Supplementary Information for

Transfer RNA genes experience exceptionally elevated mutation rates

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

Todd M. Lowe and Russell B. Corbett-Detig
E-mail: tmjlowe@ucsc.edu or rucorbet@ucsc.edu

This PDF file includes:

Supplementary text
Figs. S1 to S10
Captions for Databases S1 to S5
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Databases S1 to S5

Supporting Information Text

Methods. We aligned the hg19 human reference genome to the Macaca mulatta reference genome (rheMac2; 1), both from the UCSC Genome Browser (2). We also compared the mouse (*Mus musculus*, mm10) and rat (*Rattus norvegicus*, rn6) genomes, and the *A. thaliana* (TAIR10) and *A. lyrata* (v.1.0) genomes (3) using the same methods. For *D. melanogaster*, we used an alignment of the dm6 and droYak2 (*D. yakuba*) genomes (4). 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 used a Jukes-Cantor correction (5).

The DFE estimation 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 skews caused by demography. We used a two-epoch model of demographic change and estimated the DFEs for tRNAs and inner 3' flanking regions for each species. Each class 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.

To estimate the mutation rate at tRNA loci, we used the approach defined in Messer, 2009 (6; Equation 22). We calculate the maximum-likelihood estimator for θ (4Nu) from low-frequency sites for untranscribed reference regions and flanking regions of active tRNA genes. The ratios of $\theta_{flank} : \theta_{reference}$ should then provide estimates for the the ratio of the mutation rates in these regions. If we then assume the reference regions have a mutation rate equal to 1.45e-8 per site per generation (7), multiplying by this ratio yields an estimate of the per site per generation mutation rate at tRNA loci. To calculate U_{tRNA}, or the contribution of mutations at active tRNA loci to the genome-wide rate of deleterious mutation per diploid genome, we multiply the human genome-wide mutation rate per nucleotide per haploid genome (1.45e-8; 7), times 2 to correct for diploidy, times the number of nucleotides in active human tRNAs (25,852), times the ratios of $\theta_{flank} : \theta_{reference}$ (7.24 for inner 3', 10.36 for inner 5'). Using these ratios, we estimate that U_{tRNA} is between 0.0054 and 0.0078.

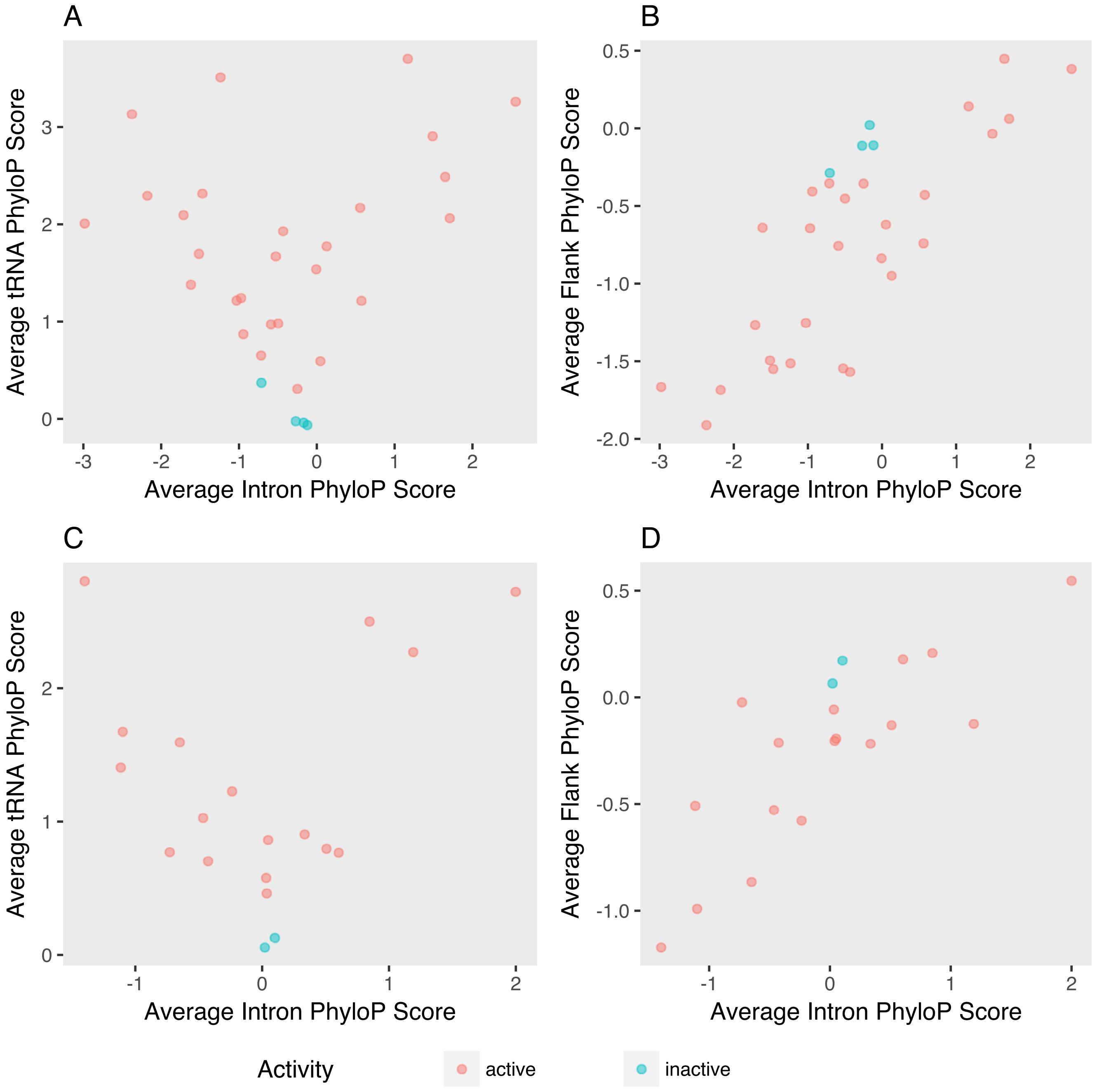

Fig. S1. Conservation of tRNA introns follows a similar pattern to conservation of tRNA flanking regions. A: The average PhyloP score across each mature human tRNA sequence is plotted against the average PhyloP score of each tRNA intron. **B**: The average PhyloP score of the flanking regions (the 40 bases up and downstream of each tRNA) of human tRNAs is plotted against the average PhyloP score of the corresponding tRNA intron. **C**: The average PhyloP score across each mature mouse tRNA sequence is plotted against the average PhyloP score of each tRNA intron. **D**: The average PhyloP score of the flanking regions (the 40 bases up and downstream of each tRNA) is plotted against the average PhyloP score of the corresponding intron for each mouse tRNA. The tRNA loci are colored by their activity classification (see Methods).

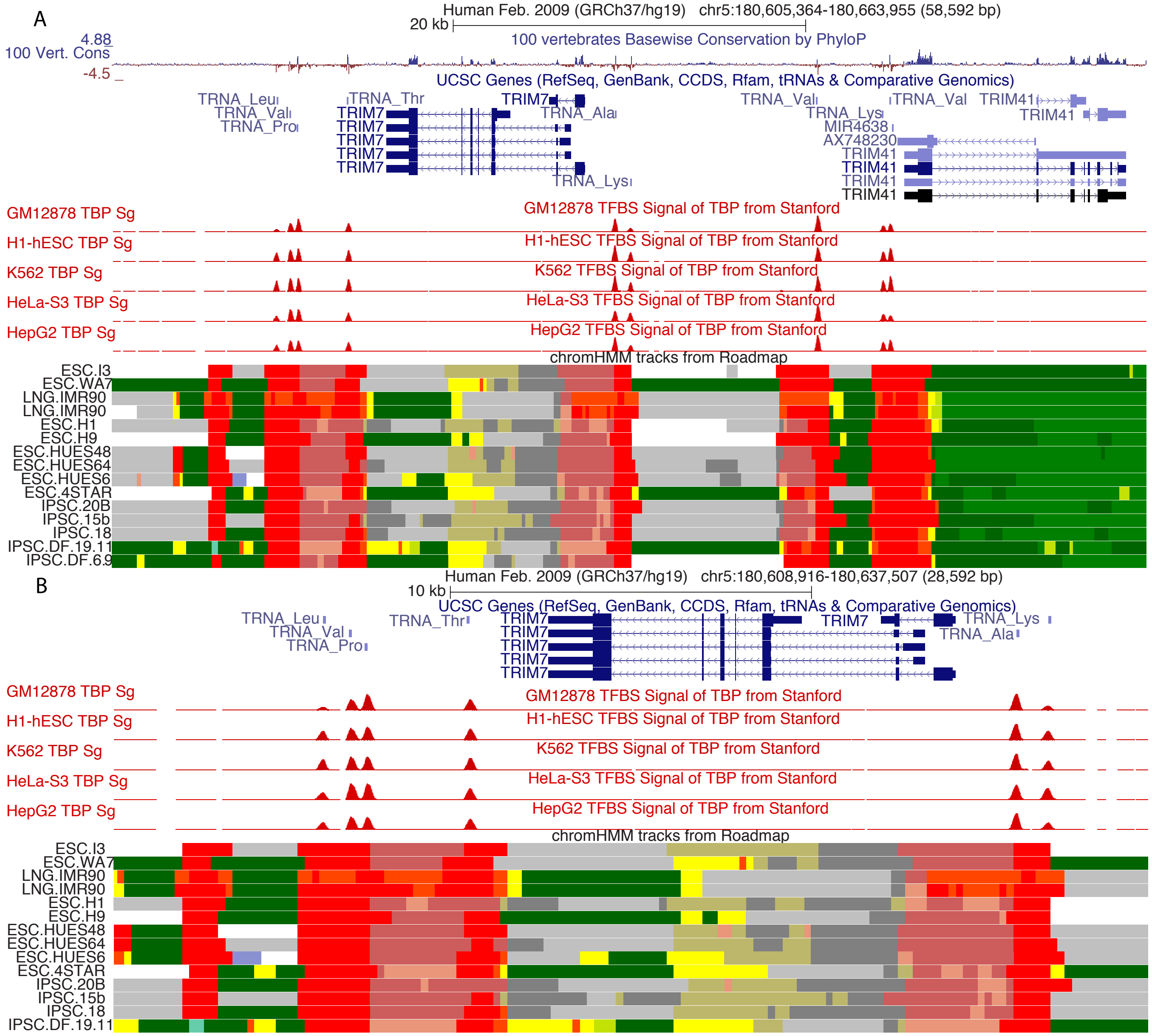
A
Human Feb. 2009 (GRCh37/hg19) chr5:180,605,364-180,663,955 (58,592 bp)
20 kb
4.88
100 Vert. Cons
-4.5
100 vertebrates Basewise Conservation by PhyloP
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
TRNA_Leu
TRNA_Val
TRNA_Pro
TRNA_Thr
TRIM7
TRNA_Ala
TRNA_Lys
TRNA_Val
TRNA_Lys
MIR4638
AX748230
TRIM41
GM12878 TBP Sg
GM12878 TFBS Signal of TBP from Stanford
H1-hESC TBP Sg
H1-hESC TFBS Signal of TBP from Stanford
K562 TBP Sg
K562 TFBS Signal of TBP from Stanford
HeLa-S3 TBP Sg
HeLa-S3 TFBS Signal of TBP from Stanford
HepG2 TBP Sg
HepG2 TFBS Signal of TBP from Stanford
chromHMM tracks from Roadmap
ESC.I3
ESC.WA7
LNG.IMR90
LNG.IMR90
ESC.H1
ESC.H9
ESC.HUES48
ESC.HUES64
ESC.HUES6
ESC.4STAR
IPSC.20B
IPSC.15b
IPSC.18
IPSC.DF.19.11
IPSC.DF.6.9
B
Human Feb. 2009 (GRCh37/hg19) chr5:180,608,916-180,637,507 (28,592 bp)
10 kb
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
TRNA_Leu
TRNA_Val
TRNA_Pro
TRNA_Thr
TRIM7
TRNA_Lys
TRNA_Ala
GM12878 TBP Sg
GM12878 TFBS Signal of TBP from Stanford
H1-hESC TBP Sg
H1-hESC TFBS Signal of TBP from Stanford
K562 TBP Sg
K562 TFBS Signal of TBP from Stanford
HeLa-S3 TBP Sg
HeLa-S3 TFBS Signal of TBP from Stanford
HepG2 TBP Sg
HepG2 TFBS Signal of TBP from Stanford
chromHMM tracks from Roadmap
ESC.I3
ESC.WA7
LNG.IMR90
LNG.IMR90
ESC.H1
ESC.H9
ESC.HUES48
ESC.HUES64
ESC.HUES6
ESC.4STAR
IPSC.20B
IPSC.15b
IPSC.18
IPSC.DF.19.11

Fig. S2. Intergenic regions in clusters of active tRNAs appear more divergent than intergenic regions in clusters of inactive tRNAs. A: A screenshot of the UCSC Genome Browser (2) showing bases chr5:180,605,364-180,663,955 in the hg19 assembly. The tRNAs surrounding the TRIM7 and TRIM41 genes are in regions of active transcription, indicated by the red regions in the chromHMM tracks (bottom) and the ChIP-Seq data shown in the Transcription-Factor Binding Site (TFBS) TATA-Binding Protein (TBP) tracks (8–10). Notably, the PhyloP track shows marked divergence flanking these tRNA genes. Importantly, this screenshot shows roughly 65kb, indicating that the divergence surrounding the tRNAs extends well past the 40 nucleotide flanking regions shown in Figure 1. **B**: Here, we have zoomed in on a subset of the area shown in A, showing bases chr5:180,608,916-180,637,507. Note that the intergenic region between the Pro and Thr tRNA genes show noticeable divergence, coinciding with the red regions in the chromHMM tracks (bottom). These screenshots are all set to the same scale for PhyloP (11), ranging from -4.88 to 4.5.

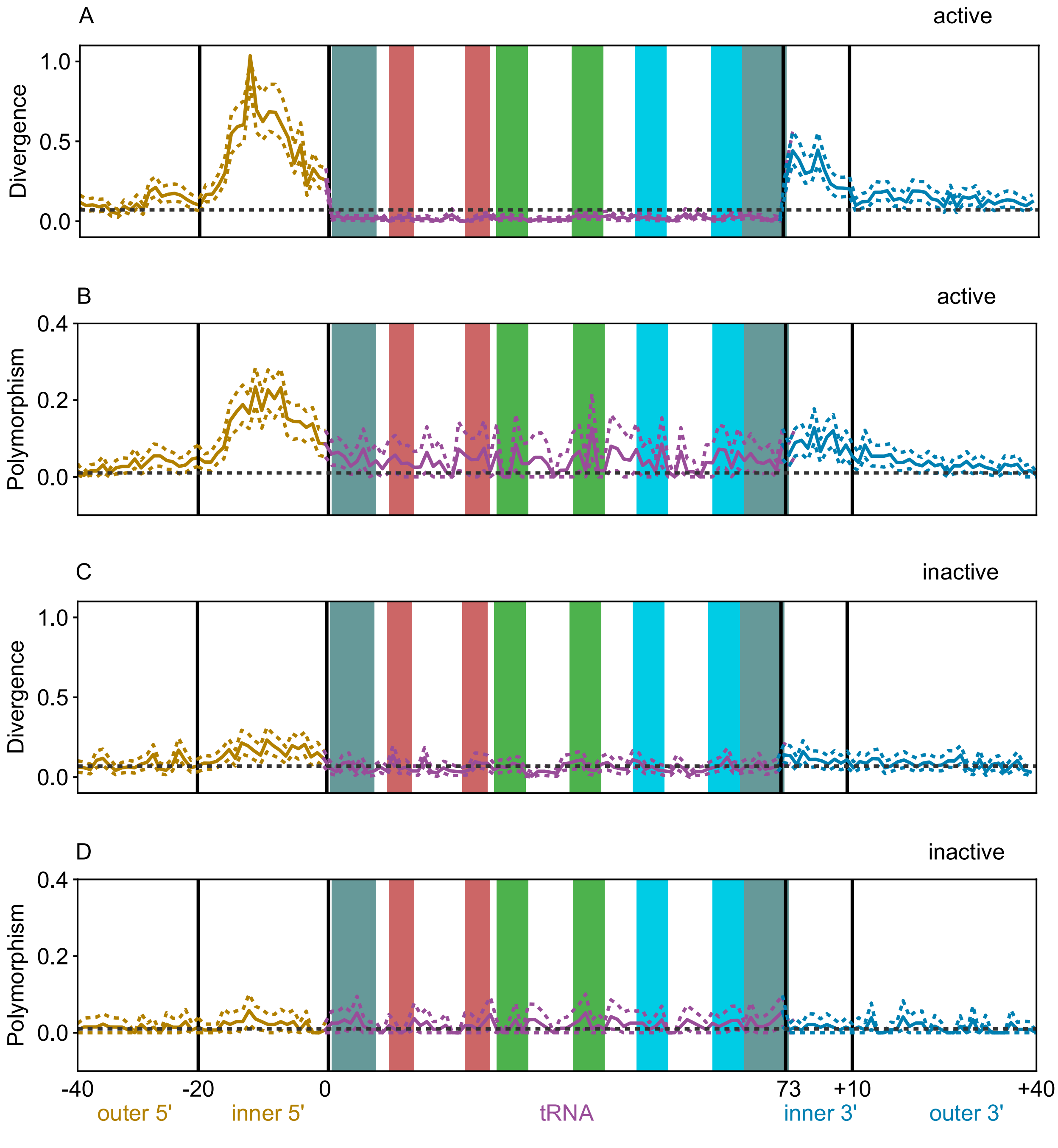

Fig. S3. Flanking region variation, tRNA polymorphism and tRNA conservation are more pronounced at active human tRNA loci. Divergence at non-gap alignments between the hg19 and rheMac2 genomes at each position within active (**A**) and inactive (**C**) tRNAs and their flanking regions. The frequency at which each position within active (**B**) and inactive (**D**) tRNA loci has a low-frequency SNP (minor allele frequency less than or equal to 0.05). 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), anticodon stem (green) and T-stem (blue) are highlighted within the tRNA (12), as in Figure 1. 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 plot depict 95% confidence intervals, calculated by bootstrapping by tRNA loci.

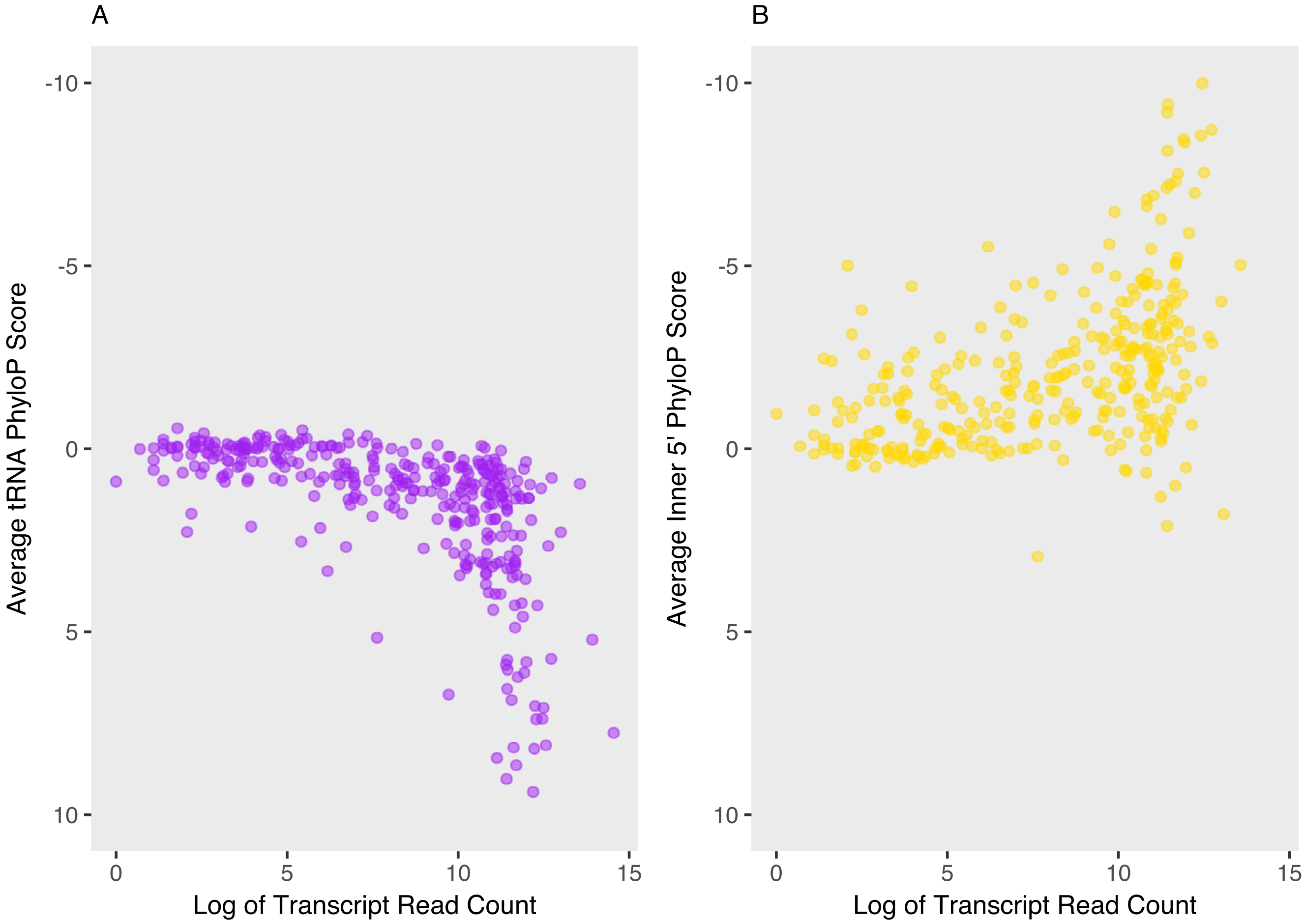

Fig. S4. Sets of human tRNA genes that produce more tRNAs in HEK293T cells tend to have highly conserved gene sequences and highly divergent flanking regions. A: The sum of the average PhyloP scores for all tRNAs corresponding to the same mature tRNA sequence is plotted against the log of the HEK293T cell RNA-seq read count for that tRNA (13). **B**: The sum of the average PhyloP scores across the flanking regions for all tRNAs corresponding to the same mature tRNA sequence is plotted against the log of the HEK293T cell RNA-seq read count for that tRNA (13). These plots differ from Figure 2C and 2D in that those figures excluded read counts for tRNAs encoded at multiple loci. Here, we are adding the average PhyloP scores for these tRNAs to account for their implied joint contribution to these read counts.

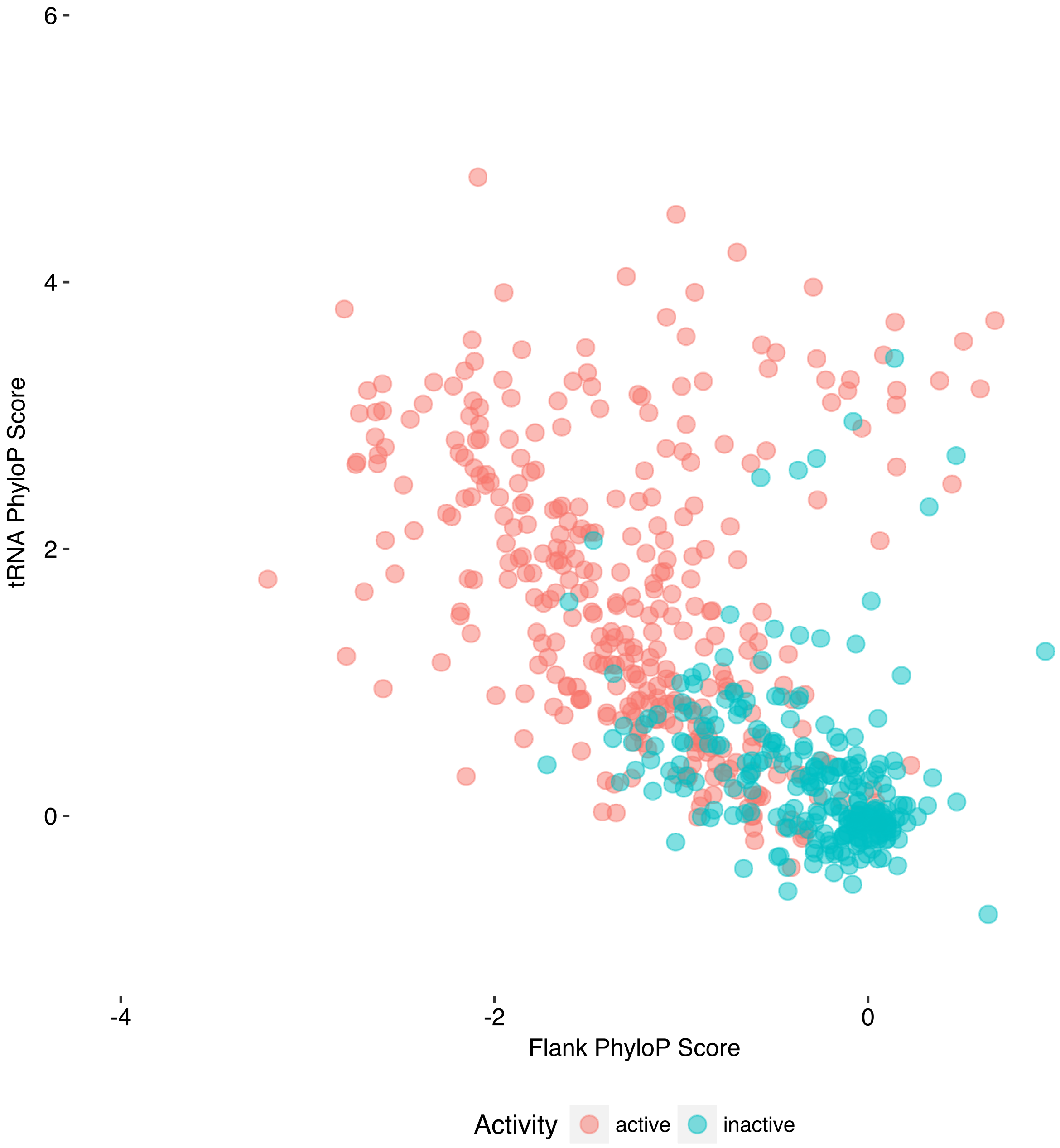

Fig. S5. Active and inactive human tRNAs can be distinguished by their PhyloP scores. Average PhyloP scores across the gene sequence is plotted against the average PhyloP score across the flanking regions (40 bases up and downstream of each gene) for each human tRNA locus. Inactive tRNA loci (blue) tend to have PhyloP scores near 0 for both their flanking regions and gene sequences, and active tRNA loci (red) tend to have negative PhyloP scores in their flanking regions and positive PhyloP scores in their gene sequences (11).

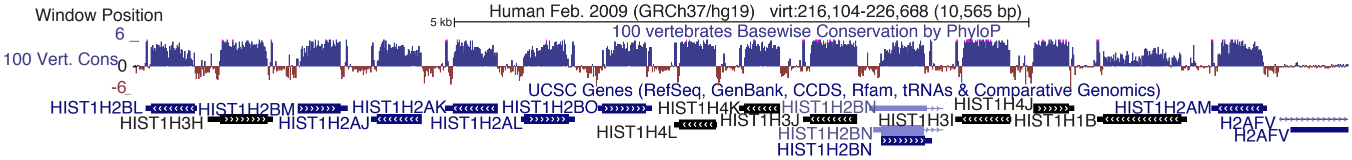

Fig. S6. Loci encoding histone proteins show qualitatively similar conservation patterns to tRNA loci. A genome browser screenshot using the multi-region viewer, depicting the PhyloP scores across the loci of several representative single-exon histone protein coding genes in the hg19 genome (2). The arrows on each gene point in the 5' to 3' direction. The PhyloP scale for this screenshot ranges from -6 to 6.

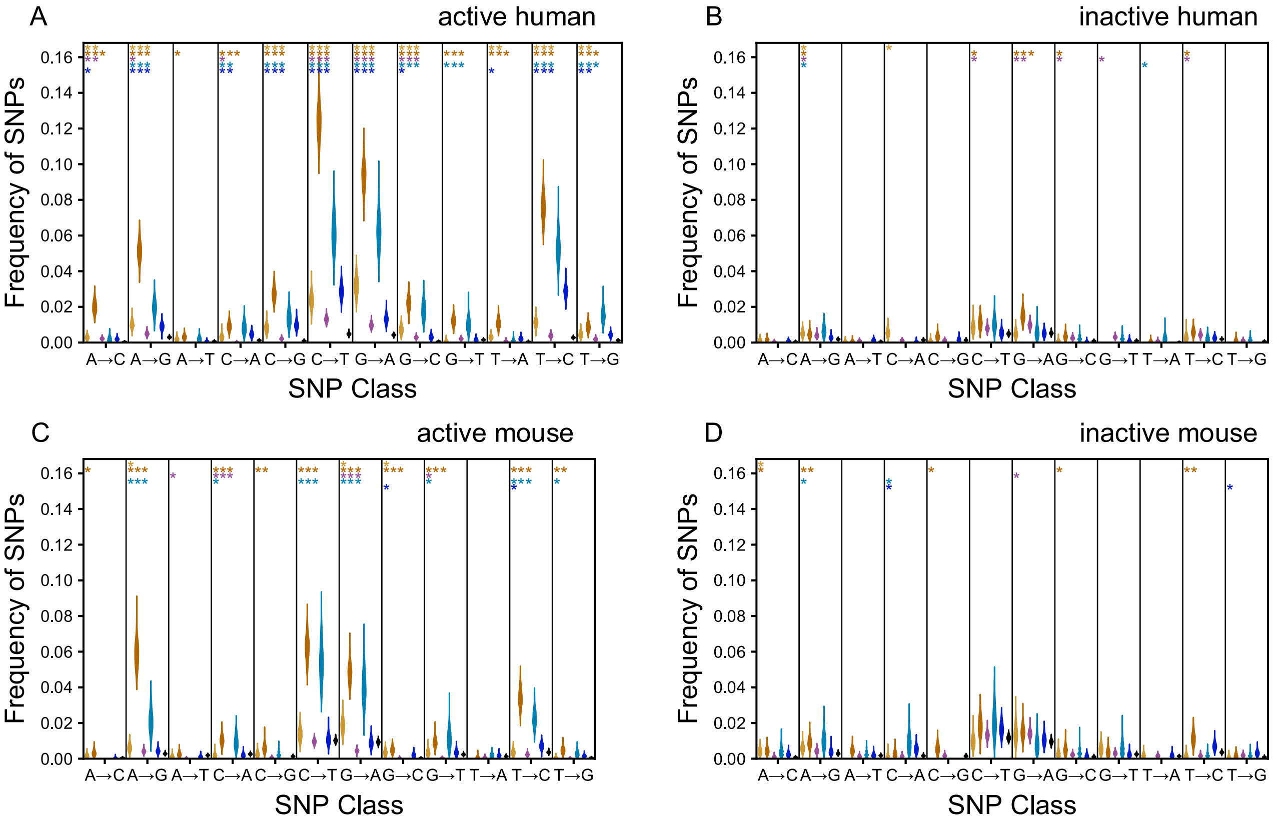

Fig. S7. Excess SNP classes consistent with TAM are more common in active tRNAs in both human and mouse. The distribution of each class of low-frequency polymorphisms, defined as a SNP with a minor allele frequency less than or equal to 0.05, is shown by region across active human tRNAs (**A**), inactive human tRNAs (**B**), active mouse tRNAs (**C**) and inactive mouse tRNAs (**D**). As in Figure 3, 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 at the top of each panel. One star represents a p value ≤ 0.05, two stars represents a p value ≤ 0.005, and three stars represents a p value ≤ 0.0005.

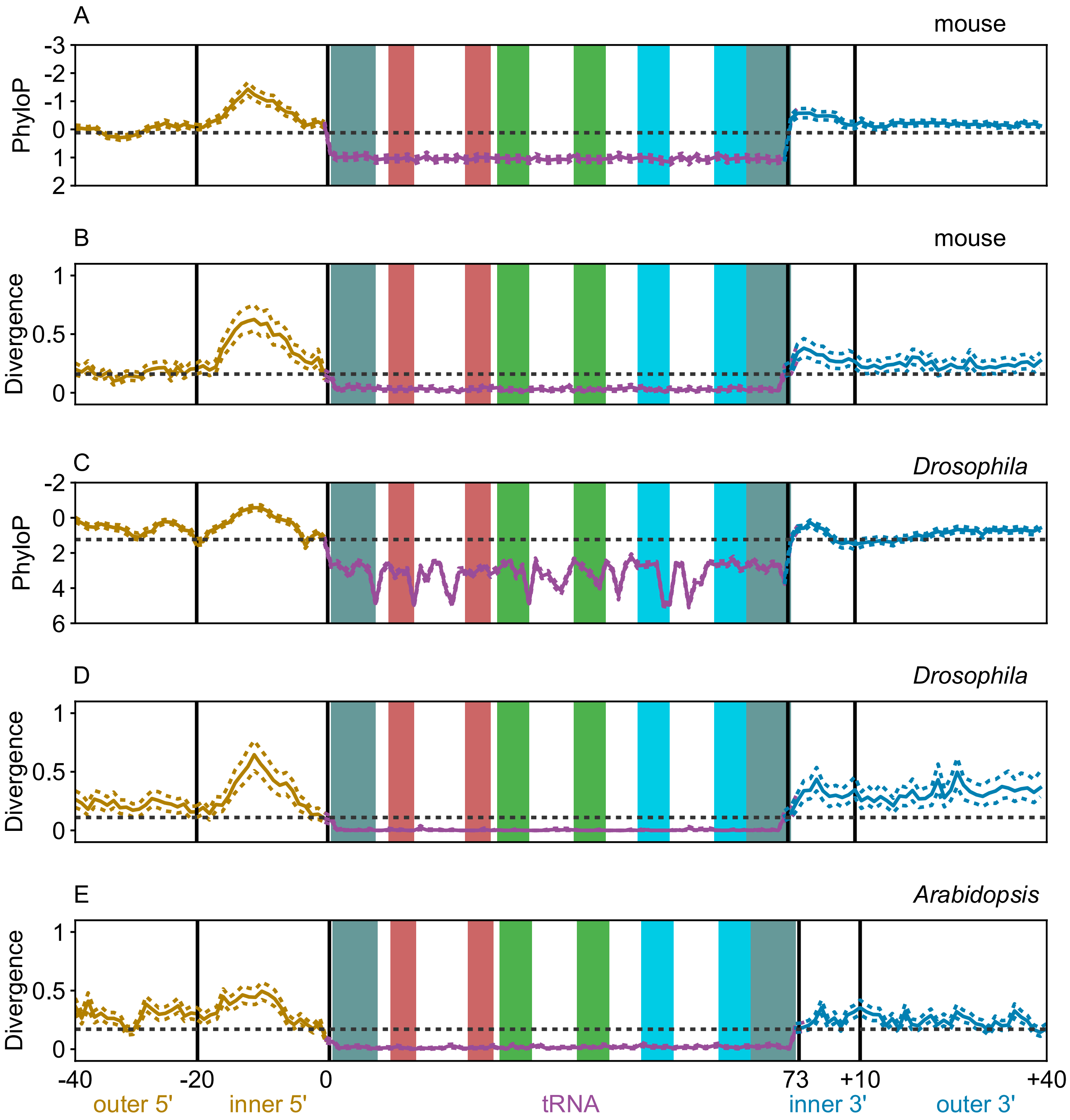

Fig. S8. Patterns of conservation and divergence at tRNA loci are consistent across *M. musculus*, *D. melanogaster*, and *A. thaliana*. The average PhyloP score is plotted for each position within the tRNA and flank, across all tRNA loci in mouse (**A**; comparing mouse to 60 vertebrate species) and *Drosophila melanogaster* (**C**; comparing *D. melanogaster* to 27 insect species). The frequency at which each position within tRNAs and flanks differs between mouse and rat (**B**), *D. melanogaster* and *D. yakuba* (**D**), and *A. thaliana* and *A. lyrata* (**E**), across all mouse, *D. melanogaster* and *A. thaliana* tRNA loci, respectively. No PhyloP information was available for *A. thaliana.* 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), anticodon stem (green) and T-stem (blue) are highlighted within the tRNA (12), as in Figure 1. 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 plot depict 95% confidence intervals, calculated by bootstrapping by tRNA loci.

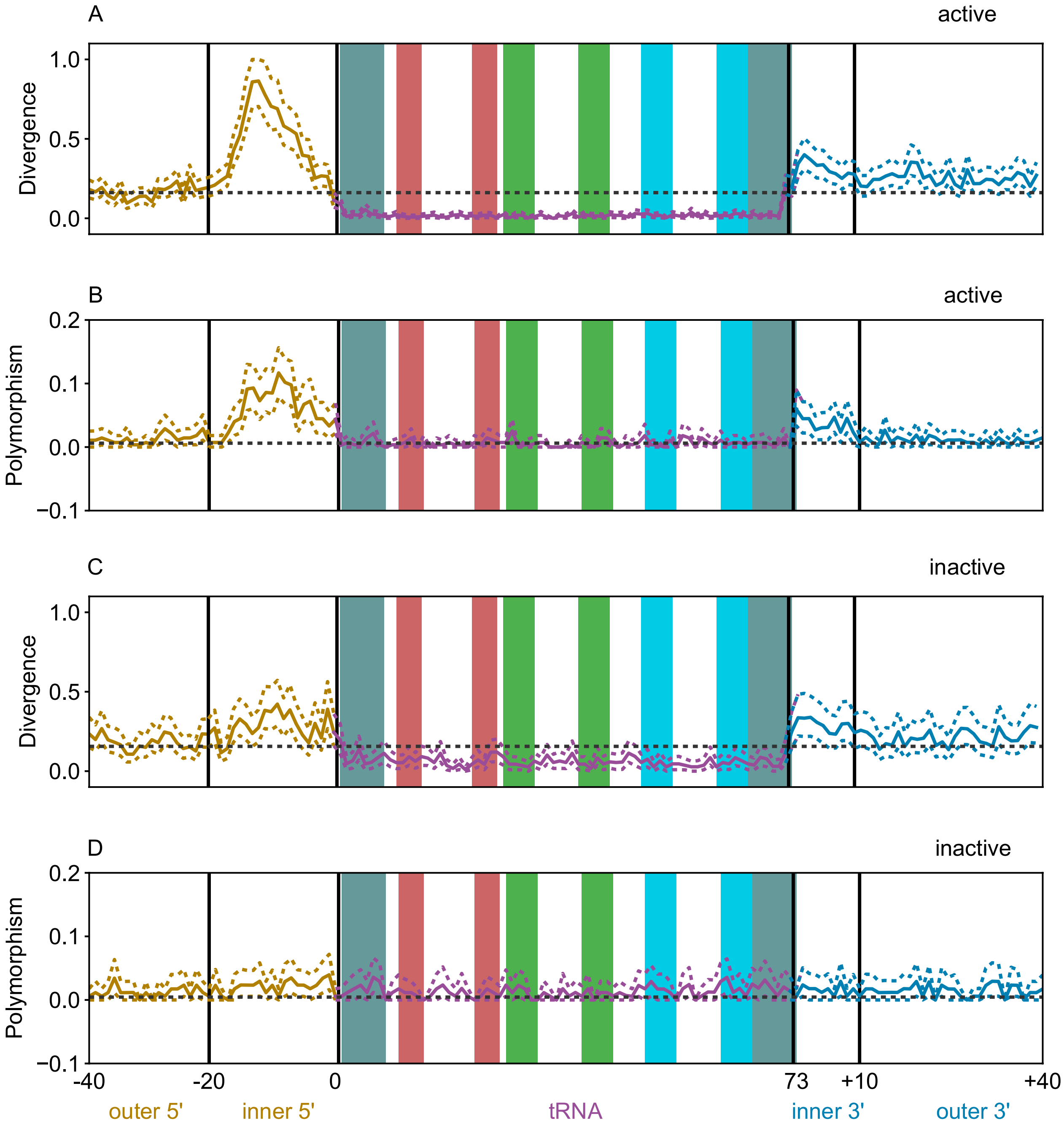

Fig. S9. Patterns of tRNA locus variation in active and inactive mouse tRNA loci are consistent with the patterns observed in human tRNA loci. Divergence at non-gap alignments between the mm10 (mouse) and rn5 (rat) genomes at each position within active (**A**) and inactive (**C**) mouse tRNAs and their flanking regions. The frequency at which each position within tRNAs and flanks has a low-frequency SNP (minor allele frequency less than or equal to 0.05) across active (**B**) and inactive (**D**) mouse 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), anticodon stem (green) and T-stem (blue) are highlighted within the tRNA (12), as in Figure 1. 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 plot depict 95% confidence intervals, calculated by bootstrapping by tRNA loci.

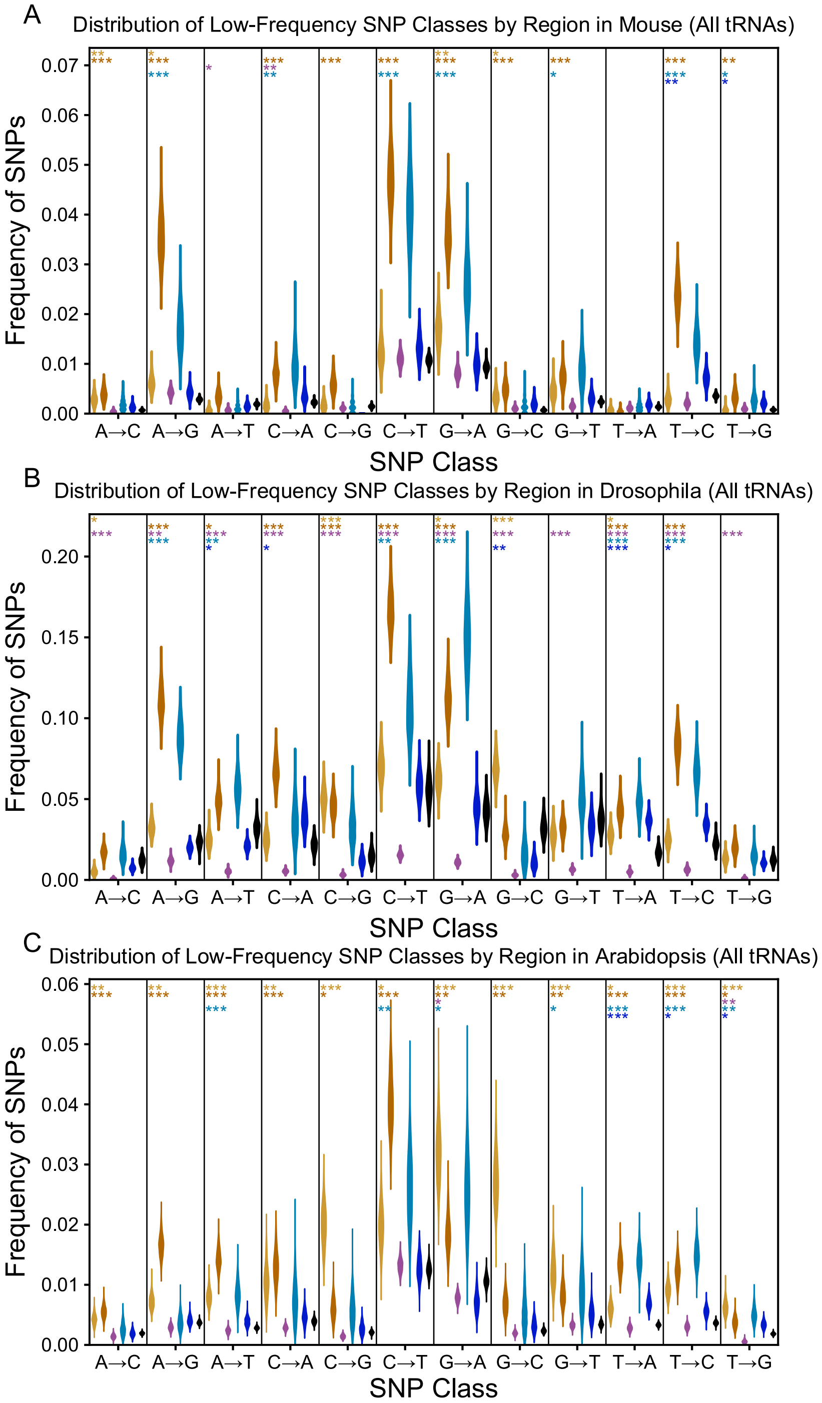

Fig. S10. SNP classes frequencies follow the same patterns in all model organisms studied. The distribution of each class of low-frequency polymorphisms, defined as a SNP with a minor allele frequency less than or equal to 0.05, is shown by region across all mouse (**A**), *D. melanogaster* (**B**) and *A. thaliana* (**C**) 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 ≤ 0.05, two stars represents a p value ≤ 0.005, and three stars represents a p value ≤ 0.0005.

Additional data table S1 (DatasetS1.xlsm)

Many other RNA species have similar PhyloP averages to tRNA loci across their gene sequences and flanking regions. We have compared the average PhyloP scores of several Pol2- and Pol3-transcribed RNA genes to the initiator methionine tRNA gene family, which is included as a general representative of tRNA genes.

Additional data table S2 (DatasetS2.csv)

Coordinates and activity classifications for all human tRNA gene sequences, flanking regions and untranscribed reference regions.

Additional data table S3 (DatasetS3.csv)

Coordinates and activity classifications for all mouse tRNA gene sequences and untranscribed reference regions.

Additional data table S4 (DatasetS4.csv)

Coordinates for all *Arabidopsis thaliana* tRNA gene sequences and untranscribed reference regions.

Additional data table S5 (DatasetS5.csv)

Coordinates for all *Drosophila melanogaster* tRNA gene sequences and untranscribed reference regions.

References

1. Gibbs RA, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *science* 316(5822):222–234.
2. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res.* 12(6):996–1006.
3. Kersey PJ, et al. (2015) Ensembl genomes 2016: more genomes, more complexity. *Nucleic acids research* 44(D1):D574–D580.
4. Consortium DG, , et al. (2007) Evolution of genes and genomes on the drosophila phylogeny. *Nature* 450(7167):203.
5. Jukes TH, Cantor CR (1969) Evolution of protein molecules. *Mammalian Protein Metabolism* pp. 21–132.
6. Messer PW (2009) Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* 182(4):1219–1232.
7. 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.
8. Myers RM, et al. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9(4):e1001046.
9. Dunham I, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
10. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* 26(12):1351–1359.
11. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20(1):110–121.
12. 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.
13. Zheng G, et al. (2015) Efficient and quantitative high-throughput tRNA sequencing. *Nat. Methods* 12(9):835–837.