Structural and Genetic Determinants of Convergence in the *Drosophila* tRNA Structure-Function Map

Julie Baker Phillips[1,2] and David H. Ardell[*,1,3]
[1]Quantitative and Systems Biology Program, University of California, Merced, CA 95343
[2]Department of Biology, Cumberland University, 1 Cumberland Square, Lebanon, TN 37087
[3]Department of Molecular and Cell Biology, University of California, Merced, CA 95343
***Corresponding author:** E-mail: dardell@ucmerced.edu
Associate Editor: Xxxx Xxxxxxx

Abstract

The evolution of tRNA multi-gene families remains poorly understood, exhibiting unusual phenomena such as functional conversions of tRNA genes through anticodon shift substitutions. We reannotated tRNA genes from twelve *Drosophila* species, incorporating previously identified ortholog sets to compare substitution rates across tRNA bodies at single-site resolution. All rapidly evolving sites fell in the same metal-ion-binding pocket at the interface of the two tRNA stacked helical domains. We applied our tRNA Structure-Function Mapper (tSFM) method independently to each *Drosophila* species and one outgroup species *Musca domestica*. We found that, although predicted tRNA structure-function maps are generally highly conserved in flies, one tRNA Class-Informative Feature (CIF) within the rapidly-evolving ion-binding pocket — Cytosine 17 (C17), ancestrally informative for lysylation identity — independently gained asparaginylation identity and substituted in parallel across $tRNA^{Asn}$ paralogs at least three times in evolution of the genus. In *D. melanogaster*, approximately 15 $tRNA^{Lys}$ and $tRNA^{Asn}$ genes co-occur in heterologous clusters, suggesting that both heterologous gene conversion and structural similarities of the closely related asparaginyl-tRNA synthetase (AsnRS) and lysyl-tRNA synthetase (LysRS) proteins may have played a role in these changes. A previously identified Asn-to-Lys anticodon shift substitution in *D. ananassae* may have arisen to compensate for the convergent and parallel gains of C17 in $tRNA^{Asn}$ paralogs of the *D. ananassae* lineage. Our results underscore the functional and evolutionary relevance of our tRNA structure-function map predictions and illuminate multiple genomic and structural factors contributing to rapid, parallel and compensatory evolution of tRNA multi-gene families.

Key words: class-informative feature; ion-binding pocket; parallel substitutions; convergent evolution; structure-function map

Article

Introduction

Transfer RNAs (tRNAs) were the first family of RNAs to be directly sequenced (Holley, 1965) and the first to be solved by X-ray crystallography (Holley *et al.*, 1965). Historically, algorithms to estimate phylogeny and substitution patterns were tested on tRNA genes (Cedergren *et al.*, 1981; Eigen *et al.*, 1989). Early in the genome era, it was reported that tRNA

© The Author 2013. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please email: journals.permissions@oup.com

genes can evolve to switch their functional (aminoacylation) identities through anticodon shift substitutions (Saks *et al.*, 1998). However, the small sizes and high similarities of tRNA genes pose obstacles to inferring their orthology, which is needed to better understand the evolutionary processes underlying functional turnover of tRNA genes. An important step forward came from the "micro-syntenic" approach to infer tRNA gene orthology using flanking sequences, first applied in *Drosophila* (Rogers *et al.*, 2010) and later to other eukaryotes (Rogers and Griffiths-Jones, 2014). These studies revealed that functional turnover of tRNA genes through anticodon shift substitutions is more frequent and widespread than previously known. However, Rogers and Griffiths-Jones (2014) were unable to discern whether anticodon shift substitutions occur more often within or between the two conserved and ancient superfamilies of aminoacyl-tRNA synthetases (aaRSs), called Classes I and II (Eriani *et al.*, 1990). The two superfamilies may be further divided into three subclasses each (Cusack, 1997), all of which pre-date the divergence of bacteria, archaea and eukarotes as exemplified by the consistency with which aaRSs may be used to root the statistical Tree of Life (Brown and Doolittle, 1995).

More recent advances in orthology estimation for tRNAs exploit positional homology and the organization of tRNA genes as repeated elements in tandem gene arrays, and revealed both the great extent of functional turnover and important role of gene conversion in tRNA evolution (Velandia-Huerto *et al.*, 2016). It has also been shown that initiator tRNAs, the distinct functional class of tRNA initiating protein synthesis at start codons, may evolve through anticodon shift substitutions and functional conversion from elongator tRNAs (Bhattacharyya and Varshney, 2016). The advances in ortholog estimation for tRNA genes have made it possible for the first time to undertake a detailed analysis of substitution rates and patterns in tRNA genes, which is one of two requirements to understand how tRNA genes evolve to switch functions. The second requirement is a means to predict the functional significance of tRNA sequence features outside the anticodon — what we call the "tRNA structure-function map."

In earlier work, we developed an approach to predict the tRNA structure-function map from genome sequences alone, based on structure-conditioned functional information statistics (Freyhult *et al.*, 2006). Our algorithm recovered known tRNA identity elements and predicted new ones subsequently verified as functionally important in tRNA biogenesis and aminoacylation (Bailly *et al.*, 2006; Sethi *et al.*, 2009). Later, in a study of alphaproteobacterial evolution, we reported that the features predicted by our algorithm, which we call tRNA Class-Informative Features (CIFs), are phylogenetically informative and robust to compositional

convergence and horizontal gene transfer (Amrine *et al.*, 2014). We further developed this approach into a machine learning "phyloclassifier" to robustly predict the cyanobacterial progenitor of plastids (Lawrence *et al.*, 2019). More recently we have shown that that tRNA structure-function maps, including Class-Informative Base-Pairs (CIBPs) and Class-Informative Mis-Pairs (CIMPs), are highly conserved both within trypanosomes and betweeen trypanosomes and humans, and also that tRNA CIFs can be used to predict functional differences between homologous enzymes from trypanosomes and humans (Kelly *et al.*, 2020). We also showed evidence that tRNA gene clusters are conserved in trypanosomes and show clear evidence of duplication, deletion, and rearrangements (Kelly *et al.*, 2020).

By combining data from multiple ortholog sets of microRNA genes, the *Drosophila* Twelve Genomes Consortium yielded one of the first analyses comparing evolutionary rates over different structurally-defined categories of sites in a non-coding RNA family (Drosophila 12 Genomes Consortium, 2007). In the present work, we extend this approach, exploiting the conservation and high conformity (Wolfson *et al.*, 2001) of tRNA genes to estimate and compare evolutionary rates across different structural components and at individual single- and paired-sites in tRNA genes from the *Drosophila* Twelve Genomes. We developed a strategy to optimize the selection of species to maximize the length of concatenated alignments across tRNA ortholog sets. Analyzing our optimized alignment using MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist *et al.*, 2012), we discovered that one of several metal ion-binding pockets in *Drosophila* tRNAs has evolved rapidly, and that this rapid evolution is directly related to repeated convergent conversion of a tRNA CIF from one tRNA family to another, in parallel over multiple tRNA genes on different chromosomes. We were able to identify multiple structural and genomic factors that are likely to have contributed to this convergent evolution. Our results suggest for the first time that anticodon shift substitutions may play a compensatory role in evolution of the tRNA structure-function map.

Results

D. simulans Contains a Non-Canonical Initiator tRNA Gene Created by an Anticodon Shift Substitution

We built a custom database of reannotated tRNA genes from *Drosophila* twelve genomes FlyBase release 2008-07 (Tweedie *et al.*, 2009) downloaded from FlyBase on October 19, 2011, which integrates orthology annotations from Rogers *et al.* (2010), initiator tRNA gene predictions from TFAM 1.3 (Tåquist *et al.*, 2007) and which filters gene records by COVE scores from tRNAscan-SE 1.3.1 (Lowe and Eddy, 1997) as described in Methods below. Our initiator tRNA gene predictions were consistent with experimentally based annotations in *D. melanogaster* and across species, in that if TFAM annotated any gene

in any ortholog set from Rogers and Griffiths-Jones (2014); Rogers *et al.* (2010) as an initiator tRNA gene, then every other gene in that ortholog set would also be independently annotated by TFAM as an initiator tRNA gene. We annotated between five and seven initiator tRNA genes per *Drosophila* species we examined, for a total of 70 annotated initiator tRNA genes across all species, of which 66 belonged to the ortholog sets from Rogers *et al.* (2010). Our tRNA gene annotations are provided in Supplementary Data.

We found that one alloacceptor anticodon shift substitution (which converts the functional identity of a tRNA from one amino acid to another), which was previously reported and validated as $\text{tRNA}^{\text{Met}}_{\text{CAU}} \rightarrow \text{tRNA}^{\text{Thr}}_{\text{CGU}}$ (or in their notation, CAT:M→CGT:T) in *Drosophila simulans* (Rogers and Griffiths-Jones, 2014; Rogers *et al.*, 2010), in fact represents the evolution of a non-canonical initiator tRNA gene in *D. simulans* (non-canonical in the sense that its predicted tRNA product does not contain the methionine anticodon CAU). This non-canonical initiator tRNA gene in *D. simulans* with FlyBase gene ID gn0256165 and transcript ID tr0296323, evolved as part of Ortholog Set 183 from Rogers *et al.* (2010). This ortholog set contains initiator tRNA orthologs in all eight species of Sophophora excluding *D. willistoni*. The *D. simulans* ortholog carries only a single A→G anticodon shift substitution at Sprinzl position 35 (Sprinzl coordinates are a standardized coordinate system for the consensus universal secondary structure of tRNAs as well as conserved, more sub-function-specific structures like the long variable arms of tRNA^{Leu} and tRNA^{Ser} (Sprinzl *et al.*, 1998)). The non-canonical initiator tRNA gene in *D. simulans* is also identical in sequence to all other initiator tRNA genes we annotated in *D. simulans*, except for the previously noted anticodon shift substitution at Sprinzl position 35.

Optimizing Species Exclusion to Maximize Gap-Free Alignment Length of Concatenated tRNA Orthology Sets

We next prepared to undertake a global analysis of site-specific substitution rates and patterns in *Drosophila* tRNA genes. To do this we analyzed a total of 753 orthologous sets from Rogers *et al.* (2010) encompassing 3218 unique FlyBase tRNA transcript IDs. We found the representation of species to be uneven across the Rogers *et al.* orthology sets. Only 47 sets, about 6%, contained orthologs from all 12 *Drosophila* species (Table 1). One species in particular, *D. willistoni*, was represented in very few sets. We proceeded to remove this species from our subsequent substitution rate analyses. We also removed orthology sets that contained isoacceptor or alloacceptor anticodon shift substitutions, uncertain functional annotations, or predicted pseudogenes. We wished to integrate data from the largest number of orthology sets that we could in order to compare substitution rates across sites. However, we also wished to avoid

Table 1. Statistics on alignments of concatenated tRNA orthology sets from subsets of *Drosophila* species

Species Subset	Species	Ortho. Sets[a]	Sites (gap-free)	Var. Sites (Partitions)[b]	P.I. Sites (Partitions)[c]
101111110000[d]	7	147	10878 (10598)	2729 (68)	27 (15)
110011100111[e]	8	85	6290 (6134)	1589 (67)	19(15)
111111110111[f]	11	83	6142 (5991)	1554 (67)	20 (15)
111111111111	12	47	3478 (3389)	896 (58)	16 (12)

[a] Number of tRNA orthology sets from Rogers *et al.* (2010)
[b] Number of variable gap-free sites (or variable site-partitions excluding gap-free sites, of 74) over concatenated alignment
[c] Number of parsimoniously informative gap-free sites (or partitions excluding gap-free sites, of 74) over concatenated alignment
[d] Sophophora excluding *D. simulans* and *D. willistoni*
[e] All species excluding *D. sechellia D. yakuba*, *D. persimilis* and *D. willistoni*
[f] All species excluding *D. willistoni*

having a large number of missing or unidentified orthologs in the concatenated alignments that we made from the Rogers *et al.* ortholog sets, so as to reduce the influence of missing data associated with large blocks of gap characters on our results. We therefore computed statistics of 562 species subsets containing seven or more of the eleven *Drosophila* species excluding *D. willistoni*, in order to find the species set that resulted in concatenated alignments of greatest gap-free length.

We found that a single species subset, labelled as "101111110000" in Table 1 and representing Sophophora excluding *D. simulans* and *D. willistoni*, was superior to others by several metrics. This species subset yielded the longest gap-free alignment concatenating the most orthology sets, the greatest numbers of variable and parsimoniously informative sites, and the greatest average number of pairwise differences per site among all alignments we examined. We proceeded to focus on this alignment for downstream analysis as well as two others for

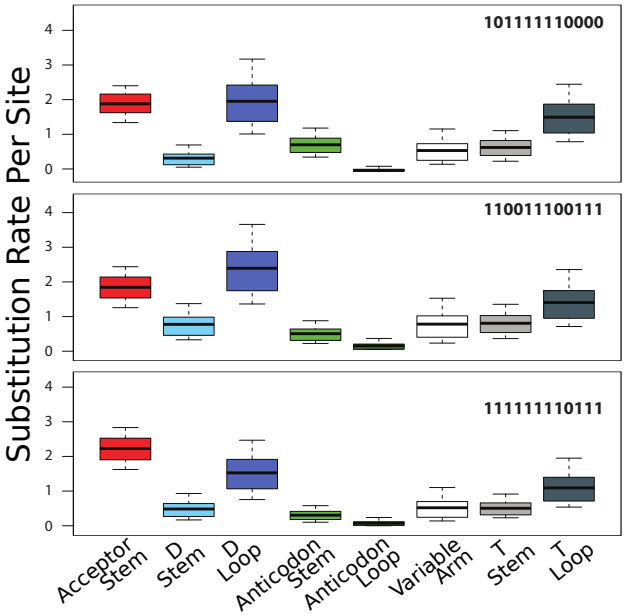

FIG. 1. Credible intervals, interquartile ranges and medians of substitution rates by secondary structural elements in *Drosophila* tRNAs as calculated in MrBayes 3.2.1 using the GTR+I/Double(GTR+I) models for loops/stems and the partitioned concatenated alignments indicated by bitsets as defined in Table 1.

the sake of comparison: a randomly picked one with a complementary pattern of species exclusion ("110011100111"), and one with nearly all species included, namely all eleven species except *D. willistoni* ("111111110111").

Substitution Patterns in *Drosophila* tRNAs

We compared substitution rates across different site-partitions of our concatenated alignments corresponding either to tRNA secondary

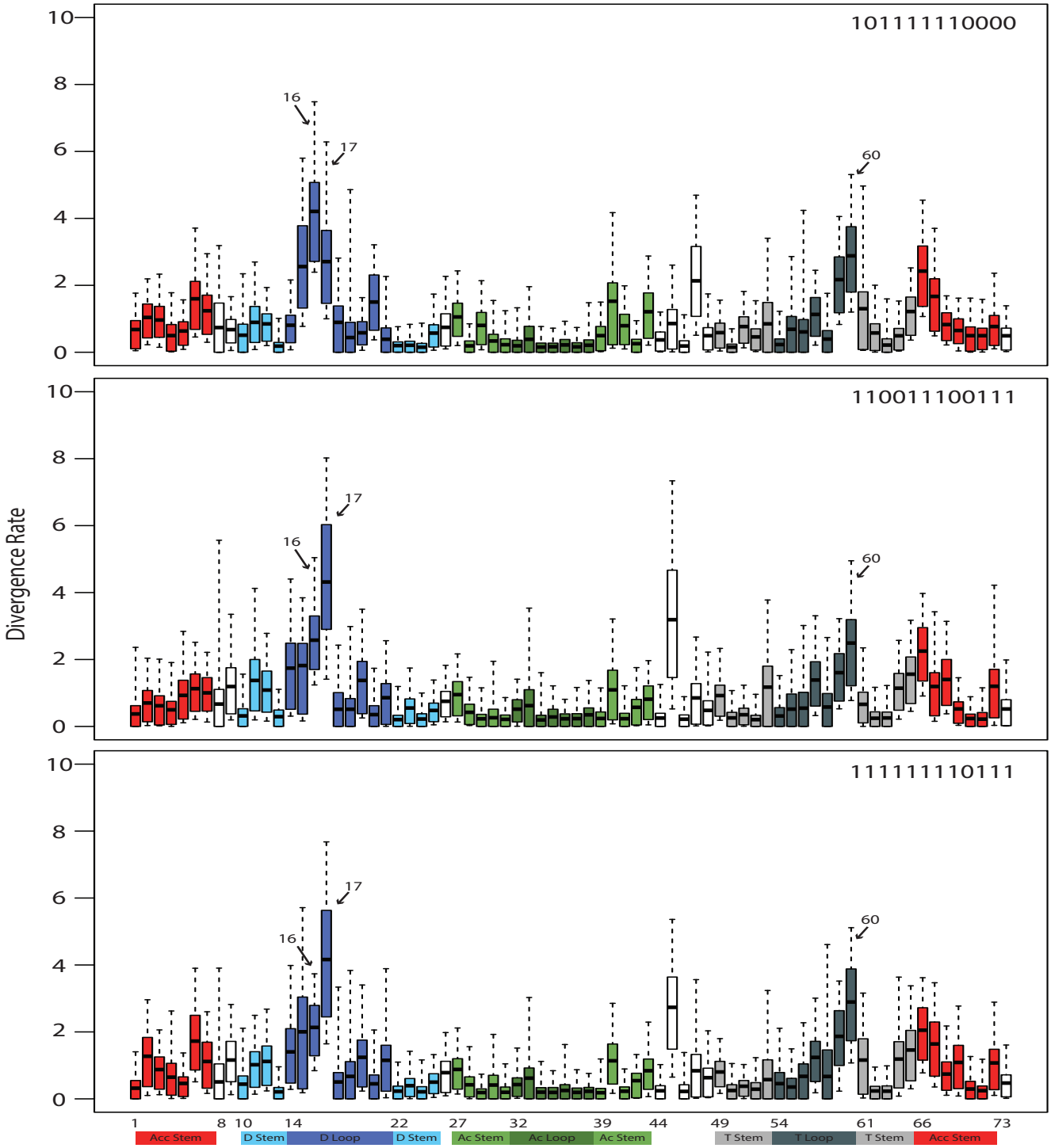

FIG. 2. Credible intervals, interquartile ranges and medians of substitution rates by individual site/Sprinzl coordinate in *Drosophila* tRNA genes as calculated in MrBayes 3.2.1 using the GTR+I models for individual sites and the partitioned concatenated alignments indicated by bit-sets as defined in Table 1.

structural elements or to individual sites defined by Sprinzl coordinates (Sprinzl and Vassilenko, 2005), using MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001; Ronquist *et al.*, 2012), aggregating over different functional classes of tRNA gene orthologs and using the fixed known species tree (*Drosophila* 12 Genomes Consortium, 2007). We fitted unpaired sites and loops with the

General Time Reversible (GTR) model (Lanave *et al.*, 1984; Rodríguez *et al.*, 1990; Tavaré, 1986) with invariant sites (GTR+I), and paired-sites and stems to Double(GTR)+I (Ronquist *et al.*, 2012). We obtained very similar results using other models including the Hasegawa-Kishino-Yano (HKY) model (Hasegawa *et al.*, 1985) and using Gamma-distributed site-rate variation.

In our coarse-grained analysis of substitution rates by secondary structural elements we observed that, regardless of which alignments or models we used, the largest substitution rates occur in the D-loop and acceptor stem followed closely by the T-loop (Figure 1). Considering the relative constraint that acceptor stems are expected to have from identity-driven interactions with proteins (Giegé *et al.*, 1998), the acceptor stem shows a surprisingly high rate of evolution. Furthermore, regardless of which alignments or models we used, in our analysis of substitution rates by individual sites, we found that three sites corresponding to Sprinzl coordinates 16, 17, and 60 showed markedly higher rates of substitution than any other site except site 45 (Figure 2).

Sprinzl coordinates 16, 17 and 60 are among eight sites that form an ion-binding pocket between the D- and T-arms (Sprinzl coordinates 15, 16, 17, 18, 19, 20, 59, and 60), at the interface of the two tRNA helical domains (Behlen *et al.*, 1990), as shown in Figure 3. This pocket was first described as site 1 in the original orthorhombic crystal form (Holbrook *et al.*, 1977) and site 3 in the monoclinic structure (Jack *et al.*, 1977). This ion-binding pocket was shown to potentially bind different metal ions including magnesium, cobalt, manganese, and lead (Behlen *et al.*, 1990; Holbrook *et al.*, 1977; Jack *et al.*, 1977; Shi and Moore, 2000). More recent structural work improving the resolution of the tRNA structure better resolved the structure of Sprinzl site 16, which is a fairly conserved uracil among eukaryotes (Marck and Grosjean, 2002), most likely post-translationally modified to dihydrouridine and demonstrated that this site is sensitive to ionic conditions (Shi and Moore, 2000). The hypergeometric probability that three sites of elevated substitution rate randomly occur within the eight sites that form the ion-binding pocket out of 74 possible sites is $\frac{\binom{8}{3}\binom{66}{0}}{\binom{74}{3}} = 8.6 \times 10^{-4}$.

In the alignment with the most species (111111110111), we identified that 45 orthology groups spanning 14 functional classes carry substitutions in sites 16, 17 and 60 of this ion-binding pocket, as shown in Table S1 of Supplementary Materials. One set showed signs of covariance between sites 16 and 17, and another between sites 17 and 60.

Chromosomal gene location was significantly associated with substitutions in pocket sites 16, 17 and 60 ($X^2 = 56.3266$, $N = 692$, $p = 6.96 \times 10^{-11}$), and extended pocket sites ($X^2 = 87.4858$, $N = 726$, $p < 2.2 \times 10^{-16}$, Supplementary Materials, Table S2. Substitutions in sites 16, 17

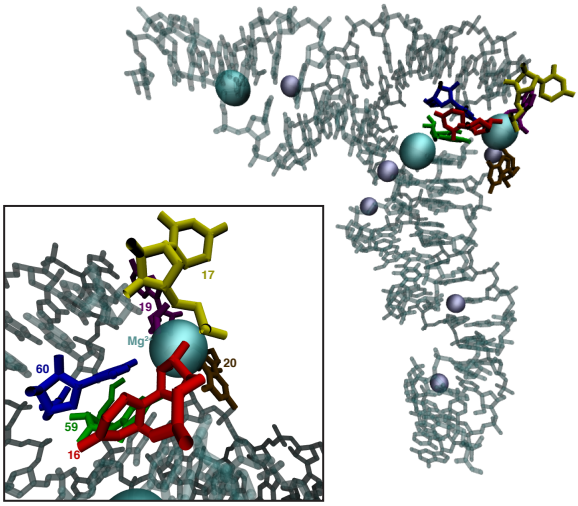

FIG. 3. Rapidly-evolving sites in Drosophila tRNA are part of a conserved, structurally plastic ion-binding pocket near the interface of the two tRNA helical domains. This image visualizes PDB structure accession ID 1EHZ (Shi and Moore, 2000) and was rendered in VMD 1.9.2 (Humphrey *et al.*, 1996).

or 60 of the pocket were not associated with alloacceptor class switches overall ($X^2=20.9616$, $N=672$, $p=0.52311$), nor were substitutions in the extended pocket ($X^2=23.2573$, $N=703$, $p=0.3873$. Ortholog sets containing anticodon shift substitutions more often exhibited substitutions in pocket sites 16,17, and 60 ($X^2=8.8797$, $N=713$, $p=0.002884$), but not extended pocket sites ($X^2=2.7304$, $N=696$, $p=0.09846$).

We provide estimates of transition and transversion rates of *Drosophila* tRNAs in Supplemetary Materials Table S3.

Class-Informative Base-Pairs and Base-Mispairs in *Drosophila melanogaster*

We re-implemented our previously published algorithm to estimate the structure-conditioned functional information statistic of Freyhult *et al.* (2006) in Python and extended it to compute the information for all sixteen possible base-pair or base mis-pair features in the consensus secondary structure of tRNAs, given a structural alignment and structural and functional annotations of tRNA gene data. The software, called tSFM for "tRNA Structure-Function Mapper," uses a permutation-based approach to calculate the significance of the tRNA CIFs that it predicts, and corrects for multiple comparisons by computing various measures of the Family-Wise Error Rate (FWER) or False Discovery Rate, including that of Benjamini and Hochberg (Benjamini and Hochberg, 1995).

We filtered our reannotated tRNA gene data from *Drosophila melanogaster* for tRNA genes with tRNAscan-SE 1.3.1 COVE scores of at least 50 bits and removed tRNA genes of indeterminate function and selenocysteine tRNA genes, leaving 288 tRNA genes remaining. We then computed tRNA Class-Informative Base Pairs and Base Mis-Pairs from these data, with only those features with a Benjamini-Hochberg False Discovery Rate of 5% shown in Fig. 4. A great deal of functional information appears to reside in paired features. Several base-mispairs are functionally informative, including U13:U22 in the D-arm associated to both $tRNA^{Pro}$ and $tRNA^{Val}$ (contained in 32 tRNA genes of 32 total), U31:U39 in the C-arm contained in all six $tRNA^{Met}$ genes, C50:U64 in the T-arm contained in nine of 11 $tRNA^{Ile}$ genes, G13:A22 in the D-arm contained in all 42 $tRNA^{Ser}$ and $tRNA^{Leu}$ genes, and a G4:G69 mis-pair in the acceptor stem associated with nine of 16 $tRNA^{Thr}$ genes. We show a plot of the

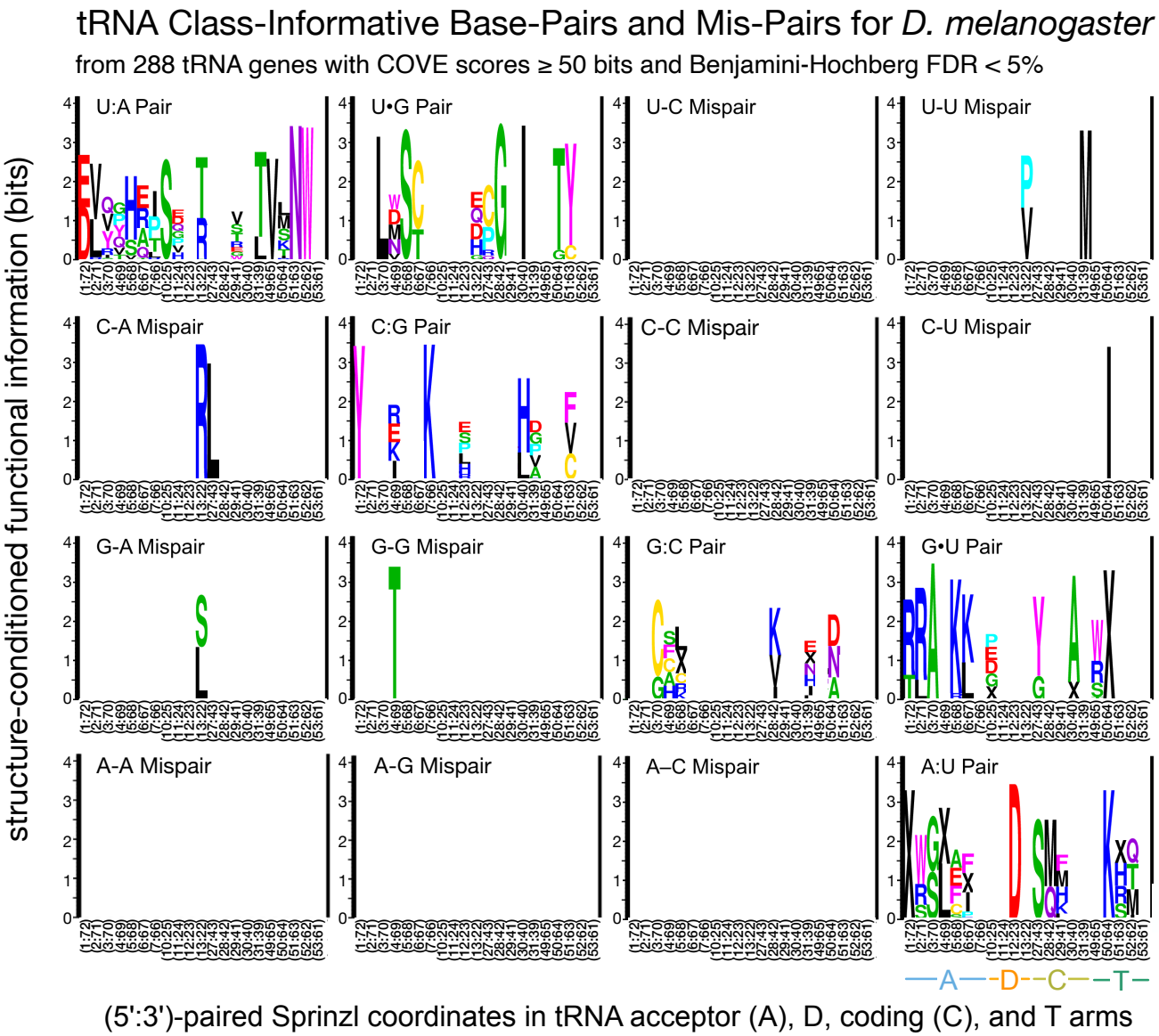

FIG. 4. Class-Informative Base-Pairs and Base Mis-Pairs from 288 reannotated tRNA genes in *Drosophila melanogaster* with COVE scores of at least 50 bits as computed in tSFM v1.0 with a Benjamini-Hochberg False Discovery Rate of 5% (CIFs not meeting this significance threshhold are not shown). The total height of a stack of letters quantifies information potentially gained about the functional type of a tRNA by a tRNA-binding protein if it specifically recognizes the paired features indicated. The letters within each stack symbolize functional types of tRNAs, wherein IUPAC one-letter amino acid codes represent elongator tRNA aminoacylation identities and X symbolizes initiator tRNAs. The relative heights of letters within each stack quantifies the over-representation of tRNA functional types carrying that feature relative to the background frequency determined by the frequencies of genes of various functional types (as calculated through normalized log-odds).

significance of paired-site CIFs as a function of CIF information in Supplementary Figure S1. A text-file containing all statistics for paired-site and single-site CIFs in *D. melanogaster* is provided in the Supplementary Code and Data.

Conservation of Class-Informative Features in Drosophila tRNAs

The wealth of information we have on *Drosophila* evolution provides an unprecedented opportunity to understand how tRNA Class-Informative Features evolve. To examine the conservation of tRNA CIFs in *Drosophila*, we computed tRNA

CIFs in each species independently. In order to polarize evolutionary changes in tRNA CIFs we additionally analyzed tRNA gene annotations in the outgroup *Musca domestica*, which diverged from *Drosophila* between 20–80 million years ago (Wiegmann *et al.*, 2003). As shown in Figure 5, tRNA CIFs are very highly conserved in *Drosophila* and flies generally. Broadly speaking, our predictions are very nearly identical in all fly genomes we analyzed, with one visibly clear exception standing out: C17, which occurs within the metal ion-binding pocket previously identified as rapidly evolving. As shown in Figure 6, CIF C17 is ancestrally associated with $tRNA^{Lys}$ in flies but it became associated with all or nearly all $tRNA^{Asn}$ paralogs in at least three phylogenetically distinct *Drosophila* lineages, one of which, leading to (*D. ananassae*), also underwent the Asn-to-Lys anticodon shift previously reported in Rogers *et al.* (2010).

Separate logo figures predicting tRNA CIFs in each species that we analyzed are provided in the Supplementary Online code and data. In Figure S2 also in Supplementary Materials, we show that, as expected, the relative average information of single-site tRNA CIFs is inversely related to its relative compositional frequency in *D. melanogaster*. The relative frequency of bases in *D. melanogaster* tRNAs decreases from G to C to U to A, while relative average information increases correspondingly.

Discussion

In an analysis of tRNA gene ortholog sets from flies, primates and other groups, Rogers and Griffiths-Jones (2014) reported no evidence of preferential anticodon shift substitutions within or between functional classes of tRNAs charged by either of the two aminoacyl-tRNA synthetase superfamilies, Class I or Class II. However, they did report that an Asn-to-Lys alloacceptor anticodon shift substitution is one of only two types of anticodon shift substitutions that occurred more than once out of a total of 30 in their dataset, once in *Drosophila* and twice independently in primates. In this work, we show that in flies, the previously reported Asn-to-Lys anticodon shift substitution occurs together with co-option of CIF C17 from $tRNA^{Lys}$ to $tRNA^{Asn}$ in one lineage, that this co-option of CIF C17 from $tRNA^{Lys}$ to $tRNA^{Asn}$ occurred at least three times independently in *Drosophila* evolution, and that when it occurred, it was associated with parallel substitutions across all or most $tRNA^{Asn}$ genes. We note that AsnRS and LysRS are both Class IIb synthetases, suggesting that perhaps co-option of tRNA CIFs occurs more frequently within aaRS sub-classes, perhaps because aaRSs of the same sub-class have more similar binding interfaces on tRNAs (Giegé *et al.*, 2008). In our recently published theory for the evolution of aaRS-tRNA interaction networks, tRNA mutations can influence interactions with multiple aaRSs when

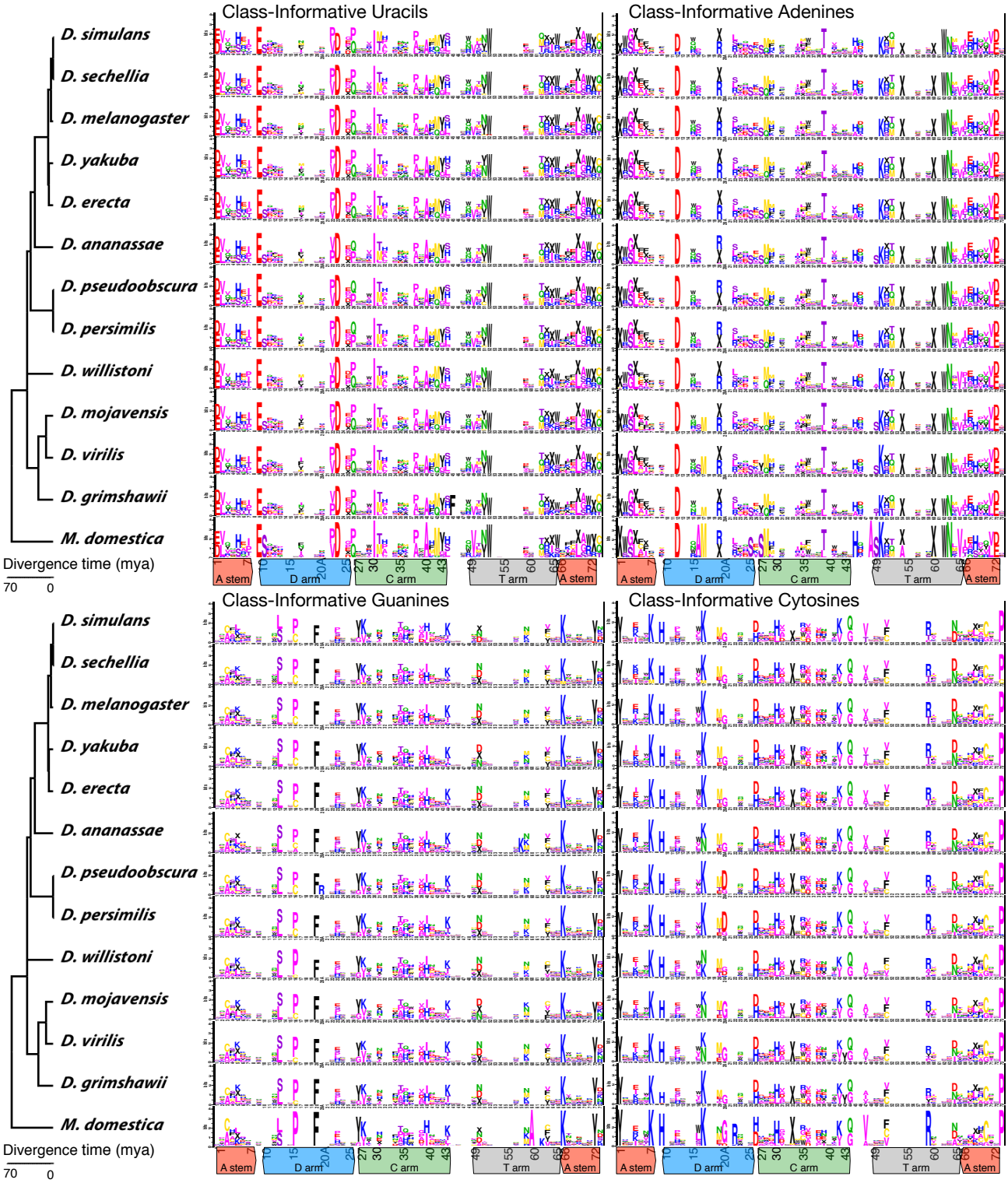

FIG. 5. Conservation of tRNA single-site CIFs in *Drosophila* with *Musca domestica* as an outgroup. Tree topology from *Drosophila* 12 Genomes Consortium (2007), with divergence dates from Tamura *et al.* (2003) and Hennig *et al.* (1981).

they occur inside shared interfaces (Collins-Hed and Ardell, 2019). The pattern we report here is consistent with an interpretation that the Asn-to-Lys anticodon shift substitution evolved to

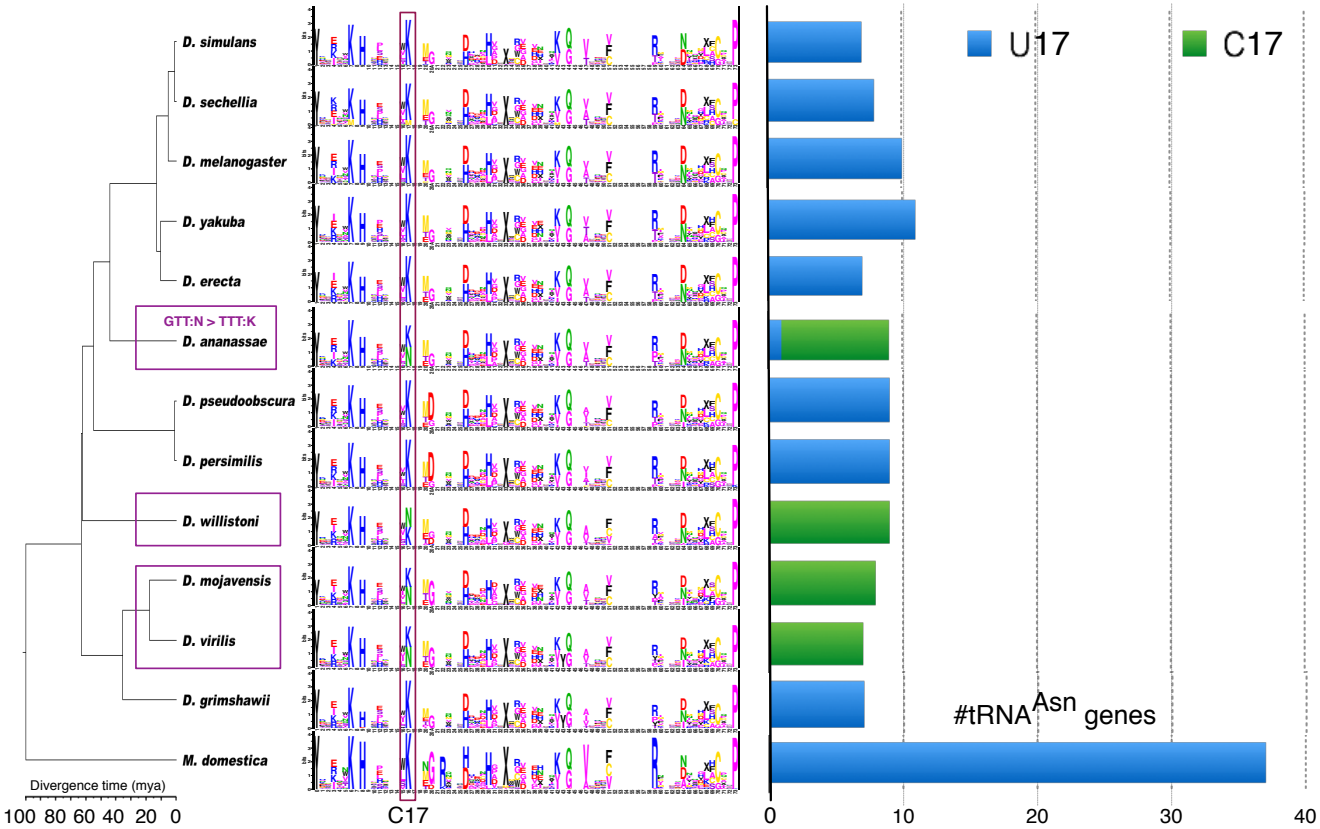

FIG. 6. Evolution of functional association of CIF C17 in flies. Cytosine single-site function logos are from Fig. 5, the tree topology is from *Drosophila* 12 Genomes Consortium (2007), and divergence dates are from Tamura *et al.* (2003) and Hennig *et al.* (1981). The three clades in which C17 became associated with $tRNA^{Asn}$ genes are indicated by purple boxes on the tree, one of which (*D. ananassae*), also underwent the Asn-to-Lys anticodon shift previously reported in Rogers *et al.* (2010) as indicated in top purple box at figure left. The stacked bar graph at figure right shows the frequencies of U17 or C17 across all $tRNA^{Asn}$ genes in each genome showing that C17 substituted in parallel across all $tRNA^{Asn}$ genes in these three lineages.

compensate the co-option of C17 from Lys to Asn in the *D. ananassae* lineage.

As examined in the UCSC Genome Browser (Kent *et al.*, 2002) for *Drosophila melanogaster*, eight $tRNA^{Asn}$ genes and four $tRNA^{Lys}$ genes co-cluster together within 70 Kb on Chromosome 2R, while at least one $tRNA^{Asn}$ gene occurs on Chromosome 4. This suggests that gene conversion from illegitmate recombination may have played a role in the convergent and/or parallel evolution of the tRNA structure-function map in *Drosophila*, but that a simple scenario of concerted gene evolution (Nei and Rooney, 2005) is not sufficient to explain the parallel substitutions of C17 in $tRNA^{Asn}$ genes that we observed. Rapid evolution of tandemly arranged tRNA genes by rearrangements appears to be universal in eukaryotes (Velandia-Huerto *et al.*, 2016) and prokaryotes (Tremblay-Savard *et al.*, 2015). This point should be revisited in long-read resequenced genome data for the *Drosophila* genus. Our results suggest that interactions within at least three different levels of biological structure may contribute to rapid evolution and functional turnover of tRNA genes: of physically

interacting sites within tRNAs, of co-clustered tRNA genes within genomes, and of overlapping interfaces for different tRNA-binding proteins.

Materials and Methods

Data, Orthology Sets and Alignments

We obtained tRNA gene sequences and annotations for 12 species of *Drosophila* from FlyBase (2008_07 release, McQuilton *et al.*, 2012) on October 16, 2011. For further functional re-annotation, we re-folded and re-annotated these tRNA sequences using tRNAscan-SE 1.3.1 (Lowe and Eddy, 1997) and ARAGORN 1.2.34 (Laslett and Canback, 2004). The genome for *Musca domestica* was downloaded from NBCI on October 8th, 2013 (Scott *et al.*, 2014). The genome was annotated using predictions from tRNAscan-SE 1.3.1 (Lowe and Eddy, 1997).

We identified Initiator tRNA genes using TFAM 1.3 (Tåquist *et al.*, 2007). We aligned all tRNAs using Infernal 1.1 (Nawrocki *et al.*, 2009) with the RFAM covariance model for tRNAs (RF00005) (Burge *et al.*, 2012). We edited alignments manually using Seaview 4.3.4 (Gouy *et al.*, 2010) to produce a final alignment of length 74, and mapped each site manually to Sprinzl coordinates (Sprinzl and Vassilenko, 2005). We verified our coordinate mapping using tRNAdb (Jühling *et al.*, 2009). We retained Sprinzl coordinate 20A and removed the majority of the variable arm, except Sprinzl coordinates 45 through 49, for subsequent analysis.

We downloaded ortholog sets of tRNA genes on October 18, 2011 from supplementary data available at `http://gbe.oxfordjournals.org/content/2/467/suppl/DC1` (Rogers *et al.*, 2010). To optimize ortholog sets, we calculated lengths of pruned species trees using the Bio::TreeIO module in BioPerl 1.4.0 (Stajich *et al.*, 2002) and calculated numbers of variable and parsimoniously informative sites of alignments, as well as average fraction of pairwise differences, using the Bio::PopGen modules (Stajich and Hahn, 2005) in BioPerl 1.4.0.

Analysis of Substition Rates

All subsets of data were curated into concatenated alignments by previously published ortholog sets (Rogers *et al.*, 2010). Ortholog sets with anticodon shift substitutions, genes of indeterminate function, or pseudogenes were removed from all analyses of substitution rates.

We estimated substitution rates with MrBayes 3.2.1 (Huelsenbeck and Ronquist, 2001; Ronquist *et al.*, 2012) using the fixed known species tree (*Drosophila* 12 Genomes Consortium, 2007). For all runs we constrained change to the tree topology by setting rates of stochastic TBR and branch multipliers to zero probability. All Bayesian analyses were run with two simultaneous chains for 4×10^6 iterations, monitoring convergence of split frequency standard deviations and saved parameters every 500 iterations.

We computed evolutionary rates by assigning sites or site-pairs into corresponding data partitions. For structural data partitions, we considered nine structural tRNA components: acceptor stem (Sprinzl sites in *Drosophila*: 1 – 7, 67 – 73), D stem (10 – 13, 23 – 26), D loop (14 – 22), anticodon stem (28 – 32, 40 – 44), anticodon loop (33 – 39), variable arm (45 – 49), T stem (50 – 54, 62 – 66), T loop (55 – 61), and "other sites" (8, 9, 27, 74). Sites in "other" are not involved in base-pairs or considered part of loop structures in tRNA. We estimated rates using the General Time Reversible (GTR) substitution model (Lanave *et al.*, 1984; Rodríguez *et al.*, 1990; Tavaré, 1986) with Invariant sites, allowing stationary state frequencies and all other substitution model parameters to be independent across partitions. Alignment data and MrBayes initialization scripts with corresponding alignments are provided in supplementary data. The first 25% of parameter calculations were discarded as burn-in for statistical analysis. Parameter posterior probabilities were imported to R (R Core Team, 2013). We used the coda package (Plummer *et al.*, 2006) for Markov Chain Monte Carlo simulations diagnostics and the lattice package (Sarkar, 2008) for multivariate analysis. All data from MCMC simulations is presented with 95% Bayesian credible intervals.

Detection of Class-Informative Features

We identified Class-Informative features (CIFs) for all 12 species of *Drosophila* using tSFM v1.0, available from `https://github.com/tlawrence3/tSFM`, using the Nemenman-Shafee-Bialek (NSB) Bayesian entropy estimator (Nemenman *et al.*, 2002) for features in two or more sequences and an exact estimator (Schneider *et al.*, 1986) otherwise. To compute the significance of paired-site CIFs in *D. melanogaster*, we computed permutation p-values (permuting functional class assignments over sequences that contain a CIF) for the total information of paired-sites only, and then computed Benjamini-Hochberg False Discovery Rates (Benjamini and Hochberg, 1995) from those p-values.

Acknowledgments

The authors thank Carolin Frank, Michael Cleary, Michael Colvin and Josh Phillips for helpful discussion. DHA was supported by a National Science Foundation grant (INSPIRE-1344279), a National Institutes of Health grant (AI127582), and a Julius Kuhn Guest Professorship at Martin Luther University in Halle-Wittenberg, United States. Computational research in this report was performed on the MERCED HPC cluster supported by the National Science Foundation, United States (ACI-1429783).

References

Amrine, K. C. H., Swingley, W. D., and Ardell, D. H. 2014. tRNA signatures reveal a polyphyletic origin of SAR11 strains among alphaproteobacteria. *PLoS computational biology*, 10(2): e1003454.

Bailly, M., Giannouli, S., Blaise, M., Stathopoulos, C., Kern, D., and Becker, H. D. 2006. A single tRNA base pair mediates bacterial tRNA-dependent biosynthesis of asparagine. *Nucleic Acids Res*, 34(21): 6083–6094.

Behlen, L. S., Sampson, J. R., DiRenzo, A. B., and Uhlenbeck, O. C. 1990. Lead-catalyzed cleavage of yeast tRNAPhe mutants. *Biochemistry*, 29(10): 2515–2523.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x.

Bhattacharyya, S. and Varshney, U. 2016. Evolution of initiator trnas and selection of methionine as the initiating amino acid. *RNA Biology*, 13(9): 810–819.

Brown, J. R. and Doolittle, W. F. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci U S A*, 92(7): 2441–2445.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., and Bateman, A. 2012. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*, 41(D1): D226–D232.

Cedergren, R. J., Sankoff, D., LaRue, B., and Grosjean, H. 1981. The evolving tRNA molecule. *CRC Critical Reviews in Biochemistry*, 11(1): 35–104.

Collins-Hed, A. I. and Ardell, D. H. 2019. Match fitness landscapes for macromolecular interaction networks: Selection for translational accuracy and rate can displace tRNA-binding interfaces of non-cognate aminoacyl-tRNA synthetases. *Theor Popul Biol*.

Cusack, S. 1997. Aminoacyl-tRNA synthetases. *Current Opinion in Structural Biology*, 7(6): 881–889.

Drosophila 12 Genomes Consortium 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167): 203–18.

Eigen, M., Lindemann, B. F., Tietze, M., Winkler-Oswatitsch, R., Dress, A., and von Haeseler, A. 1989. How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science*, 244(4905): 673–679.

Eriani, G., Delarue, M., Poch, O., Gangloff, J., and Moras, D. 1990. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature*, 347(6289): 203–206.

Freyhult, E., Moulton, V., and Ardell, D. H. 2006. Visualizing bacterial tRNA identity determinants and antideterminants using function logos and inverse function logos. *Nucleic Acids Research*, 34(3): 905–916.

Giegé, R., Sissler, M., and Florentz, C. 1998. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res*, 26(22): 5017.

Giegé, R., Touzé, E., Lorber, B., Théobald-Dietrich, A., and Sauter, C. 2008. Crystallogenesis Trends of Free and Liganded Aminoacyl-tRNA Synthetases. *Crystal Growth & Design*, 8(12): 4297–4306. Publisher: American Chemical Society.

Gouy, M., Guindon, S., and Gascuel, O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2): 221–224.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2): 160–174.

Hennig, W. *et al.* 1981. *Insect phylogeny.* John Wiley & Sons Ltd.

Holbrook, S. R., Sussman, J. L., Warrant, R. W., Church, G. M., and Kim, S. H. 1977. RNA-ligand interactions:(I) magnesium binding sites in yeast tRNAPhe. *Nucleic Acids Research*, 4(8): 2811–2820.

Holley, R. W. 1965. Structure of an alanine transfer ribonucleic acid. *Journal of the American Medical Association*, 194(8): 868–871.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and

Zamir, A. 1965. Structure of a ribonucleic acid. *Science*, 147(3664): 1462–1465.

Huelsenbeck, J. P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8): 754–755.

Humphrey, W., Dalke, A., and Schulten, K. 1996. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1): 33–8– 27–8.

Jack, A., Ladner, J. E., Rhodes, D., Brown, R. S., and Klug, A. 1977. A crystallographic study of metal-binding to yeast phenylalanine transfer RNA. *Journal of Molecular Biology*, 111(3): 315–328.

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., and Pütz, J. 2009. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research*, 37(Database issue): D159–62.

Kelly, P., Hadi-Nezhad, F., Liu, D. Y., Lawrence, T. J., Linington, R. G., Ibba, M., and Ardell, D. H. 2020. Targeting tRNA-synthetase interactions towards novel therapeutic discovery against eukaryotic pathogens. *PLoS Negl Trop Dis*, 14(2): e0007983.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, a. D. 2002. The Human Genome Browser at UCSC. *Genome Res.*, 12(6): 996–1006. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Lanave, C., Preparata, G., Saccone, C., and Serio, G. 1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1): 86–93.

Laslett, D. and Canback, B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, 32(1): 11–16.

Lawrence, T. J., Amrine, K. C., Swingley, W. D., and Ardell, D. H. 2019. tRNA functional signatures classify plastids as late-branching cyanobacteria. *BMC Evol. Biol.*, 19(1): 224.

Lowe, T. M. and Eddy, S. R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5): 955–964.

McQuilton, P., St Pierre, S. E., Thurmond, J., and FlyBase Consortium 2012. FlyBase 101– the basics of navigating flybase. *Nucleic Acids Research*, 40(Database issue): D706–14.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10): 1335–1337.

Nei, M. and Rooney, A. P. 2005. Concerted and Birth-and-Death Evolution of Multigene Families. *Annual Review of Genetics*, 39(1): 121–152.

Nemenman, I., Shafee, F., and Bialek, W. 2002. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems 14 - Proceedings of the 2001 Conference, NIPS 2001*, Advances in Neural Information Processing Systems. Neural information processing systems foundation.

Plummer, M., Best, N., Cowles, K., and Vines, K. 2006. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1): 7–11.

R Core Team 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodríguez, F., Oliver, J. L., Marín, A., and Medina, J. R. 1990. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142(4): 485–501.

Rogers, H. H. and Griffiths-Jones, S. 2014. tRNA anticodon shifts in eukaryotic genomes. *RNA*, 20(3): 269–281.

Rogers, H. H., Bergman, C. M., and Griffiths-Jones, S. 2010. The evolution of tRNA genes in Drosophila. *Genome Biol Evol*, 2: 467–477.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. 2012. MrBayes 3.2: efficient bayesian phylogenetic inference and model

choice across a large model space. *Systematic Biology*, 61(3): 539–542.

Saks, M. E., Sampson, J. R., and Abelson, J. 1998. Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science*, 279: 1665–1670.

Sarkar, D. 2008. *Lattice: Multivariate Data Visualization with R*. Springer, New York.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188(3): 415–431.

Scott, J. G., Warren, W. C., Beukeboom, L. W., Bopp, D., Clark, A. G., Giers, S. D., Hediger, M., Jones, A. K., Kasai, S., Leichter, C. A., Li, M., Meisel, R. P., Minx, P., Murphy, T. D., Nelson, D. R., Reid, W. R., Rinkevich, F. D., Robertson, H. M., Sackton, T. B., Sattelle, D. B., Thibaud-Nissen, F., Tomlinson, C., van de Zande, L., Walden, K. K. O., Wilson, R. K., and Liu, N. 2014. Genome of the house fly, Musca domestica L., a global vector of diseases with adaptations to a septic environment. *Genome Biol.*, 15(10): 466.

Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. 2009. Dynamical networks in tRNA:protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(16): 6620–6625.

Shi, H. and Moore, P. B. 2000. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: a classic structure revisited. *RNA*, 6(8): 1091–1105.

Sprinzl, M. and Vassilenko, K. S. 2005. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, 33(Database issue): D139–40.

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

Stajich, J. E. and Hahn, M. W. 2005. Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*, 22(1): 63–73.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10): 1611–1618.

Tamura, K., Subramanian, S., and Kumar, S. 2003. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular Biology and Evolution*, 21(1): 36–44.

Tåquist, H., Cui, Y., and Ardell, D. H. 2007. TFAM 1.0: an online tRNA function classifier. *Nucleic Acids Research*, 35(Web Server issue): W350–3.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures in Mathematics of Life Sciences*, 17: 57–86.

Drosophila 12 Genomes Consortium 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167): 203–218.

Tremblay-Savard, O., Benzaid, B., Lang, B. F., and El-Mabrouk, N. 2015. Evolution of trna repertoires in bacillus inferred with orthoalign. *Mol Biol Evol*, 32.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., and Zhang, H. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res*, 37(suppl_1): D555–D559. Publisher: Oxford Academic.

Velandia-Huerto, C. A., Berkemer, S. J., Hoffmann, A., Retzlaff, N., Romero Marroquín, L. C., Hernández-Rosales, M., Stadler, P. F., and Bermúdez-Santana, C. I. 2016. Orthologs, turn-over, and remolding of trnas in primates and fruit flies. *BMC Genomics*, 17(1): 617.

Wiegmann, B. M., Yeates, D. K., Thorne, J. L., and Kishino, H. 2003. Time flies, a new molecular time-scale for brachyceran fly evolution without a clock. *Systematic Biology*, 52(6): 745–756.

Wolfson, A. D., LaRiviere, F. J., Pleiss, J. A., Dale, T., Asahara, H., and Uhlenbeck, O. C. 2001. tRNA Conformity. In *The Ribosome*, pages 185–193. CSHL Press.