Title

Evolution of Eukaryotic Transfer Ribonucleic Acid

Author

Phillips, Julie Baker

Publication Date

2013

UNIVERSITY OF CALIFORNIA, MERCED

Evolution of Eukaryotic Transfer Ribonucleic Acid

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Quantitative and Systems Biology

by

Julie Baker Phillips

Committee in charge:

Professor Michael E. Colvin, Chair
Professor David H. Ardell
Professor Michael Cleary
Professor Carolin Frank

2013

The dissertation of Julie Baker Phillips is approved:

____

____

____

____
Chair

University of California, Merced

2013

DEDICATION

I dedicate this dissertation to my God, and my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my advisor Prof. David Ardell for his constant support and advocacy on my behalf and I would like to acknowledge helpful discussions with Prof. Michael Colvin (UC Merced), Prof. Carolin Frank (UC Merced), and Prof. Michael Cleary (UC Merced). I would like to thank Joshua L. Phillips, PhD (Los Alamos National National Laboratory) for his assistance with molecular dynamics simulations of tRNA. Finally, I would like to thank former members of the Ardell lab Katherine Amrine, PhD (UC Davis) and Professor Wesley Swingley (NIU) for their helpful discussions over the years.

VITA

2002	Bachelor of Science in Mathematical Science, Middle Tennessee State University
2004	Master of Science in Biological Science, Middle Tennessee State University
2012	Doctor of Philosophy in Quantitative and Systems Biology, University of California, Merced

PUBLICATIONS

Phillips, J.B., and M. Klukowski (2008) Influence of season and adrenocorticotropic hormone on corticosterone in free-living female Eastern fence lizards (Sceloporus undulatus). Copeia 2008 (3): 570-578.

Cain, C.J., Conte, D.A., Garca-Ojeda, M.E., Daglio, L.G., Johnson, L., Lau, E.H., Manilay, J.O., **Phillips, J.B.**, Rogers, N.S., Stolberg, S.E., Swift, H.F. and M.N. Dawson (2008) What systems biology is (not yet). Science 320: 1013-1014.

PUBLISHED ABSTRACTS

Phillips, J.B., Phillips, J.L. and D. Ardell (2013) . Molecular Evolution and Molecular Dynamics of Drosophila tRNAs. Bay Area RNA Club. San Francisco, California. (talk)

Phillips, J.B., Phillips, J.L. and D. Ardell (2013). Molecular Evolution of Eukaryotic Nuclear tRNAs. 21st Annual Conference of the Society for Molecular Biology and Evolution. Chicago, IL. (poster)

Phillips, J.B. and D. Ardell (2012). Molecular Evolution of Nuclear tRNAs. 24th tRNA Conference. Olmu, Chile. (poster)

Phillips, J.B. and D. Ardell (2012). Evolution of tRNA in Drosophila. Joint Congress on Evolutionary Biology. Ottawa, Canada. (poster)

FIELDS OF STUDY

Major Field: Quantitative and Systems Biology

Studies in Computational Biology
Professor David H. Ardell

Studies in Physiology
Professor Rudy M. Ortiz

Studies in Herpetology and Behavioral Endocrinology
Professor Matthew Klukowski

ABSTRACT

Evolution of Eukaryotic Transfer Ribonucleic Acid

by

Julie Baker Phillips
Doctor of Philosophy in Quantitative and Systems Biology

University of California, Merced, 2013

Professor Michael E. Colvin, Chair

Patterns of substitution rates footprint functional constraints and highlight potential adaptive evolutionary change in macromolecules. The vast majority of molecular evolutionary studies have been on proteins. Even though tRNAs were the first RNAs to be sequenced and structurally solved, substitution rate analysis of tRNAs has not yet been published. In this dissertation, we advance the knowledge of tRNA evolution in eukaryotes, using sequence data from the twelve species of *Drosophila*. First I introduce background concepts covering the sequence, structure and function of tRNA, as well as describe early work studying the evolution of this molecule. In the second chapter, I describe divergence rates for tRNA sites, structures, and alloacceptor classes. I also discuss the evolution of the tRNA-protein interaction network in *Drosophila* with evidence of changes in "Class Informative Features" (CIFs) between species and clades of flies. In the third chapter, I discuss how work in site divergence rates lead us to investigate the effects of sequence mutations on one primary ion binding pocket through molecular dynamic simulations. Finally, I discuss some preliminary results using yeast

sequence data to examine whether results from flies is specific to this group of species, or whether our results are generalizable to other eukaryotes. I also discuss the preliminary work using *Drosophila melanogaster* population sequences to ascertain the selective pressures acting upon tRNA sequences.

Chapter 1

Introduction

1.1 Biological Relevance of Ribonucleic Acids

The central dogma of molecular biology, first described by Francis Crick in 1958 and then revised in 1970, describes the process by which genetic information flows in a biological system (Crick, 1958, 1970). The dogma describes three primary molecules, deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins. Crick defined, along with these molecules, six conceivable transfers of information. The three primary methods of transfer include DNA replication, transcription (DNA to RNA), and translation (RNA to protein). Modern biology now knows that protein coding genes, a segment of DNA, are transcribed by an enzyme, RNA polymerase, into messenger RNA (mRNA) which carries the same genetic information as the transcribed gene. mRNA is then translated into a protein by the ribosome. This process is governed by a non-coding RNA molecule, transfer RNA (tRNA), that decodes the mRNA, by base triplets, into amino acids by grammar rules of the genetic code. Crick also postulated in his early work that RNA could self-replicate, RNA could be reverse transcribed to DNA, and that DNA could be directly translated into a protein, bypassing RNA. Research confirmed Crick's 1970 revision, that RNA does self-replicate and that RNA can undergo reverse-transcription to DNA.

1.2 RNA

RNA is a polymeric molecule made by stringing together individual ribonucleotides, adding the 5′-phosphate group of one nucleotide to the 3′-hydroxyl group of the previous nucleotide. Similar to DNA, each ribonucleotide consists of a phosphate group, a sugar molecule - specifically ribose, and one of four nitrogenous bases. The nitrogenous bases are two purines, adenine (A) and guanine (G), and two pyrimidines, cytosine (C) and uracil (U). Like DNA, an individual strand of RNA has the same basic structure, composed of nitrogenous bases bound together by a sugar-phosphate backbone. In an addition to the canonical (purine-pyrimidine) Watson-Crick base pairs of DNA, RNA molecules employ a variety of non-canonical base pairing confirmations, allowing for a multitude of complex secondary and tertiary structures.

Unlike DNA, RNA is largely a single-stranded molecule. Although usually single-stranded, RNA can form double-stranded structures, which are important to the various functions of the many types of RNA. Single-stranded RNA can form many secondary structures by folding over the single stand and forming helices and loops, which are then stabilized by intramolecular hydrogen bonds. Such base-pairing of RNA is critical for RNAs' many functions. Common secondary structural elements of RNA include: hairpins, bulges, internal loops, multibranched lops. stems, and pseudoknots (Batey et al., 1999).

RNA, much like proteins, can form complex three-dimensional folds that are often essential for proper function. Of the many interesting three-dimensional motifs, for brevity, only those relevant to the current work will be reviewed. Coaxial stacking is one the fundamental tertiary motifs of RNA, conferring overall stability to the tertiary structure. Coaxial stacking occurs when two helices duplex to form a single contiguous helix, which is stabilized by base stacking at the interface of the two helices, Figure 1.1A. In addition to the formation of base pairs and helices, RNA molecules are capable of forming base triples and triplexes, Figure 1.1B. Triplexes utilize unpaired

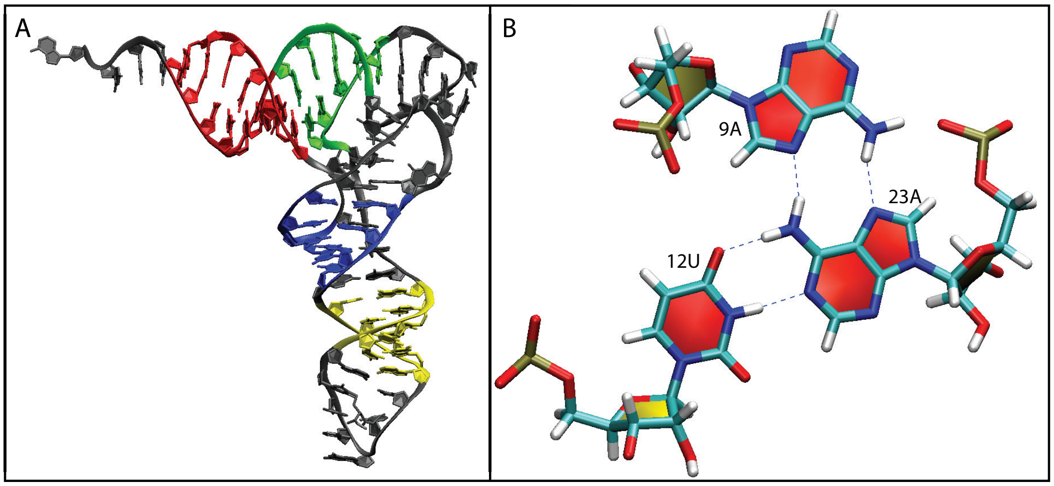

Figure 1.1: Tertiary stabilizing interactions in tRNA. (A) Coaxial stacking in transfer RNA. The helical regions of the dihydrouridine (blue) and anticodon (yellow) stems stack to form a contiguous helix, as well as the acceptor (red) and thymidine (green) stems. (B) Base triplex in tRNA: 9A-12U-23A. The 9-adenosine hydrogen atom bonds with a 2′-hydroxyl group on the minor groove face of a reverse-Hoogsteen A23-U12 base pair. Image were created using VMD (Humphrey et al., 1996) using data from the yeast tRNA$^{\mathrm{Phe}}$ crystal structure (PDB ID: 1EHZ). (Shi and Moore, 2000).

bases in RNA loops to form loop-loop interactions conferring the ability to create unique 3-dimensional shapes and offer further stabilization.

1.2.1 Non-coding RNA

Messenger RNA carries genetic information relevant for the coding of proteins necessary for survival of the organism. The term non-coding RNA (ncRNA) commonly describes RNA that do not encode a protein, but ncRNAs do contain information and serve important biological functions. Part of the central dogma was that RNA functions mainly as an informational messenger between DNA and its associated protein, yet evidence suggests that the developmental programming and the phenotypic difference between species and individuals is heavily influenced by many regulatory ncRNAs which are now recognized and studied (Mattick and Makunin, 2006). Transfer RNA was the first described ncRNA among this growing list of regulatory molecules, which also includes ribosomal RNA (rRNA), small nuclear RNA (snRNA), small nucleolar

RNA (snoRNA), small interfering RNA (siRNA), piwi-interacting RNAs (piRNAs), and micro RNA (miRNA).

1.3 Transfer RNA

Transfer RNA is the class of adapter molecules whose primary biological purpose is in the molecular process of translation. tRNAs are recognized by functional classes (alloacceptors) that describe the amino acid the tRNA serves to decode during translation. A group of functional classes contain isoacceptors, tRNA species that bind to alternate codons within the same alloacceptor class.

1.3.1 Discovery of tRNA

The central dogma of molecular biology was published in the same year as Hoagland and Zamecnik described a soluble RNA (Hoagland et al., 1958). Alongside Hoagland and Zemecnik, two additional labs, Ogata and Nohara (Ogata and Nohara, 1957), and Holley (Holley, 1957), also independently obtained evidence that this RNA molecule was able to transfer an amino acid to a protein in a protein synthesizing system. This soluble RNA was later renamed transfer RNA (tRNA) and has became well characterized as the molecule that mediates the translation of mRNAs on the ribosome. Less than a decade later, Holly sequenced a yeast alanine tRNA and proposed a secondary structure, making tRNA the first RNA to be sequenced and have a solved structure (Holley et al., 1965; Holley, 1965). Holley was awarded the Nobel Prize in Physiology or Medicine in 1968 for his discovery.

Along with the central dogma came the adaptor hypothesis (Crick, 1958). Crick believed that central to protein synthesis was an ‘adaptor’ molecule, whose role was to carry amino acids to the template and then ‘fit’ itself onto the mRNA. He hypothesized there would exist twenty adaptors in the simplest system, one for each amino acid.

Holley's discovery of tRNA validated Crick's adaptor hypothesis, and tRNA is one of the most well studied molecules in the RNA family.

1.3.2 Function in Translation

Translation is the conversion of a mRNA message into a protein. Proteins are composed of amino acids which are specified in triplets, a codon, by mRNA according to a genetic code. The genetic code is nearly universal across all domains of life. Translation is carried out on the ribosome, the cellular workhorse. tRNAs serve as the 'adapter' molecule that decode by codons, nucleotide triplets, into amino acids according to the code on the mRNA strand. There are 20 canonical amino acids that must be decoded accurately during the process of translation. All classes of tRNA are structurally confined by their function in protein synthesis, since each tRNA must be able to fit in the ribosome.

The adaptor function of the tRNA involves two regions, the post-transcriptionally added CCA tail at the $3'$ terminus, and the anticodon loop. The CCA tail is not encoded in nearly all eukaryotic tRNA genes (Shi et al., 1998). In eukaryotes, the maturation of the $3'$ terminus is achieved by an essential enzyme, the CCA-adding enzyme (tRNA nucleotidyltransferase), which catalyzes the post-transcriptional addition of CCA using adenosine $5'$-triphosphate (ATP) and cytidine $5'$-triphosphate (CTP) as substrates. Amino acids are covalently attached to the terminal adenosine of the CCA tail by an amino-acyl tRNA synthetase (aaRS). aaRSs are a specialized class of enzymes that catalyzes the addition of a specific amino acid or a related precursor to a compatible cognate tRNAs. There are two classes of aaRSs depending upon the tertiary structure of the enzyme. These two classes of syntheses also divide tRNAs into two classes dependent upon with aaRS charges the tRNA with the cognate amino acid.

After attachment of the cognate amino acid, the 'charged' tRNA aligns its anticodon with a codon on the mRNA template, which binds to the appropriate codon through complementary base pairing. Of the 64 (4^3) possible codons, three are stop codons which serve as the signal for the termination of translation; the remaining 61 codons encode amino acids.

The ribosome has three sites for tRNA binding, designated the P (peptidyl), A (aminoacyl), and E (exit) sites. The initiator methionyl tRNA ($\text{tRNA}^{\text{iMet}}_{\text{CAT}}$) signals the start of the elongation process once it is bound at the P site. The second tRNA is escorted to the ribosome by an elongation factor (eEF-1α in eukaryotes) complexed with GTP, the tRNA binds to the A site by pairing with the next codon of the mRNA. GTP is hydrolyzed to GDP and the tRNA is released from eEF-1α, this release is the rate limiting step in elongation. The extended time is used as a proof-reading step in elongation before the a peptide bond is formed between amino acids on the growing polypeptide chain. During translocation, another elongation factor (eEF-2 in eukaryotes) coupled to GTP aids the positioning of a new codon of mRNA in the empty A site. The 'charged' tRNA is translocated from the A site to the P site, and the uncharged tRNA from the P site translocated to the E site. The binding of a new 'charged' tRNA to the A site induces release of the uncharged tRNA from the E site. This process will continue until a stop codon is read on the mRNA.

1.3.3 Non-Translational Functions

Increasing amounts of evidence support that the ribosome is not the only target of tRNAs for delivering amino acids, non-translational processes include processes of lipid modification and antibiotic biosynthesis. In addition, recent evidence in the past decade has suggested a regulatory role for tRNAs similar to that of miRNA (Pederson, 2010). Recently described tRNA-derived RNA fragments (tRFs) introduce a new aspect of tRNA biology. tRFs are a novel class of small RNAs that are second in abundance only to miRNAs (Lee et al., 2009). These fragments tRFs vary in length from ranges of 17 – 26 and 30 – 35. In addition to the proposed regulatory roles, several tRNAs can be used in other amino acid addition pathways. These non-translational reactions are a result of tRNAs interactions with several types of acceptor molecules including membrane lipids, peptidoglycan precursors, proteins and intermediates for the biosynthesis of antimicrobial molecules (Giegé, 2008; Banerjee et al., 2010).

1.3.4 Sequence and Structure

The primary structure of tRNA is a sequence of 73 – 93 nucleotides, depending upon the presence or absence of the variable arm along with the variable length of the D-loop. tRNA often contain numerous post-transcriptionally modified nucleotides. The secondary structure of tRNA is described as being a clover leaf shape due to the nature of base pairing between nucleotides, which creates four stem structures, consisting of four to seven Watson-Crick type base pairs, and four loop structures, Figure 1.2. The four stems (arms) of a tRNA molecule are the acceptor stem, dihydrouridine (D) stem, anticodon stem and the thymidine (T or TψC) stem. Three of the loop structures are named the D-loop, T-loop and anticodon loop describing the stem from which the loop originates. The fourth loop in the variable loop, so named because the length of the loop in variable depending upon the functional class of the tRNA.

The tertiary interactions result in compact "L-shape" structure. The stems of the D and anticodon arms are stacked upon each other, as do the stems of the T-arm and acceptor arm, Figure 1.1A. These two coaxial stacks are oriented perpendicularly with respect to one another by tertiary interactions between the D and T-loops to yield the overall canonical L-shape. Only 41 of 76 bases are involved in the classic helical structures of the tRNA stems, yet 72 bases are involved in stacking interactions (Kim et al., 1974). Triplexes between the variable loop and the major groove of the D-arm add additional stability to the overall tertiary structure.

1.3.5 Ion Binding

Ion binding pockets in tRNA were first described in the original tRNA crystallography papers in both the orthorhombic crystal form (Holbrook, Sussman,

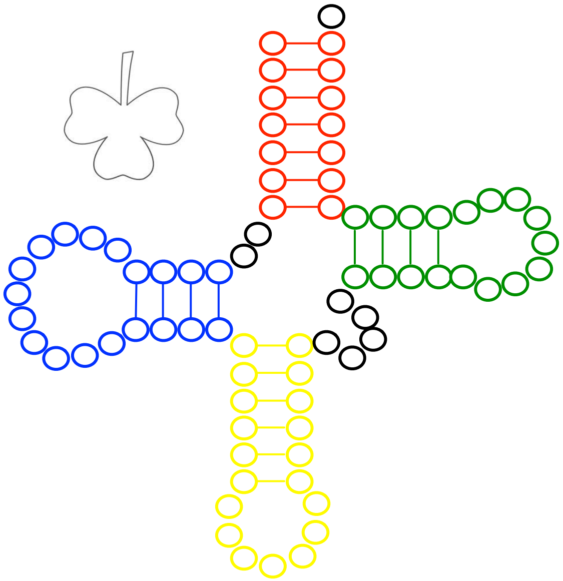

Figure 1.2: The secondary structures of tRNA forms a clover leaf shape. The four stems and the associated loop structures: the acceptor stem (red), D-stem and loop (blue), anticodon stem and loop (yellow) and T-stem and loop (green).

Warrant, Church, and Kim, Holbrook et al.) and the monoclinic structure (Jack et al., 1977). Since then the importance of the ion binding pockets have been the subject of many scientific endeavors. The neutralization of backbone phosphate charges by metal ion binding are an integral part in RNA folding (Pan et al., 1993; Hermann and Westhof, 1998). In addition to stabilizations of helical structures, there are positions within a folded structure where single-stranded phosphates are constrained in a closed space, thereby forming binding "pockets" where metal ions bind even more tightly than to a helix structure (Pan et al., 1993). Formation of such metal ion-binding "pockets" is expected to be strictly dependent on complete folding of the RNA.The tertiary structure of tRNA contains numerous metal ion binding pockets. Research has suggested that 3 to 4 strong and more than 20 weak magnesium cation (Mg^{2+}) binding sites exist in tRNAs (Holbrook, Sussman, Warrant, Church, and Kim, Holbrook et al.; Danchin, 1972; Jack et al., 1977). In addition to interactions with Mg^{2+}, metal ion-binding pockets involving both ribose phosphates and bases have the potential to interact with many varieties of metal ions.

1.4 Early studies in tRNA evolution

A vast majority of evolutionary studies have been on proteins. tRNA evolution poses a slightly more complex problem. As described, the primary biological role of tRNA is in the process of translation, which alone dictates that each tRNA molecule interact successfully with other coevolving components of the machinery of protein synthesis. tRNAs have been documented to participate in many biological processes aside from translation which adds to the complexity of documenting the evolution of this important molecule. Studying the patterns of substitution rates provides the blueprints for the functional constraints on tRNA and highlights potential adaptive evolutionary changes. Due to tRNA's central role in translation and the nearly universal genetic code, evidence suggests that tRNA genes are likely among the earliest genes to arise (Eigen et al., 1989). Research has long been interested in identifying the primordial origin of modern tRNA and the events surrounding the transition of this primordial RNA into the canonical tRNA structure (Di Giulio, 1992).

Some of the first work in tRNA evolution aimed to look at the ancestry of tRNA alloacceptors and isoacceptors. A reasonable assumption is that isoacceptors arose by gene duplication from a common ancestor having the same amino acid identity, but early work suggested that phenylalanine and glycine tRNAs for prokaryotes and eukaryotes derived from common ancestry while, tyrosine tRNAs had independent origins (Cedergren et al., 1980). Cedegren et al. concluded that one of the earliest events in tRNA evolution was the divergence of the $\mathrm{tRNA}^{\mathrm{Val}}_{\mathrm{GAG}}$ from the glycine tRNA family (Cedergren et al., 1980). Saks et al. demonstrated that isoacceptors can evolve through gene recruitment events in which a point mutation in the anticodon recruits a tRNA from one isoaccepting group to another (Saks et al., 1998).

1.5 tRNA interaction network

As previously described, there are 64 possible codons in the genetic code that encode for 20 amino acids. Not every combination of possible codons is utilized in every molecular system, but each codon is linked to a standard genetic code. Some codons are used preferentially in mRNA sequences resulting in a codon bias, where the presence of codons that encode for the same amino acid are used with varying frequency in mRNA. For all translation processes, there is a standard initiator codon (AUG) that codes for a very unique methionine amino acid (i-Met) and three stop codons (UAG, UAA and UGA). Stop codons do not have an associated tRNA.

tRNAs serving in the translation process are confined structurally to the bounds of the ribosome structure, and thus the tertiary structure is well conserved across the classes of tRNA. In addition to structurally conformity, each tRNA must be recognized by elongation factors and modification enzymes. Despite the necessity for similarity, each tRNA must be distinguished correctly by its cognate aaRS as to not incorporate incorrect amino acids into polypeptide chains. The nature of the recognition and discrimination of tRNA in this semi-closed system generates a network of tRNA-protein interactions, Figure 1.3. Theories rooted in statistics have allowed for the detection of specific nucleotides that allow to distinguish tRNAs and provide a measure of importance attached to that distinction (Freyhult et al., 2006; Ardell, 2010). Function logos recover known tRNA identity elements, features that govern the recognition and discrimination of tRNAs in the tRNA-protein interaction network (Giegé et al., 1998). Function logos contain a predicted set of tRNA identity elements that confer recognition to the correct aaRS; these predictions are now termed Class Informative Features (CIFs) (Amrine et al., ress). From this theoretical framework, network maps can be created to assess profiles of tRNA recognition and discrimination and potential describe the evolution of this network.

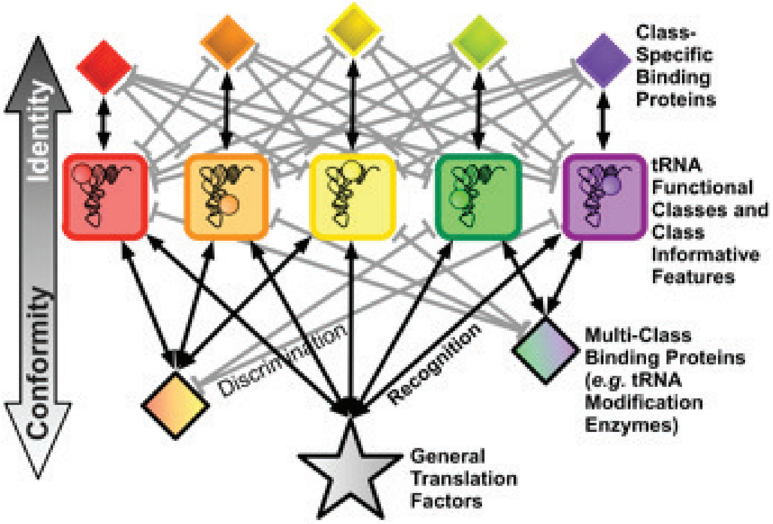
Identity
Conformity
Class-Specific Binding Proteins
tRNA Functional Classes and Class Informative Features
Multi-Class Binding Proteins (e.g. tRNA Modification Enzymes)
Discrimination
Recognition
General Translation Factors

Figure 1.3: A schema of the tRNA-protein interaction network. (Amrine et al., ress)

Chapter 2

Molecular Evolution of Eukaryotic Cytosolic tRNA in 12 Species of *Drosophila*

2.1 Introduction

Patterns of substitution rates footprint functional constraints and highlight potential adaptive evolutionary change in macromolecules. Historically these studies have been primarily on protein evolution. To identify functional constraints and adaptive changes, we need to identify what type of mutations occur, where they occur and at what rates. The sequencing of the *Drosophila* genomes have made it possible to investigate rates of interspecies divergence in a group of organism that vary considerably in their morphology, ecology and behavior. Studies using the twelve genomes have discovered a wide variety of evolutionary signatures in proteins and miRNA (*Drosophila* 12 Genomes Consortium, 2007; Stark et al., 2007; Hahn et al., 2007; Lu et al., 2008; Larracuente et al., 2008; Sella et al., 2009).

Even though tRNAs were the first RNAs to be sequenced and structurally solved (Hoagland et al., 1958; Holley et al., 1965), substitution rate analysis of tRNAs has not yet been published. Using the twelve *Drosophila* genomes, the first evolutionary rates in a non-coding RNA, miRNA, were shown to have signatures of positive selection through analysis of divergence and polymorphism patterns (*Drosophila* 12 Genomes Consortium, 2007). Additionally, the *Drosophila* 12 Genomes Consortium described a pattern of structural evolution whereby unpaired sites evolve more rapidly than paired sites in mature miRNA transcripts, which was not conserved on the complementary sequence (*Drosophila* 12 Genomes Consortium, 2007).

Transfer RNA genes are the most abundant family of ncRNA genes in all 12 genomes, with 297 tRNAs in *D. melanogaster* and 261 – 484 tRNA genes in the other species (*Drosophila* 12 Genomes Consortium, 2007). Early studies in tRNA evolution have demonstrated that various regions of the tRNA differ sharply in their fixation rates (Cedergren et al., 1981). The early works of Cedergren et al. (Cedergren et al., 1981) described a cluster of 26 sites from the acceptor, T, and anticodon stems that constitute the best mutational targets in tRNA. They conclude mutational "hot spots" have seen repetitive cycles of mutation and have reached mutational equilibrium, as much as possible given functional constraints. In addition to this early work in tRNA molecular evolution, further work in tRNA evolution has examined: the ancestry of tRNA alloacceptors and isoacceptors (Cedergren et al., 1980; Saks et al., 1998); the evolution of tRNA recognition by amino-acyl synthetases (Woese et al., 2000; Saks and Sampson, 2013); and the evolutionary origins of this ancient molecule (Di Giulio, 1992, 1995; Widmann et al., 2005). Recently published work in tRNA evolution in *Drosophila* suggests that tRNA genes are under substantial flux, using tRNAs from conservative orthologous sets, the combined rate of tRNA gene turnover (gains and losses) within the *Drosophila* genus is 2.18 per million years (Rogers et al., 2010). The average rate of gains is 1.30 per million years, with 0.88 losses per million years. These rates of tRNA gene turnover are more than 2-fold higher than the rates published using bacterial tRNA (Withers et al., 2006).

The structure and function of tRNA genes have been extensively studied, yet the general mechanisms controlling tRNA gene family evolution remain unclear, due to challenges in distinguishing paralogs from orthologs that are highly similar in sequence. Publications of carefully curated tRNA orthology sets in Drosophila (Rogers et al., 2010) and yeast (Byrne and Wolfe, 2005; Conant and Wolfe, 2008), in combination with modern computational tools has opened a new avenue of research to address the evolution of tRNA structure and function. The current work presents molecular substitution rates that have yet to be described for tRNA. I present here an analysis of the evolutionary rates of individual sites, structural elements, and alloacceptros of tRNA in twelve species of *Drosophila*. We confirm our site and structural results with data from yeast.

2.2 Materials and Methods

2.2.1 Data

tRNA sequences and annotations for 12 species of *Drosophila* were obtained from FlyBase (2008_07 release) (McQuilton et al., 2012) on October 16, 2011, see Appendix A Table 2.1. FlyBase 2008_07 release is presently missing some annotations of tRNA functional class and anticodons; thus tRNAs were annotated using the union of predictions from tRNAscan-SE 1.3.1 (Lowe and Eddy, 1997) and ARAGORN 1.2.34 (Laslett and Canback, 2004). Initiator methionine classifications were made using TFAM 1.3 (Tåquist et al., 2007), see Appendix A Table A.1. A total of 3504 tRNA were aligned using infernal 1.1 (Nawrocki et al., 2009) using the RFAM covariance model for the tRNA family (RF00005) built with infernal 1.1 (Burge et al., 2012). Alignments were edited manually using SeaView 4.3.4 (Gouy et al., 2010) to produce a final alignment 74 nucleotides in length that was then manually mapped to Sprinzl coordinates (Sprinzl and Vassilenko, 2005); coordinates were verified using the transfer RNA database (Jühling et al., 2009). Sprinzl coordinate 20A was retained for all subsequent analysis. The majority of the variable arm was removed; only Sprinzl coordinate 45 through 49 were retained for further analysis.

Orthology sets were downloaded on October 18, 2011 from http://gbe.oxfordjournals.org/content/2/467/suppl/DC1 (Rogers et al., 2010). A total of 753 orthologous sets are available encompassing 3218 unique tRNA transcript ids. Orthology sets that included all 12 species are limited to 47. To ensure that data sets were sufficiently large for analysis, all possible subsets of species were combined, in total 4096 (2^{12}) subsets were examined. Each subset was assigned a bit code for presence (1) or absence (0) of a species. All presented subset bit codes are ordered to reflect the phylogenetic tree published in Rogers *et al* (Rogers et al., 2010). Due to an overwhelming lack of data for *D. willistoni*, this species was removed from all subsequent divergence analyses. Subsets with seven or more species were ranked by number of orthology sets, length of resulting phylogentic tree, the number of segregating sites, and the number of parsimoniously informative sites. Tree lengths of pruned trees were calculated using the Bio::TreeIO module in BioPerl 1.4.0 (Stajich et al., 2002; Stajich and Hahn, 2005). The number of segregating and parsimoniously informative sites was calculated along the length of the alignment; these sites were further processed with a modulo 74 calculation. This calculation allowed identification of which Sprinzl coordinates are segregating and/or parsimoniously informative. Eighteen sets were chosen representing the largest alignment lengths balanced with largest numbers of segregating sites modulo 74, parsimoniously informative sites modulo 74, and phylogenetic distances (Table 2.2).

tRNA sequences and orthologies for twenty species of yeast were obtained from the Yeast Genome Order Browser (YGOB) (http://ygob.ucd.ie) on August 24, 2012 (Byrne and Wolfe, 2005). Annotations of tRNAs were confirmed using the same methods as described for *Drosophila*, described above, see Appendix A Table A.2. A total of 4730 tRNA were aligned using infernal 1.1 (Nawrocki et al., 2009) using the RFAM covariance model for the tRNA family (RF00005) built with infernal 1.1 (Burge et al., 2012). Alignments were edited manually using SeaView 4.3.4 (Gouy et al., 2010) to produce a final alignment 75 nucleotides in length that was then manually mapped to Sprinzl coordinates (Sprinzl and Vassilenko, 2005); coordinates were verified using the transfer RNA database (Jühling et al., 2009). Sprinzl coordinates 20A and 20B were retained for all subsequent analysis. The majority of the variable arm was removed; only Sprinzl coordinate 45 through 49 were retained for further analysis.

Yeast orthologies contained 1360 defined orthologies encompassing 4741 tRNA gene transcripts. Yeast orthologies include two tracks for each yeast species that underwent a gene duplication event resulting in two columns associated to each duplicated yeast species. Thus, one species could be represented twice in one orthology. To ensure that data sets were sufficiently large for analysis, all possible subsets of species were combined, in total 1048576 (2^{20} minus all subsets containing fewer than six species) subsets were examined. Each subset was assigned a bit code for presence (1) or absence (0) of a species, where the duplicated tracks are assigned a 1 if and only if one tRNA gene is present from the species. Duplicated tracks for post-genome duplication species pose a separate evolutionary question. Therefore, to avoid possible confounds for data analysis, an orthology is only relevant when a single tRNA gene copy is available for a post-duplication species. All presented subset bit codes are ordered to reflect the phylogenetic tree published with the YGOB database (Byrne and Wolfe, 2005). Similar to *Drosophila* orthologies, there is an overwhelming lack of orthologies including ample species for divergence analysis. Approximately 25% of the defined orthologies include 10 or fewer species. Data subsets were ranked and chosen by number of orthology sets and number of species represented.

2.2.2 Divergence Rate Analysis

Phylogenetic analysis was performed by MrBAYES 3.2.1 (Ronquist et al., 2012; Altekar et al., 2004; Huelsenbeck and Ronquist, 2001) using the general time reversible (GTR) substitution model (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990) with equal-distributed rate variation across sites and a proportion of invariable sites. Two simultaneous runs for $4 * 10^6$ generations were performed with diagnostics calculated every 500 generations. All subsets of data were curated into concatenated alignments by

Table 2.1: The number of tRNA genes, unique tRNAs, genome size (MB) and tRNA genes per MB are provided for twelve species of *Drosophila* for the tRNA annotation of the 2008_07 FlyBase release.

Species	**tRNA gene**	**Unique tRNA**	**Genome Size (MB)**	**tRNA Genes per MB**
D. simulans	266	90	85	3.13
D. sechellia	299	102	166	1.80
D. melanogaster	314	111	139	2.26
D. yakuba	328	89	165	1.99
D. erecta	284	85	152	1.87
D. ananassae	307	98	230	1.33
D. pseudoobscura	294	92	152	1.93
D. persimilis	305	104	188	1.62
D. willistoni	296	105	235	1.26
D. mojavensis	264	98	193	1.37
D. virilis	277	89	206	1.34
D. grimshawi	260	99	200	1.30

previously published orthologies (Byrne and Wolfe, 2005; Rogers et al., 2010). Thus, all comparisons were made across orthology sets eliminating the need to separate data by functional class. In an effort to reduce possible noise in the data, all orthologies that were identified to involve a class switch, either isoaccepter, or putative alloaccepter changes, or a mixture of functional and pseudogene predictions, were removed from all divergence analysis.

For single site data partitions, concatenated alignments of selected subsets of data (Table 2.2) were partitioned into 74 partitions for *Drosophila* and 75 for yeast. Each Sprinzl coordinate was defined as a separate partition. The general time reversible substitution model (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990) was used with no rate variation across sites. The nucleotide models were allowed to be unique for each partition, thereby allowing stationary state frequencies and all other substitution model parameters to be independent across partitions. For stationary state frequencies, a flat Dirichlet prior was used, Dirichlet(1,1,1,1). For the prior on topology, we assumed the published tree for the twelve *Drosophila* species (*Drosophila* 12 Genomes Consortium, 2007) and yeast (Byrne and Wolfe, 2005). Topology and branch lengths were constrained by setting the stochastic TBR mechanism and branch

Table 2.2: Data partitioning results for eighteen sets used in Mr. Bayes analysis. Sets were ranked primarily by largest alignment lengths balanced with largest numbers of segregating sites modulo 74, parsimoniously informative sites modulo 74, and phylogenetic distances.

Bitcode	**No. Species**	**Alignment Length**	**Gap-free Aln Length**	π	**Percent Id**	**Tree Length**	**No. Segregating Sites (mod 74)**	**No. Parsimoniously Informative Sites (mod 74)**
011111110000	7	9842	9590	46.0000	99.52	0.0174	2469 (68)	23 (14)
100011100111	7	6734	6567	38.4762	99.43	0.0232	1696 (66)	19 (15)
100111100101	7	7178	6994	41.3810	99.42	0.0231	1813 (66)	21 (15)
101111110000	7	10878	10598	54.3048	99.50	0.0176	2729 (68)	27 (15)
101111110111	10	6512	6350	33.1167	99.50	0.0266	1647 (66)	20 (15)
110011100101	7	6660	6489	40.7143	99.39	0.0238	1686 (67)	19 (15)
110011100111	8	6290	6134	37.5714	99.40	0.0249	1589 (67)	19 (15)
110111100101	8	6660	6489	37.9643	99.43	0.0248	1686 (67)	20 (15)
110111110000	7	9916	9661	47.7333	99.51	0.0178	2500 (68)	23 (14)
111011110000	7	9694	9445	48.4000	99.50	0.0178	2436 (68)	23 (14)
111101110000	7	9990	9734	46.8762	99.53	0.0174	2500 (69)	24 (14)
111110100101	8	7030	6849	37.4286	99.47	0.0241	1787 (68)	22 (15)
111110110000	7	10286	10023	41.2571	99.60	0.0171	2582 (68)	21 (12)
111111010000	7	10064	9807	43.4286	99.57	0.0167	2526 (68)	16 (11)
111111100000	7	9990	9730	45.0952	99.55	0.0179	2511 (68)	19 (12)
111111110000	8	9694	9445	44.6071	99.54	0.0188	2435 (68)	24 (14)
111111110110	10	6660	6492	33.8278	99.49	0.0239	1688 (68)	22 (16)
111111110111	11	6142	5991	31.8344	99.48	0.0278	1554 (67)	20 (15)

multiplier to a zero probability. We used Metropolis-coupled MCMC (Metropolis et al., 1953; Hastings, 1970), as implemented in MrBayes 3.0 (Huelsenbeck and Ronquist, 2001; Ronquist et al., 2012), to estimate the posterior probability distribution. All Bayesian analyses were run for $4 * 10^6$ generations saving rate multipliers every 500 generations.

For structural data partitions, concatenated alignments of selected subsets of data (Table 2.2) were partitioned into 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.

In addition to the data partitions by Sprinzl coordinate and structural components, divergence analyses were run using partitions of alloacceptors, anticondon dependent-independent aminoacylation, and variable arm length (Type I and Type II) in *Drosophila*. For alloacceptor data partitions, data was partitioned into the standard twenty amino acids with the additional classes of initiator methionine and selenocysteine. For divergence analysis of anticodon dependence/independence, Ala, Leu and Ser tRNAs were partitioned separately from anticodon dependent tRNAs – all classes excluding Ala, Leu and Ser. Type II tRNAs, those with long variable arms (Leu, Ser and Tyr), and Type I tRNAs, those with short variable arms (all classes excluding Leu, Ser and Tyr) are partitioned separately. Finally we partitioned the data according to aminoacyl synthetase class: class I (Arg, Cys Glu, Gln, Ile, Leu, Met, Tyr, Trp and Val) and class II (Ala, Asn, Asp, Gly, His, Lys, Phe, Pro, Ser, and Thr).

2.2.3 Statistical Analysis of Divergence 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. Results presented in the main text are for three data subsets of *Drosophila*: (i) the set with the largest number of species (111111110111); (ii) the set with the second largest tree length (110011100111, second only to the set with largest number of species); and (iii) the set with the largest number of included orthologies (101111110000).

2.2.4 Evolution of Class Informative Features

We identified putative class informative features (CIFs) for all 12 species of *Drosophila* using previously described methods (Freyhult et al., 2006) and a custom perl script (bplogofun courtesy of D. Ardell), which in addition to single site CIFs also calculates CIFs for base pairs of the stem regions in tRNA and p-values adjusted with a Benjamini-Yekutieli false discovery rate (Benjamini and Yekutieli, 2001). Unlike earlier work by Freyhult et al (Freyhult et al., 2006), each individual *Drosophila* species provides ample tRNA for the generation of individual function logos eliminating the need to combine tRNA across species. The combination of the individual state logos for each of the 4 nucleotides creates a tRNA profile for the tRNA interaction network.

To determine evolution of CIFs, a custom perl script that parses bplogofun output and describes the evolutionary potential of site-nucleotide states for a given CIF was created. First the information for each CIF in a stack of letters for a given site-nucleotide is calculated by multiplying the total information of the site-nucleotide state by the Gorodkin fraction for a given amino acid class within the stack. CIFs are filtered if the total information for the site-nucleotide state is zero. All remaining CIFs are assigned a bit code representing the species that share the CIF. CIFs are annotated by bit code with conservation statistics for each CIF.

For this analysis it is no longer necessary to confine the data to tRNAs present in the Rogers' orthologies (Rogers et al., 2010), but rather we used the entire set of annotated tRNAs for all twelve species (Table 2.1). The number of annotated tRNA genes in FlyBase does not imply that all tRNAs are functional. Cove scores for all functional classes were examined using a cove score cut-off below 60 bits. 655 tRNAs were removed from the CIF evolution analysis. In an effort to filter potential noise from our data, after low cove scoring tRNAs are removed, CIF analysis was re-run and conservation statistics re-calculated.

To aid in parsimony mapping of CIF evolution, an outgroup was added to the analysis. The genome for *Musca domestica* was downloaded from NBCI, ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/invertebrates/ Musca_domestica/Musca_domestica-2.0.2/, on October 8th, 2013. *M. domestica* and the twelve *Drosophila* species belong to the section Schizophora in the order Diptera. The estimated time of divergence between *M. domestica* and *Drosophila* is between 20 – 80 million years ago (Wiegmann et al., 2003). The genome was annotated using predictions from tRNAscan-SE 1.3.1 (Lowe and Eddy, 1997). Initiator methionine classifications were made using TFAM 1.3 (Tåquist et al., 2007). A total of 969 tRNA were aligned using infernal 1.1 (Nawrocki et al., 2009) using the RFAM covariance model for the tRNA family (RF00005) built with infernal 1.1 (Burge et al., 2012). Alignments were edited manually using SeaView 4.3.4 (Gouy et al., 2010) to produce a final alignment 74 nucleotides in length that was then manually mapped to Sprinzl coordinates (Sprinzl and Vassilenko, 2005); coordinates were verified using *D. melanogaster* genes from the transfer RNA database (Jühling et al., 2009). Sprinzl coordinate 20A was retained for all subsequent analysis. The majority of the variable arm was removed; only Sprinzl coordinate 45 through 49 were retained for further analysis.

2.3 Results

2.3.1 Divergence Analysis

Divergence Rate by Alloaceptor Partitions

Glutamic acid and the initiator methionine class show the highest rate of divergence but the signal tends to dissipate with the addition of more orthology sets, Figure 2.1. The initiator methionine data includes only one orthology set, and results should be interpreted with caution.

Alloacceptor classes can be broadly grouped based on the mechanism of amino-acylation by their amino-acyl synthetase (aaRS). Class I amino-acyl synthetases amino-acylate tRNAs on the $2'$-OH of an adenosine nucleotide and include alloacceptor classes: Arg, Cys Glu, Gln, Ile, Leu, Met, Tyr, Trp, and Val. Class II tRNA are amino-acylated on the $3'$-OH of the same adenosine as class 1 and include alloacceptor class: Ala, Asn, Asp, Gly, His, Lys, Phe, Pro, Ser, and Thr. We do not detect any difference in patterns of divergence when tRNAs are partitioned according to their amino-acyl synthetase class, Figure 2.2.

Although tRNAs share a similar cloverleaf-like secondary and L-shaped tertiary structure, they can be divided into two classes according to the length of the variable arm. Class I tRNAs have a short variable arm with 45 nucleotides, while class II tRNAs have a long variable arm with $>$10 nucleotides. Divergence patterns between these classes of tRNAs show higher divergence rates in class I tRNAs.

The majority of aaRS recognize their cognate tRNA and interact with the anitcodon, but a small group of aaRS do not interact with their respective tRNA through the anticodon: alanine, leucine and serine. Alloacceptor classes were partitioned based upon anticodon independent (Ala, Leu and Ser) recognition and anticodon dependent recognition (all other tRNA classes). Anticodon dependent recognition tRNAs show elevated rates of divergence compared to anticodon independent recognition.

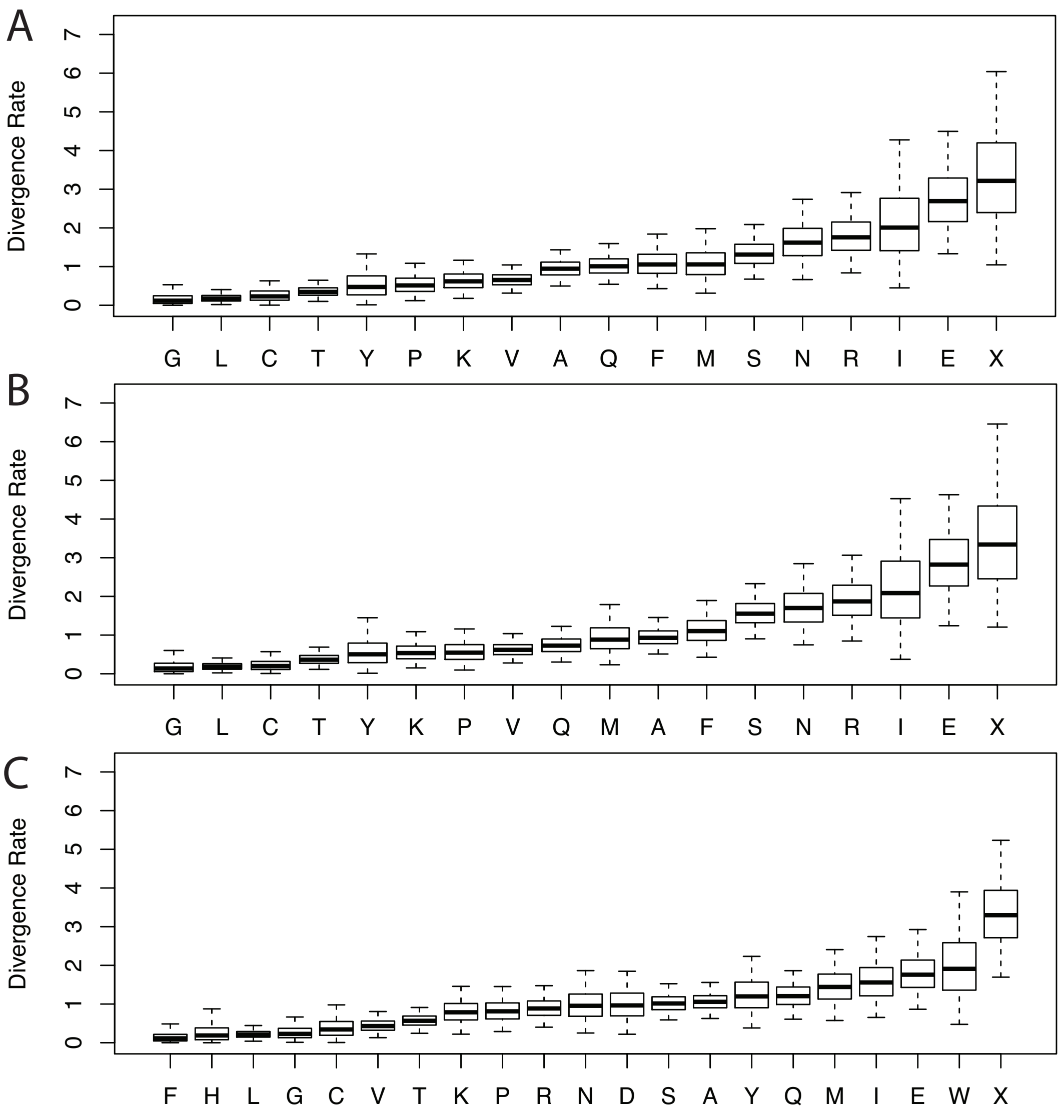

Figure 2.1: Alloacceptor divergence rates for three Mr.Bayes analyses show elevated rates in glutamic acid and initiator methionine. Each data subset shows similar patterns of site rates. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

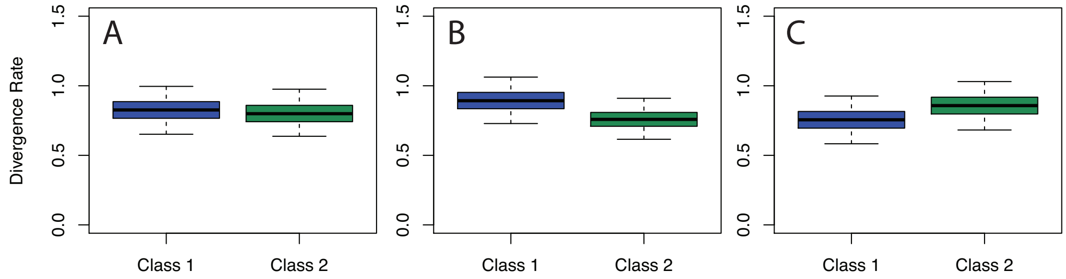

Figure 2.2: Divergence rates for three Mr.Bayes analyses indicate no distinguishable divergence rates between Class I and Class II amino-acyl synthetases. Each data subset shows similar patterns of site rates. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

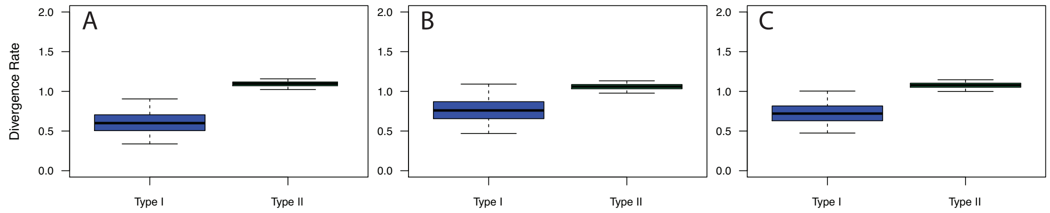

Figure 2.3: Class I and II divergence rates for three Mr.Bayes analyses show elevated divergence in class I tRNAs. Each data subset shows similar patterns of site rates. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

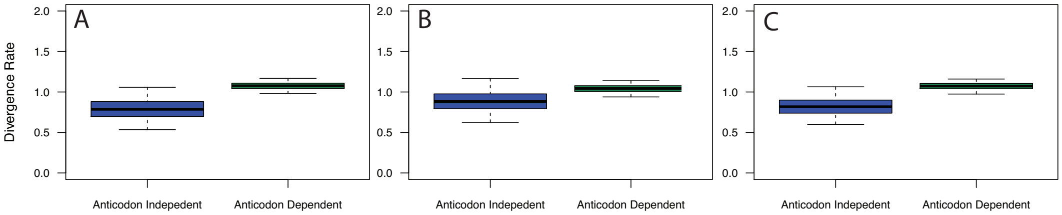

Figure 2.4: Divergence rates for anticodon-dependent/independent recognition for three Mr.Bayes analyses show elevated rates of divergence in anticodon dependent recognition tRNAs. Each data subset shows similar patterns of site rates. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

Divergence Patterns in tRNA Structure

In all structural divergence analyses, the D-loop shows the greatest evolutionary rate, followed by the T-loop, Figure 2.5. Considering the relative constraint that acceptor stems are expected to have from identity-driven interactions with proteins, the acceptor stem shows a surprisingly high rate of evolution, Figure 2.5. This pattern of divergence is dissimilar to the one observed in yeast, Figure 2.5. Yeast show elevated divergence in the acceptor and anticodon stems.

Divergence Patterns by Sprinzl Coordinate

In all *Drosophila* site divergence analyses, three sites show elevated rates in Sprinzl coordinates 16, 17, and 60, Figure 2.7. Sites in yeast show a conserved elevation in Sprinzl coordinate 17, but not in sites 16 and 60, Figure 2.8. The contribution of individual orthologies and their related sequence patterns were mapped to functional classes and chromosomal locations for *D. melanogaster* (Appendix A, Table A.3). In total, using the data subset containing the largest number of species (111111110111), we found 23 orthology groups that contributed to the divergence signals in sites 16, 17

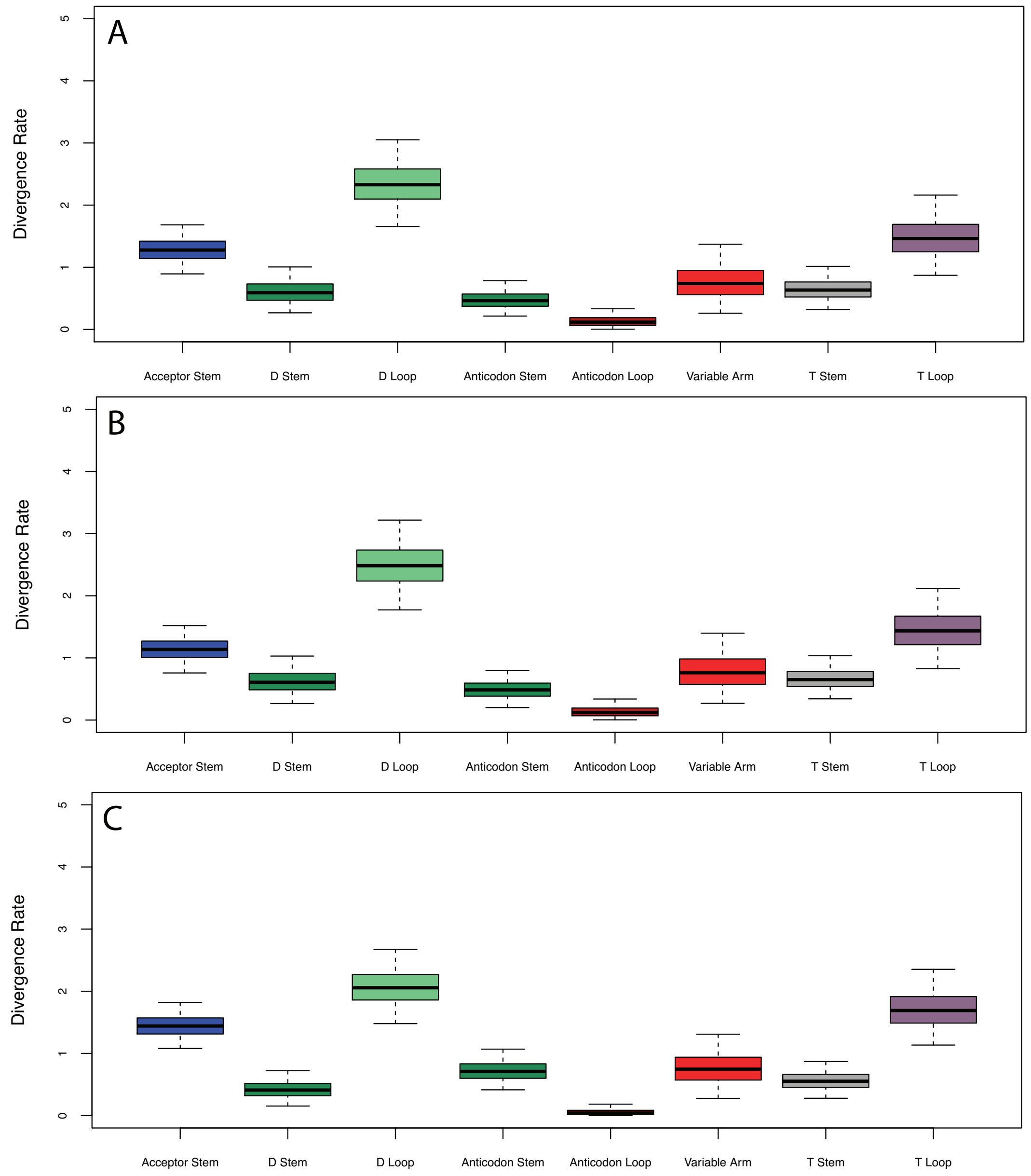

Figure 2.5: Structural divergence rates for three Mr.Bayes analyses. The D-loop shows the greatest evolutionary rate, followed by the T-loop. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets; (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets.

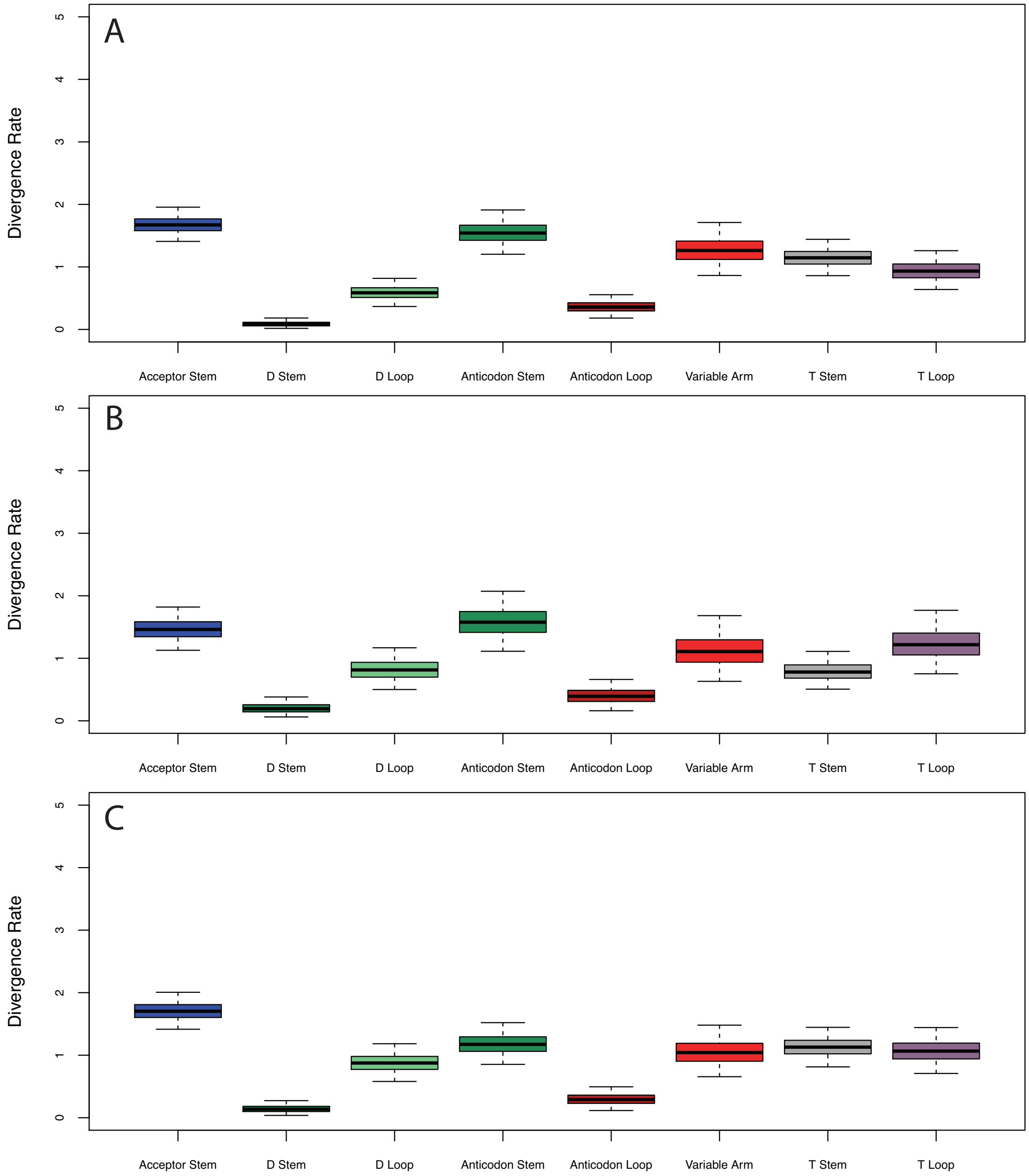

Figure 2.6: Structural divergence rates for three Mr.Bayes analyses. The anticodon and acceptor stems show the greatest evolutionary rates. (A) 4 post-duplication species (*Saccharomyces uvarum, Saccharomyces kudriavzevii, Saccharomyces mikatae, Saccharomyces cerevisiae*) and 2 pre-duplication species (*Torulaspora delbrueckii, Lachancea kluyver*); 51 orthologous sets; (B) 4 post-duplication species (*Saccharomyces uvarum, Saccharomyces kudriavzevii, Saccharomyces mikatae, Saccharomyces cerevisiae*) and 2 pre-duplication species (*Zygosaccharomyces rouxii, Torulaspora delbrueckii*); 49 orthologous sets; and (C) 3 post-duplication species (*Saccharomyces uvarum, Saccharomyces mikatae, Saccharomyces cerevisiae*) and 3 pre-duplication species (*Torulaspora delbrueckii, Lachancea kluyveri, Lachancea thermotolerans*); 42 orthologous sets.

and 60. Only 2 orthologies showed signs of covariance between two sites; 1 for 16-17, and 1 for 17-60. This subset of orthologies covered twelve alloaceptor classes (Ala, Arg, Asn, Ile, Leu, iMet, Met, Phe, Pro, Sel, Ser, Val) and were found in Muller elements A – E using gene locations in *D. melanogaster*. This diversity is expanded to include two more alloacceptor classes (Gln, The) when looking across the 18 data subsets examined in this work.

Sites 16, 17 and 60 are three sites known to belong to an ion binding pocket formed by a total of eight residues in the D and T loops – Sprinzl coordinates 15, 16, 17, 18, 19, 20, 59, and 60 (Behlen et al., 1990). This pocket was first described as site 1 in the original orthorhombic crystal form (Holbrook, Sussman, Warrant, Church, and Kim, Holbrook et al.) and site 3 in the monoclinic structure (Jack et al., 1977). This ion binding site has been shown to bind many metal ions including: magesium, cobalt, manganese, and lead (Jack et al., 1977; Holbrook, Sussman, Warrant, Church, and Kim, Holbrook et al.; Shi and Moore, 2000; Behlen et al., 1990). Even though the patterns of site elevations are slightly different in yeast, the elevated site 59 in yeast is also known to belong to the aforementioned binding pocket.

We expanded the data past the orthologies used in the site divergence rate analysis to include all orthologies with changes found in sites with elevated site rates, and also changes in the expanded set of ion binding pocket sites (15, 18, 19, 20, and 59), Appendix A, Tables A.4 – A.5. Chromosomal gene location showed a significant relationship with mutations in pocket sites 16, 17 and 60 ($X^2 = 56.3266$, $N = 692$, $p = 6.96 * 10^{-11}$), additionally the extended pocket sites also show a significant relationship with chromosomal gene location ($X^2 = 87.4858$, $N = 726$, $p < 2.2 * 10^{-16}$, Appendix A, Table A.4). Sequence mutations in sites 16, 17 or 60 of the pocket were not associated with alloacceptors ($X^2 = 20.9616$, $N = 672$, $p = 0.52311$), nor were sequence mutations in the extended pocket associated to alloacceptor ($X^2 = 23.2573$, $N = 703$, $p = 0.3873$, Appendix A, Table A.5). Orthologies containing class switching tRNAs more often exhibit changes in the pocket sites with elevated divergence rates ($X^2 = 8.8797$, $N = 713$, $p = 0.002884$), but not in the extended pocket sites ($X^2 = 2.7304$, $N = 696$, $p = 0.09846$).

Divergence using Other Evolutionary Models

Phylogenetic analysis using the general time reversible (GTR) substitution model (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990) with equal-distributed rate variation across sites and a proportion of invariable sites was the primary evolutionary model used for divergence analysis. The GTR evolutionary model assumes a symmetric substation matrix. To add value to our analyses we ran additional analyses using the HKY model (Hasegawa et al., 1985). The HKY model distinguishes between the rate of transitions and transversions and allows unequal base frequencies. We also tested the HKY model with a gamma-distributed rate variation across sites. In each analysis, two simultaneous runs for $4 * 10^6$ generations were performed with diagnostics calculated every 500 generations. We tested the additional evolutionary models in the structure and site partitioned data.

2.3.2 Patterns of Evolution in Class Informative Features

Individual comparison of function logos across twelve species, Figures 2.13 – 2.16, indicate the presence of CIF evolution among these species. For example, *D. melanogaster* experiences an individual CIF gain for histidine at U23 in the comparative logo, Figure 2.14. To do this task by eye can be tedious and also lead to possible missed signals. For the first time, we have analytically described the evolutionary patterns of CIFs using tRNA gene sets from individual species. *Drosophila* tRNAs show a surprising amount of CIF evolution, especially considering the short divergence time separating species, 62.9 MYA (Tamura et al., 2003). In total, we identified 1920 invariant CIFs with the inclusion of the *M. domestica* outgroup, an additional 19 are invariant within the ingroup. There are 280 single species gains and 59 single species

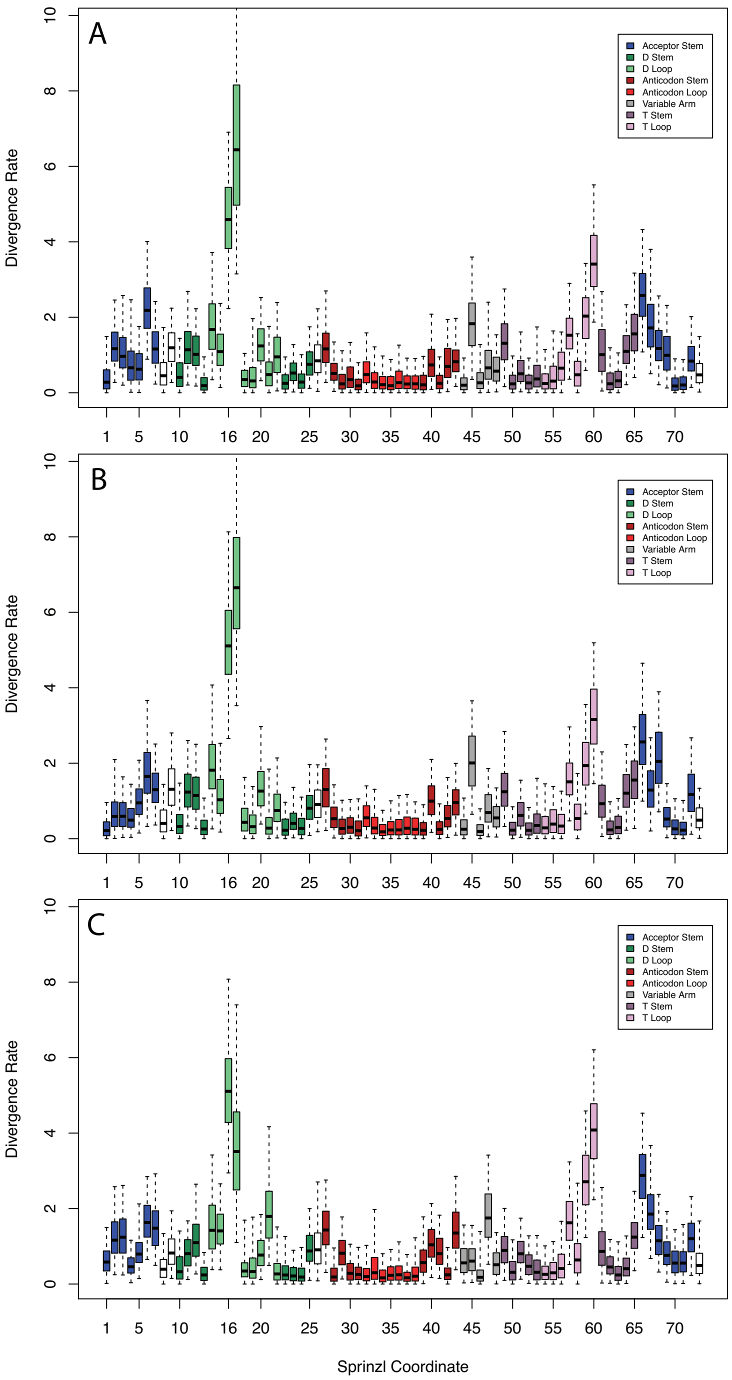

Figure 2.7: Site divergence rates for three Mr.Bayes analyses show elevated site rates in Sprinzl coordinates 16, 17, and 59. Each data subset shows similar patterns of site rates. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

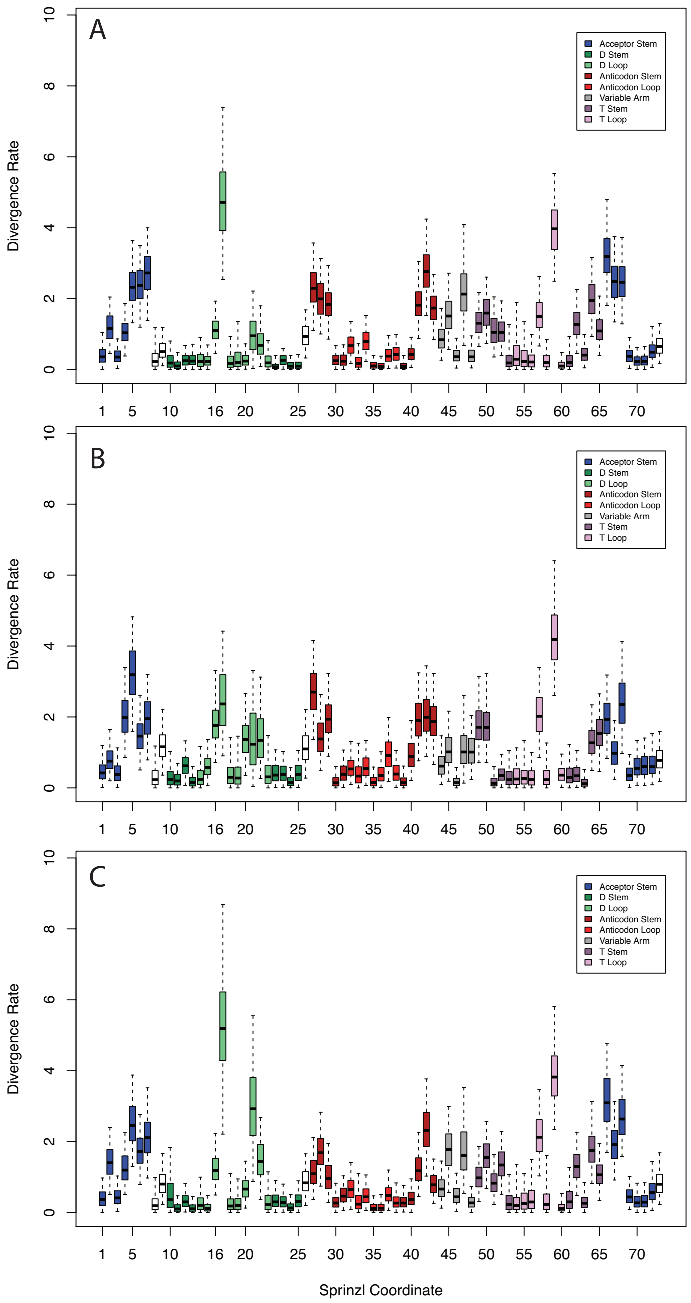

Figure 2.8: Site divergence rates for three Mr.Bayes analyses show individual Sprinzl coordinates 17 and 59. Many sites in the acceptor stem are also elevated. (A) 4 post-duplication species (*Saccharomyces uvarum, Saccharomyces kudriavzevii, Saccharomyces mikatae, Saccharomyces cerevisiae*) and 2 pre-duplication species (*Torulaspora delbrueckii, Lachancea kluyver*); 51 orthologous sets; (B) 4 post-duplication species (*Saccharomyces uvarum, Saccharomyces kudriavzevii, Saccharomyces mikatae, Saccharomyces cerevisiae*) and 2 pre-duplication species (*Zygosaccharomyces rouxii, Torulaspora delbrueckii*); 49 orthologous sets; and (C) 3 post-duplication species (*Saccharomyces uvarum, Saccharomyces mikatae, Saccharomyces cerevisiae*) and 3 pre-duplication species (*Torulaspora delbrueckii, Lachancea kluyveri, Lachancea thermotolerans*); 42 orthologous sets.

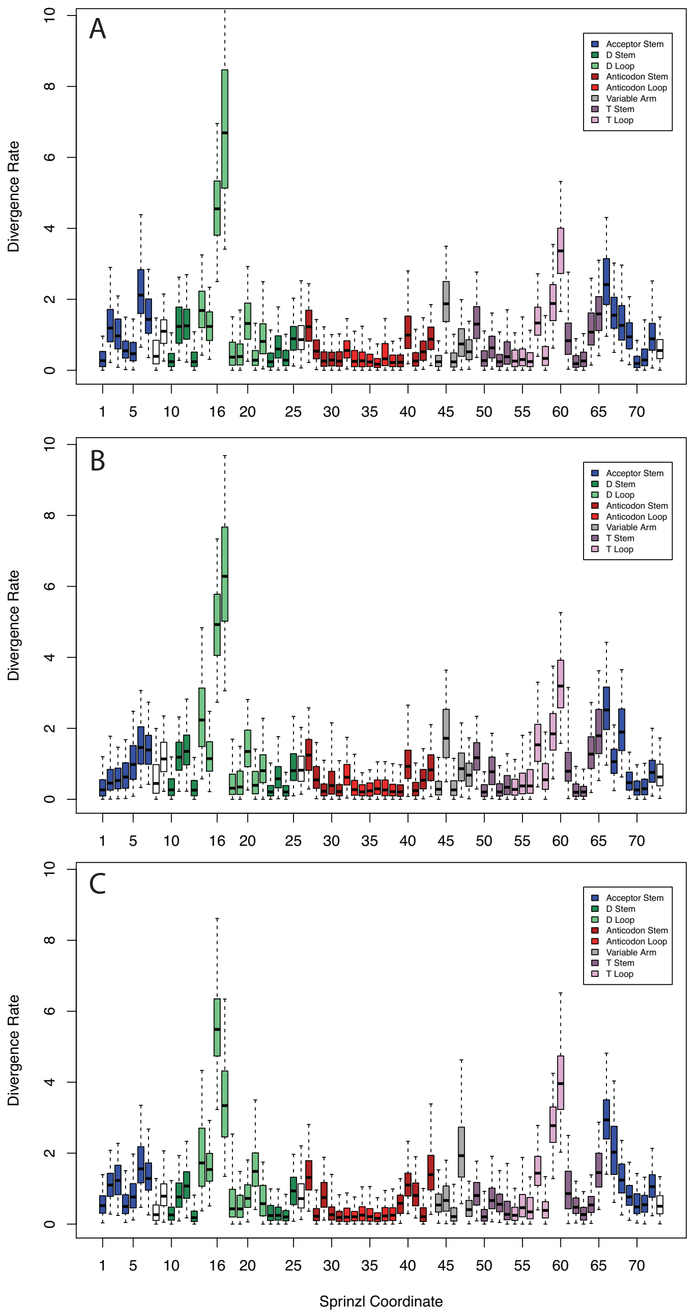

Figure 2.9: Site divergence rates for three Mr.Bayes analyses show confirmatory results to the GTR model when simulations are run with a HKY evolutionary model. Sites shows similar patterns to the structural divergence rates in the GTR model data. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

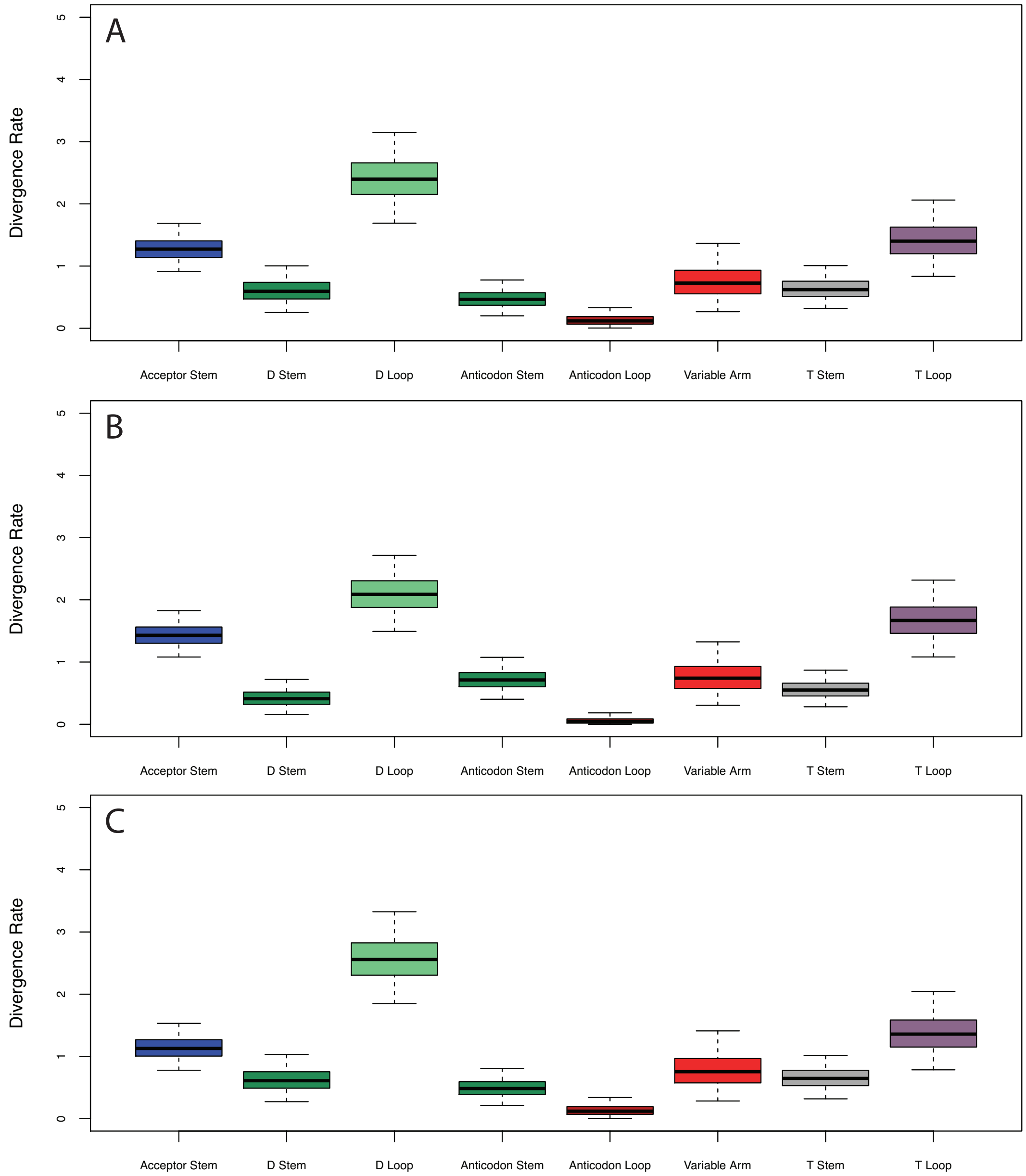

Figure 2.10: Structural divergence rates for three Mr.Bayes analyses show confirmatory results to the GTR model analysis. Each data subset shows similar patterns to the structural divergence rates in the GTR model data.. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

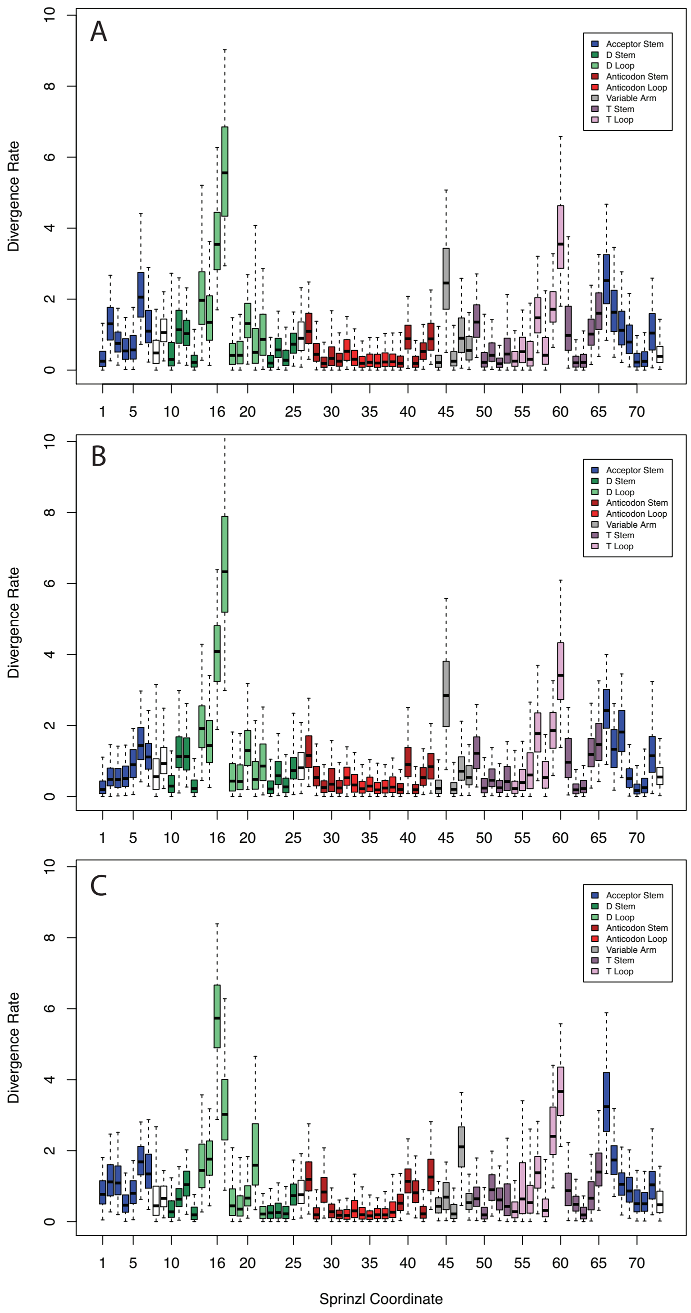

Figure 2.11: Site divergence rates for three Mr.Bayes analyses show confirmatory results to the GTR model when simulations are run with a HKY+γ evolutionary model. Sites show similar patterns to the structural divergence rates in the GTR model data.. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

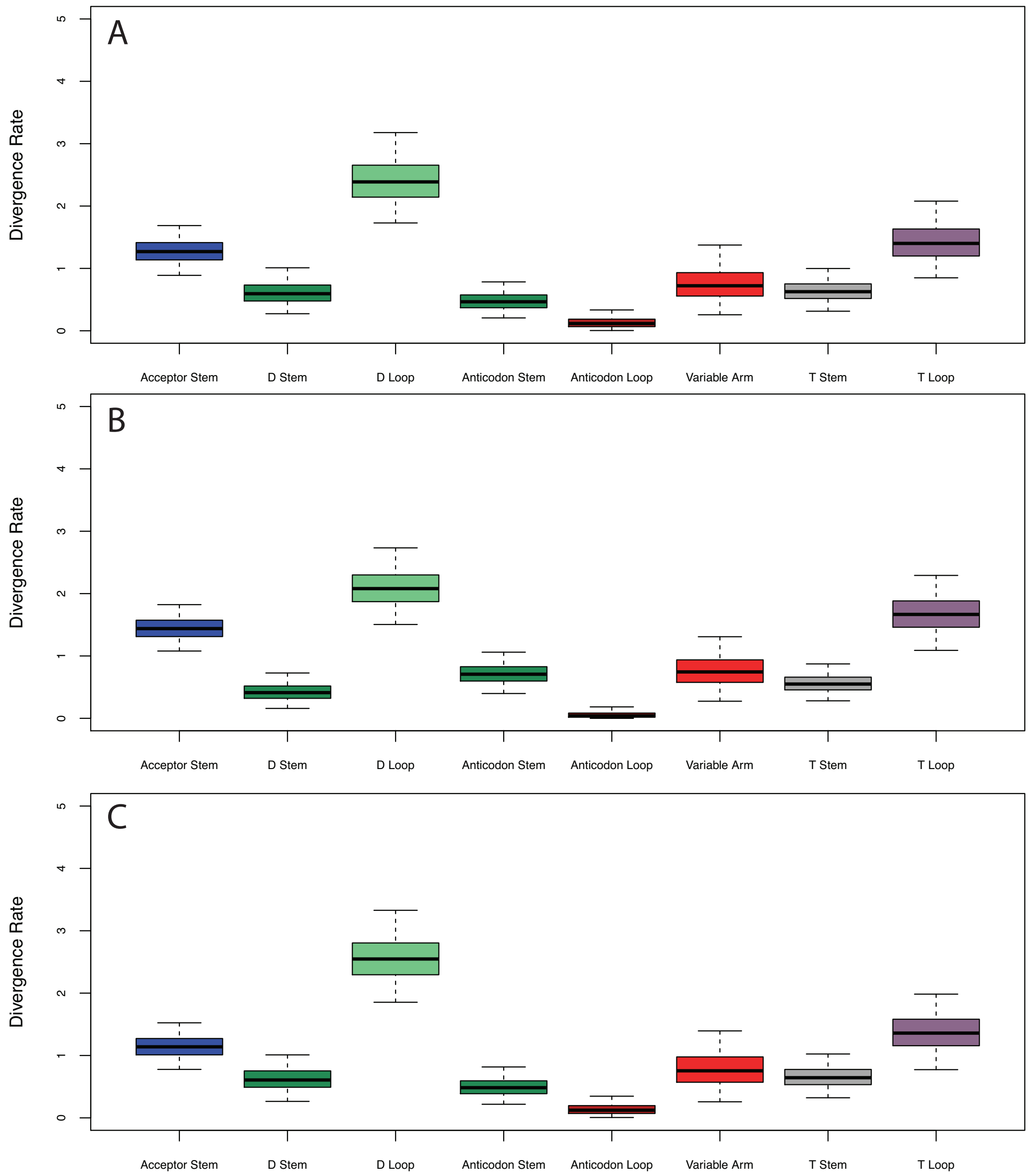

Figure 2.12: Structural divergence rates for three Mr.Bayes analyses show confirmatory results to the GTR model when simulations are run with a HKY+γ evolutionary model. Each data subset shows similar patterns to the structural divergence rates in the GTR model data.. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets); (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets

losses, Figure 2.17. Additionally, we are able to track the evolution of base paired CIFs using our new data pipeline, which accounted for approximately 25% our of total data.

Since this is the first time such an analysis has been performed it is challenging to distinguish possible noise from true signal. We have filtered the data using cove score cut-offs, total site information, total CIF information as well as correcting for false discovery rate (FDR) using a Benjamini-Yekutieli correction. The data presented in Figure 2.17 represents a filtered set of CIFs that we believe to be informative and true signals at nodes in the *Drosophila* phylogeny.

2.4 Discussion

We have implemented a method to determine substitution rate analysis of tRNAs and analyzed the evolutionary rates of alloacceptors, structural components and Sprinzl coordinates. Additionally, we have characterized patters of evolution in Class Informative Features. From this data, we have described: (i) elevated rates in the D and T-loop, along with surpassingly high divergence rates in the acceptor stem; (ii) elevated divergence in three individual sites known to belong to an ion binding pocket; and (iii) patterns of CIF evolution in a set of closely related species.

2.4.1 Elevated Divergence in the Aceptor Stem

As expected, the D-loop shows the greatest evolutionary rate, followed by the T-loop (Cedergren et al., 1981). The D-loop is highly variable in length across tRNA class and broadly across the domains of life. Prior work in miRNA transcripts has described a pattern of structural evolution whereby unpaired sites evolve more rapidly than paired in mature miRNA transcripts (*Drosophila* 12 Genomes Consortium, 2007). Taken together along with results from early tRNA evolution (Cedergren et al., 1981) these lines of evidence suggest that the elevated rates in the D-loop are not surprising, yet many sites contained within the D-loop serve for overall tRNA stability through triplexes

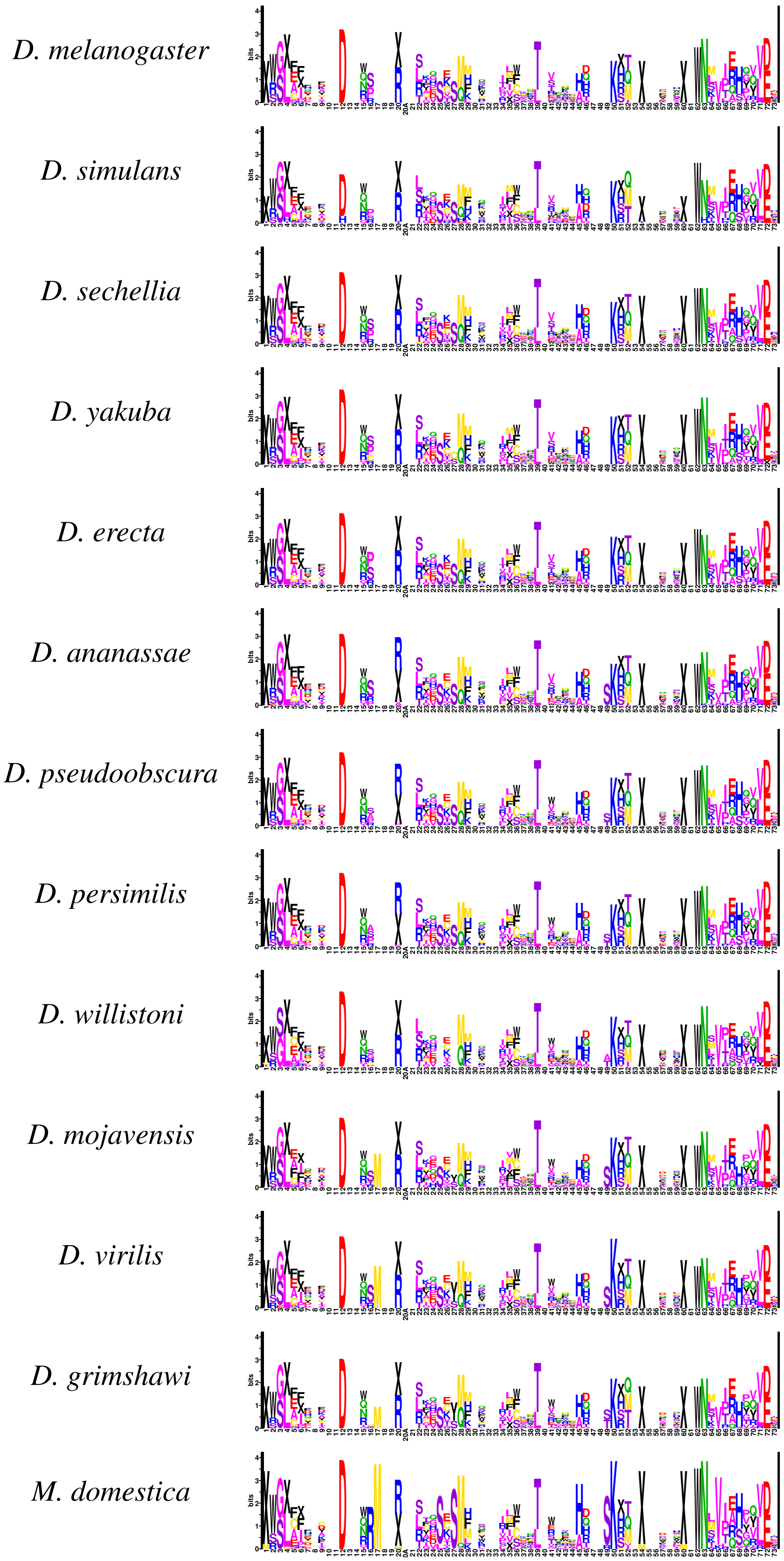

Figure 2.13: Comparative Function Logos across *Drosophila* with outgroup *M. domestica* for Adenine

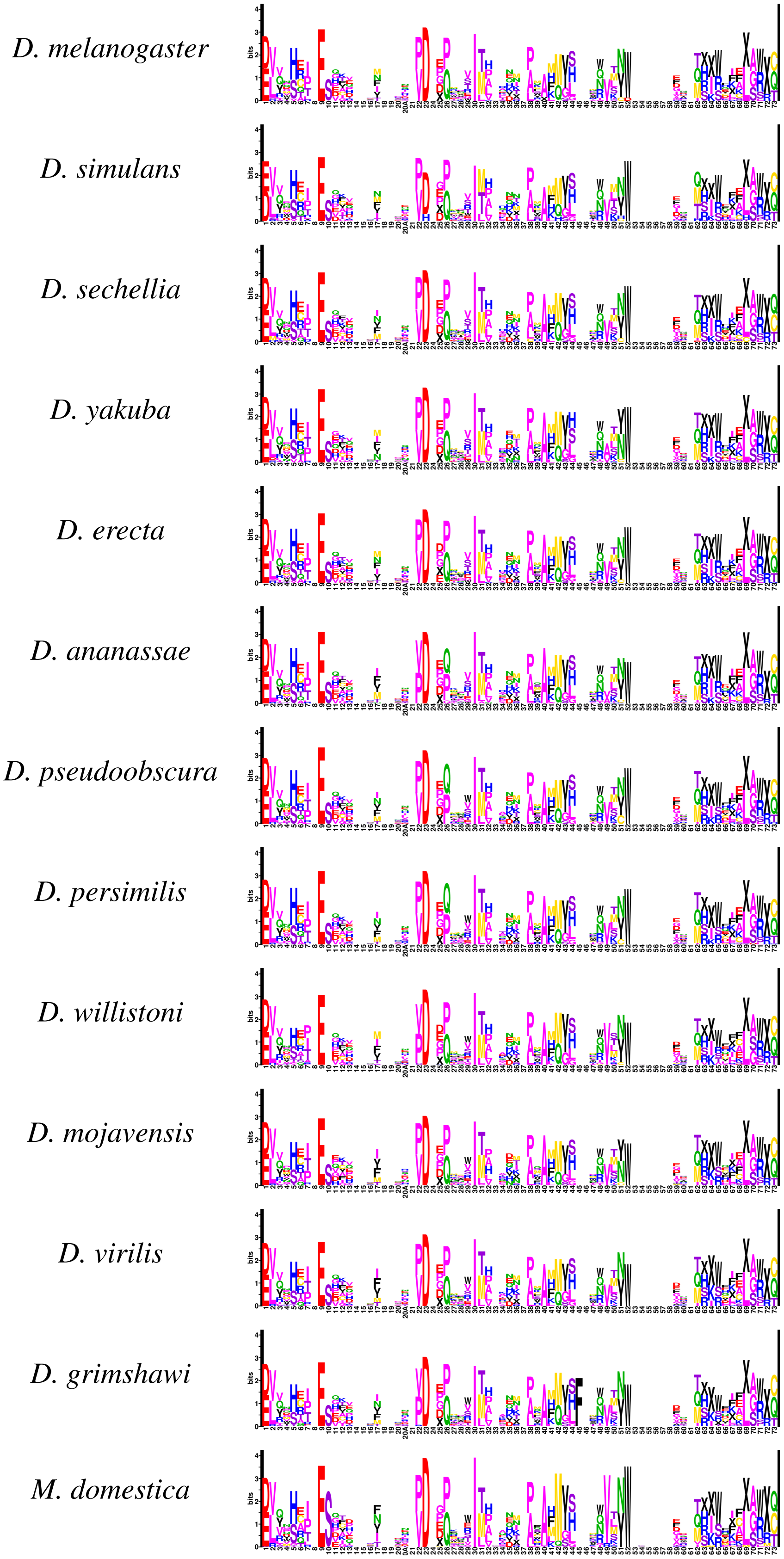

Figure 2.14: Comparative Function Logos across *Drosophila* with outgroup *M. domestica* for Uracil

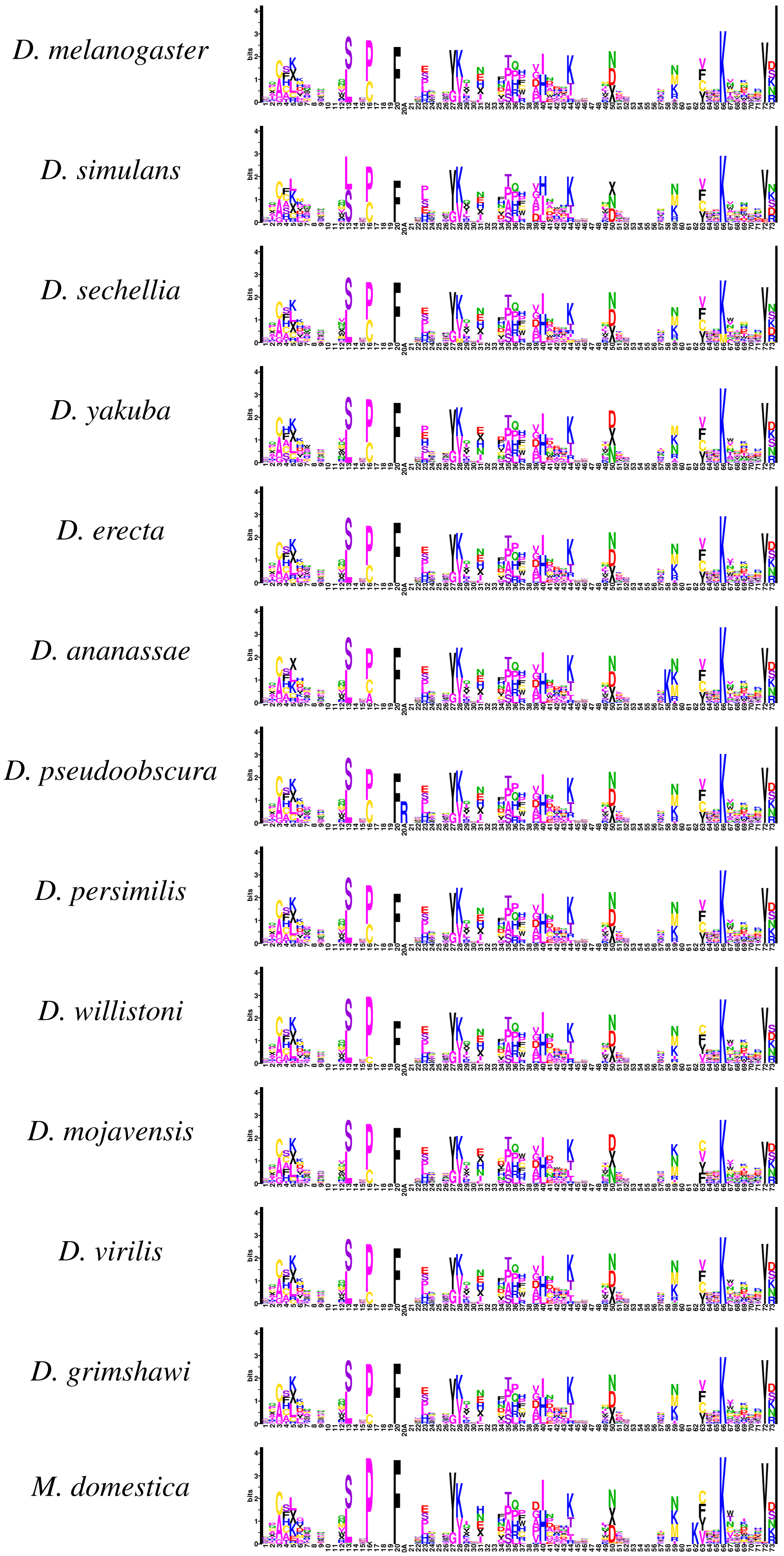

Figure 2.15: Comparative Function Logos across *Drosophila* with outgroup *M. domestica* for Guanine

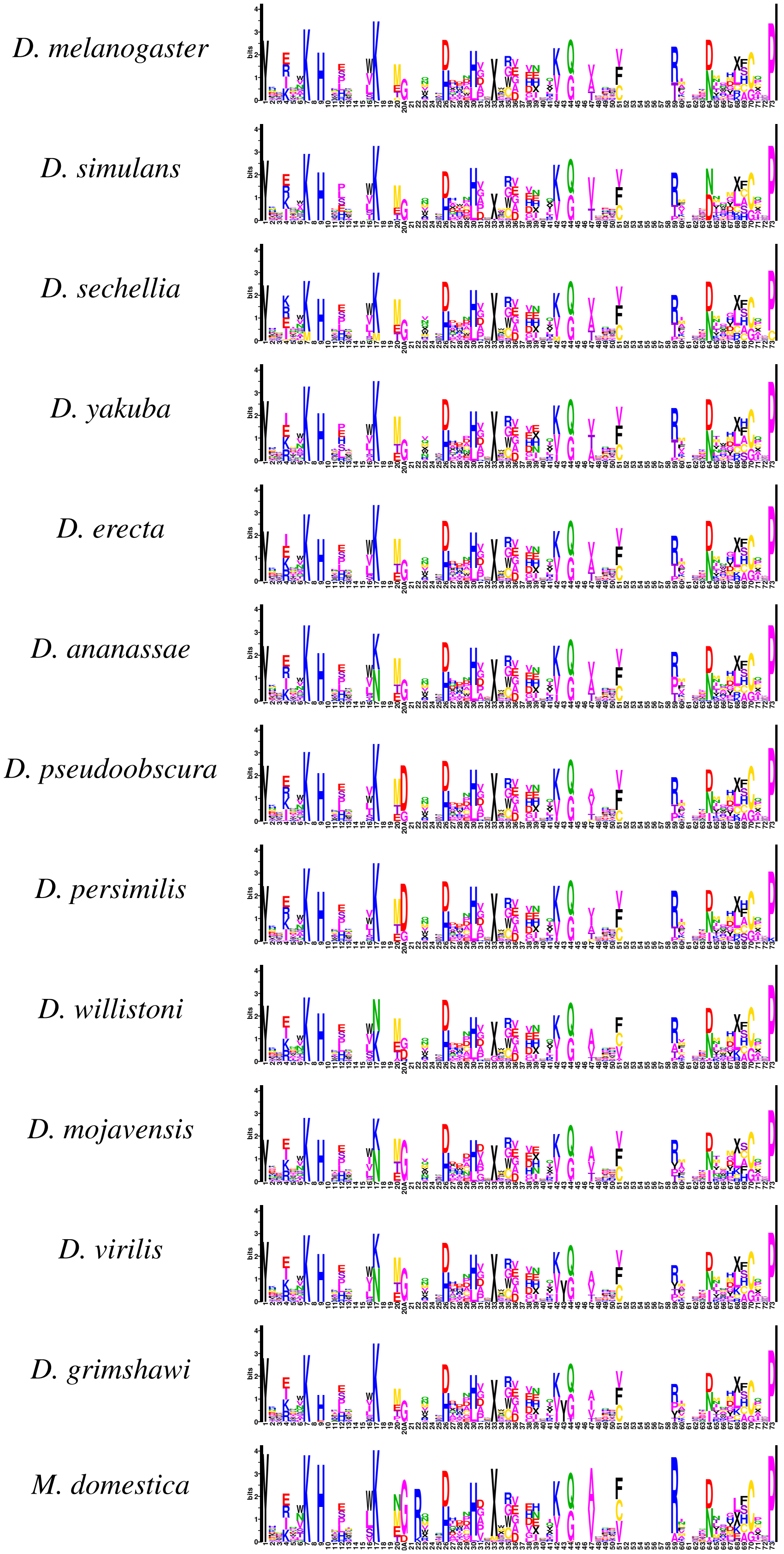

Figure 2.16: Comparative Function Logos across *Drosophila* with outgroup *M. domestica* for Cytosine

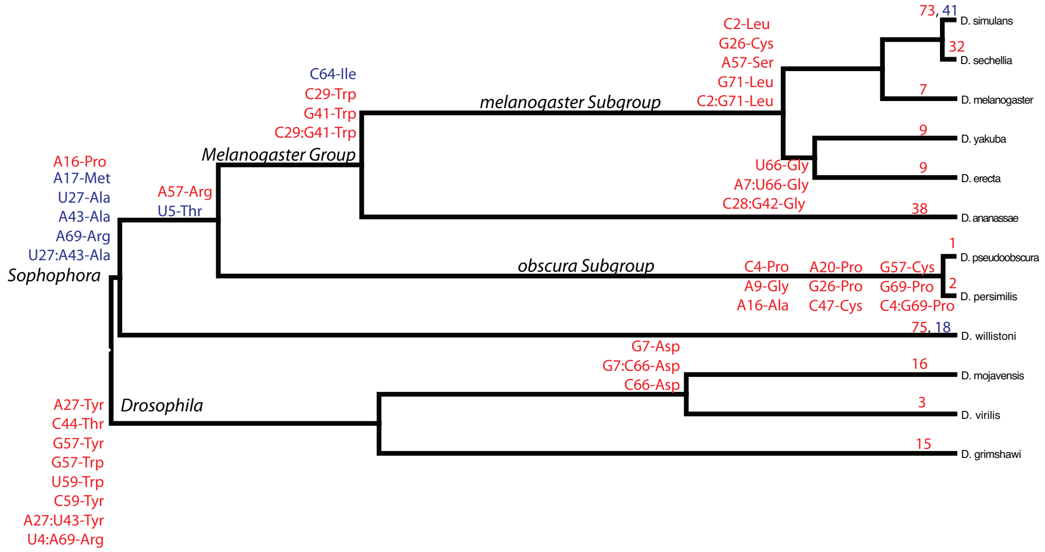

Figure 2.17: Inferred tRNA gains (red) and losses (blue). Individual gains and losses of CIFs are represented numerical at tree tips. Tree topology taken from Clark et al. (*Drosophila* 12 Genomes Consortium, 2007). Branch lengths are scaled according to divergence times found in Tamura et al. (Tamura et al., 2003).

with the variable arm and through stability conferred through ion binding (Marck and Grosjean, 2002; Feig and Uhlenbeck, 2005). Marck and Grosjean report highly conserved base pair mismatching with the variable arm suggestive that the D-loop plays an important role on the structural and functional properties of tRNA molecules (Marck and Grosjean, 2002).

tRNA identity elements are an experimentally verified set of recognition sites on tRNAs which allow for recognition and discrimination by aaRS. In tRNA belonging to both classes of synthetates, tRNA identity elements lie predominately at the distal ends of the tRNA in the acceptor stem. Considering this relative constraint that the acceptor stem is expected to have from identity-driven interactions with proteins, the acceptor stem shows a surprisingly high rate of evolution among *Drosophila* and yeast tRNAs.

2.4.2 Elevated Divergence Rate in Three Sites of an Ion Binding Pocket

Our results indicate a surprisingly elevated divergence in three Sprinzl coordinates,16, 17 and 60, as well as 59 in yeast. These sites belong to an important ion binding pocket formed by a total or eight residues in the D and T loops (Behlen et al., 1990). This particular ion binding site has been shown to bind many metal ions including: magnesium, cobalt, manganese, and lead (Jack et al., 1977; Holbrook, Sussman, Warrant, Church, and Kim, Holbrook et al.; Shi and Moore, 2000; Behlen et al., 1990). Recently a crystallography paper has improved the resolution of the tRNA structure (Shi and Moore, 2000). In the resolved structure only one site's resolution was distinguishable from its lower resolution antecedents: Sprinzl site 16. Sprinzl site 16 is typically a fairly conserved uracil among eukaryotes (Marck and Grosjean, 2002), which is most likely post-translationally modified dihydrouridine. Shi and Moore demonstrated that this site is sensitive to ionic conditions (Shi and Moore, 2000).

2.4.3 Substantial Evolution in Class Informative Features

Prior work from this lab has developed function logos as a method to predict, at the level of individual nucleotides before post-transcriptional modification, the template of information in tRNA gene sequences associated to specific functional identity classes (Freyhult et al., 2006). Function logos contain a predicted set of tRNA identity elements that confer recognition to the correct aaRS; these predictions are now termed Class Informative Features (CIFs) (Amrine et al., ress). Patterns of CIF evolution and turnover have not been previously published.

At the root of the subgenera *Sophophora* and *Drosophila*, we find evidence for fourteen gains and losses of CIFs, two of which (A16 and A17) are in site with elevated divergence rates. The pattern of CIF evolution within such a short divergence time is a surprising and novel result. Additionally, the recently diverged species pair, *D. pseudoobscura* and *D. persimilis*, have lost a CIF for alanine in site 16. Sites 16, 17 and 60 are associated to 11 individuals gains among seven species (*D. sechellia, D. melanogaster, D. ananassae, D. willistoni, D. mojavensis, and D. grimshawi*), and 1 loss in *D. simulans*. Additionally, sites 16, 17 and 60 are present in 54 invariant CIFS and 5 additional invariant CIFs within the in-group. We propose that the elevated divergence rates described in these sites may be related to the process of CIF evolution in *Drosophila*.

In general, there is broad conservation in tRNA identity among ortholog sets across *Drosophila* species, yet within the defined orthologies for *Drosophila* there were 22 observed changes in tRNA function, the majority of which involve mutations in anticodons (Rogers et al., 2010). Seven of these shifts are mapped to changes within the *D. simulans* – *D. sechellia* sisters. We also see elevated gains and losses of CIFs within this sister taxa. We have been able to map some of the class switches to gains and losses of CIFs in our analysis. *D. simulans* is reported to have experienced an anticodon shift from tRNA$^{\mathrm{Met}}_{\mathrm{CAT}}$ to tRNA$^{\mathrm{The}}_{\mathrm{CGT}}$ (Rogers et al., 2010). In site 33, we see perfect conservation of a CIF for initiator methionine if cytosine is present, yet we identify C33-Threonine as one of the single gains in *D. simulans* corresponding to the anticodon shift reported in Rogers et al. (Rogers et al., 2010). The C33-Threonine CIF is a highly significant CIF in *D. simulans*.

There is evidence for high rates of CIF turnover in *D. simulans*, *D. ananassae* and *D. willistoni*. One possible reason for the evaluated rates in *D. simulans* is that the genome assembly is a mosaic of several different low coverage assemblies (Begun et al., 2007; *Drosophila* 12 Genomes Consortium, 2007). In *D. ananassae* and *D. willistoni*, there are elevated tRNA gene counts likely due to elevated pseudo-tRNA gene predictions (*Drosophila* 12 Genomes Consortium, 2007). CIF turnover rates in these two species may indicate some sort of tRNA identity flux in a gene system with potential extra tRNA gene copies.

2.4.4 Conclusions

We have implemented a method to determine substitution rate analysis of tRNAs and analyzed the evolutionary rates of alloacceptors, structural components and Sprinzl coordinates. Prior work from this lab has developed function logos as a method to predict, at the level of individual nucleotides before post-transcriptional modification, the template of information in tRNA gene sequences associated to specific functional identity classes (Freyhult et al., 2006). Additionally, we have applied novel methods to describe patterns of evolution in Class Informative Features (CIFs) in twelve closely related species.

Chapter 3

Molecular Dynamics of Ion Binding in *Drosophila* tRNA

3.1 Introduction

Divalent metal ions are known to be very effective in the stabilization of RNA molecules. Additionally, nucleoside modifications participate collectively in the stabilization of the tRNA molecule. A recent crystallography paper has improved the resolution of the tRNA structure (Shi and Moore, 2000) from the original structures published in the 1970's (Holbrook, Sussman, Warrant, Church, and Kim, Holbrook et al.; Jack et al., 1977) . In this work, there are eleven identified divalent cation-binding sites in tRNA. In the entire molecule, aside from the metal ion binding sites, only one site in the higher resolution structure was indistinguishable from its lower resolution antecedents. In particular the position of a single nucleotide, Sprinzl site 16, was called into question by the higher resolution structure, a Sprinzl coordinate that is known to be often post-translationally modified to dihydrouridine. Shi and Moore demonstrated that this site is particularly sensitive to ionic conditions.

Prior work (Chapter 2) identified three rapidly evolving sites inside this major ion binding pocket, Sprinzl coordinate 16, 17, and 60. Research suggests that the ion binding capacity of this pocket plays an important role in tRNA biology (for a review please see Feig and Uhlenbeck (2005)). Numerous divalent metal ions are known to stimulate the cleavage of RNA, including lead (Pb^{2+}), europium (Eu^{3+}), magnesium (Mg^{2+}), and manganese (Mn^{2+}) (Krzyzosiak et al., 1988; Ciesiołka et al., 1989; Marciniec et al., 1989; Deng and Termini, 1992; Pan et al., 1993; Wrzesinski et al., 1995). The best example of this process has been described in yeast $tRNA^{Phe}$ bound to Pb^{2+} which results in rapid and specific hydrolysis of the RNA chain between residues H2U17 (dihydrouridine) and G18 in the D loop, both in solution and in the crystal (Brown et al., 1985; Krzyzosiak et al., 1988; Behlen et al., 1990). Research has suggested the presence of a U59-C60 sequence in the TψC-loop was required for highly efficient and specific lead-induced cleavage (Krzyzosiak et al., 1988). This speculation has been revised since a G59-C60 sequence and an A59-C60 in yeast tRNAs is cleaved even faster (Ciesiołka et al., 1989). Therefore, the common sequence feature of tRNAs cleaved by lead with high efficiency is the presence of C60 in the TψC-loop, suggesting an overall importance for C60 in the formation of the strong lead binding sites. The catalytic RNA cleavage reactions based on the yeast $tRNA^{Phe}$ have no apparent biological significance; however they suggest that there may exist folding motifs for other RNAs that can utilize metal ions to perform site-specific cleavages (Deng and Termini, 1992). Ion specific induced cleavage of tRNA could reflect a general property of RNA which have evolved into highly specific self-cleaving RNA processing events (Deng and Termini, 1992).

Crystal structures are representations of thermodynamically stable conformations, and cannot provide information concerning the molecular movement of these structures under various conditions. Molecular dynamic simulations capitalize on crystal structures to theoretically predict the dynamic behavior of molecules. As previously mentioned, tRNA was the first RNA to be sequenced (Holley et al., 1965; Holley, 1965) and have a solved crystal structure (Holbrook, Sussman, Warrant, Church, and Kim, Holbrook et al.; Jack et al., 1977). tRNA was also the first RNA computationally studied using molecular dynamic simulations (Harvey et al., 1985;

Prabhakaran et al., 1985). Long-time simulations of tRNA have been performed with tRNA free in solution (Li and Frank, 2007), bound to aaRS (Sethi et al., 2009), bound to EF-Tu (Eargle et al., 2008), and bound to the ribosome (Sanbonmatsu et al., 2005). Molecular dynamics are able to describe the effects of modified nucleosides and ions in stabilizing RNA structure since there are no general rules to describe the effects of modified nucleosides on tRNA structure and dynamics (Alexander et al., 2010).

In an effort to better understand results from our earlier divergence studies, we wanted to investigate the dynamic behavior of sites in the ion binding pocket. Unfortunately no crystal structure is available for *Drosophila* tRNAs. Our first aim was to recapitulate the structural flexibility published in the Shi and Moore crystal structure in *Saccharomyces cerevisiae*, and to be able to investigate the dynamic motion of sites in the ion pocket in the presence of two cations, magnesium and manganese. Additionally, we sought to model *D. melanogaster* tRNAs using homology modeling along a *S. cerevisiae* crystal structure background. Finally we were interested in visualizing *D. melanogaster* tRNAs in the presence of two cations, magnesium and manganese, with naturally occurring pocket mutations (described in Chapter 2, also see Table A.5).

3.2 Materials and Methods

3.2.1 Pocket Dynamic in *Saccharomyces cerevisiae* tRNA$^{\mathrm{Phe}}$

Initial atomic coordinates for the *S. cerevisiae* tRNA$^{\mathrm{Phe}}$ structure were obtained from the RCSB Protein Data Bank, PDB ID: 4TRA, wildtype pocket sequence: D16-D17-C60 (Westhof et al., 1988). Parameter and topology files were prepared using the tleap utility of AmberTools12 (Case et al., 2012). Parameters for all standard bases were obtained from Amber ff99, and parameters for all modified bases were obtained from the SantaLucia lab modified nucleic acids website (Aduri et al., 2007). Base mutations were performed using in-house scripts which aligned the substituted base backbone and sidechain center of mass with the original bases. No steric clashes were observed using this approach. Coordinates and topology files were converted to GROMACS (Hess et al., 2008) format using the amb2gmx.pl script (Mobley et al., 2006). All structures were then placed in a 10nm^3 periodic box, and solvated using TIP3P water, 20 mMol $MgCl_2$, 20mMol $MnCl_2$, and 110 mMol NaCl. Lennard-Jones parameters for all ions were taken from Amber ff99, except for Mn^{2+} which were taken from the study by Bradbrook et al. (Bradbrook et al., 1998). In order to facilitate initial ion binding, the locations of the three of the Mn^{2+} ions were set to match those found in the Mn-bound *S. cerevisiae* tRNA^{Phe} structure, PDB ID: 1EHZ, wildtype pocket sequence: D16-D17-C60 (Shi and Moore, 2000). 10000 steps of steepest-descents energy minimization were performed on each system, followed by 100ps of simulation at constant temperature (300K) and pressure (1atm) for equilibration using GROMACS 4.5.5. Finally, 10ns of NVT-ensemble production simulation were performed for each system at 300K.

3.2.2 Modeling Homologous and Mutated tRNAs in *Drosophila*

Initial atomic coordinates for the *S. cerevisiae* tRNA^{Phe} structure were obtained from the RCSB Protein Data Bank, PDB ID: 1EHZ (Shi and Moore, 2000) and PDB ID: 4TRA (Westhof et al., 1988). Sequence information for the *D. melanogaster* tRNA^{Phe} (wild type pocket sequence: D16-D17-C60) and *D. melanogaster* tRNA^{Val} (wild type pocket sequence: C16-Gap17-C60) were obtained from the transfer RNA database (Jühling et al., 2009). Base mutations were performed on the 1EHZ structure using in-house scripts which aligned each necessary base substitute (best fit of backbone atoms and sidechain center of mass) to create both of the *D. melanogaster* tRNAs above. Additionally, two mutants of tRNA^{Val} (C16D and C16U) were created using the same approach. Fortunately, no steric clashes were observed using this approach. Parameter and topology files were prepared using the tleap utility of AmberTools12 (Case et al., 2012). Parameters and template structures for all standard RNA bases were obtained from Amber ff99. Similarly, parameters and template structures for all modified RNA bases were obtained from the SantaLucia lab modified nucleic acid website (Aduri et al., 2007). Lennard-Jones parameters for Mg^{2+} and Mn^{2+} ions were taken from Li et al. (Li et al., 2013). Coordinates and topology files were converted to GROMACS (Hess et al., 2008) format using the amb2gmx.pl script (Mobley et al., 2006).

In order to facilitate initial ion binding, a "Mn Bound" condition was prepared for the *D. melanogaster* structures. The locations of three Mn^{2+} ions, and six Mg^{2+} ions were set to match those found in the *S. cerevisiae* tRNA^{Phe} structure, 1EHZ. In order to prepare a "Mg Bound" condition lacking Mn^{2+}, the *D. melanogaster* tRNA structures were updated by replacing two of the Mn^{2+} ions (those closest to the pocket) with Mg^{2+} since these locations were observed to accommodate Mg^{2+} by Shi and Moore. The third Mn^{2+} ion was simply removed, and no replacement was provided as Shi and Moore found no Mg^{2+} binding in this location (Shi and Moore, 2000).

A total of 10 structures were prepared using the above protocol: wildtype *S. cerevisiae* tRNA^{Phe} from 1EHZ (Mn), wildtype *S. cerevisiae* tRNA^{Phe} from 4TRA (Mg), wildtype *D. melanogaster* tRNA^{Phe} from 1EHZ (Mn), wildtype *D. melanogaster* tRNA^{Phe} from 1EHZ (Mg), wildtype *D. melanogaster* tRNA^{Val} from 1EHZ (Mn), wildtype *D. melanogaster* tRNA^{Val} from 1EHZ (Mg), *D. melanogaster* tRNA^{Val} (C16D) from 1EHZ (Mn), *D. melanogaster* tRNA^{Val} (C16D) from 1EHZ (Mg), *D. melanogaster* tRNA^{Val} (C16U) from 1EHZ (Mn) and *D. melanogaster* tRNA^{Val} (C16U) from 1EHZ (Mg). All structures were then placed in a $10nm^3$ periodic box, and solvated using TIP3P water, 10 mMol MgCl2, and 110 mMol NaCl. 10000 steps of steepest-descents energy minimization were performed on each system, followed by 500ps of simulation at constant temperature (300K) and pressure (1atm) for equilibration using GROMACS 4.6.3, with restraints (1000 kJ/mol) on the tRNA, and weak restraints (100 kJ/mol) on the Mg^{2+} and Mn^{2+} ions. Finally, NVT-ensemble production simulation were performed for each system at 300K, with frames saved every 1ps. The total simulation time for each system differs at the time of writing, as the simulations are currently ongoing:

Mg Bound		Mn Bound	
D. melanogaster, tRNA$^{\mathrm{Phe}}$ (wt)	20ns	*D. melanogaster*, tRNA$^{\mathrm{Phe}}$ (wt)	20ns
D. melanogaster, tRNA$^{\mathrm{Val}}$ (wt)	30ns	*D. melanogaster*, tRNA$^{\mathrm{Val}}$ (wt)	10ns
D. melanogaster, tRNA^{Val}(C16D)	30ns	*D. melanogaster*, tRNA$^{\mathrm{Val}}$ (C16D)	7.5ns
D. melanogaster., tRNA$^{\mathrm{Val}}$ (C16U)	30ns	*D. melanogaster*, tRNA$^{\mathrm{Val}}$ (C16U)	9.5ns

In order to assess the structural differences observed in each of the simulations, snapshots from every 10ps of the last 5ns of each simulation were extracted. Structural differences between the pocket and overall tRNA structure were calculated by extracting the relevant backbone atoms. For the pocket, backbone atoms from bases 15, 16, 17, 18, 19, 20, 59, and 60, were extracted, and all frames were concatenated into one trajectory file (both Mg Bound and Mn bound conditions). The 1EHZ and 4TRA structures were included as well to facilitate direct comparison to the crystal structures. Metric scaling (Gower and Legendre, 1986) was performed on the resulting combined trajectory by computing root-mean-squared distances on all pairs of structures using MDSCTK v1.2 (Phillips et al., 2008), and post-processing using in-house scripts. The pocket motion was projected onto the two largest principal components, accounting for 50.56% of the total variance in the data. Scatter plots of these projections were produced in R v.3.0.2 (R Core Team, 2013), and points referring to different tRNAs were color- and shape-coded to facilitate comparison of the results. The same approach was taken using the backbone atoms from all bases to compare the global impact of base mutations, where the first two principal components accounted for 61.38% of the total variance in the data.

Representative pocket structures for each tRNA in both the Mg bound and Mn bound conditions were selected using a kernel density estimate approach. Bivariate gaussian kernels were placed on each data point in the space spanned by the first two principal components of the metric scaling analysis. The variance of the gaussians along the first and second PCs was determined using a common rule-of-thumb, $4.24 \times \sigma_{PC} \times N^{-1/5}$, where σ_{PC} is the variance of the data along the principal component and N is the number of samples, yet the results were highly insensitive to perturbations in these parameter values. The density of the space was determined by summing the kernel values at each point in a 200 by 200 grid spanning the space of the two PCs. The tRNA pocket frame closest to the point with the highest density, representing the most common conformational state, was selected as the representative structure, and this process was repeated to obtain representative structures for all tRNAs examined. Representative structures can be found in Figure 3.3.

3.3 Results

3.3.1 Recapitulation of Pocket Dynamic in *S. cerevisiae* tRNA$^{\mathrm{Phe}}$

At 10 nanoseconds, the molecular dynamics simulation of *S. cerevisiae* tRNA$^{\mathrm{Phe}}$ are similar in pocket structure to the published to 1EHZ structure (Shi and Moore, 2000), thus we are able to capture through simulations the structural conformation in a Mn^{2+} environment. The simulations indicate that the rotation of Sprinzl site 60 is largely responsible for the positional changes in Sprinzl site 59. Sprinzl site 17 plays an important role in guiding the Mn^{2+} ion to the binding pocket.

3.3.2 Homology of *Drosophila* tRNA from a *Saccharomyces* Crystal Structure

In our analysis, *S. cerevisiae* tRNAs, the pocket backbones show a distinct pattern of placement along the first principle component when bound to different cations, Figure 3.2 A and B. The pocket conformation of the wildtype *D. melangaster* tRNA$^{\mathrm{Phe}}$ and *D. melangaster* tRNA$^{\mathrm{Val}}$ have similar shift in conformational states going from a magnesium to a manganese bound state compared to *S. cerevisiae* indicating successful homology modeling of *D. melangaster* tRNAs from a *S. cerevisiae* crystal structure backbone. In the principle component analysis, we see the same shift along the first principle component between the pocket backbone of *D. melangoaster* wildtype tRNAs and the natural shift between the pocket backbone of *S. cerevisiae* crystal

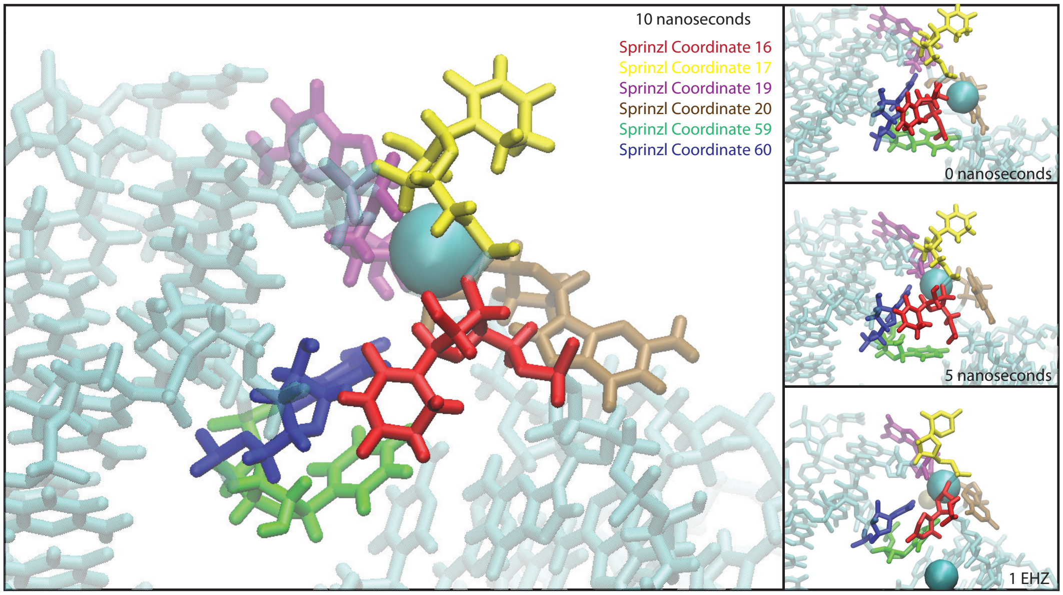

Figure 3.1: Molecular dynamics simulations indicates the rotation of Sprinzl site 60 is largely responsible for the positional changes in Sprinzl site 59. Sprinzl site 17 plays an important role in guiding the Mn^{2+} ion to the binding pocket. The 10 ns frame closely resembles the 1EHZ structure (Shi and Moore, 2000).

structures in their ion bound states.

In our analysis of the complete tRNA backbone of *S. cerevisiae* tRNAs, the pocket backbone shows no change in placement along the first or second principle component when bound to different cations. The overall backbone conformation of the wildtype *D. melangaster* $tRNA^{Phe}$ shows the same lack of conformational shift in the transition from a magnesium to manganese bound state indicating that the crystal structure from *S. cerevisiae* is likely sufficient for homology modeling of the *D. melangaster* $tRNA^{Phe}$. In comparison, the wildype *D. melangaster* $tRNA^{Val}$ experiences a shift along the first principle component, suggestive that there are distinct overall structural differences between *D. melangaster* $tRNA^{Phe}$ and *D. melangaster* $tRNA^{Val}$ when using the methods applied in the current work.

Visual comparison of the pocket backbone to the wildtype *S. cerevisiae* $tRNA^{Phe}$ to wild type *D. melangaster* $tRNA^{Phe}$ indicates a similar structure in the presence of each cation, Figure 3.3 A and E. In comparison, the wildype *D. melangaster* $tRNA^{Val}$ is also similar to the *S. cerevisiae* backbone, Figure 3.3 B and F, though the conformation is not as close as that of wildtype *D. melangaster* tRNA$^{\mathrm{Phe}}$.

3.3.3 Impact of Sequence Mutation in Ion Bound Structures

Although the mutated *D. melangaster* tRNA$^{\mathrm{Val}}$ (C16D) shows a similar shift along the first principle component in the pocket backbone when moving from a magnesium to a manganese bound state, we see additional shifting along the second principle component, a behavior unlike wildtype *S. cerevisiae* and *D. melanogaster* tRNAs. If the mutated uracil in site 16 of the *D. melangaster* tRNA$^{\mathrm{Val}}$ (C16U) is not post-translational modified to a dihydrouridine, we see an backward shift along the first principle component compared to wildtype *S. cerevisiae* and *D. melanogaster* tRNAs very unlike wildtpe *S. cerevisiae* and *D. melanogaster* tRNA behavior.

Visual comparison of the pocket backbone to the wildtype *S. cerevisiae* tRNA$^{\mathrm{Phe}}$ to mutated *D. melangaster* tRNA$^{\mathrm{Val}}$ indicate mutations in the presence of either cation alter the pocket backbone, Figure 3.3 C, D and G, H. The mutation of a cytosine to an unmodified uracil exhibits a more drastically altered backbone compared to wildtype *S. cerevisiae* tRNA$^{\mathrm{Phe}}$. The pocket backbone structure shares a more similar conformation in the presence of Mn^{2+} compared to Mg^{2+}, and mutations of cytosine to a dihyrouridine restore the mutated structure to a more similar state to wildtype *S. cerevisiae* tRNAs.

The overall backbone structure of mutated *D. melangaster* tRNA$^{\mathrm{Val}}$ (C16D) and tRNA$^{\mathrm{Val}}$ (C16U) show a different structural backbone compared to the wildtype *S. cerevisiae* tRNA$^{\mathrm{Phe}}$ in a magnesium bound state. Interestingly, both mutated *D. melangaster* tRNAs shift to share a similar overall backbone structure in the presence of manganese.

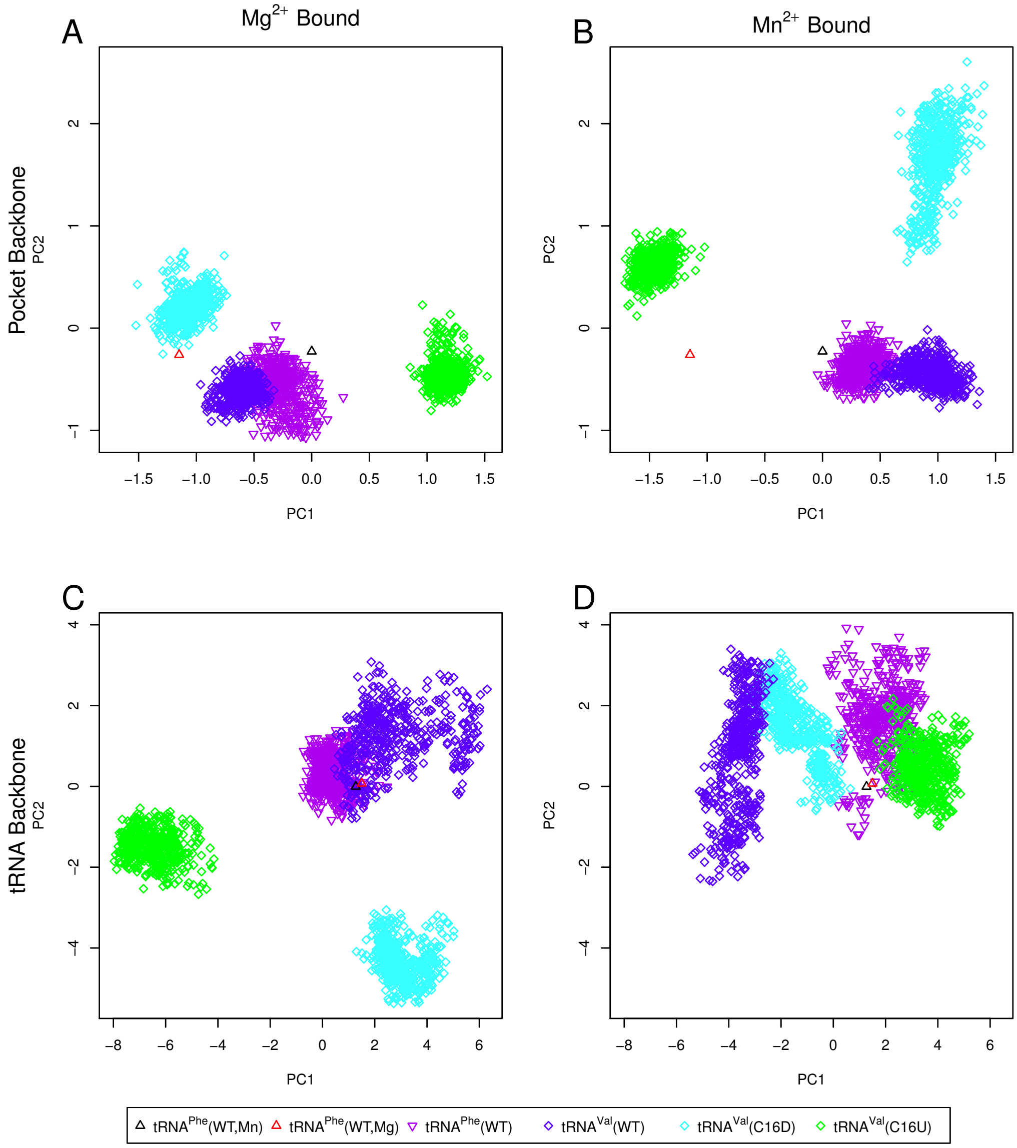

Figure 3.2: A principle component analysis of the ion binding pocket (panels A and B) and the overall structural backbone (panels C and D) of *S. cerevisiae* tRNA$^{\mathrm{Phe}}$, *D. melangaster* tRNA$^{\mathrm{Phe}}$, *D. melangaster* tRNA$^{\mathrm{Val}}$, and two mutations of *D. melangaster* tRNA$^{\mathrm{Val}}$ in the presence of Mg^{2+} (panels A and C), and Mn^{2+} (panels B and D).

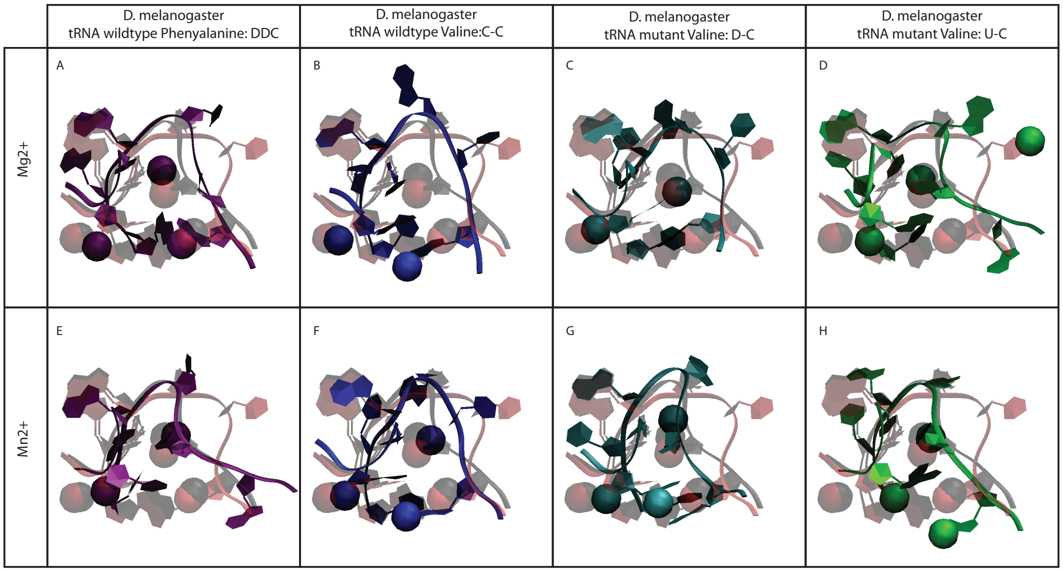
D. melanogaster tRNA wildtype Phenyalanine: DDC
D. melanogaster tRNA wildtype Valine:C-C
D. melanogaster tRNA mutant Valine: D-C
D. melanogaster tRNA mutant Valine: U-C
Mg2+
Mn2+
A
B
C
D
E
F
G
H

Figure 3.3: The dependence of the pocket backbone structure on the cation identity. The *S. cerevisiae* tRNA$^{\mathrm{Phe}}$ in the presence of Mg^{2+} (red) and in the presence of Mn^{2+} (black) are shown with *D. melangaster* wildtpe and mutated tRNAs. The pocket backbone is dependent upon both ion and sequence states.

3.4 Discussion

Our preliminary results have demonstrated the ability to recapitulate the different crystal structures found when the ion pocket is bound to either magnesium or manganese ions. The position of site 16 relative to site 59 is contingent upon the divalent ion present in the binding pocket, which contains sites 14 through 19 in the D-loop, and 59 and 60 in the T-Loop. Since the work recapitulating the structural differences noted in Shi and Moore (Shi and Moore, 2000), we have identified improved ion parameters to better reflect relative hydration free energies, ion-oxygen radial distribution functions and water coordination numbers (Li et al., 2013). These improved parameters were used in the homology and mutational studies that followed the preliminary results recapitulating the most recently published crystal structure.

Our results are suggestive that homology modeling of *Drosophila* tRNAs from *S. cerevisiae* tRNA crystal structures produces a well behaved tRNA for molecular dynamic simulations. To our knowledge, these may be the first molecular dynamic simulations to use homology modeling to simulate tRNA dynamics between two species. We feel confident that our homology modeled tRNAs are producing reasonable predictions.

Modification of the nucleotides that construct the pocket show overall gross structural changes seen in the tRNA backbone. The modification to unmodified uracil results in the most dramatic structural difference. This result is not unexpected. Although tRNAs are highly stable molecules, pathways to turnover damaged tRNAs or normal tRNAs when cells are in stressed conditions have been described (Houseley and Tollervey, 2009). One such pathway is through the TRAMP complex, where loss of m1A in tRNA$^{\mathrm{Met}}$ causes altered interactions between the D and T loops and leads to tRNA turnover; hence post-translation modifications serve a function to assure that only appropriately structured tRNAs are delivered to the protein synthesis machinery. Modifications also participate collectively in the stabilization of the tRNA molecule; some important structural functions include restriction of nonfunctional alternative folding, cooperative binding of Mg^{2+}, and thermal stabilization (El Yacoubi et al., 2012). The modification of uracil to dihydrouridine is the single most common form of post-transcriptional modification in tRNA from bacteria and eukaryotes out of 70+ modifications known to occur in these two phylogenetic domains, but unlike other modification that offers structural stability, dihydrouridine allows structural flexibility in tRNAs (Dalluge et al., 1996). Although we have no data to suggest that Sprinzl coordinate 16 is always modified, there is great support in the literature that it is a highly likely modification. Our data suggests that the flexibility conferred by the dihydrouridine is likely necessary for proper tRNA-ion interactions in the presently studied ion pocket.

3.5 Future Directions

We plan to continue this line of investigation by calculating binding free energies for various ions in the highly divergent pocket and make comparison to other tRNA ion binding pockets that bind solely to magnesium. Additionally, we plan to quantify the changes in the ion pocket shape and size due to various nucleotide substitutions.

Chapter 4

Exploratory Analysis of Selective Pressures in *Drosophila melanogaster* tRNA and Flanking Regions

4.1 Introduction

The neutral theory of molecular evolution was introduced independently by many researchers in the 1960s. Neutral evolution theory claims that most substitutions have no influence on the survival of a genotype in the population. A fundamental paper in population genetics, written by Kimura and Ohta in 1971, was the first to develop the population genetics associated with the theory of neutral evolution (Kimura and Ohta, 1971). The paper demonstrated that most mutations are neutral and these neutral mutants are a small fraction of the total mutants at the time of occurrence. These neutral mutations after fixation take the form of a polymorphism in the population. Ohta would later adapt the neutral theory to the nearly neutral theory, by suggesting that some mutations are slightly advantageous or slightly deleterious (Ohta, 1973). Ohta argued that mutation in a single site of base pair was slightly deleterious by weakening the bond between the two nucleotides, but that this slightly deleterious mutation could results in a mutation in the base paired nucleotide to restore the binding, making this mutation slightly advantageous (Ohta, 1973).

Many of the early studies in neutral evolution theory focused on protein evolution due to the high degree of protein sequence similarity between diverged species. Kreitman was the first to describe sequence variation in a sample of alleles obtained from natural populations of *D. melanogaster* in a milestone paper for evolutionary genetics (Kreitman, 1983). Kreitman described the presence of many silent polymorphisms in both introns and exons of the alcohol dehydrogenase locus. He proposed that the number of silent mutations in the coding region were a result of the strongly deleterious selective pressure, but that polymorphisms in the non-coding regions suggested their conservation was not necessary for proper gene transcription. Many of these early works concluded that regulatory evolution may be considerably more import in protein evolution compared to evolution in non-coding regions. Mutations in coding sequences have the potential to alter or disrupt protein folding and function. As a result, selective pressures acting upon protein coding genes have been a primary research focus. Still, most of the typical eukaryotic genome is comprised of non-coding DNA and the current body of knowledge concerning evolution forces acting on these regions is comparatively little to that of proteins.

Recently, numerous papers have been published indicating that noncoding regions in various *Drosophila* species have slower rates of evolution and higher levels of selective constraints in introns and intergenic sequences compared to synonymous sites in coding regions (Bergman and Kreitman, 2001; Halligan et al., 2004; Andolfatto, 2005; Haddrill et al., 2008). The selective constraints on non-coding regions is assumed to be related to the presence of *cis*-regulatory elements or conserved secondary structures of RNA (Carlini et al., 2001; Bergman et al., 2005; Casillas et al., 2007). These slow evolving non-coding regions include introns, untranslated regions (UTRs) and intergenic DNA, and the rates of these slowly evolving regions are considerably higher than previously estimated (Andolfatto, 2005). Based on these lines of evidence, non-coding regions are likely subject to the action of positive selection as well as negative selection, if they are indeed functionally important. The signatures of these different types of selection can only be distinguished by combining within-species polymorphism data with between-species measures of divergence (McDonald and Kreitman, 1991).

Research on selective constraints in non-coding RNA, specifically tRNA, is largely unexplored. To address this gap in knowledge, I aimed to elucidate the selective pressure of highly divergent sites in the ion binding pocket. I have leveraged a collection of genomes from a population of *D. melanogaster* to better describe the evolutionary forces that give rise to the high between-species divergence rates reported in chapter 2. Using the Drosophila Genetic Reference Panel (Mackay et al., 2012), a collection of genomes from a population consisting of 192 inbred lines of *D. melanogaster*, I present an exploratory analysis to begin elucidating the type of selection acting upon highly divergent sites.

4.2 Methods

4.2.1 Population Genome Data

Whole genomes for 192 inbred lines of *D. melanogaster* were downloaded from http://dgrp.gnets.ncsu.edu/data/ on May 2, 2013. The *D. melanogaster* reference genome was downloaded from FlyBase (2008_10 release) (McQuilton et al., 2012) on October 16, 2011. *D. simulans* (2012_13 release) and *D. yakuba* (2008_10 release) reference genomes were also downloaded from FlyBase on May 2, 2013. We chose to use the most recent genome annotation for *D. simulans* since this genome has been resequenced since the twelve genome paper (*Drosophila* 12 Genomes Consortium, 2007). Population genomes were not annotated for gene features, but chromosome lengths were equal to the reference sequence.The reference genome tRNAs, as well as *D. simulans* and *D. yakuba*, were annotated using the union of predictions from tRNAscan-SE 1.3.1 (Lowe and Eddy, 1997) and ARAGORN 1.2.34 (Laslett and Canback, 2004). Initiator methionine classifications were made using TFAM 1.3 (Tåquist et al., 2007). A custom perl script was then used to cut tRNA sequences from the population genomes using the coordinates from annotated tRNAs in the reference genome. Annotations for the population data were assumed to be the same as the reference genome. All tRNAs (49059) were aligned using Infernal 1.1 (Nawrocki et al., 2009) using the RFAM covariance model for the tRNA family (RF00005) built with Infernal 1.1 (Burge et al., 2012). Alignments were edited manually using SeaView 4.3.4 (Gouy et al., 2010) to produce a final alignment 74 nucleotides in length that was then manually mapped to Sprinzl coordinates (Sprinzl and Vassilenko, 2005); coordinates were verified using the transfer RNA database (Jühling et al., 2009). Sprinzl coordinate 20A was retained for all subsequent analysis. The majority of the variable arm was removed; only Sprinzl coordinate 45 – 49 were retained for further analysis. Flanking regions, 500 nucleotides in length, were cut directly before and after tRNA coordinates according to the reference genome. For individual tRNA genes, $5'$ and $3'$ regions were aligned using MUSCLE v3.8.31 (Edgar, 2004). Regions were aligned with and without the reference genomes.

The population genomes contained numerous ambiguity codes, therefore data was filtered to reduce potential noise signals in downstream analysis. All regions, tRNA, $5'$ flank and $3'$ flank, were subject to three thresholds, 1%, 5%, and 10%, for ambiguity tolerance. We examined the percentage of data lists through calculations of means and medians, Figures 4.1. For all further analysis a data filtering of 5% ambiguity tolerance was employed. After the initial filtering, any sequence with ternary ambiguities or more than a singleton binary ambiguity were removed. Remaining sequences where treated as haploid sequences. Sequences were then 'doubled' and binary ambiguities translated into nucleotides, creating two sequences for every gene with less than 5% ambiguity codes for any nucleotide (N) and no binary or ternary ambiguity codes. After filtering, all sequences were realigned using the protocols described above.

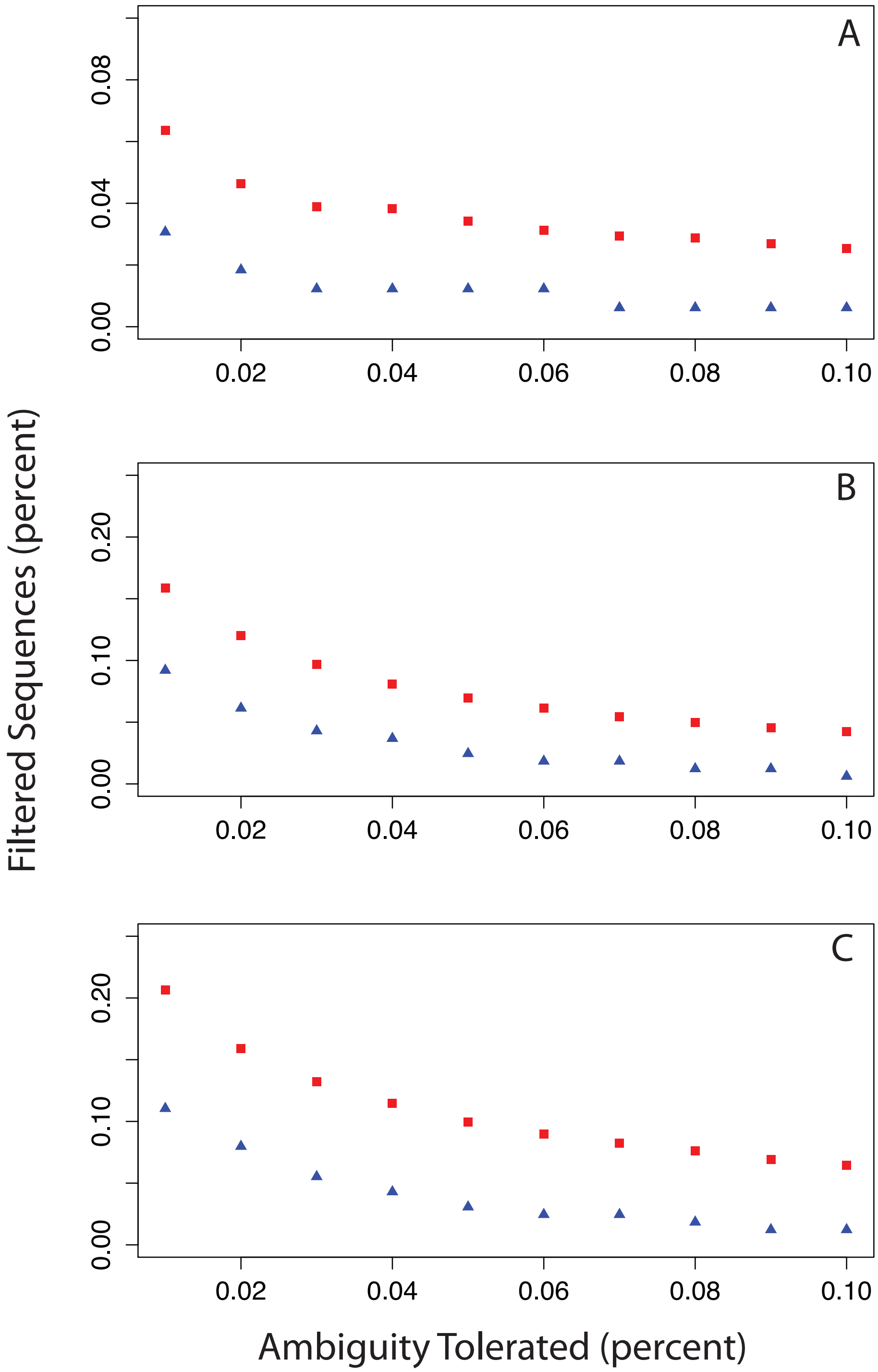

Figure 4.1: Ambiguity codes were filtered by removal of (A) tRNA, (B) 5′ flanking region and (C) 3′ flanking region sequences with more than 1%, 5% and 10% ambiguity codes along the sequence. We analyzed the amount of data reduction by each threshold using calculations of means (red) and medians (blue).

4.2.2 Calculations of Tajima's D

Population tRNA genes and flanking regions were separated by gene and calculations of Tajima's D (Tajima, 1993, 1989) and statistical significance calculations were made using DNA Sequence Polymorphism 5.10 (DnaSP) (Rozas and Rozas, 1999). Additionally, genes and flanking regions were broken intp site partitions based upon alignment position and then calculations of Tajima's D were made using DnaSP 5.10. All genes and regions were analyzed in batch mode and statistics were calculated using the total number of segregating sites. DnaSP calculates statistical significance of Tajima's D and other measures of polymorphism through coalescent simulations through a neutral infinite-sites model without recombination and assumes a large constant population size.

4.2.3 Divergence Analysis of Flanking Regions

tRNA sequences for 12 species of *Drosophila* were obtained from FlyBase (2013_05 release) (McQuilton et al., 2012) on October 25, 2013. For this analysis, we were restricted to tRNA genes in published orthologies (Rogers et al., 2010) were necessary. Orthology sets were downloaded on October 18, 2011 from http://gbe.oxfordjournals.org/content/2/467/suppl/DC1. A total of 3193 tRNA were aligned using infernal 1.1 (Nawrocki et al., 2009) using the RFAM covariance model for the tRNA family (RF00005) built with infernal 1. (Burge et al., 2012). Alignments were edited manually using SeaView (Gouy et al., 2010) to produce a final alignment 74 nucleotides in length. Flanking regions of 500 nucleotides, upstream ($5'$ end) and downstream ($3'$ end), were cut based on annotated coordinates of tRNA genes. Some flanking regions contained ambiguity code N, for divergence analysis all N nucleotides were treated as gaps. MrBayes analyzes gaps as missing data and not as character state.

Phylogenetic analysis was performed by MrBAYES 3.2.1 (Ronquist et al., 2012; Altekar et al., 2004; Huelsenbeck and Ronquist, 2001) using the general time reversible (GTR) substitution model (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990) with equal-distributed rate variation across sites and a proportion of invariable sites. Two simultaneous runs for $4 * 10^6$ generations were performed with diagnostics calculated every 500 generations. All subsets of data were curated into concatenated alignments by previously published orthologies (Rogers et al., 2010), thus all comparisons were made across orthology sets. In an effort to reduce possible noise in the data, all orthologies that were identified to involve a class switch, either isoacceptor or putative alloacceptor changes, or a mixture of functional and pseudogene predictions were removed from all divergence analysis.

Concatenated alignments from three of selected subsets in Table 2.2 – the set with the largest number of species (111111110111), the set with the second largest tree length (110011100111) and the set with the largest number of included orthologies (101111110000) – were partitioned into 3 partitions – tRNA, upstream region, and downstream region. Subsets of data were taken from earlier work (see Section 2.2) to increase the amount of data per partitions, and maximize discovery potential. tRNA genes and flanking regions were included in concatenated alignments if and only if both flanking regions were of length 500. The general time reversible substitution model (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990) was used with no rate variation across sites. The nucleotide models were allowed to be unique for each partition, thereby allowing stationary state frequencies and all other substitution model parameters to be independent across partitions. For stationary state frequencies, a flat Dirichlet prior was used, Dirichlet(1,1,1,1). For the prior on topology, we assumed the published tree for the twelve *Drosophila* species (*Drosophila* 12 Genomes Consortium, 2007). Topology and branch lengths were constrained by setting the stochastic TBR mechanism and branch multiplier to zero probability. We used Metropolis-coupled MCMC (Metropolis et al., 1953; Hastings, 1970), as implemented in MrBayes 3.0 (Huelsenbeck and Ronquist, 2001; Ronquist et al., 2012), to estimate the posterior probability distribution. All Bayesian analyses were run for $4 * 10^6$ generations saving rate multipliers every 500 generations.

4.3 Results & Discussion

4.3.1 Selective Forces Acting upon tRNA Genes and Associated Flanking Regions

tRNA genes show no variation from Tajima's D values of zero. Thus, with the current sparse data, tRNA appear to be evolving neutrally, Figure 4.2 A. In flanking regions associated with tRNA genes, an excess of low-frequency polymorphism is found relative to neutral expectation indicated by the negative estimates of Tajimas D statistic, possibly suggesting flanking regions are under purifying selection: approximately 31% (n = 272) of 5′ regions (Figure 4.2 B) and approximately 33% (n = 269) in 3′ regions (Figure 4.2 C). Flanking regions failed to correlate with any particular functional class of tRNA (Figure 4.2 B & C) or chromosomal location (Figure 4.3 B & C).

These results agree with previously published results in *D. melanogaster*, which found negative estimates of Tajima's D across non-coding regions when examining an entire chromosomal arms (Mackay et al., 2012). By using polymorphism data and divergence data between *D. melanogster* and *D. yakuba*, Mackay et al. suggest that UTRs, introns and intragenic regions are primarily under a weakly deleterious selection regime (Mackay et al., 2012). Research comparing *D. melanogaster* to *D. simulans* demonstrated for all classes of noncoding sequence, when compared to synonymous sites, reduced levels of polymorphism and divergence, high selective constraints, and a skew of mutations toward rare variants (Andolfatto, 2005). Taken together, published data suggest non-coding regions, at least in *D. melanogaster*, are under negative selective pressure.

Characterization of the minor allele frequencies for each tRNA Sprinzl coordinate indicate that even though a small handful of sites are highly divergent, sites largely exhibit low levels of polymorphism, Figure 4.4. This result again suggests the likelihood that tRNA sites are evolving neutrally. The Tajima's D calculations reported here are similar to reported values for Tajima's D in synonymous sites in

D. melanogaster (Andolfatto, 2005). Many studies investigating polymorphism and divergence do not restrict sequence lengths to such a small window as required in the analysis of tRNA genes, for this analysis only 74 nucleotides, thus the conclusions drawn here are from limited data.

4.3.2 Divergence rates in Flanking Regions

The cis-acting elements of most Pol III transcription units are located within the transcribed region in many eukaryotic tRNAs. The intragenic transcription promoter is typically comprised of two binding regions: the A box, 12 – 20 bp downstream of the transcription start site (TSS), and the B box, located 30 – 60 bp downstream of the A box. The A box has a consensus sequence TRGYnnAnnnG and starts at position +8 of the mature tRNA (Dieci et al., 2013). The TSS is most frequently located between between 10 – 12 nucleotides upstream of the start of mature tRNA coding sequence between 18 – 20 nucleotides upstream of the thymine that marks the beginning of the A box. In yeast eukaryotic genomes, the TSS is surrounded by a core promoter element identified by the consensus sequence tCAAca, but this has not been documented in *Drosophila* (Dieci et al., 2013). The TFIIIC-B box interaction is the main determinant of both selectivity and stability of TFIIIC-DNA complexes, while the A box is involved in TFIIIB recruitment and transcription initiation (Orioli et al., 2012). The B box is a highly conserved sequence, GGTTCGANTCC starting at position 52 on the T-stem (Sharp et al., 1985).

Within the set of genes for a single tRNA isoacceptor, there is little 5′ or 3′ flanking sequence homology, except that 3’-flanking sequences are AT-rich. Oligothymidylate stretches in the 3′-flanking sequences likely serve as termination signals for RNA polymerase III (Sharp et al., 1985). Recent evidence suggests that even in absence of sequence homology, the 5′-flanking regions of tRNA genes of *D. melanogaster* are dominated by a TTTGGC motif located between -15 and -20 with respect to the TSS, whose function remains largely unexplored (Orioli et al., 2012).

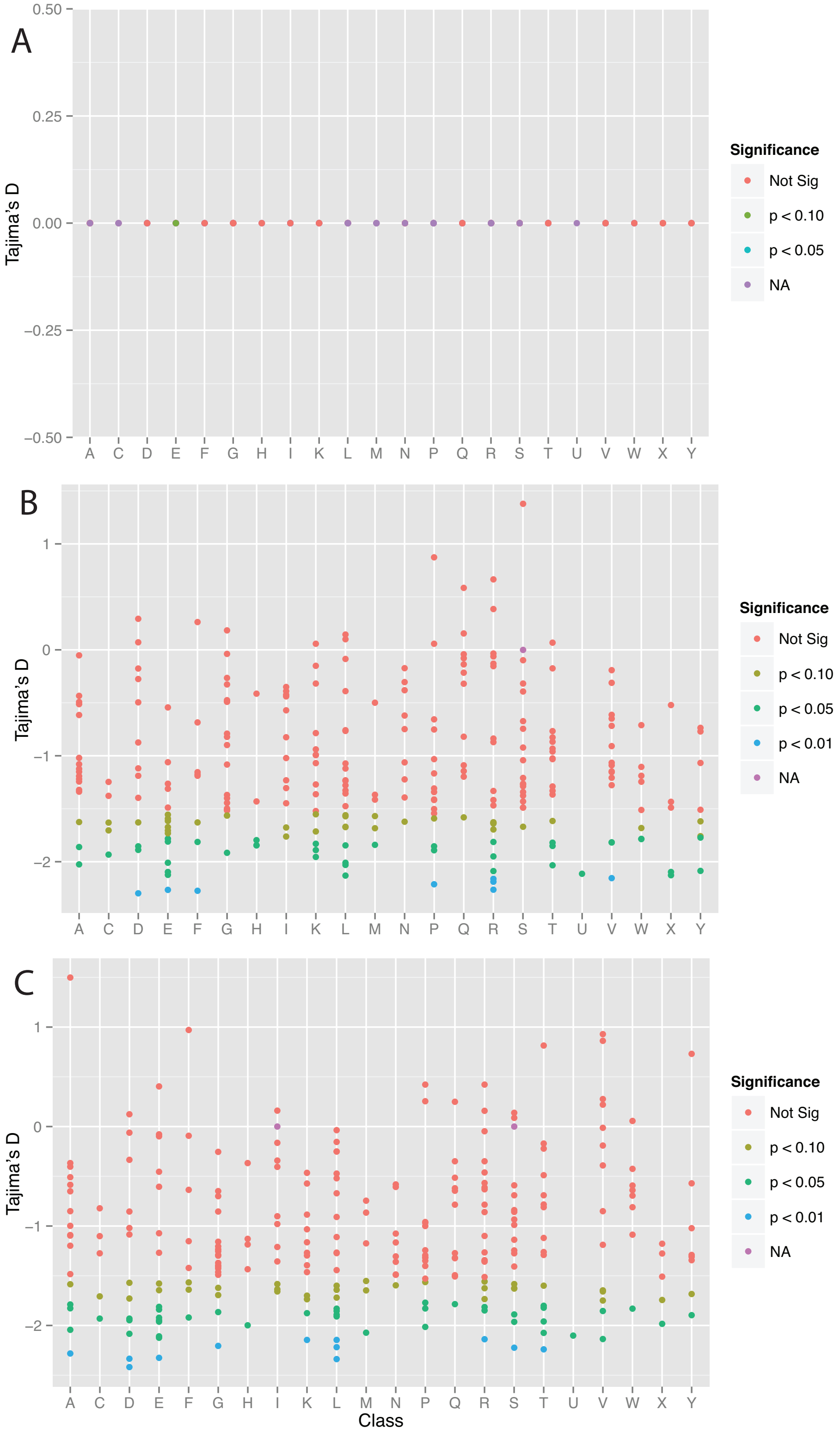

Figure 4.2: Tajima's D was calculated for by (A) tRNA genes along with flanking regions, both (B) 5′ and (C) 3′. tRNA genes show no variation from zero, while flanking regions have significant negative values for Tajima's D. Significant negative values of Tajima's D show no association to functional class in flanking regions.

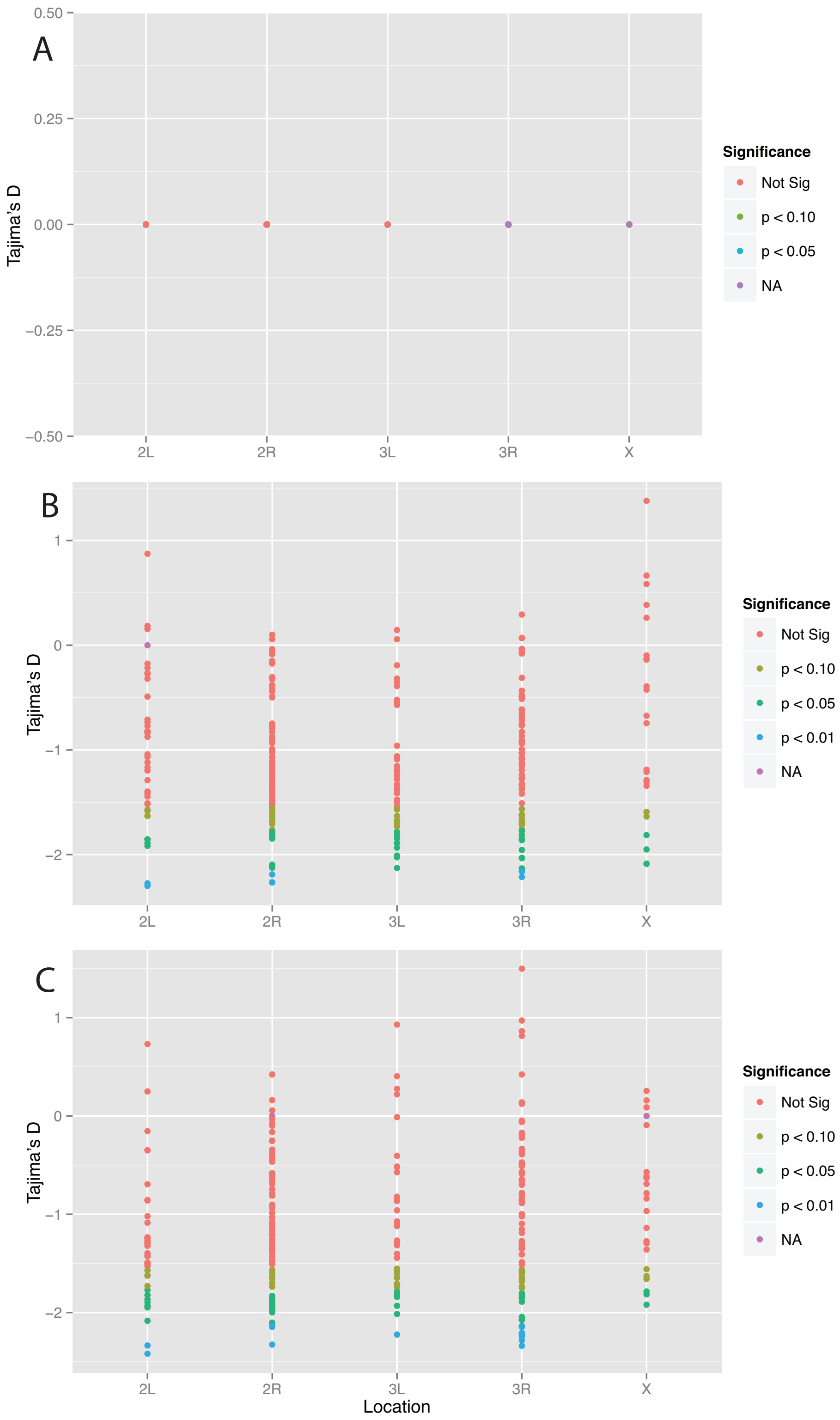

Figure 4.3: Tajima’s D was calculated for by (A) tRNA genes along with flanking regions, both (B) 5′ and (C) 3′. tRNA genes show no variation from zero, while flanking regions have significant negative values for Tajima’s D. Significant negative values of Tajima’s D show no association to chromosomal location in flanking regions.

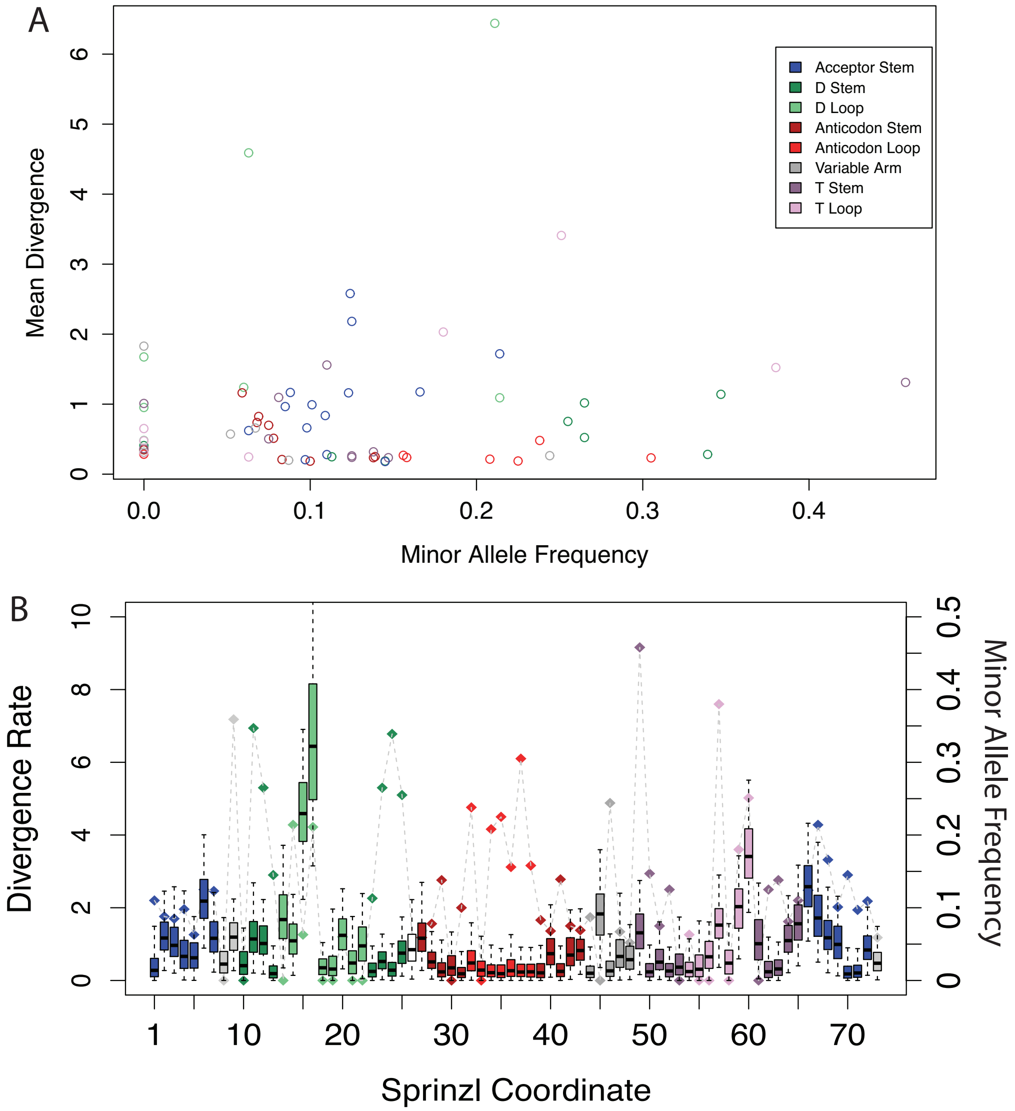

Figure 4.4: (A) Minor allele frequencies were calculated by Sprinzl coordinate and plotted against site divergence rates. (B) MAF was also plotted by site with divergence rates. Highly divergent sites have low levels of polymorphism.

Divergence analysis suggests that the 5′ flanking region has lower levels of divergence compared to both the tRNA gene and 3′ flanking region, Figure 4.5. The 3′ flanking region has the highest levels of divergence, Figure 4.5. This trend was conserved across all tested subsets of the data. The low divergence rate in the 5′ flanking region is likely a result of the poorly understood transcription factors in the 5′ flanking region. Studies in *Drosophila* and *Bombyx* have confirmed reduced transcriptional efficiency when the 5′ flanking region is mutated or truncated (Sprague et al., 1980; Schaack et al., 1984; Bertling et al., 1987; Horvath and Spiegelman, 1988). Sequences affecting transcription factor binding extend more than 60 base pairs into the 5′ flank, and approximately 35 base pairs into the 3′ flank (Schaack et al., 1984). Our results echo that the 5′ flanking region is likely to contain transcription regulatory sequences that are under heavy selective pressure, resulting in a reduced divergence rate. The higher divergence rate in the 3′ flanking region may be a result of fewer regulatory sites, thereby reducing the strong signal seen in the 5′ flank.

With the limited data currently available, we are unable to detect any signature for selective force upon the tRNA genes. Tajima's D is indistinguishable from zero and the overall divergence for tRNA genes is low suggesting that tRNA genes are evolving neutrally. In addition, the polymorphism data also suggests tRNA sites are evolving neutrally. In contrast, the low divergence rate in the flanking region and the negative Tajima's D suggests negative selection is acting upon at least some subset of the flanking regions associated with tRNA genes.

4.4 Future Directions

One of the motivations for this exploratory analysis was to use flanking regions to detect selective forces resulting in the divergence pattern described in Chapter 2. I had anticipated using flanking regions as linked neutral sites for a McDonald-Kreitman test to determine the selective forces acting upon tRNA as opposed to the steady accumulation of mutations by neutral evolution (McDonald and Kreitman, 1991). The

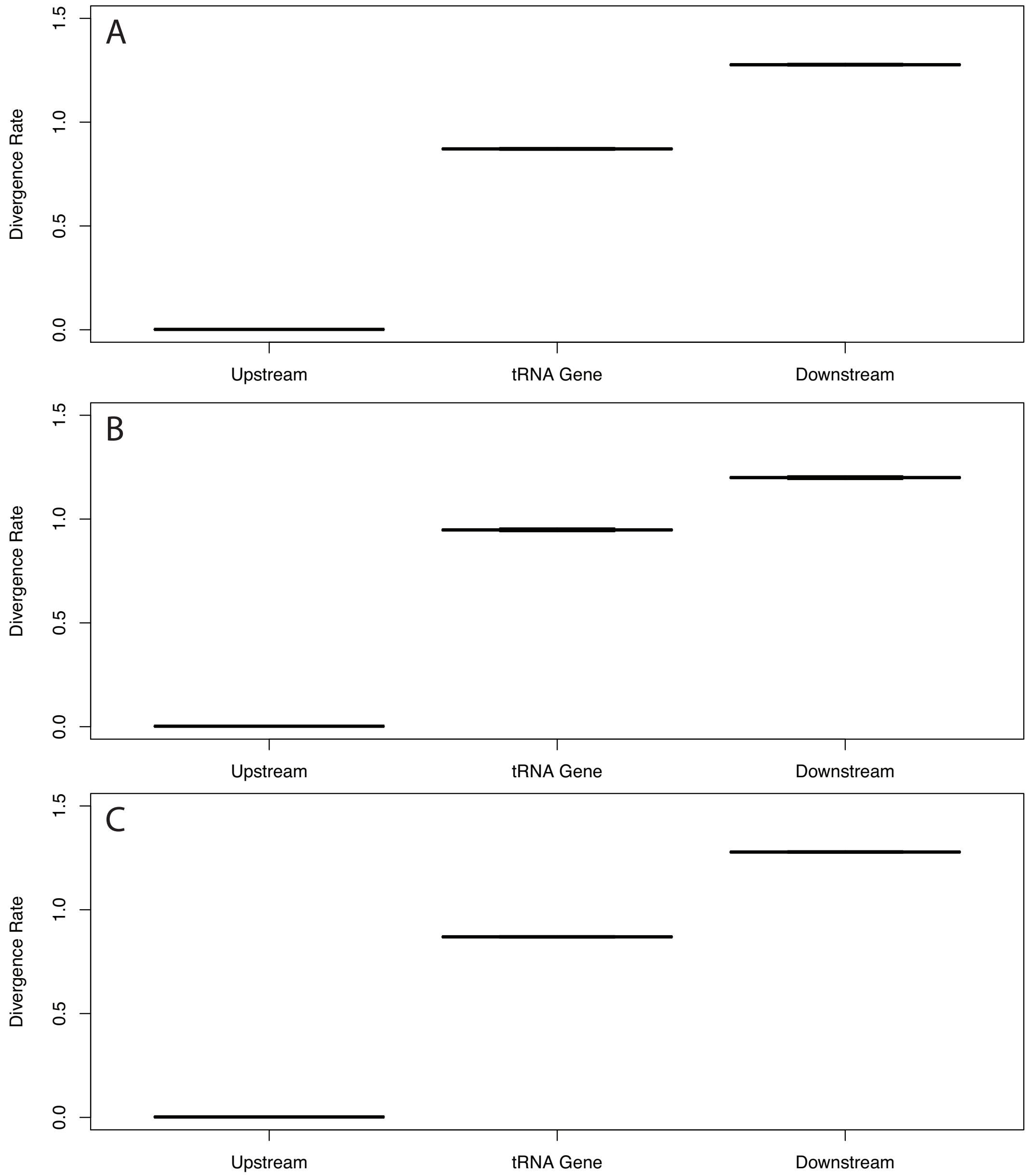

Figure 4.5: 3′ flanking regions have lowest rates of divergence suggesting possible conservation of upstream regulatory regions. (A) the set with the largest number of species (111111110111, missing *D. willistoni*); 84 orthologous sets; (B) the set with the second largest tree length (110011100111, second only to the set with largest number of species, set contains *D. melanogaster, D. simulans, D. erecta, D.ananassae, D pseudoobscura, D. mojavensis, D. virilis, D. grimshawi*); 86 orthologous sets; and (C) the set with the largest number of included orthologies (101111110000, *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D.ananassae, D. pseudoobscura, D. persimilis*); 155 orthologous sets.

McDonald-Kreitman test compares the amount of intraspecies variation to interspecies variation at two types of sites, neutral and non-neutral. Given the limited data, this analysis proved challenging, but with continued effort through distinguishing promotors and TSSs, a McDonald-Kreitman test may be able to distinguish selective forces acting upon highly divergent tRNA sites.

Chapter 5

Conclusions

tRNAs afford a unique opportunity to carry out molecular evolutionary analysis at single-site resolution because of their relatively high structural conservation and copy numbers in eukaryotic genomes. Still one of the limitations is the curation of well defined, conservative orthologies. Additionally, tRNA studies often involve data size limitations. In this work, I have presented novel data partitioning techniques which enable the analysis of limited data sets, while maintaining species diversity. Using sequence data from the twelve species of *Drosophila* as well as yeast, I have estimated divergence rates for tRNA sites, structures, and alloacceptor classes. Through this, I have found evidence of an unusual, previously undescribed, elevated evolutionary rate associated to sites in one of the several ion-binding pockets in *Drosophila* tRNAs near the ”core.”

tRNAs are confined structurally to the bounds of the ribosome structure, and each tRNA must be recognized by elongation factors and modification enzymes. Despite the necessity for similarity, each tRNA must be distinguished correctly by its cognate aaRS through recognition and discrimination in a network of tRNA-protein interactions. In this work, I have describe the first detailed bioinformatic predictions of tRNA features governing tRNA-protein interactions in any eukaryotic genomes, particularly in the context of a group of related genomes. Surprisingly, evidence suggests that such features have evolved and changed even within drosophilids. Some of the elevated divergence rates described in this work may also be associated to turnover of features that potentially govern tRNA-protein interactions in flies.

RNA structural stability is linked intimately with ion binding and post-translational modifications. From our results, we hypothesize that a post-translational modification of Sprinzl coordinate 16 in a primary ion binding pocket is likely necessary for proper flexibility and function of tRNA. This ion binding site has been shown to bind many metal ions including: magnesium, cobalt, manganese, and lead. We have used published tRNA crystal structures in yeast to successfully simulate molecular dynamics of *Drosophila* tRNAs, specifically the effect of various cations binding tRNA, magnesium and manganese. Additionally, simulations of mutant tRNAs predict unusual plasticity in the previously described ion binding pocket showing elevated divergence.

In an effort to elucidate the selective pressure of highly divergent sites in the ion binding pockets, a collection of genomes from a population of *D. melanogaster* was used to identify evolutionary forces that give rise to the high between-species divergence rates reported in chapter 2. For each tRNA Sprinzl coordinate, even though a small handful of sites are highly divergent, sites largely exhibit low levels of polymorphisms suggesting that tRNA sites are under neutral selection. Future work should leverage the McDonald-Kreitman test to compare the amount of intraspecies variation to interspecies variation between tRNA genes and their flanking regions.

Appendix A

Supplementary Figures

This appendix contains additional figures and tables for the tRNA evolution in *Drosophila* found in Chapter 2. Please see Section 2.2.1 for details on data presented in the following figures and tables, and Section 2.2.4 for details on the construction of functional logos and CIF evolution.

Table A.1: The total number of tRNAs per functional class as predicted by tRNAscan-SE, ARAGORN and TFAM for twelve species of *Drosophila* from the 2008_07 FlyBase release. For comparison, data from the Genomic tRNA database for *D. melanogaster* is included.

	D. simulans	*D. sechellia*	*D. melanogaster*	*D. melanogaster*[a]	*D. yakuba*	*D. erecta*	*D. ananassae*	*D. pseudoobscura*	*D. persimilis*	*D. willistoni*	*D. mojavensis*	*D. virilis*	*D. grimshawi*
A	14	19	17	17	25	17	20	18	18	18	16	15	14
F	8	9	8	8	9	9	8	7	7	6	8	8	9
G	23	19	20	20	23	21	25	23	22	18	22	21	22
I	8	12	11	11	12	11	13	12	13	14	10	11	10
L	23	23	22	23	24	21	23	22	22	23	18	20	19
M	7	9	6	12	9	6	7	5	5	8	6	7	5
P	16	17	17	17	19	16	16	14	16	16	14	12	13
V	15	18	15	15	19	16	16	15	14	17	14	15	14
W	9	6	8	8	8	8	6	8	8	6	6	6	6
X	6	6	6		7	6	5	6	7	6	7	5	5
C	7	6	7	7	8	7	7	11	8	11	8	8	6
N	7	9	10	12	11	8	9	9	9	9	8	7	7
Q	12	11	12	12	14	13	14	12	16	12	11	19	16
S	16	20	20	20	22	21	24	21	21	18	16	18	15
T	16	19	16	18	19	17	19	20	20	17	20	18	19
Y	10	9	10	9	10	9	8	9	9	10	7	9	10
H	8	13	5	5	5	6	5	5	7	7	5	5	5
K	15	16	19	19	22	16	24	17	18	17	19	24	18
R	20	23	26	26	26	23	25	21	24	24	19	19	18
D	9	14	14	14	15	14	13	14	18	15	12	13	12
E	15	18	20	25	20	18	19	23	20	22	17	16	16
U	1	1	1	1	1	1	1	1	1	1	1	1	1
SUBTOTAL	265	297	290	299	328	284	307	293	303	295	264	277	260
Undetermined	1	2	2	1	0	0	0	1	2	1	0	0	0
Pseudo	0	0	0	4	0	0	0	0	0	0	0	0	0
TOTAL	266	299	292[b]	304	328	284	307	294	305	296	264	277	260

[a] These class predictions were taken from the Genomic tRNA database (http://gtrnadb.ucsc.edu/) for comparison to the predications made by tRNAscan-SE, ARAGORN and TFAM.

[b] The annotations for D. melanogaster contained 22 annotations for tRNA in the mitrochondrion genome. The addition of 22 to the total is 314; the number of annotations found in the tRNA files.

Table A.2: The total number of tRNAs per functional class as predicted by tRNAscan-SE, ARAGORN and TFAM for twenty species of yeast. Asterisk ($*$) denotes a pre-duplication event species.

	*E. gossypii**	*C. glabrata*	*E. cymbalarie**	*K. africana*	*K. lactis**	*K. naganishii*	*V. polyspora*	*L. thermotolerans**	*L. waltii**	*N. castellii*	*N. dairensis*	*S. cerevisia*	*L. kluyveri**	*S. kudraivzevii*	*S. mikatae*	*S. uvarum*	*T. blattae*	*T. delbrueckii**	*T. phaffii*	*Z. rouxii**
A	13	15	9	17	10	9	14	15	16	17	15	16	17	17	17	17	24	12	12	21
C	3	3	3	5	3	4	4	4	4	4	5	4	5	4	4	3	5	3	4	4
D	10	9	6	13	8	8	12	10	10	14	13	16	13	13	16	16	19	9	12	14
E	11	12	8	17	10	9	14	17	17	14	14	16	16	14	17	16	17	12	12	14
F	6	6	6	9	5	6	8	7	8	9	5	10	8	10	11	10	12	6	9	8
G	15	15	10	22	10	12	17	20	20	18	16	21	19	20	21	21	27	14	16	24
H	5	6	3	6	4	4	5	5	5	8	7	7	6	7	8	6	7	4	5	6
I	10	11	6	15	8	9	12	11	12	14	14	15	13	14	13	15	19	9	12	13
K	12	15	9	20	13	11	17	18	17	20	17	21	19	21	22	21	24	15	15	21
L	14	16	12	20	13	12	17	18	10	21	15	21	19	20	20	21	24	16	14	20
M	4	3	3	5	3	2	4	4	7	5	5	5	5	6	6	5	5	4	4	5
N	6	9	5	10	6	6	8	7	8	10	9	10	9	12	11	10	11	7	8	9
P	8	8	6	11	8	7	9	10	9	8	7	12	12	12	13	12	14	7	8	13
Q	8	8	7	10	7	6	10	10	10	10	11	10	11	10	10	10	11	8	8	13
R	13	15	10	17	12	11	16	15	15	18	18	19	19	18	21	19	18	13	13	16
S	15	15	11	20	11	12	18	16	18	20	17	18	18	16	19	23	28	13	14	22
T	11	13	6	15	9	10	16	12	15	16	18	16	14	15	17	15	18	12	12	14
V	14	13	8	16	10	10	15	15	18	17	18	18	18	19	18	18	21	15	13	18
W	5	5	4	6	4	4	5	5	5	6	3	6	5	6	6	7	7	4	4	6
X	4	4	3	4	3	2	3	4	3	4	4	5	5	4	6	5	6	3	4	4
Y	4	6	6	9	5	5	8	6	6	7	3	8	7	7	8	8	9	7	6	7
Pseudo	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Total	191	207	142	267	162	159	232	230	233	260	234	275	258	265	284	278	326	193	205	272

Table A.3: Sequence patterns, orthology sets, functional classes and locations (in *D. melanogaster*) contributing to elevated site rates. Results are provided for the largest subset of data partitions (X indicates the species is excluded from that partition). Species string: *D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta, D.ananassae, D pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis, D. grimshawi*.

Species String	**Nucleotide**	**Orthology Set (Class Anitcodon) Sequence**	**Location in *D. melanogaster***
111111110111	16-17	72 (U TCA):TGGGGGGG0GGG:-TTTTTTTTTT	2R:complement(7245069..7245155)
111111110111	16-60	2 (R TCT):TTTTTCCC0TCC:TTTTTCCCTCC	2L:1965426..1965498
111111110111	16	92 (L AAG):CCCCCCCC0CCT	2R:complement(10871832..10871913)
		112(I AAT):TTTTTTTT0CCC	2R:15603070..15603143
		114(X CAT):TTTTTTCT0TTT	2R:15613410..15613481
		136(S TGA):TTTTTCTT0TTT	2R:complement(18959542..18959623)
		137(S TGA):TTTTTCTT0TTT	2R:18960104..18960185
		188(P CGG):AAAAAAAA0GGG	3L:18611490..18611561
		192(R TCG):AAAAAAAA0TTT	3R:1213950..1214022
		226(V AAC):TTTTTTCT0TTT	3R:12147412..12147484
		251(V CAC):CCCCCCTC0TCC	3R:15615844..15615916
		285(S AGA):CCCCCCCC0ACC	X:complement(13919142..13919223)
111111110111	17	46 (N GTT):TTTTTCTT0CCT	2R:complement(2040108..2040181)
		77 (M CAT):TTTTTTTT0ATT	2R:complement(7548068..7548140)

Table A.3 – continued from previous page

Species String	Nucleotide	Orthology Set (Class Anitcodon) Sequence	Location in *D. melanogaster*
111111110111	17	102(F GAA):T-TTTTTT0TTT	2R:complement(13492040..13492112)
		138(N GTT):TTTTTCTT0CCT	2R:20261586..20261659
		186(M CAT):TTTTTTTT0AAA	3L:16343017..16343089
		205(N GTT):TTTTTCTT0CCT	3R:complement(3965255..3965328)
111111110111	60	231(A AGC):GGGGGAGG0GGG	3R:13445460..13445532
		240(A AGC):GGGGGAGG0GGG	3R:complement(13471029..13471101)
		245(A AGC):GGGAGAGG0AGG	3R:complement(13484103..13484175)
		246(A AGC):GGGGGAGG0AGG	3R:13493909..13493981
		252(A AGC):GGGGGAGG0GGG	3R:15616694..15616766
11111111011X	16	260(S GCT):TTTTTATT0AAX	3R:complement(18222902..18222983)
11111111011X	60	221(T TGT):CCCCCCTT0TCX	3R:complement(8148285..8148356)
		242(A AGC):GGGGGAGG0GGX	3R:13472090..13472162
1X111111011X	16	258(S GCT):AXAAATAA0TTX	3R:18222250..18222331
		259(S GCT):AXAAAAAA0TTX	3R:complement(18222521..18222602)
111111110XXX	16	83 (I AAT):TCCCTTCC0XXX	2R:complement(9317787..9317860)
		85 (I AAT):TTTCTTCC0XXX	2R:9318489..9318562
		272(Q TTG):CCCTTTTT0XXX	X:complement(3321485..3321556)

Table A.3 – continued from previous page

Species String	Nucleotide	Orthology Set (Class Anitcodon) Sequence	Location in *D. melanogaster*
111111110XXX	17	193(M CAT):TTTTTAAA0XXX	3R:complement(2321761..2321833)
111111110XXX	60	234(A AGC):GGGGGAGG0XXX 238(A AGC):GGGGGAGG0XXX 244(A AGC):AGGGGAGG0XXX	3R:13448268..13448340 3R:complement(13456764..13456836) 3R:13482867..13482939
11XX111X0111	16	287(S AGA):TTXXTTCX0CCC	X:13964674..13964755
11X1111X01X1	16	292(P CGG):AAXAAGTX0TXT	X:18459797..18459868
1111X1110XXX	17	61 (N GTT):TTTTXCTT0XXX	2R:complement(2077634..2077707)
1XXX111X0111	60	166(A AGC):GXXXGAGX0GGG	3L:complement(8021646..8021718)
111111X10XXX	16	273(Q CTG):TTTTTTXC0XXX 294(I AAT):TTTTTTXC0XXX	X:3713732..3713803 X:complement(21102352..21102426)
1XX1111X01X1	16	220(T TGT):TXXTTTTX0TXA	3R:8032297..8032368
1X1111110XXX	16	274(P CGG):GXGAAGTT0XXX 289(R TCG):TXTTTTCC0XXX	X:3721655..3721726 X:complement(13997752..13997824)
1X1111110XXX	17	47 (N GTT):TXTTTCTT0XXX	2R:2040691..2040764

Table A.4: Sequence changes in ion binding pockets sites by orthology for chromosome arms (Muller element A – E), when location data was available for *D. melanogaster*. The X chromosome (Muller element A) has a larger fraction of change compared to other chromosome arms.

Chromosome Location	**Orthologies with Changes in Sites 15, 18, 19, 20, or 59 (% of total)**	**Orthologies with Changes in Sites 16, 17 or 60 (% of total)**
2L	5 (0.119)	4 (0.095)
2R	10 (0.096)	23 (0.221)
3L	6 (0.113)	7 (0.132)
3R	6 (0.073)	21 (0.256)
X	9 (0.273)	11 (0.333)
Not Available	14 (0.030)	18 (0.039)

Table A.5: Sequence changes in ion binding pockets sites by orthology for each alloacceptor.

Alloacceptor	**Orthologies with Changes in Sites 15, 18, 19, 20, or 59 (% of total)**	**Orthologies with Changes in Sites 16, 17 or 60 (% of total)**
A	2 (0.044)	6 (0.133)
C	1 (0.042)	1 (0.042)
D	2 (0.047)	3 (0.070)
E	2 (0.035)	3 (0.053)
F	1 (0.067)	2 (0.133)
G	4 (0.062)	4 (0.062)
H	0 (0.000)	1 (0.083)
I	1 (0.031)	3 (0.094)
K	3 (0.058)	5 (0.096)
L	1 (0.020)	7 (0.140)
M	1 (0.077)	1 (0.077)
N	1 (0.067)	4 (0.267)
P	6 (0.176)	6 (0.176)
Q	1 (0.033)	3 (0.100)
R	5 (0.077)	4 (0.062)
S	5 (0.109)	8 (0.174)
T	3 (0.083)	5 (0.139)
U	0 (0.000)	0 (0.000)
V	3 (0.079)	4 (0.105)
W	0 (0.000)	0 (0.000)
X	0 (0.000)	3 (0.176)
Y	2 (0.091)	1 (0.045)
Pseduo	2 (0.111)	3 (0.167)

References

Aduri, R., B. T. Psciuk, P. Saro, H. Taniga, H. B. Schlegel, and J. SantaLucia (2007). Amber force field parameters for the naturally occurring modified nucleosides in RNA. *Journal of Chemical Theory and Computation 3*(4), 1464–1475.

Alexander, R. W., J. Eargle, and Z. Luthey-Schulten (2010). Experimental and computational determination of tRNA dynamics. *FEBS Letters 584*(2), 376–386.

Altekar, G., S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist (2004). Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference. *Bioinformatics 20*(3), 407–415.

Amrine, K. C. H., W. D. Swingley, and D. H. Ardell (in press). tRNA signatures reveal polyphyletic origins of streamlined SAR11 genomes among the alphaproteobacteria.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in *Drosophila*. *Nature 437*(7062), 1149–1152.

Ardell, D. H. (2010). Computational analysis of tRNA identity. *FEBS Letters 584*(2), 325–333.

Banerjee, R., S. Chen, K. Dare, M. Gilreath, M. Praetorius-Ibba, M. Raina, N. M. Reynolds, T. Rogers, H. Roy, S. S. Yadavalli, and M. Ibba (2010). tRNAs: cellular barcodes for amino acids. *FEBS Letters 584*(2), 387–395.

Batey, R. T., R. P. Rambo, and J. A. Doudna (1999). Tertiary motifs in RNA structure and folding. *Angewandte Chemie International Edition 38*(16), 2326–2343.

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y.-P. Poh, M. W. Hahn, P. M. Nista, C. D. Jones, A. D. Kern, C. N. Dewey, L. Pachter, E. Myers, and C. H. Langley (2007). Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology 5*(11), e310.

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

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*(4), 1165–1188.

Bergman, C. M., J. W. Carlson, and S. E. Celniker (2005). *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics 21*(8), 1747–1749.

Bergman, C. M. and M. Kreitman (2001). Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Research 11*(8), 1335–1345.

Bertling, W., T. Dingermann, and M. Kaiserwerth (1987). Comparative study of 5' flanking sequences of eukaryotic genes: possible functional implications. *International Journal of Biological Macromolecules 9*(2), 63–70.

Bradbrook, G. M., T. Gleichmann, S. J. Harrop, J. Habash, J. Raftery, J. Kalb (Gilboa), J. Yariv, I. H. Hillier, and J. R. Helliwell (1998). X-ray and molecular dynamics studies of concanavalin-A glucoside and mannoside complexes relating structure to thermodynamics of binding. *Journal of the Chemical Society, Faraday Transactions 94*(11), 1603–1611.

Brown, R. S., J. C. Dewan, and A. Klug (1985). Crystallographic and biochemical investigation of the lead(II)-catalyzed hydrolysis of yeast phenylalanine tRNA. *Biochemistry 24*(18), 4785–4801.

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

Byrne, K. P. and K. H. Wolfe (2005). The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research 15*(10), 1456–1461.

Carlini, D. B., Y. Chen, and W. Stephan (2001). The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics 159*(2), 623–633.

Case, D. A., T. A. Darden, T. E. I. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Goetz, I. Kolossv a ry, K. F. Wong, F. Paesani, J. Vanicek, R. M. Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, M. J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. A. Kollman (2012). AMBER 12. Technical report, University of California, San Francisco.

Casillas, S., A. Barbadilla, and C. M. Bergman (2007). Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Molecular Biology and Evolution 24*(10), 2222–2234.

Cedergren, R. J., B. LaRue, D. Sankoff, G. Lapalme, and H. Grosjean (1980). Convergence and minimal mutation criteria for evaluating early events in tRNA evolution. *Proceedings of the National Academy of Sciences of the United States of America 77*(5), 2791–2795.

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

Ciesiołka, J., J. Wrzesinski, P. Górnicki, J. Podkowiński, and W. J. Krzyzosiak (1989). Analysis of magnesium, europium and lead binding sites in methionine initiator and elongator tRNAs by specific metal-ion-induced cleavages. *European Journal of Biochemistry 186*(1-2), 71–77.

Conant, G. C. and K. H. Wolfe (2008). Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics 179*(3), 1681–1692.

Crick, F. H. (1958). On protein synthesis. *Symp. Soc. Exp. Biol. 12*, 138–163.

Crick, F. H. (1970). Central dogma of molecular biology. *Nature 227*(5258), 561–563.

Dalluge, J. J., T. Hashizume, A. E. Sopchik, J. A. McCloskey, and D. R. Davis (1996). Conformational flexibility in RNA: the role of dihydrouridine. *Nucleic Acids Research 24*(6), 1073–1079.

Danchin, A. (1972). tRNA structure and binding sites for cations. *Biopolymers 11*(7), 1317–1333.

Deng, H. Y. and J. Termini (1992). Catalytic RNA reactions of yeast tRNA(Phe) fragments. *Biochemistry 31*(43), 10518–10528.

Di Giulio, M. (1992). On the origin of the transfer RNA molecule. *Journal of Theoretical Biology 159*(2), 199–214.

Di Giulio, M. (1995). Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule? *Journal of Theoretical Biology 177*(1), 95–101.

Dieci, G., A. Conti, A. Pagano, and D. Carnevali (2013). Identification of RNA polymerase III-transcribed genes in eukaryotic genomes. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms 1829*(3-4), 296–305.

Eargle, J., A. A. Black, A. Sethi, L. G. Trabuco, and Z. Luthey-Schulten (2008). Dynamics of recognition between tRNA and elongation factor Tu. *Journal of Molecular Biology 377*(5), 1382–1405.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*(5), 1792–1797.

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

El Yacoubi, B., M. Bailly, and V. de Crécy-Lagard (2012). Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annual Review of Genetics 46*(1), 69–95.

Feig, A. L. and O. C. Uhlenbeck (2005). The role of metal ions in RNA biochemistry. In *The RNA World*, pp. 1–34. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

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

Giegé, R. (2008). Toward a more complete view of tRNA biology. *Nature Structural & Molecular Biology 15*(10), 1007–1014.

Giegé, R., M. Sissler, and C. Florentz (1998). Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Research 26*(22), 5017–5035.

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

Gower, J. C. and P. Legendre (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification 3*(1), 5–48.

Haddrill, P. R., D. Bachtrog, and P. Andolfatto (2008). Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Molecular Biology and Evolution 25*(9), 1825–1834.

Hahn, M. W., M. V. Han, and S. G. Han (2007). Gene family evolution across 12 *Drosophila* genomes. *PLoS Genetics 3*(11), e197.

Halligan, D. L., A. Eyre-Walker, P. Andolfatto, and P. D. Keightley (2004). Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Research 14*(2), 273–279.

Harvey, S. C., M. Prabhakaran, and J. A. McCammon (1985). Molecular-dynamics simulation of phenylalanine transfer RNA. I. Methods and general results. *Biopolymers 24*(7), 1169–1188.

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

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Hermann, T. and E. Westhof (1998). Exploration of metal ion binding sites in RNA folds by brownian-dynamics simulations. *Structure 6*(10), 1303–1314.

Hess, B., C. Kutzner, D. van der Spoel, and E. Lindahl (2008). GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation 4*(3), 435–447.

Hoagland, M. B., M. L. Stephenson, J. F. Scott, L. I. Hecht, and P. C. Zamecnik (1958). A soluble ribonucleic acid intermediate in protein synthesis. *The Journal of Biological Chemistry 231*(1), 241–257.

Holbrook, S. R., J. L. Sussman, R. W. Warrant, G. M. Church, and S. H. Kim. RNA-ligand interactions:(I) magnesium binding sites in yeast tRNAPhe.

Holley, R. W. (1957). An alanine-dependent, ribonuclease-inhibited conversion of AMP to ATP, and its possible relationship to protein synthesis. *Journal of the American Chemical Society*.

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

Holley, R. W., J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir (1965). Structure of a ribonucleic acid. *Science 147*(3664), 1462–1465.

Horvath, D. and G. B. Spiegelman (1988). Sequences between the internal control regions of tRNAArg of *Drosophila melanogaster* influence stimulation of transcription of the 5' flanking dna. *Nucleic Acids Research 16*(6), 2585–2599.

Houseley, J. and D. Tollervey (2009). The many pathways of RNA degradation. *Cell 136*(4), 763–776.

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

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

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

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

Kim, S. H., F. L. Suddath, G. J. Quigley, A. McPherson, J. L. Sussman, A. H. Wang, N. C. Seeman, and A. Rich (1974). Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science 185*(4149), 435–440.

Kimura, M. and T. Ohta (1971). Protein polymorphism as a phase of molecular evolution. *Nature 229*(5285), 467–469.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature 304*(5925), 412–417.

Krzyzosiak, W. J., T. Marciniec, M. Wiewiorowski, P. Romby, J. P. Ebel, and R. Giegé (1988). Characterization of the lead(II)-induced cleavages in tRNAs in solution and effect of the Y-base removal in yeast tRNAPhe. *Biochemistry 27*(15), 5771–5777.

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

Larracuente, A. M., T. B. Sackton, A. J. Greenberg, A. Wong, N. D. Singh, D. Sturgill, Y. Zhang, B. Oliver, and A. G. Clark (2008). Evolution of protein-coding genes in *Drosophila*. *Trends in Genetics 24*(3), 114–123.

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

Lee, Y. S., Y. Shibata, A. Malhotra, and A. Dutta (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & Development 23*(22), 2639–2649.

Li, P., B. P. Roberts, D. K. Chakravorty, and K. M. Merz (2013). Rational design of particle mesh ewald compatible lennard-jones parameters for +2 metal cations in explicit solvent. *Journal of Chemical Theory and Computation 9*(6), 2733–2748.

Li, W. and J. Frank (2007). Transfer RNA in the hybrid P/E state: Correlating molecular dynamics simulations with cryo-EM data. *Proceedings of the National Academy of Sciences of the United States of America 104*(42), 16540–16545.

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

Lu, J., Y. Shen, Q. Wu, S. Kumar, B. He, S. Shi, R. W. Carthew, S. M. Wang, and C.-I. Wu (2008). The birth and death of microRNA genes in *Drosophila*. *Nature Genetics 40*(3), 351–355.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs (2012). The *Drosophila melanogaster* genetic reference panel. *Nature 482*(7384), 173–178.

Marciniec, T., J. Ciesiołka, J. Wrzesinski, and W. J. Krzyzosiak (1989). Identification of the magnesium, europium and lead binding sites in *E. coli* and lupine tRNAPhe by specific metal ion-induced cleavages. *FEBS Letters 243*(2), 293–298.

Marck, C. and H. Grosjean (2002). tRNomics: analysis of tRNA genes from 50 genomes of eukarya, archaea, and bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA 8*(10), 1189–1232.

Mattick, J. S. and I. V. Makunin (2006). Non-coding RNA. *Human molecular genetics 15*(1), R17–29.

McDonald, J. H. and M. Kreitman (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature 351*(6328), 652–654.

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

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087.

Mobley, D. L., J. D. Chodera, and K. A. Dill (2006). On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *The Journal of Chemical Physics 125*(8), 084902.

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

Ogata, K. and H. Nohara (1957). The possible role of the ribonucleic acid (RNA) of the pH 5 enzyme in amino acid activation. *Biochimica et biophysica acta 25*(3), 659–660.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature 246*(5428), 96–98.

Orioli, A., C. Pascali, A. Pagano, M. Teichmann, and G. Dieci (2012). RNA polymerase III transcription control elements: Themes and variations. *Gene 493*(2), 185–194.

Pan, T., D. M. Long, and O. C. Uhlenbeck (1993). 12 divalent metal ions in RNA folding and catalysis. pp. 271–302. Cold Spring Harbor, NY: Cold Spring Harbor Monograph Archive.

Pederson, T. (2010). Regulatory RNAs derived from transfer RNA? *RNA 16*(10), 1865–1869.

Phillips, J. L., M. E. Colvin, E. Y. Lau, and S. Newsam (2008). Analyzing dynamical simulations of intrinsically disordered proteins using spectral clustering. In *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, Philadelphia, PA, pp. 17–24. IEEE.

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

Prabhakaran, M., S. C. Harvey, and J. A. McCammon (1985). Molecular-dynamics simulation of phenylalanine transfer RNA. II. Amplitudes, anisotropies, and anharmonicities of atomic motions. *Biopolymers 24*(7), 1189–1204.

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

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

Rogers, H. H., C. M. Bergman, and S. Griffiths-Jones (2010). The evolution of tRNA genes in *Drosophila*. *Genome Biology and Evolution 2*(0), 467–477.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology 61*(3), 539–542.

Rozas, J. and R. Rozas (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics 15*(2), 174–175.

Saks, M. E. and J. R. Sampson (2013). Evolution of tRNA recognition systems and tRNA gene sequences. *Journal of Molecular Evolution 40*(5), 509–518.

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

Sanbonmatsu, K. Y., S. Joseph, and C.-S. Tung (2005). Simulating movement of tRNA into the ribosome during decoding. *Proceedings of the National Academy of Sciences of the United States of America 102*(44), 15854–15859.

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

Schaack, J., S. Sharp, T. Dingermann, D. J. Burke, L. Cooley, and D. Söll (1984). The extent of a eukaryotic tRNA gene. 5'- and 3'-flanking sequence dependence for transcription and stable complex formation. *The Journal of biological chemistry 259*(3), 1461–1467.

Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto (2009). Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics 5*(6), e1000495.

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

Sharp, S. J., J. Schaack, L. Cooley, D. J. Burke, and D. Söll (1985). Structure and transcription of eukaryotic tRNA genes. *CRC Critical Reviews in Biochemistry 19*(2), 107–144.

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

Shi, P. Y., N. Maizels, and A. M. Weiner (1998). CCA addition by tRNA nucleotidyltransferase: Polymerization without translocation? *The EMBO Journal 17*(11), 3197–3206.

Sprague, K. U., D. Larson, and D. Morton (1980). 5' flanking sequence signals are required for activity of silkworm alanine tRNA genes in homologous in vitro transcription systems. *Cell 22*(1 Pt 1), 171–178.

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

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

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

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, H. F. curators, B. D. G. Project, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, M. Kellis, M. A. Crosby, B. B. Matthews, A. J. Schroeder, L. Sian Gramates, S. E. St Pierre, M. Roark, K. L. Wiley, Jr, R. J. Kulathinal, P. Zhang, K. V. Myrick, J. V. Antone, W. M. Gelbart, J. W. Carlson, C. Yu, S. Park, K. H. Wan, and S. E. Celniker (2007). Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature 450*(7167), 219–232.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics 123*(3), 585–595.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics 135*(2), 599–607.

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

Tåquist, H., Y. Cui, and D. H. Ardell (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.

Westhof, E., P. Dumas, and D. Moras (1988). Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Crystallographica Section A: Foundations of Crystallography 44*(2), 112–124.

Widmann, J., M. Di Giulio, M. Yarus, and R. Knight (2005). tRNA creation by hairpin duplication. *Journal of Molecular Evolution 61*(4), 524–530.

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

Withers, M., L. Wernisch, and M. dos Reis (2006). Archaeology and evolution of transfer RNA genes in the escherichia coli genome. *RNA 12*(6), 933–942.

Woese, C. R., G. J. Olsen, M. Ibba, and D. Söll (2000). Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiology and Molecular Biology Reviews 64*(1), 202–236.

Wrzesinski, J., D. Michałowski, J. Ciesiołka, and W. J. Krzyzosiak (1995). Specific RNA cleavages induced by manganese ions. *FEBS Letters 374*(1), 62–68.