4):1229–1241.

628 47. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE (2016) A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. *Mol.*
629 *Biol. Evol.* 33(12):3308–3313.

630 48. Fredrik Pettersson BM, Ardell DH, Kirsebom LA (2005) The length of the 5' leader of Escherichia coli tRNA precursors influences bacterial
631 growth. *J. Mol. Biol.* 351(1):9–15.

632 49. Bogu GK, et al. (2015) Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. *Mol. Cell. Biol.* 36(5):809–819.

633 50. Consortium TGP (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74.

634 51. Waterston RH, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

635 52. authors listed N (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814):796–815.

636 53. Corbett-Detig RB, Hartl DL, Sackton TB (2015) Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol.*
637 13(4):e1002112.

638 54. Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide
639 polymorphisms in five populations. *Am. J. Hum. Genet.* 66(1):216–234.

640 55. Afgan E, et al. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids*
641 *Res.* 44(W1):W3–W10.

642 56. Gibbs RA, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316(5822):222–234.

643 57. Kersey PJ, et al. (2016) Ensembl genomes 2016: more genomes, more complexity. *Nucleic Acids Research* 44(D1):D574–D580.

644 58. Rosenbloom KR, et al. (2015) The UCSC genome browser database: 2015 update. *Nucleic Acids Research* 43(D1):D670–D681.

645 59. Clark AG, et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450(7167):203–218.

646 60. Jukes TH, Cantor CR (1969) Evolution of protein molecules. *Mammalian Protein Metabolism* pp. 21–132.

647 61. Cozen AE, et al. (2015) ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments.
648 *Nat. Methods* 12(9):879–884.

649 62. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography
650 based on nucleotide polymorphism frequencies. *Genetics* 177(4):2251–2261.

651

652

653

654

655

656