Transfer RNA genes experience exceptionally elevated mutation rates

Bryan P. Thornlow[a], Josh Hough[a], Jacquelyn M. Roger[a], Henry Gong[a], Todd M. Lowe[a,b,1], and Russell B. 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: tmjlowe@ucsc.edu or rucorbet@ucsc.edu

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

The authors declare no conflict of interest.

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 seven and ten 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 new method for estimating individual tRNA gene activity. Collectively, this study illuminates the extreme competing forces in tRNA gene evolution, and implies that mutations at tRNA loci contribute disproportionately to mutational load and have unexplored fitness consequences in human populations.

Keywords: tRNA | transcription | mutagenesis | TAM | computational prediction

Significance Statement

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 seven- to tenfold 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 implicate that prediction of tRNA gene expression based on sequence variation data is possible.

\body

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 approximately 10- to 20-fold as many tRNA genes encoded in their nuclear chromosomes, which are required for cytosolic protein translation (2, 4). In spite of 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 complementary DNA 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 non-canonical DNA structures, which can hinder repair pathways and promote errors by the polymerase (11). The RNA strand can also re-anneal 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 induces 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 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 1000-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 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, DNA 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 2-5 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 similar mutation rates to 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 & 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 conservation of each human genomic position across 100 vertebrate species (28), by position within each tRNA locus (see Methods). Positive PhyloP scores indicate strong conservation and negative scores indicate accelerated evolution. To study the effects of evolution on a shorter time-scale, we also estimated sequence divergence between human and *Macacca mulatta* at each tRNA locus. Mature tRNA sequences are highly conserved across all positions, based on both average PhyloP score (28; Fig. 1A; SI Appendix, Dataset S1) 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 find 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 outer 3’ flank (11-40 bases downstream) 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 single nucleotide polymorphisms (SNPs; minor allele frequency < 0.05%) for each tRNA locus. Low-frequency SNPs are evolutionarily young, and 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 to 15 bases upstream of the start of an active tRNA gene, indicating disproportionate enrichment (hypergeometric test, $p < 1.65e-48$) 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 (33). Based on these data, we classified human tRNA genes as "active" if they are located in expressed regions in several cell lines and otherwise "inactive" (Fig. 2; see Methods). 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, $p < 8.40e-53$), and the flanking regions of active tRNAs are significantly more *divergent* than the flanking regions of inactive tRNAs ($p < 7.98e-61$). 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,F). Active tRNAs also have significantly more low-frequency SNPs per site in human populations than inactive tRNAs across the entire locus, including the tRNA and flanking regions ($p < 3.72e-36$; 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 to 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 (SI Appendix, Fig. 3A). 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 (34), few studies have shown that flanking regions affect expression in higher eukaryotes (35). Transcription initiation is long relative to elongation (36, 37), 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 non-uniform 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 is varies among tRNA genes (2, 38), 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 approximately 25 nucleotides upstream of the tRNA (39), 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 (SI Appendix, 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 (40–42) and conservation of the mature tRNA sequence (Spearman's rho = 0.64, $p < 2.2e-16$) across all human tRNAs and the opposite relationship in the flanking regions (Spearman's rho = -0.64, $p < 2.2e-16$; Fig. 2). TBP ChIP-Seq data directly reflects transcriptional activity for each locus, as its occupancy is significantly correlated with and required for transcription (20, 43–47). Second, mature tRNA sequence read counts are strongly correlated with tRNA conservation (Spearman's rho = 0.18, $p < 0.001$) and flanking region divergence (Spearman's rho = -0.61, $p < 2.2e-16$; Fig. 2; SI Appendix, Fig. S4). These read counts were collected from a single human embryonic kidney cell line by Zheng et al. (32) 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 are 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 (SI Appendix, Dataset S1). For example, ribosomal proteins are very highly transcribed (48) and have well conserved exons, but their introns and flanking regions are not as divergent as tRNA flanking regions (28, 49). 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 transcription-associated mutagenesis (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, relative to untranscribed reference regions (Fisher's exact test, $p < 0.05$ for all comparisons; Fig. 3). Removal of CpG sites (50) does not significantly affect these results. The relative excesses of these SNPs are much more pronounced in active tRNA loci than in inactive tRNA loci (SI Appendix, Fig. S7A,B). These results 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 (9, 17, 19).

It is difficult to discern whether this increased prevalence is due to TAM or selection to preserve structural integrity of the tRNA. To preserve tRNA secondary structure, we 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, as they should disrupt stem helices less often. However, the mutational skew expected of regions affected by TAM is stronger in regions flanking tRNAs. Transcription initiation is relatively long compared to elongation (36, 37), which might contribute to increased mutagenesis by APOBEC enzymes, more collisions (14) or double-stranded breaks. However, divergence at tRNA flanking regions is correlated with divergence at introns in both human (SI Appendix, Fig. S1B, Spearman's rank, rho = 0.734, $p < 5.58e-6$) and mouse (SI Appendix, Fig. S1D, rho = 0.733, $p < 5.24e-4$), indicating similar mutation rates across tRNA loci. Our results therefore suggest that TAM drives the excess of transitions among low-frequency SNPs across tRNA loci.

tRNA flanking region variation in other model organisms is consistent with variation observed in humans. To confirm that our results are not restricted to humans, we also analyzed tRNAs in *Mus musculus*, *Drosophila melanogaster* and *Arabidopsis thaliana*. We find similar patterns of sequence conservation of tRNA loci in each, when measuring PhyloP or divergence to outgroups (SI Appendix, Fig. S8). The 5' flanks are consistently more divergent than the 3' flanks, and the most divergent sites are roughly 10-15 bases upstream of the tRNA in all species. We also used chromatin-IP data across nine mouse tissues to classify mouse tRNAs based on their expression (51). Active mouse tRNAs are more strongly conserved than their inactive counterparts (Mann-Whitney U, $p < 1.81e-19$), and their flanks are more divergent ($p < 7.04e-22$, Fig. 2G,H), consistent with our results from the human data (Fig. 2E,F). Active mouse tRNAs also have more low-frequency SNPs in their flanking regions than inactive mouse tRNAs ($p < 2.23e-4$; SI Appendix, Fig. S9). Such consistency suggests that a shared underlying molecular mechanism drives these patterns of sequence variation.

Low-frequency SNPs in the tRNA gene sequences also follow similar qualitative patterns to the human data. We observe excess transitions in all species studied (SI Appendix, Fig. S10), and active mouse tRNAs show a greater excess of low-frequency transitions than do inactive mouse tRNAs (SI Appendix, Fig. S7C,D). However, these patterns vary across species (Fig. 4B-D). For example, in mouse, tRNA genes have more low-frequency SNPs than the untranscribed reference regions (Fig. 4B), but the opposite is true in *D. melanogaster* (Fig. 4D). Low-frequency SNPs are thought not to be strongly affected by selection (29), but selection is more

efficient in species with greater effective populations sizes (Fig. 4A). Effective population size (52–55) and tRNA copy number vary across species, and because the sample sizes and data quality differ between population samples, these differences may be attributable to differences in the impact of selection or in ascertainment of low-frequency variation.

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 analyzed species. In contrast, regions flanking tRNAs are inferred to be either neutral or subject to weak selection ($N_eS < 10$; Fig. 4A). Our estimates of the proportions of new mutations falling into each N_eS range of the DFE for tRNAs indicate 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 the human or mouse populations (Fig. 4A). Given that estimates of N_e in humans (7,000; 52) and mouse (25,000-120,000; 53) are substantially lower than in *A. thaliana* (300,000; 54) and *D. melanogaster* (> 1,000,000; 55), this difference in strength of selection may partially reflect differences in N_e, and might explain the differences in low-frequency SNPs in tRNA loci across species (Fig. 4B-D). In turn, this might indicate that the strength of purifying selection, independent of effective population size, at tRNA loci is consistent across diverse species.

The strength of selection across species may also reflect the number of unique tRNA gene sequences in each genome. For example, roughly half of all human tRNA genes have unique sequences, but the majority of *D. melanogaster* tRNAs have identical copies (2). tRNAs with the same anticodon but different sequences may have different functions, and that 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 multi-copy 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 multi-copy tRNA genes. However, little SNP data is available for multi-copy tRNAs compared to 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 (56, 57). To estimate the relative mutation rates at active tRNA loci, we calculated the average ratios of θ for the inner 3' and 5' flanking regions of active human tRNA genes to the untranscribed reference regions using the approach of Messer (29; see Methods). We estimate θ in the flanking regions instead of the tRNAs because strong selection can cause underestimation of θ (29), and our results indicate that active human tRNAs are subject to strong selection while the flanking regions are likely selectively neutral (Fig. 4A). We therefore estimate that the mutation rate is between 7.24 (inner 3'; 95% confidence interval 7.12-7.33) and 10.36 (inner 5'; 10.16-10.41) times greater at tRNA loci than the genome-wide average. Given that there are 25,852 base pairs of active human tRNA sequence, and using 1.45e-8 as the genome-wide mutation rate (58), we estimate that U (the genome-wide rate of deleterious mutation per diploid genome) contributed by tRNAs is between 0.0054 and 0.0078. Since active tRNAs make up only 0.0009% of the human genome (2), this implies that mutations in tRNAs contribute disproportionately to

mutational load. Our findings highlight that mutations at tRNA loci are likely an important source of fitness and disease variation in human populations.

Conclusions

Our findings demonstrate the exceptional transcription rates of tRNA genes causes a similarly substantial increase in mutation rates through TAM. 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, and may be a good predictor of relative tRNA gene transcription levels. The conflict between extreme TAM and consequent strong purifying selection at tRNA loci is potentially an unappreciated source of genetic disease, and may have a profound impact on human fitness, yet to be fully addressed.

Materials and Methods

Defining tRNA loci and flanking regions. We used tRNA coordinates from GtRNAdb (2) for the human, *M. musculus*, *D. melanogaster*, and *A. thaliana* genomes. For each species, we defined untranscribed reference regions by searching 10 kilobases upstream of each tRNA and selecting a 200-nucleotide tract. If this tract was within a highly transcribed region of the genome (based on genome-wide chromatin-IP data; 33), overlapped a conserved element (defined as a region with a phastCons log odds score greater than 0; 28), was within 1,000 nucleotides of a known gene (49), or overlapped a 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 known genes, previously assigned reference regions, and conserved elements. 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 (59, 60) due to the high functional densities of these species' genomes.

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 gene on the coding strand, and the outer 5' flank as the 20 bases directly upstream of the inner 5' flank. The inner 3' flank refers to the 10 bases downstream of the tRNA gene, 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 have less variation. Transcription usually ends about 10 bases downstream of tRNA genes (38).

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 chromatin state across the genome (33). We analyzed the regions surrounding each tRNA in each epigenome sample, and used clustering to classify each genomic region according to its most common epigenomic state. We classified all human tRNAs based on the epigenomic state annotation in the genome. In the corresponding model, regions in state 1 are likely to be transcribed. The 342 tRNAs in state 1 in at least four of the 127 tissues analyzed are "active tRNAs", and we consider the remaining 254 tRNAs "inactive". To classify mouse tRNAs, we used a 15-state Hidden Markov Model based on chromatin-IP data in which states 5 and 7 corresponded to regions near active promoters (51). We considered the 272 tRNAs in genomic regions annotated as state 5 or 7 in at least 3% of tissues "active" and the remaining 188 tRNAs "inactive".

Aligning tRNAs. We aligned all tRNAs across all species using covariance models (61) and assigned coordinates to each position in each tRNA and flank based on the Sprinzl numbering system (31). 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 analyses. If a conserved element (regions with a phastCons log odds score greater than 0; 28) was present 4-10 bases up 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 (59, 60). We obtained *M. musculus* and *A. thaliana* data from Waterston, et al. (63) and the Arabidopsis Genome Initiative (64), respectively. All non-human data were aligned and genotypes curated as described in Corbett-Detig, et al. (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 UCSC Genome Browser (49) and calculated the average score for each position within the tRNAs and flanking regions. No PhyloP data was 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 the SI Appendix.

Transcription factor binding. The ENCODE Project Consortium used ChIP-Seq data to identify binding regions for regulatory factors (40–42), including the TATA-binding protein (TBP) and Pol3 transcription factors in the human genome (20). These data were taken from the UCSC Genome Browser (49). 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 also calculated the average PhyloP score across the flanking regions for each tRNA (28), and used Spearman's rank correlation test on these data.

Correlating variation to cell-line read counts. Zheng, et al. (32) used demethylation sequencing to detect tRNAs within human embryonic kidney (HEK293T) cells (32, 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. (32) 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. 2C,D; 32, 34). Separately, we summed the average PhyloP scores at these loci and correlating to total tRNA read counts (SI Appendix, Fig. S4).

Estimating the distribution of fitness effects. We estimated the distribution of fitness effects (DFE) for each species using the method of Keightley et al. (69) and the DFE-α software. See SI Appendix for details.

Estimating mutation rate in active tRNA genes. We used the equation $\theta(k) = kG^k$ defined in Messer 2009 (29) to estimate the mutation rate at active tRNA loci. We calculated the ratios of θ in active tRNA flanking regions to θ in the reference regions for k = 1,2,3 and bootstrapped by tRNA loci to calculate 95% confidence intervals. 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, and Andrew Holmes for assisting with mouse data analysis. We thank the Corbett and Lowe Labs for suggestions and feedback. This work was supported by a grant from the National Human Genome Research Institute, National Institutes of Health (2R01HG006753-04A1 to T.L.). B.T. was funded by an NIH training grant (T32 HG008345).

References

1. Tang DT, Glazov EA, McWilliam SM, Barris WC, Dalrymple BP (2009) Analysis of the complement and molecular evolution of tRNA genes in cow. *BMC Genomics* 10:188.

2. Chan PP, Lowe TM (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* 44(D1):D184–189.

3. Suzuki T, Nagao A, Suzuki T (2011) Human mitochondrial tRNAs: biogenesis, function, structural aspects, and diseases. *Annu. Rev. Genet.* 45:299–329.

4. Schimmel P (2018) The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. *Nat. Rev. Mol. Cell Biol.* 19(1):45–58.

5. Kutter C, et al. (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat. Genet.* 43(10):948–955.

6. Ishimura R, et al. (2014) RNA function. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science* 345(6195):455–459.

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

8. 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.

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

10. 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.

11. Gaillard H, Aguilera A (2016) Transcription as a threat to genome integrity. *Annual Review of Biochemistry* 85(1):291–317.

12. Kim N, Jinks-Robertson S (2012) Transcription as a source of genome instability. *Nat. Rev. Genet.* 13(3):204–214.

13. Aguilera A, García-Muse T (2013) Causes of genome instability. *Annual review of genetics* 47:1–32.

14. 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.

15. 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.

16. 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.

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

18. Taylor BJ, Wu YL, Rada C (2014) Active rnap pre-initiation sites are highly mutated by cytidine deaminases in yeast, with aid targeting small rna genes. *elife* 3.

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

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

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

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

23. 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.

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

25. 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.

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

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

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

29. Messer PW (2009) Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* 182(4):1219–1232.

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

31. 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.

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

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

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

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

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

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

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

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

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

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

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

43. Roberts DN, Stewart AJ, Huff JT, Cairns BR (2003) The RNA polymerase III transcriptome revealed by genome-wide localization and activity-occupancy relationships. *Proc. Natl. Acad. Sci. U.S.A.* 100(25):14695–14700.

44. Mason PB, Struhl K (2003) The FACT complex travels with elongating RNA polymerase II and is important for the fidelity of transcriptional initiation in vivo. *Mol. Cell. Biol.* 23(22):8323– 8333.

45. Kuras L, Kosa P, Mencia M, Struhl K (2000) TAF-Containing and TAF-independent forms of transcriptionally active TBP in vivo. *Science* 288(5469):1244–1248.

46. Zanton SJ, Pugh BF (2004) Changes in genomewide occupancy of core transcriptional regu lators during heat stress. *Proc. Natl. Acad. Sci.*

U.S.A. 101(48):16843–16848.

47. Li XY, Virbasius A, Zhu X, Green MR (1999) Enhancement of TBP binding by activators and general transcription factors. *Nature* 399(6736):605–609.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

64. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814):796–815.

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

66. Blanchette M, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14(4):708–715.

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

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

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

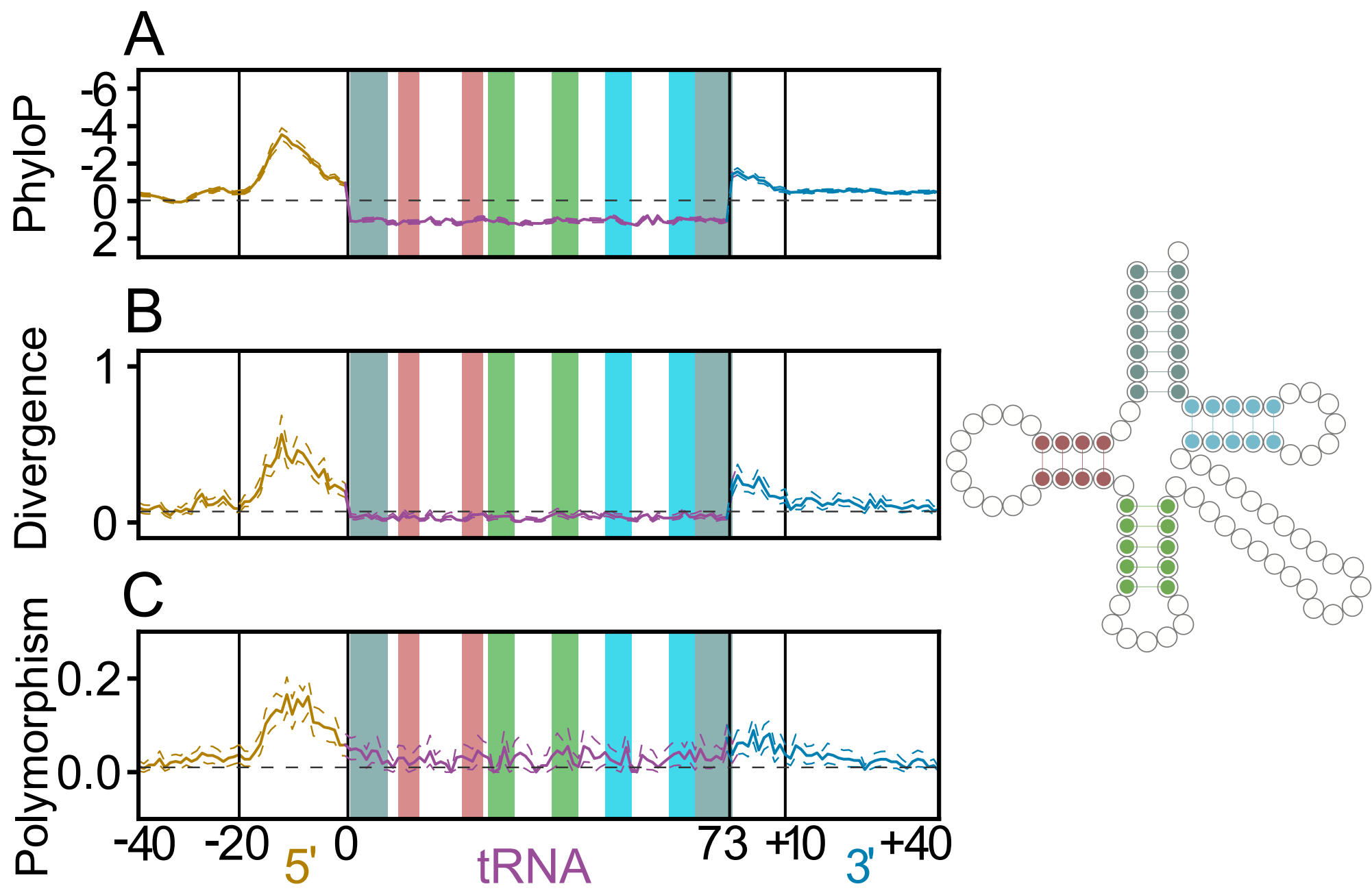
Figure Legends

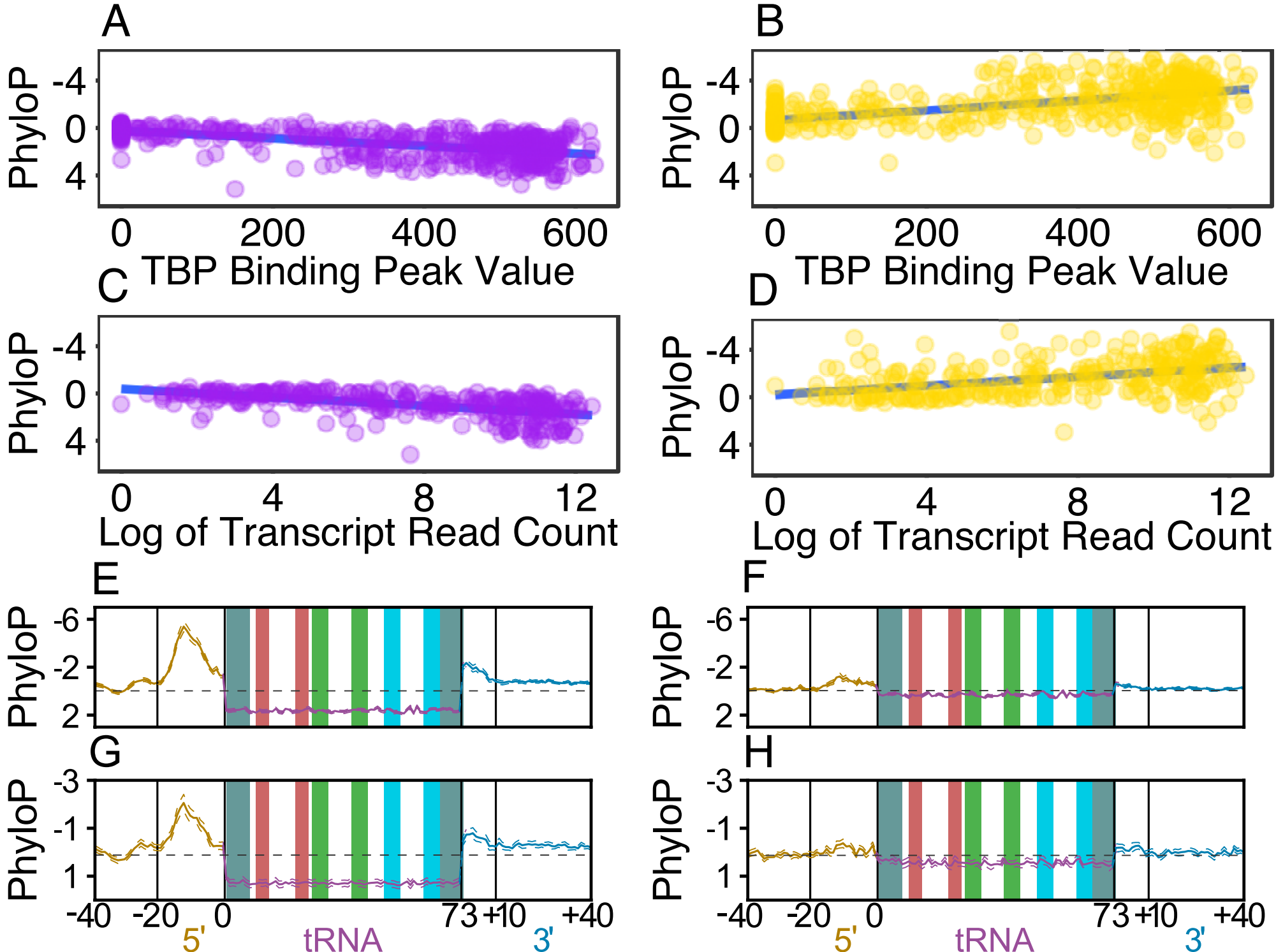
Fig. 1. Strong pattern of variation in regions flanking human tRNA genes relative to vertebrates, upon comparison to Rhesus macaque, and within the human population. (A) Average PhyloP score (comparing humans to 100 vertebrate species) is plotted for each position within the tRNA and flank, across all human tRNAs. (**B**) Divergence between the human and *M. mulatta* tRNA genes and their flanking regions. (**C**) Frequency of low-frequency SNPs (minor allele frequency less than or equal to 0.05) across all human tRNAs. The acceptor stem (gray), D-stem (red), anticodon stem (green) and T-stem (blue) are highlighted within the tRNA, both in the linear plots and in the 2D structure legend to the right (2, 31). Nucleotide numbering below plot relative to mature tRNA boundaries, with inner and outer flanks demarcated by a shift in mutation rate (see Methods). Dotted lines surrounding plots depict 95% confidence intervals, calculated by non-parametric bootstrapping by tRNA loci.

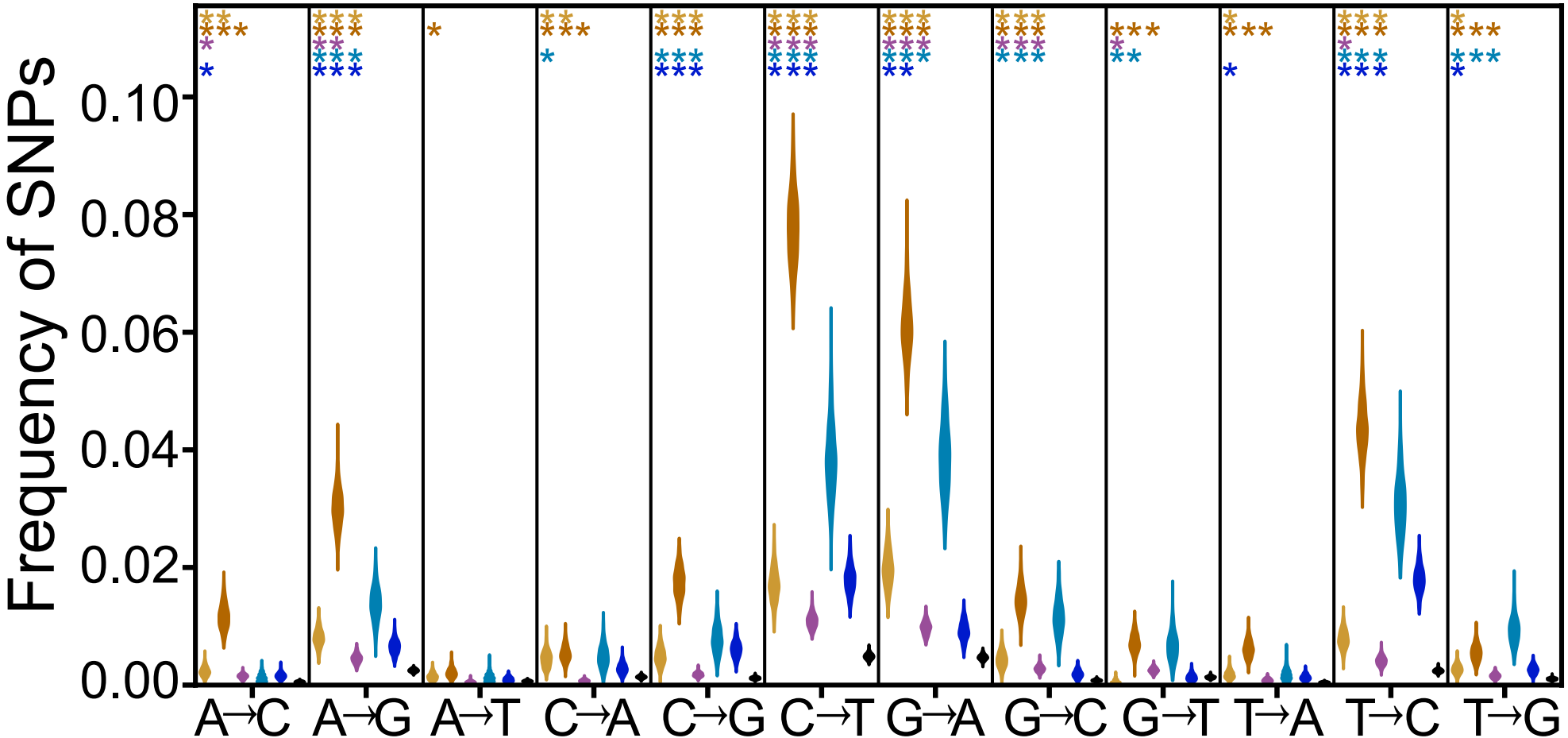
Fig. 2. tRNA expression is significantly correlated to both tRNA conservation and flanking region divergence. TATA-Binding Protein (TBP) peak value (expression) plotted versus PhyloP score (conservation) for each (**A**) mature tRNA and (**B**) adjacent inner 5' flanking region. Log of the HEK293T cell DM-tRNA-seq read count (expression; 32) plotted versus PhyloP score (conservation) for each (**C**) gene encoding a *unique* mature tRNA sequence and (**D**) corresponding inner 5' flanking region. Both TBP occupancy and transcript abundance are greater for highly conserved mature tRNA loci (**A,C**) and those with the most divergent flanks (**B,D**). Plotted as in Fig. 1, human tRNA loci that are separated into (**E**) active versus (**F**) inactive groups show characteristic differences seen in (**A-D)**. Mouse tRNA loci split into (**G**) active versus (**H**) inactive groups show a strikingly similar pattern as seen in human (**A-F**).

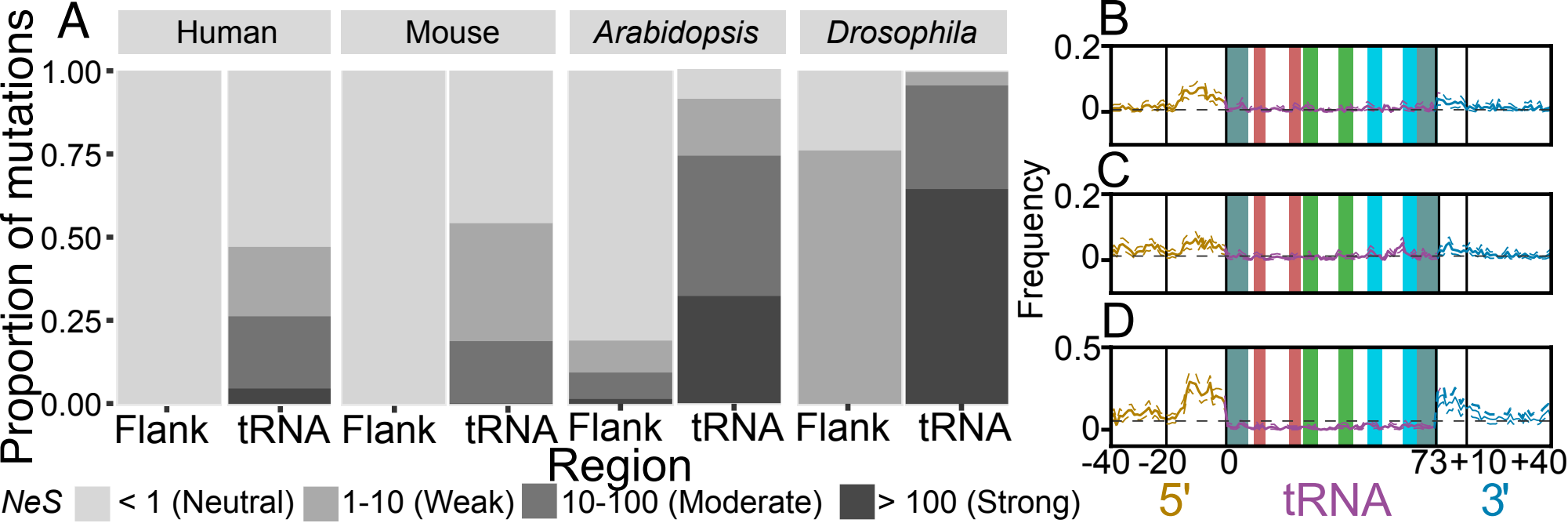
Fig. 3. SNP classes most common in regions affected by TAM are also most common at tRNA loci. Distribution of each class of low-frequency polymorphisms by region across all human tRNAs. Significance levels of Fisher's exact tests comparing the SNP distribution within each region of the tRNA and flank (outer 5' flank in yellow, inner 5' flank in orange, tRNA in purple, inner 3' flank in cyan, outer 3' flank in blue) to that of the untranscribed reference region (black) represented by stars. One star represents a p value ≤ 0.05, two stars ≤ 0.005, and three stars ≤ 0.0005.

Fig. 4. Estimated Distribution of Fitness Effects (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 shown for each bin of purifying selection strength, estimated on a scale of N_eS, where N_e is the effective population size and S is the strength of selection. Species are arranged by increasing N_e. Low-frequency SNPs plotted as in Fig. 1C for (**B**) mouse, (**C**) *A. thaliana*, and (**D**) *D. melanogaster*.


A
PhyloP
-6
-4
-2
0
2
B
Divergence
1
0
C
Polymorphism
0.2
0.0
-40
-20
0
73
+10
+40
5'
tRNA
3'


A
B
C
D
E
F
G
H
PhyloP
-4
0
4
0
200
400
600
TBP Binding Peak Value
0
4
8
12
Log of Transcript Read Count
-6
-2
2
-3
-1
1
-40
-20
0
73
+10
+40
5'
tRNA
3'


Frequency of SNPs
0.00
0.02
0.04
0.06
0.08
0.10
A→C
A→G
A→T
C→A
C→G
C→T
G→A
G→C
G→T
T→A
T→C
T→G


A
Human
Mouse
Arabidopsis
Drosophila
Proportion of mutations
1.00
0.75
0.50
0.25
0.00
Flank
tRNA
Region
B
C
D
Frequency
0.2
0
0.5
-40 -20 0 73 +10 +40
5'
tRNA
3'
NeS
< 1 (Neutral)
1-10 (Weak)
10-100 (Moderate)
> 100 (Strong)