Exploring codon-anticodon adaptation in eukaryotes

Anna Van Weringh

Supervisor: Xuhua Xia

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for a
Master's degree from the
Ottawa-Carleton Institute of Biology

Abstract

tRNA genes have the fundamental role of translating the genetic code during protein synthesis. Beyond solely a passive decoding role, the tRNA pool exerts selection pressures on the codon usage of organisms and the viruses that infect them because processing codons read by rare tRNAs can be slow or even erroneous. To better understand the interactions of codons and anticodons in eukaryotic species, we first investigated whether tRNAs packaged into HIV-1 particles may relate to the poor codon usage of HIV-1 genes. By comparing the codon usage of HIV-1 genes with that of its human host, we found that tRNAs decoding poorly adapted codons are overrepresented in HIV-1 virions. Because the affinity of Gag-Pol for all tRNAs is non-specific, HIV packaging is most likely passive and reflects the tRNA pool at the time of viral particle formation. Moreover, differences that we found in the codon usage between early and late genes suggest alterations in the tRNA pool are induced late in viral infection. Next, we tested whether a reduced tRNA anticodon pattern, which was called into question by predicted tRNA datasets, is maintained across eukaryotes. tRNA prediction methods are prone to falsely identifying tRNA-derived repetitive sequences as functional tRNA genes. Thus, we proposed and tested a novel approach to identify falsely predicted tRNA genes using phylogenetics. Phylogenetic analysis removed nearly all the genes deviating from the anticodon pattern, therefore the anticodon pattern is reaffirmed across eukaryotes.

Résumé

Les gènes d'ARN de transfert (ARNt) interviennent lors de la traduction du code génétique pendant la synthèse des protéines. L'interprétation de codons traduits par des ARNt rares peut être lente et même parfois infidèle. Par conséquent, la disponibilité d'ARN entraîne une préférence d'usage des codons chez les organismes et les virus qui les infectent. Afin de mieux comprendre les interactions entre codons et anticodons dans les espèces eucaryotes, nous avons d'abord cherché à savoir si les ARNt contenus dans les particules de VIH-1 aident à la traduction des gènes du VIH. En comparant l'usage des codons des gènes du VIH-1 avec celui de son hôte humain, nous avons constaté que les ARNt mal adaptés sont surreprésentés dans les virions. L'affinité de Gag-Pol pour les ARNt est non-spécifique. Il est donc probable que les ARNt empaquetés reflètent les ARNt disponibles lors de la formation des particules de VIH. Par ailleurs, les différences observées dans l'usage des codons entre les gènes précoces et tardifs suggèrent que la disponibilité des ARNt évolue au cours du cycle viral. Ensuite, nous avons testé si un modèle d'usage réduit des ARNt est maintenu à travers les eucaryotes. Ce modèle a été remis en cause par la prédiction d'ARNt. Cependant, les méthodes de prédiction ARNt ont tendance à identifier les ARNt dérivés de séquences répétitives comme étant des gènes d'ARNt fonctionnels. Ainsi, nous avons proposé et testé une nouvelle méthode phylogénétique pour éliminer les gènes d'ARNt faussement prédits. L'analyse phylogénétique a éliminé presque tous les gènes s'écartant du modèle décrit. Nous réaffirmons donc la validité de ce modèle à travers les eucaryotes.

Acknowledgments

Many thanks to all who have supported me during the completion of this thesis. Although I cannot name everyone, I would like to explicitly thank my supervisor, Dr Xuhua Xia, my committee members, Dr Marcel Turcotte and Dr Michel Dumontier, and Dr Stéphane Aris-Brosou for their guidance. I would also like to thank my past and present lab mates, in particular Manon Ragonnet with whom I have muddled through from start to finish (with much editing along the way). I would like to thank the University of Ottawa and NSERC for their generous funding, which allowed me the opportunity to complete this work. Lastly, I would like to thank my friends and family for helping me keep a smile on my face.

List of abbreviations

A – adenosine
C – cytosine
G – guanine
T – thymine (DNA only)
U – uracil (RNA only)
I – inosine
Y – pyrimidines: U/T and C
R – purines: A and G
B – C, U/T and G
N – A, C, U/T and G

wN – wobble nucleotide (incl. wA, wG, wI, wC, etc)
RNA – ribonucleic acid
DNA – deoxyribonucleic acid
tRNA – transfer RNA
mRNA – messenger RNA
aaRS – aminoacyl tRNA synthetases

IUPAC amino acids abbreviations:
Gly – glycine
Ala – alanine
Val – valine
Leu – leucine
Ile – isoleucine
Ser – serine
Thr – threonine
Asp – aspartic acid
Glu – glutamic acid
Lys – lysine
Arg – arginine
Cys – cysteine
Met – methionine
Phe – phenylalanine
Tyr – tyrosine
Trp – Tryptophan
His – histidine
Pro – proline
Gln – glutamine
Asn - asparigine

e.g. $tRNA^{Ala/GGC}$ – alanine transfer RNA with anticodon GGC

MCMC – Markov Chain Monte Carlo
PP – Posterior Probability
GTR – General Time Reversible

CAI – codon anticodon index
RSCU – Relative Synonymous Codon Usage

Mbp – Mega (10^6) base pair
aa – amino acid
sec – second

HIV – Human Immunodeficiency Virus
AIDS – Acquired Immune Deficiency Syndrome
CD4+ T-cell – T lymphocyte cell possessing the CD4 surface molecule
GagVLP – Virus-Like Particles lacking the Gag-Pol protein
SIV – Simian Immunodeficiency Virus
HTLV – Human T-Lymphotropic Virus

gtRNAdb – Genomic tRNA Database
wA_d – deviating wA
wG_d – deviating wG
MSA – multiple sequence alignment

List of Tables

List of figures

List of Chapters

1 Introduction

1.1 tRNAs in translation

1.1.1 Protein synthesis

Synthesizing the proteins encoded by genes is a fundamental process, common to all living organisms. The genetic blueprints for all proteins are, simply put, strings of nucleotide triplets called codons, where each codon determines which amino acid is added. Genes are *transcribed* into messenger RNA (mRNA), which are then *translated* into proteins by the ribosome, a large enzyme complex that catalyzes protein synthesis, found in high numbers within all living cells. Transfer RNA (tRNA) genes are the answer keys to each codon; tRNAs read codons within the ribosome and deliver the corresponding amino acids. Codon decoding is accomplished by triplets of nucleotides within the RNA sequence of the tRNA, called the anticodon, which base pair with the three nucleotides of the codon. Upon correct base pairing between codon and anticodon, the amino acid brought in by the tRNA is added to the growing peptide chain.

As the ribosome reads through an entire mRNA sequence, tRNAs move through the ribosome, entering at the A-site, moving to the P-site and exiting by the E-site. Translation begins with the three nucleotides of the initiation codon, methionine, sitting exposed in the bottom of A-site of the ribosome. Numerous incorrect, non-cognate tRNA molecules may enter into the A-site, but only the corresponding, cognate $tRNA^{Met}$ is accepted. Integral to the specificity of translation, acceptance is a highly specific process. After entry of a tRNA, the ribosome changes conformation from an 'open' to a 'closed' state, bringing nucleotides within the 16S RNA (an RNA component of the ribosome) to physically interact with the minor groove of the codon-anticodon helix; these interactions test that the geometry of the base pairing is correct (Ogle et al. 2001; Ogle et al. 2002). Only after the cognate tRNA is accepted does the ribosome translocate to read the next codon: the $tRNA^{Met}$ is transferred to the P-site in the ribosome, leaving the next codon exposed in the pocket of the A-site. After entry and acceptance of the next cognate tRNA into the A-site, a peptide bond is made between the amino acids of the P-site and of the new tRNA in the A-site, to begin building the peptide. Here, the growing peptide chain remains attached to the tRNA in the A-site. When the ribosome moves to read the next codon, the growing peptide chain and the tRNA it is tethered to are transferred to the P-site, while the $tRNA^{Met}$, now 'empty', moves to the last, E-site and is released. This iterative process comprises the elongation phase of protein synthesis; translation continues until the entire mRNA sequence has been read through.

Translation is an extremely fast process. The rate of amino acid addition in elongation is estimated to approach 15 aa/sec in prokaryotes and a slower 2 aa/sec in eukaryotes (Cannarozzi et al. 2010). Although, the rate of synthesis in eukaryotes has been shown to vary between 1-10 aa/sec depending on the transcript and physiological conditions of the cell (Cannarozzi et al. 2010). A rate-limiting step of elongation is the delivery of cognate tRNAs to the A-site (Fluitt, Pienaar, and Viljoen 2007). If the cognate tRNA is found at low concentrations, it takes longer to enter the A-site and thus

slows the progression of peptide synthesis. If a cognate tRNA is not at all available or is found at such low concentrations that is takes too long to enter the A-site, it becomes more likely that a non-cognate tRNA is accepted, causing addition of the wrong amino acid to the growing peptide chain (Scorer, Carrier, and Rosenberger 1991).

1.1.2 The genetic code

The genetic code (Figure 1.1) is the set of rules dictating which amino acid each codon denotes.

Figure 1.1 – The nuclear genetic code of vertebrate organisms, taken from a lecture slide of Xuhua Xia.

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

In total, there are 64 possible codons. All codons are designated to either 1 of 20 amino acids or a stop signal (stop codon). With more codons than signals needing to be encoded, the genetic code is redundant, with redundancies primarily found at the 3rd codon position. Codon families are formed by codons that encode the same amino acid, nearly all of which share the same 1st and 2nd position nucleotides. Certain amino acids are encoded only by a single codon. Then, there are numerous

codon families with only 2 members, called 2-fold degenerate families or 2-codon families. The 2-codon families tend to differ only at the 3rd codon position and are either A and G-ending (R-ending) or are C and U-ending (Y-ending). There is one 3-fold degenerate family, Ile, which encodes all but the G-ending codon. Next, there are 4-codon families, which are encoded by all the possible 3rd codon position nucleotides. Lastly, there are 6-codon families, which are amalgamations of a 4-codon and a 2-codon family. The 6-codon families are the only families where codons that differ at a position other than the 3rd codon position can still encode the same amino acid. In the standard vertebrate code there are 3 stop codons.

The genetic code is able to evolve and is slightly variable between organisms (Osawa et al. 1990). Minor changes in the genetic code have been acquired across large classes of species, including bacteria, archea, fungi, invertebrates, vertebrates (also referred to as the standard genetic code), and the intracellular organelles with their own genomes, mitochondria and chloroplasts.

1.1.3 The tRNA molecule

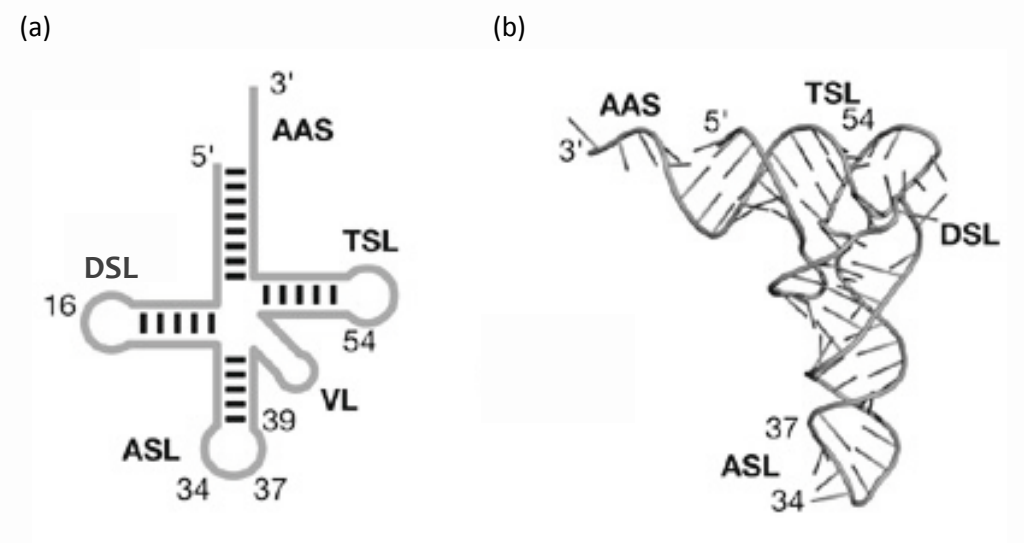
tRNAs are non-coding RNA (ncRNA) molecules, *i.e.* their genes are not translated into proteins, but rather function as RNA molecules.

Many ncRNAs form complex, three-dimensional structures, which are indispensable for proper functioning. RNA structures are described with three hierarchical levels. The primary, most basic level of organization of an RNA molecule is the sequence itself: a string of A, C, G and U nucleotides, as dictated by the genetic sequence. Then, base pairs between nucleotides within RNA molecule can be formed. These base pairs hold together the secondary structure, which is composed of stretches of base pairs called stems as well as loops of unpaired sequences between the stems. The allowable base pairs in RNA are the Watson-Crick pairings, AU and GC, along with the GU base pair that is unique to RNA. Secondary structures are quite strong, as base pairing interactions are held

together by hydrogen bonds, the different pairings differ in the number of hydrogen bonds they form. From strongest to weakest: GC pairs form 3 hydrogen bonds, AU pairs form 2 hydrogen bonds and GU pairs form only a single hydrogen bond. The final level of organization of RNA molecules, the tertiary structure, involves larger interactions between the stems and loops of the secondary structure and describes the three-dimensional shape a ncRNA takes on in the cellular milieu. Such interactions include stacking of multiple base paired regions, held together by van der Waals forces, and the formation of additional base pairing between unpaired nucleotides of loops, internal loops or bulges (Westhof and Auffinger 2000). The stability of both tertiary and secondary structures are dependent upon the temperature and surrounding environment, for example salts are required to counter-act the repulsion between the negatively charged phosphate backbone of each nucleotide (Westhof and Auffinger 2000).

tRNA genes are on average 73-93 nucleotides long. tRNAs across all domains of life share a common secondary structure called the cloverleaf structure (Figure 1.2a). The cloverleaf forms three stem loops with one additional, terminal stem that brings together the 3' and 5' ends. The terminal stem is called the acceptor stem. Amino acids tether to the terminal 3' end of tRNA molecules; the 3' end has an overhang of four nucleotides, ending with the conserved sequence 5'CCA3'. The three stem loops formed, from 5' to 3' are the D-loop, the anticodon loop, and the T-Ψ-C loop. The center 3 nucleotides of the anticodon loop compose the anticodon. The cloverleaf structure then folds into the final tertiary structure of the tRNA, the L-shape (Figure 1.2b). There are numerous invariant nucleotides along the cloverleaf structure that are required for proper formation of the L-shape (Garrett and Grisham 2005). In the final three-dimensional tRNA molecule, held at one end is the anticodon loop and at the other the accepting stem. This conformation is required for function in the ribosome, where the codon-anticodon interactions occur in the bed of the A-site and the growing peptide chain extends away from the ribosome.

Figure 1.2 – Conserved (a) cloverleaf secondary structure and (b) tertiary structure of a tRNA, the L-shape. The following abbreviations are explained: AAS – Amino acid Accepting Stem, DSL – D Stem Loop, TSL – T-Ψ-C Stem Loop, ASL – Anticodon Stem Loop and VL – Variable Loop. Figure adapted with permission from (Agris 2008).



1.1.4 tRNA loading

Amino-acyl tRNA synthetases (aaRS) are the enzymes that catalyze the formation of the bond between the acceptor stem of the tRNA and its correct amino acid, a process is called tRNA loading. For each of the 20 amino acids there is a unique aaRS. Therefore, to maintain agreement between loaded amino acid and the anticodon, each aaRS must recognize all the tRNAs within its family and, crucially, none of the other 19 tRNA families. Were this specificity lost and tRNAs incorrectly loaded, the fundamental ability to build a protein as per the genetic instructions would be lost. Thus, each aaRS possesses a uniquely high affinity for its amino acid and its tRNAs. Moreover, the majority of aaRS also have an additional proof-reading function to help ensure the correct amino acid has been added to the tRNA (Berg, Tymoczko, and Stryer 2002). The recognition of the tRNA relies upon specific identity sites, which are sequences unique to each tRNA family along their contact surface with their aaRS. These identity sites include but are not limited to the anticodon

sequence and they vary between aaRS. For example, for $aaRS^{Met}$ and $aaRS^{Val}$ the anticodon is the sole determinant for tRNA recognition. Thus, mutating the anticodon of other tRNAs to the methionine or valine anticodon allows them to be loaded by $aaRS^{Met}$ and $aaRS^{Val}$, respectively (Garrett and Grisham 2005). Additionally, some modified nucleosides are positive determinants for amino-acylation (Sylvers et al. 1993).

1.1.5 Wobble base pairing

Another interaction, key to maintaining the specificity of translation, occurs between the codon and anticodon. Yet there is some flexibility in the decoding of tRNA genes, called wobble base pairing, between the third codon position and the first, wobble position of the anticodon. Consequently, single anticodons are capable of decoding multiple codons, to a maximum of four codons to one anticodon. This flexibility reflects the redundancy of the genetic code at the 3rd codon position, reducing the number of tRNAs required to decode all possible codons. Nevertheless, in 6-codon families a single tRNA is incapable of reading all the codon members, since they differ at position other than the 3rd position. For example Leucine, the codons encoded CTN and TTR must each have tRNA family, $tRNA^{Leu/NAG}$ and $tRNA^{Leu/YAA}$, respectively, to decode them.

In 1966 Francis Crick first proposed the Wobble Hypothesis. Crick identified base pairs other than the traditional A-U and G-C Watson-Crick pairings. These wobble pairings were shown to be capable of a similar geometry as Watson-Crick base pairs (Crick 1966), and thus acceptable. The possible base pairs were also extended to a modified base found at the wobble position, inosine (wI, wobble I). wI residues are post-transcriptionally modified wA (wobble A) residues (Crick 1966; Auxilien et al. 1996). Crick proposed that wI anticodons may decode U, C and A-ending codons (although it was later shown that I reads A-ending codons rather poorly (Curran 1995)). Crick also

proposed the base pairing between G and U residues. In total, Crick proposed that wU decodes A and G-ending codons, wC G-ending, wA U-ending and wG U and C-ending codons.

However, with increased experimental data on decoding capacity and the discovery of many more modified bases at the wobble site, numerous updates have since been made to these original rules. For example, unmodified wU residues are capable of decoding all 4-codon families (Yokoyama et al. 1985). In following, it was proposed that the numerous modifications made to wU residues must limit their liberal decoding, so as to retain specificity in 2 and 3-fold degenerate families (Agris 1991; Lim 1994). Thus, modifications at the wobble site not only expand, but also limit the decoding capacity of a nucleotide.

Additionally, the *in vivo* decoding capacity of numerous wI and modified wU anticodons in *S. cerevisiae* were recently studied (Johansson et al. 2008), showing that the decoding capacity of the different modifications actually varied between codon families. For example, one particular wU modification, ncm^5U, is shared between numerous 4-fold degenerate families in *S. cerevisiae*. In the valine family, the ncm^5U wobble nucleotide decoded A and G-ending codons equally well, while in serine and in threonine the ncm^5U wobble nucleotide decoded the A-ending codon mostly, and the G-ending rather poorly. Moreover, in the proline family, the ncm^5U wobble nucleotide decoded all 4 codons of the family. Thus, the decoding rules for each wobble nucleotide are in reality much more difficult to assign. There are unknown features, other than the wobble nucleotide, which determine the decoding capacity of each anticodon (Johansson et al. 2008).

1.1.6 tRNA decoding patterns in prokaryotes and eukaryotes

When looking at the total number of tRNA genes, it can be understood how a system of tRNA genes is used to decode all codons. From this perspective, organisms in the three domains of life have rather different tRNA decoding systems. Wobble base pairing allows the decoding of all 64

possible codons, save stop codons, with an inferior number of tRNA anticodons. In bacterial and archeal species, the number of total tRNA genes is constrained by smaller genomes. Typically, they possess only a few tRNAs per family; in the more extreme cases, sometimes only one (Chan and Lowe 2009). Conversely, the large genomes of eukaryotes allow for very high numbers of tRNA genes, on average between 10-30 per codon family. Consequently, anticodon choices and particularly anticodon diversity between prokaryotes and eukaryotes are very different. For all codons to still be decoded in prokaryotes, the best wobble choice is one that is exceptionally versatile, to read the maximum numbers of codons within the family (Xia 2008). Specifically, in bacteria with very few tRNAs the same anticodons are often chosen: the wU for 4-codon families, and the wG and modified wU (here modified to reduce decoding capacity), respectively, in NNY and NNR 2-codon families (Rocha 2004). Nevertheless, even the versatile anticodons are limited, and certain codons will not have an available tRNA. Consequently, codons with few or no tRNA to decode them are rare.

As the genomes of organisms become larger, particularly in eukaryotic species, the total number of tRNAs increase significantly. For example, the yeast *Saccharomyces cerevisiae* encodes a total of 271 tRNA genes (Chan and Lowe 2009), numbers sufficient for multiple tRNAs to decode each codon family and for no codon to be without a cognate tRNA (Johansson et al. 2008). In eukaryotes more complicated than *S. cerevisiae*, the tRNA gene numbers are further increased. Nevertheless, despite the increased tRNA gene numbers, the total possible 61 anticodons are never observed. Rather, a reduced set of anticodons is observed across numerous species with the total anticodon diversity hypothesized to never exceed 45, as extrapolated from datasets of the total tRNA content of lower eukaryotes and partial sets from a few higher eukaryotes (Percudani 2001; Marck and Grosjean 2002; Agris 2004). The limit of 45 can be attributed to limited anticodon diversity for the decoding of Y-ending codons. In all codon families, there is an observed dichotomous choice between wG and wA anticodons. In 3-fold and 4-fold degenerate families, this wA is altered to wI to

expand its decoding capacity to read both the Y-ending codons, the same decoding capacity as wG. Meanwhile, in 2-fold degenerate families, anticodons with an unmodified wA are never observed. Primarily, wG is used to decode both the Y-ending codons. It has been proposed that the unmodified wA destabilizes translation when in the P-site of the ribosome (Lim 1994; Lim 1995), a possible explanation for its absolute avoidance. Nevertheless, these claims has never been directly addressed experimentally and there are rare cases where unmodified wA anticodons are found, with no apparent cost (Chen et al. 2002).

Commonly applied when assigning the decoding capacity of each nucleotide is a rule of parsimony, which proposes that overlaps in decoding should be rare (Percudani 2001). As such, when assigning the decoding capacity of anticodons given the total anticodon content, wobble base pairing is assigned only when there is no tRNA available that can decode a codon with Watson-Crick base pairing (Percudani 2001; Cannarozzi et al. 2010). From the study of the *S. cerevisiae* tRNA genes, for the most part it does seem that there is very little overlap between the decoding of the different anticodons. Although, there are exceptional families with codons read by multiple anticodons, for example leucine CUY codons were decoded by a wG and an unmodified wU (Johansson et al. 2008).

1.2 Codon-anticodon adaptation

1.2.1 Synonymous mutations and codon-anticodon adaptation in unicellular organisms

Due to the degeneracy of the genetic code, there are many different ways to encode a single protein sequence. Synonymous mutations, as opposed to non-synonymous, are mutations that do not change the amino acid sequence, *i.e.* codons are changed to another codon in the same codon family. Although they do not affect the protein sequence, the frequencies of all codons within the genes of an organism are not uniform. Certain codons tend to be preferred and others rarely used; these

preferences are unique to each species (Grantham et al. 1980). Unequal choices between synonymous codons are called the codon usage bias. Moreover, the same codon usage bias is not always shared among genes in the same genome (Gouy and Gautier 1982; Ikemura 1985), or even across a genic sequence (Akashi 1994).

In part, codon usage bias is influenced by the mutation bias of an organism such that in GC-biased genomes, G and C-ending codons tend to be preferred and A and T-ending codons avoided; the opposite is true of AT-biased genomes. The nucleotide bias of an organism is often simply explained by a constant imbalance in mutation frequencies (Sueoka 1962). In addition to the mutation bias, there is translational selection acting upon codon usage (Sharp and Devine 1989), whose strength depends upon how frequently that gene is expressed: highly expressed or constitutively expressed genes demonstrate stronger biases in codon usage than genes expressed at low levels (Bennetzen and Hall 1982; Ikemura 1985; Sharp and Devine 1989). Translational selection on codon usage is a response to unequal concentrations of tRNAs in the cellular tRNA pool. In particular, rare tRNAs stall translation, as protein synthesis cannot progress until the cognate tRNA enters into the ribosome. Conversely, abundant tRNAs allow rapid, efficient translation of the codons they decode. In agreement, experimentally optimizing codon usage has been shown to alter translational speed (Robinson et al. 1984; Sorensen and Pedersen 1991). Moreover, codons read by rare tRNAs are associated with increased rates of amino acid misincorporation (Scorer, Carrier, and Rosenberger 1991) and frame-shifting events (Sipley and Goldman 1993).

In unicellular organisms, tRNA gene copy numbers correlate extremely well to the tRNA pool; in these cells expression of tRNA genes has little to no regulation (Ikemura 1985). Consequently, strong agreement between the codon usage bias and tRNA gene copy numbers is readily observed across bacterial and fungal organisms, including the model organisms *Escherichia coli* and *S. cerevisiae* (Ikemura 1981; Gouy and Gautier 1982; Ikemura 1982; Ikemura 1985) as well

as mitochondrial genomes (Xia 2005). Yet, there is some variation in the level of agreement between tRNA content and codon usage between species. For example, translational selection is intensified across all genes in rapidly dividing bacteria due to the high demand for protein synthesis for cell growth to proceed through the cell cycle (Rocha 2004; Kudla et al. 2009; Sharp, Emery, and Zeng 2010).

However, it is not only that codon usage adapts to the tRNA pool. Rather, there is also selection upon the anticodon content to match codon usage, *i.e.* codon usage and tRNA content co-evolve (Bulmer 1987). Since the mutation bias influences the global preferences at the 3rd codon position, changes in the GC-bias of an organism would alter codon usage bias. The tRNA anticodons are then under selection to adapt to the new general codon usage patterns (Shields 1990).

1.2.2 Codon-anticodon adaptation in multicellular organisms

In the lower eukaryotes *C. elegans* (Duret 2000) and *D. melanogaster* (Moriyama and Powell 1997), codon-anticodon adaptation still appears to be occurring at the broad level of tRNA gene numbers and general codon usage bias (Kanaya et al. 2001). However, unlike in prokaryotes and simple eukaryotes where there is a high correlation between tRNA gene numbers and the resulting tRNA concentrations (Kanaya et al. 1999), in the more complex, multicellular organisms this clear agreement is no longer observed (Kanaya et al. 2001). Yet, there is additional regulation of tRNA gene expression that results in differences in tRNA expression levels across different tissues (Dittmar, Goodenbour, and Pan 2006) and, similarly, tissue specific codon usage due to gene expression differences in different cell types (Plotkin, Robins, and Levine 2004). Thus, in multicellular organisms, it is no longer a matter of codon adaptation across a host, but rather within each tissue. Consequently, the strength of codon-anticodon adaptation in higher organisms remains

unclear, and more precise data on the tissue specific expression patterns of genes and tRNAs are required to fully address the question of tissue specific codon-anticodon adaptation.

Another factor complicating the study of codon-anticodon adaptation in higher eukaryotes is that so far tRNA gene content has only been computationally predicted from genomic sequences, save for the few model organisms where tRNA content has been experimentally determined. This becomes particularly problematic in the genomes of multicellular organisms, where there is an increased frequency of erroneously predicted genes in the tRNA datasets (Lowe and Eddy 1997; Kramerov and Vasetskii 2009). These falsely predicted tRNAs cloud studies on codon-anticodon adaptation. The topics of tRNA prediction and false prediction are further discussed in greater detail (section 1.3).

1.2.3 Viral codon adaptation

Viruses are intracellular obligate parasites; they are on the edge of the definition of life, in that they are capable of reproduction, but only when parasitizing, or infecting, a living cell. After gaining access into a new host cell, viruses must co-opt the translational machinery to express their proteins. Therefore, the tRNA pool in a cell not only exerts selection pressures upon its own codon usage, but also on the codon usage of viruses. Thus, critical for efficient translation of viral proteins, there is selection from the tRNA pool acting on the codon usage bias of viral genes; pressure to choose within synonymous codons of each family those that are read by the most abundant anticodons (Bulmer 1987; Sharp and Li 1987). Because host codon usage also co-adapts with its tRNA pool (Bulmer 1987), a well-adapted virus is therefore one that has evolved codon usage frequencies similar to those of host genes. The first observation of such viral adaptation to its host codon usage was in bacteriophages (Sharp, Rogers, and McConnell 1984), which has since been confirmed on many occasions (Sau et al. 2006; Lucks et al. 2008; Wong et al. 2010). Meanwhile, a

rare alternative to adhering to the codon usage bias is to encode and express tRNA genes, which alter the tRNA pool to suit viral needs; although only DNA viruses seem to possess this ability (Limor-Waisberg et al. 2011).

Despite selection pressures to adhere to the host tRNA pool, numerous viruses nevertheless appear to have codon usage dominated more by mutation bias, rather than translational selection (Adams and Antoniw 2004; Gu et al. 2004) which may be due to the high mutation rates of many viruses (Jenkins and Holmes 2003). RNA viruses in particular, with an AT-rich mutation bias show poor codon adaptation when infecting GC-rich mammalian hosts (Berkhout et al. 2002; Jenkins and Holmes 2003). Thus, if the mutation bias acts in opposition to the selection from the tRNA pool, codon adaptation may be overwhelmed. Yet, even for viruses that show low agreement with host codon usage, altering codon usage influences the success of viral translation. For example, worsening the codon usage of poliovirus significantly reduced its ability to replicate (Coleman et al. 2008), while codon optimizing HIV sequences improves protein production (Haas, Park, and Seed 1996).

1.3 tRNA prediction

Direct sequencing of tRNA molecules is very difficult, due to the numerous modified nucleotides that alter how these nucleotides form base pairs. When sequenced, a spectrum of nucleotides is introduced rather than the Watson-Crick pair of the unmodified nucleotide (Iida, Jin, and Zhu 2009). tRNAs are found in such high numbers though, that carefully determining which modified nucleotide is where by more intensive methods would be infeasible. Moreover, there is the additional possibility of tissue specific expression, which dramatically increases the amount of samples that would require analysis.

Thus, the identification of tRNA genes, save in model organisms, is done entirely by computational methods. Thankfully, being so highly conserved across all organisms, tRNA

molecules are well suited to computational prediction. In general, the highly conserved secondary structure is a widely targeted motif used, or the large datasets of previously experimentally sequenced tRNA genes are used to build statistical descriptions of the family to identify putative tRNA genes. Among the numerous tRNA prediction programs, tRNAscan-SE (Lowe and Eddy 1997) is one of the most successful and is widely used (Marck and Grosjean 2002; Goodenbour and Pan 2006; Tang et al. 2009).

1.3.1 tRNAscan-SE

tRNAscan-SE is an implementation of three algorithms to identify putative tRNA genes from genomic DNA sequences. All three algorithms were initially proposed independently: tRNAscan (Fichant and Burks 1991), EufindtRNA (Pavesi et al. 1994) and an implementation of a covariance model (CM) for tRNA sequences (Eddy and Durbin 1994).

tRNAscan 1.3 takes advantage of the many invariant sites in tRNA sequences, scattered along the tRNA sequence, and the typical tRNA structure. tRNAscan 1.3 begins by searching each window within the sequence for the third (5'->3' speaking) stem loop, the T-Ψ-C loop, by first identifying the many invariant nucleotides in this stem loop, and then by testing whether the stem may be folding with the surrounding sequence. The nucleotide frequencies at all sites along the tRNA sequence were characterized initially by computing a consensus matrix from known nuclear tRNA sequences, save close homologues to reduce skewing, from a curated database of experimentally identified tRNAs, the Sprinzl database (Sprinzl et al. 1987). This gave a final test set of 242 sequences from bacteria, lower eukaryotes and vertebrates. Then, the first stem loop, the D loop, is searched in the 37-120 nucleotides upstream of any T-Ψ-C loops identified. If the D loop is also found, then it is determined whether an anticodon loop between the T-Ψ-C and D loops may be formed, allowing for the presence or absence of introns. An anticodon loop is only accepted if a U

residue is found directly 5' to the anticodon, a highly conserved nucleotide in tRNA sequences. If all these features are identifiable, the sequence is predicted to be a putative tRNA sequence (Fichant and Burks 1991). Alone, tRNAscan 1.3 has a true positive (TP) rate of 95.1% and a search speed of 400 bp/s (base pairs/second)[1]. However, it predicted 0.37 false positives (FP) per Mbp (10^6 base pairs) (Lowe and Eddy 1997).

EufindtRNA does not make use of the secondary structure of tRNA genes, but rather searches for conserved promoter sequences to identify tRNA genes. This approach was designed to specifically identify eukaryotic tRNA genes, motivated by the poor performance of structure-based tRNA prediction on the few atypical tRNA structures. tRNAs are all transcribed by the host RNA polymerase III (Pol III), while protein-coding genes are transcribed by RNA polymerase II. Pol III requires two conserved promoter 'boxes', A and B boxes, and a transcription termination sequence. The sequences of each these sequence elements were characterized by weight matrices to provide a statistical description of the sequences. Alone, the sequence description had poor tRNA predictive power, but the highly conserved spacing between these boxes, particularly between the A and B boxes and between the B box and termination signal, were able to identify tRNA sequences (Pavesi et al. 1994). Against eukaryotic sequences, EufindtRNA identified 98.6% of genes, but only 88.8% against non-eukaryotic tRNA datasets. EufindtRNA is extremely fast, with a search speed of 373,000 bp/s. Similar to tRNAscan 1.3, alone it has a fairly high FP rate, predicted around 0.23 FP/Mbp (Lowe and Eddy 1997).

Last of the three algorithms that compose tRNAscan-SE is the tRNA covariance model (CM). CMs are flexible, probabilistic definitions of both the primary sequence and secondary structure of the ncRNA. CMs can be implemented for any RNA gene family, trained from sequences

[1] For a Silicon Graphics Indigo2 R4400 200MHz workstation (also applicable for subsequent search speeds provided in this section).

of the ncRNA in question. In particular, it is possible to use available multiple sequence alignments or unaligned sequences. Unaligned input sequences are iteratively aligned during the training of the CM. During optimization steps a measure of the mutual information shared by the RNA molecules that is being increased: properly identifying and/or aligning base pairs improves the CM. In other words, the program is searching for the high correlations observed between nucleotides involved in a base pair. Once built, a CM may be used to identify putative members of that gene family; a sequence is aligned by dynamic programming to the CM and given a probability score. CMs are extremely successful at predicting tRNA genes with a TP rate of 99.8% and predicting less than 0.002 FP/Mbp. However, the tRNA CM scan sequences extremely slowly, at a rate of 20 bp/s (Lowe and Eddy 1997). Thus, it is not the best suited program to scan entire genomes, particularly the larger eukaryotic genomes (Eddy and Durbin 1994).

Altogether in tRNAscan-SE, tRNAscan 1.4 and EufindtRNA are used to identify putative tRNA genes. Then, it is only these putative sequences that are then analyzed by a tRNA CM, overcoming the speed constraints of the CM. The update made to tRNAscan-SE (from version 1.3 to 1.4) rendered it able to deal with ambiguous nucleotides, such that if an ambiguous nucleotide is found in a stem, it is predicted to base pair. This modification does, however, increase the FP rate against sequences with low resolution. Nevertheless, this is not problematic for the false positive (FP) rate because putative genes have yet to be screened by the tRNA CM. To maximize true positive (TP) and minimize false negative (FN) rates, the stringency requirements of EufindtRNA and tRNAscan are reduced. Again, the FP rate is unaffected because these genes have yet to be screened by the tRNA CM. The tRNA CM used in tRNAscan-SE was trained using 1415 known tRNA genes from the Sprinzl database at the time (Sprinzl et al. 1987). Overall, tRNAscan-SE scans sequences at a rate of 30,000 bp/s, approximately 1000 to 3000 times faster than the CM alone (Lowe and Eddy 1997). Moreover the integration of these three algorithms is efficacious: tRNAscan-SE has an

extremely low FP rate against randomly simulated sequences (<0.000007) as a consequence of the stringent CM, while retaining a high TP rate of 99.5% (Lowe and Eddy 1997).

1.3.2 Repetitive sequences fool tRNA prediction programs

Although tRNAscan-SE has a low false positive rate against random sequences, genomes are in reality not composed of random sequences. Rather, non-random background sequences fool tRNA predictions, the most important source of FP being from tRNA-derived repetitive sequences (Lowe and Eddy 1997). tRNA-derived repetitive sequences include common repetitive sequences such as SINEs (Short Interspersed Nuclear Elements); SINEs are tRNA genes that have been picked up by the machinery of retro-transposons (relics of retroviruses). Retro-transposons have the enzyme to reverse-transcribe (make a DNA copy of the RNA sequence) RNA sequences in the cell and integrate the newly synthesized DNA copy into the genome. Thus, a second copy of the original genetic sequence has been created. Moreover, having once been picked up by a retro-transposon, a sequence is flanked by elements that encourage this sequence to be amplified within the genome. For a tRNA-derived repetitive sequence, this new copies of tRNA sequence will typically have lost functionality in the process, having lost the flanking sequences that ensured its proper expression. Thus, this new copy is free to accumulate mutations.

Having originated from functional tRNA gene sequences, they share high sequence similarity with tRNA genes, and thus serve as a significant source of false tRNA predictions. Plus, being found in extremely high numbers, particularly in the larger eukaryotic genomes (Kramerov and Vasetskii 2009), these can dramatically inflate the number of tRNA genes predicted. For example, *Bos taurus* has near to 4000 total tRNA genes predicted and *Danio rerio* near to 12000 total genes predicted, despite average total numbers of tRNA genes in multicellular eukaryotes closer to 450 (Chan and Lowe 2009).

1.3.3 tRNA secondary structure

To perform bioinformatic analyses appropriate for RNA, a secondary structure is often required. The CM in tRNAscan-SE predicts the secondary structure along with the tRNA sequence. Thus, the algorithm is not only predicting a sequence that is likely to be a tRNA gene, rather it is predicting a sequence with a particular structure that is likely to be a tRNA gene. CM algorithms search for conserved stems as by the identification of compensatory mutation patterns, shared across numerous sequences, and are capable of predicting high quality structures of ncRNAs (Muse 1995; Gultyaev, Franch, and Gerdes 2000). Although, predictive success of covariance models is limited for ncRNA gene families with a flexible structure, with the highly conserved tRNA structure this is not the case.

1.4 Phylogenetics

Phylogenetics is a method that attempts to reconstruct the evolutionary history of sequences. It does so by analyzing aligned sequences and reconstructing relationships based on the differences found between those sequences. The final output of phylogenetic analysis is an evolutionary tree, which visually represents the best estimate of the relationships between the analyzed sequences.

1.4.1 Bayesian phylogenetics

Bayesian phylogenetics calculates the posterior probability of reconstructed trees given the data, using the Bayes theorem equation (1.1) where *H* stands for the hypothesis and *D* stands for the data (Yang and Rannala 1997; Larget and Simon 1999; Huelsenbeck et al. 2001).

$$P(H \mid D) = \left(\frac{P(D \mid H)P(H)}{P(D)} \right) \tag{1.1}$$

More specifically in phylogenetics the hypothesis includes tree topology, branch lengths and estimated parameters of the evolutionary model, while the data is the aligned input sequences. Within

equation 1.1, are terms for the prior probability distribution $P(H)$, and the marginal probability distribution of the data $P(D)$. The probability distribution $P(D|H)$ is the likelihood of the data given the hypothesis, which describes how well the model and tree explain the data (Huelsenbeck et al. 2001). Alone, the conditional probability is the traditional, non-Bayesian approach to calculating probability.

The prior may convey expectations about the results before looking at the data. For example if *a priori* a particular hypothesis is well supported, for it to be rejected new data would need to very strongly support its rejection. Similarly, the opposite would also be true: if a hypothesis was *a priori* very unlikely, then to accept that hypothesis, the data would need to very strongly support its acceptance. Although if there is no prior expectation, an uninformative distribution may be chosen so that the prior would effectively have no influence on the resulting posterior probability.

The marginal probability however is extremely complex to calculate and so analytically solving the posterior distribution is unfeasible. Yet, the use of the MCMC (Markov Chain Monte Carlo) was proposed to circumvent the need to calculate this (Yang and Rannala 1997; Mau, Newton, and Larget 1999) and allows Bayesian phylogenetics to be completed within reasonable time scales. MCMC is a random walk algorithm, where the algorithm 'steps' between different possible trees, or states, that describe the data. After choosing an initial state (either input by the user or randomly chosen), modifications to the parameters or tree are proposed. The probability of moving to the proposed state is given by the ratio of the posterior probabilities of the new state over the current state (Larget and Simon 1999). Thus, the better the new state is compared to the current, the more likely it will be accepted. But, there remains chance of accepting a tree with a lower likelihood value. It is in taking this ratio that simplifies out the complicated marginal probability, as it is a constant shared by all states (Yang and Rannala 1997; Larget and Simon 1999). If the new state is accepted, it is then subject to perturbation to propose the next new state. Overall, most steps taken by the MCMC

are in a direction that better explain the data until an optimum is found. At this point, the likelihood values of the trees being sampled stabilizes and the chain is said to have converged or reached equilibrium. Iterations prior to reaching equilibrium are called the burn-in.

Once stabilization is reached, the MCMC continues to run, sampling a set of highly likely trees. Given long enough runs, the frequency of times each tree is sampled is theoretically equal to the posterior probability; thus, the number of times that each tree is visited by the chain after stabilization can be approximated as the posterior probability (PP) of that particular tree (Larget and Simon 1999; Jow et al. 2002). Although for sampling to be truly reflective of the posterior distribution, the MCMC must be mixing well (Huelsenbeck et al. 2001). If too few states are accepted, then the MCMC is not sampling enough of the possible states. Conversely if too many states are accepted, then the chain may not be able to converge. Trees sampled during the MCMC, save the burn-in, are used to calculate the final, consensus tree, as well as the PP values in support of each node within the tree.

Although MCMC samplers are able to accept trees with lower likelihood values, chains may nevertheless get stuck in local optima. However, due to the stochastic nature of MCMC, each run is different and may produce different results. Thus to ensure that the chains are converging to the optimal solution, independent runs, with different initiation states, should be completed and checked for consistency (Huelsenbeck et al. 2001).

1.4.2 Evolutionary models

Crucial for accurate phylogenetic reconstruction is properly measuring the evolutionary distances between sequences. Not all changes are equally likely: mutations differ in their frequencies and in how well they are tolerated within sequences. Moreover, the likelihoods of different mutations are influenced by the nucleotide composition. Evolutionary models attempt to represent these

different likelihoods, assigning one or more parameters to estimate the frequencies of nucleotides, or states, within sequences and one or more parameters for the rates of change between these states.

DNA evolutionary models are 4x4 matrices. These model the transition between the four nucleotides. The most complex evolutionary model assumes that all possible changes occur at different rates and that all nucleotides are found at different frequencies. This model has the maximum number of parameters, which can be simplified by making assumptions about sequence evolution. Yet, all evolutionary models assume that sequence evolution is time-reversible, *i.e.* that the rate of change from A to T is the same as from T to A. There are numerous nested 4x4 evolutionary models. The simplest model, JC69, assumes all nucleotides are found at equal frequencies and that all changes are equally likely. Whereas, the most complex model, GTR, has all possible (time-reversible) changes and all nucleotide frequencies estimated with different parameters. In the many models of intermediate complexity, there are many possible reductions of parameters. Many of these models are optimized for protein-coding sequences, based on mutation frequencies observed in genic sequences. For example, transitions are changes that occur either between purines (A and G) or between pyrimidines (C and T), while transversions are changes from purine to pyrimidine or *vice versa*. Thus, in 2-fold degenerate families, transversions create non-synonymous mutations, while transitions create synonymous mutations. Therefore, transversions are found at lower rates in protein coding sequences than transitions. The K80 model was proposed to account for this difference; it models transitions and transversions with different parameters, to account for the differences in rates between these two types of mutations (Kimura 1980).

1.5 Phylogenetics for RNA sequences

Traditional multiple sequence alignment (MSA) methods and evolutionary models are optimized for the analysis of protein-coding sequences, not ncRNA sequences. Both only look at

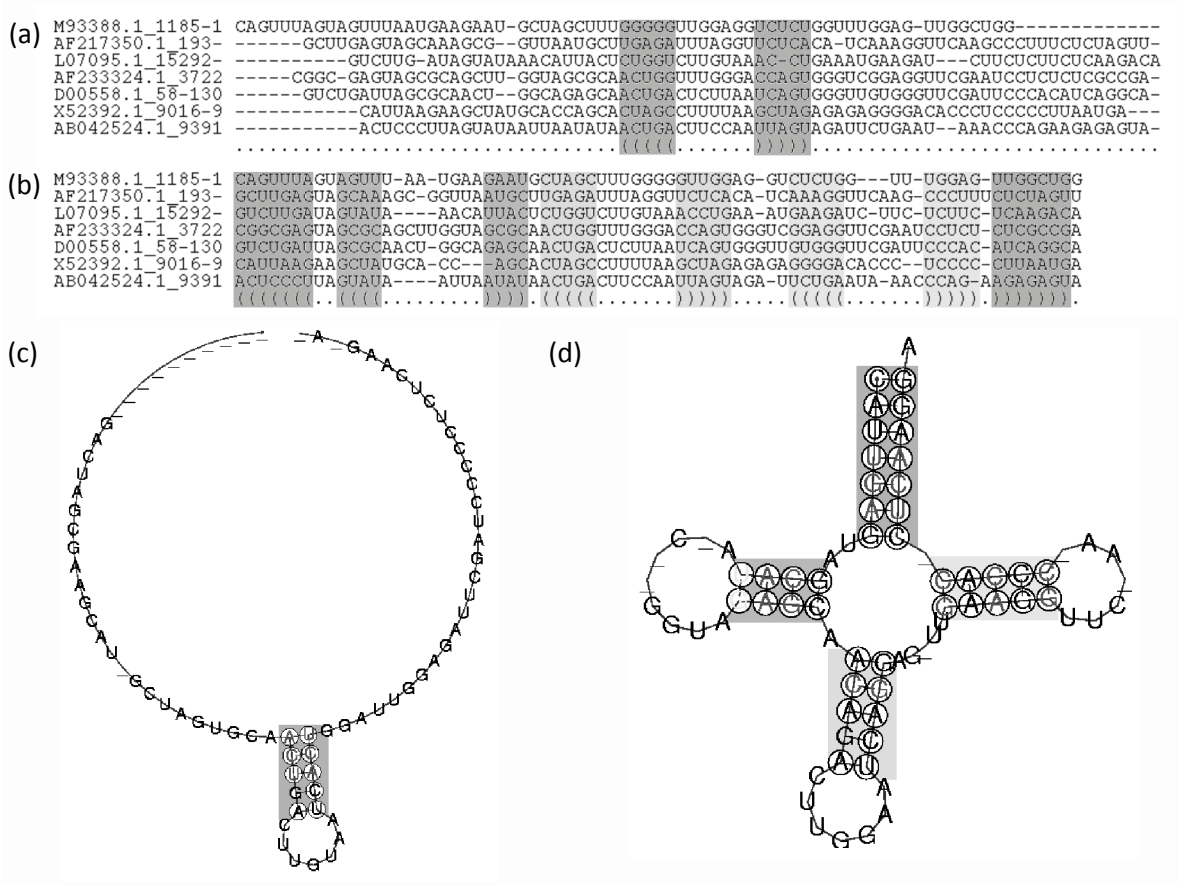
primary sequence information. However, ncRNA and protein-coding sequences differ quite significantly from one another; many of the assumptions that are appropriate for protein-coding sequences are inappropriate for ncRNA. Consequently, models and programs have been specifically designed for the analysis of ncRNA sequences.

1.5.1 Multiple sequence alignment for RNA sequences

For phylogenetics to be completed, a multiple sequence alignment (MSA) must first be built. The MSA aims to maximize similarity and minimize mismatches in a column by aligning sequences, allowing the insertion of gaps throughout each sequence. Thus, the MSA assigns which sites in each sequence will be considered as homologous. Moreover, it is from the MSA that the frequencies of exchange between different nucleotides for the evolutionary models are estimated.

Traditional methods use information from the sequence alone to build the MSA. However, these have been shown to perform quite poorly on ncRNA sequences, including tRNA genes (Hochsmann, Voss, and Giegerich 2004; Bauer, Klau, and Reinert 2007). These poor quality alignments would dramatically limit the reliability of any further analysis built off the MSA. The quality of these alignments, however, is dramatically improved by considering structural information along with the sequence during construction of the MSA (Kjer 1995; Hickson, Simon, and Perrey 2000; Xia 2000; Xia, Xie, and Kjer 2003; Hochsmann, Voss, and Giegerich 2004; Siebert and Backofen 2005), by penalizing and rewarding the loss and maintenance of base pairs, respectively, in the scoring measure used to assess and guide alterations to the MSA (Figure 1.3).

Figure 1.3 – Comparison of MSA for 7 tRNA sequences using (a) Clustal W, a traditional MSA program, and (b) LARA, a program that incorporates secondary structure information into the building of the MSA. Secondary structures were then reconstructed from these alignments with RNAalifold (c) for the Clustal W alignment and (d) for the LARA alignment, to evaluate the quality of the alignment. Base pairing regions identified by RNAalifold are shaded in each alignment. Figure adapted with permission from (Bauer, Klau, and Reinert 2007).



1.5.2 Compensatory mutations in RNA sequences and RNA evolutionary models

Due to their secondary structures, paired regions of RNA sequences evolve by compensatory mutations to avoid the loss of base pairs (Kimura 1985; Higgs 1998; Chen et al. 1999). In agreement, evidence of compensatory mutations is commonly found when analyzing ncRNA sequences (Rousset, Pelandakis, and Solignac 1991; Kirby, Muse, and Stephan 1995). The evolution of distant nucleotides in an RNA sequence is linked if nucleotides are involved in base pairs. Consider two

nucleotides *i* and *j* that pair with one another, and a mutation occurs at the position *i*. If the change at *i* causes the base pairing capacity to be lost, there is selection to restore the base pair. This is brought about either through reversion, site *i* returns to its original state, or through a second, compensatory mutation at position *j* that restores the base pair. With reversion being one of the two ways to restore a base pair, the rate of mutations in base pairing regions tends to be lower than in unpaired sequences (Kimura 1985; Stephan 1996).

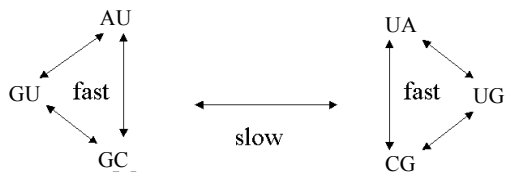
It is most likely for the first mutation to be retained if it creates a, albeit less stable, GU intermediate base pair. This slightly deleterious new base pair improves the likelihood of maintaining the first mutation, which provides time for the second mutation to restore a stronger Watson-Crick pair. Consequently, the exchange of base pairs are not equal, but are biased towards those with a GU or UG intermediate (Rousset, Pelandakis, and Solignac 1991; Higgs 1998). For example, beginning with a GC base pair, the most favorable first mutation would be changing the C to a U, creating a GU pairing. Then the subsequent mutation would be to change the G to an A, creating an AU pair. Similarly, the opposite change, CG to UA is also frequent. Whereas, changing from GC to UA or CG to AU is infrequent, having no GU intermediate to pass through (Figure 1.4a).

Due to the compensatory nature of mutations in RNA sequences, it is inappropriate to employ the single-site evolutionary models. Rather, single, unpaired sites should be modeled under a site-independent model while base pairs should be modeled with a paired site model (Jow et al. 2002; Knies et al. 2008). Numerous models have been proposed to describe the nature of compensatory mutations. A 6-state model (6x6 matrix) was the first paired evolutionary model proposed, allowing for all the pairing states: AU, UA, CG, GC, GU and UG, but not for any mismatches (Tillier and Collins 1995). The additional incorporation of mismatches was accomplished by combining all mismatches into one state, forming a 7-state model, and treating all mismatches individually, forming a 16-state model (Schoniger and von Haeseler 1994; Schoniger and von Haeseler 1999; Savill,

Hoyle, and Higgs 2001). 6-state and 7-state models were found to fit RNA data as well as the 16-state models (Higgs 2000; Savill, Hoyle, and Higgs 2001), an example 7-state model is shown below (Figure 1.4b).

Figure 1.4 – (a) Diagram of the relative likelihoods of moving between base pairs (b) Paired evolutionary model (7x7) 7D, which has three exchange parameters: alpha for the 'fast' exchanges, beta for the 'slow' exchanges and gamma for the transition between base pair and mismatch (MM). Panel (a) was obtained from Paul Higg's website: http://www.physics.mcmaster.ca/~higgsp/RNA.htm while panel (b) was taken with permission from (Savill, Hoyle, and Higgs 2001).

(a)



(b)

		1	2	3	4	5	6	7
		AU	*GU*	*GC*	*UA*	*UG*	*CG*	*MM*
1	*AU*	*	$\pi_2\alpha_s$	$\pi_3\alpha_d$	$\pi_4\beta$	$\pi_5\beta$	$\pi_6\beta$	$\pi_7\gamma$
2	*GU*	$\pi_1\alpha_s$	*	$\pi_3\alpha_s$	$\pi_4\beta$	$\pi_5\beta$	$\pi_6\beta$	$\pi_7\gamma$
3	*GC*	$\pi_1\alpha_d$	$\pi_2\alpha_s$	*	$\pi_4\beta$	$\pi_5\beta$	$\pi_6\beta$	$\pi_7\gamma$
4	*UA*	$\pi_1\beta$	$\pi_2\beta$	$\pi_3\beta$	*	$\pi_5\alpha_s$	$\pi_6\alpha_d$	$\pi_7\gamma$
5	*UG*	$\pi_1\beta$	$\pi_2\beta$	$\pi_3\beta$	$\pi_4\alpha_s$	*	$\pi_6\alpha_s$	$\pi_7\gamma$
6	*CG*	$\pi_1\beta$	$\pi_2\beta$	$\pi_3\beta$	$\pi_4\alpha_d$	$\pi_5\alpha_s$	*	$\pi_7\gamma$
7	*MM*	$\pi_1\gamma$	$\pi_2\gamma$	$\pi_3\gamma$	$\pi_4\gamma$	$\pi_5\gamma$	$\pi_6\gamma$	*

Only an approximate 2% of the base pairing regions in a dataset of rRNA sequences were made up by mismatches (Higgs 2000). The favorable dinucleotide exchanges, which pass through a GU intermediate, may also be modeled as a single exchange. The GU intermediate state is only transiently occupied (Tillier and Collins 1998; Jow et al. 2002): approximately 3% of base pairs in rRNA helical regions were GU or UG (Higgs 2000). Thus, evolutionary models have been introduced that allow for direct double substitution transitions (Higgs 2000; Savill, Hoyle, and Higgs 2001). By simulation studies comparing the performance of paired and unpaired models; paired models have been shown to more reliably fit compensatory evolution (Telford, Wise, and Gowri-

Shankar 2005). In following, Telford *et al* suggested that paired models be used whenever possible when analyzing RNA sequences.

1.6 Human Immunodeficiency Virus (HIV)

1.6.1 Pathogenicity

HIV is the etiologic agent of Acquired Immune Deficiency Syndrome (AIDS) in humans, a disease resulting from the loss of protection against pathogens due to weakening of the immune response. The virus is transmitted between hosts through sexual intercourse, blood (*e.g.* sharing needles, blood transfusion, wounds) or from an infected mother to her child during pregnancy, birthing or breast-feeding (Gayle and Hill 2001). After infection of a new host, there is a latency period lasting several years, where patients show few to no symptoms. The progression to AIDS occurs with the depletion of CD4+ T-cells, essential effectors of the adaptive immune response (Stevenson 2003). Thus, subsequent to the development of AIDS symptoms, patients rapidly succumb to opportunistic infections (Gayle and Hill 2001). There are two major families of HIV, HIV-1 and HIV-2. HIV-1 is the more virulent; symptoms progress faster and are more severe than with HIV-2. Moreover, HIV-1 is pandemic, having spread across the globe. Meanwhile, HIV-2 is endemic only to West Africa (CDC 2010). Though drugs have been developed that are capable of delaying the progression of HIV-1 infection to AIDS, there is no cure or vaccine available to abate further spread of the virus. Furthermore, HIV-1 is a rapidly evolving virus, and the development of drug resistance is an ever-present challenge, impeding efforts to control the epidemic (CDC 2010).

1.6.2 Classification, genome and lifecycle

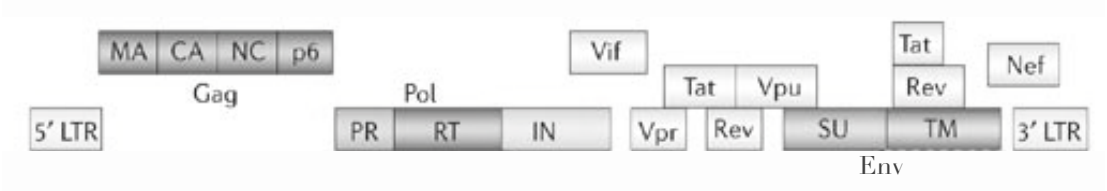
HIV is classified in the viral family *Retroviridae*, in the subfamily *Lentiviridae*; Retroviruses are characterized by an RNA genome, with a DNA intermediate as part of their lifecycle (Cullen 1991b). HIV infection begins with attachment of an HIV viral particle to a susceptible host cell, via

interactions between the envelope (Env) protein gp120, its receptor, CD4 and one of its co-receptors, CCR5 and CXCR4 (Sierra, Kupfer, and Kaiser 2005). This interaction initiates fusion of the viral and cellular membranes, providing entry of the virus into its new host cell. Following entry into the cell, a DNA copy of the RNA genome is synthesized, catalyzed by the viral polymerase, reverse-transcriptase; this newly synthesized DNA is then transported into the nucleus and becomes integrated, by the action of a viral protein integrase, into the host's genome (Sierra, Kupfer, and Kaiser 2005). The host cell's RNA polymerase then transcribes viral genes and these mRNA are subsequently translated by the host's translational machinery (Sierra, Kupfer, and Kaiser 2005). During late stages of the viral lifecycle, structural proteins are translated to high concentrations and full-length genomes are transcribed to form new viral particles, which begin assembling at the cell's outer membrane and then bud out of the cell (Cullen 1991b; Emerman and Malim 1998; Kleiman, Jones, and Musier-Forsyth 2010).

The genome of HIV-1 encodes the proteins Gag, Pol, Vif, Vpr, Tat, Rev, Vpu, Env, Nef, and is flanked by terminal repeats (LTR; Long Terminal Repeats) on either end (Figure 1.5) (Lengauer and Sing 2006). The 5' LTR promotes transcription and packaging of genomes into viral particles (Nishitsuji et al. 2001) while the 3' LTR promotes polyadenylation (Ashe et al. 1995). Many HIV-1 genes are translated as poly-proteins that are subsequently cleaved to form the smaller individual proteins, by the action of a viral protease. The proteins Gag and Env are structural, being the primary proteins within the viral particle: Gag encodes the non-transmembrane proteins including the capsid and nucleocapsid proteins, while Env encodes the transmembrane glycoprotein. Meanwhile the Pol protein includes the major viral enzymes: reverse-transcriptase, integrase and protease. The smaller proteins, on the other hand, have various regulatory functions within the cell. For example, Tat promotes recruitment of a cellular factor cyclin-T1, which strongly induces transcription of viral genes, while Rev promotes export of HIV-1 RNA from the nucleus into the cytoplasm (Stevenson

2003). Tat and Rev are also responsible for timing the switch into the late stages of the viral lifecycle including the initiation of viral particle production, by inducing the expression of structural proteins (Cullen 1991b). Additionally, viral proteins may often serve to counteract the canonical anti-viral defense mechanisms unleashed by the host cell. For example, the host protein APOBEC3, which mutagenizes the HIV genome, is targeted by the viral protein Vif (Sierra, Kupfer, and Kaiser 2005), while the viral protein Vpu interferes with the targeting of viral proteins for degradation induced by the host protein TRIM5α (Goujon and Malim 2010).

Figure 1.5 – Organization of genes in the HIV-1 genome. The entire HIV-1 genome is ~9kb (kilobases) long; genes are scaled to their respective lengths. The following abbreviations are explained: LTR – long terminal repeats, MA – matrix, CA – capsid, NC – nucleocapsid, PR – protease, RT – reverse-transcriptase, IN – integrase, SU – surface protein, TM – transmembrane protein. Figure adapted with permission from (Lengauer and Sing 2006).



1.7 Overview of subsequent chapters

For over 20 years, it has been known that codon usage and the transfer RNAs (tRNAs) that decode them co-evolve. However, in complex, multicellular eukaryotes the extent and strength of this codon-anticodon adaptation remains less clear for many reasons. In eukaryotes, tRNA genes are found in very high numbers, with high anticodon diversity. Moreover, tRNAs are predicted, yet even the most successful tRNA prediction programs are prone to falsely predicting tRNA-derived repetitive sequences as tRNA genes. We here explore the theme of codon-anticodon adaptation in eukaryotes first from the perspective of codon usage and then from the perspective of tRNA anticodon content. In chapter 2, we delve into the tRNA pool available during HIV-1 infection, investigating whether

tRNAs packaged into HIV-1 virions may help the translation of viral genes, whose codon usage is typically thought of as poorly adapted to its host. From this hypothesis generating study, we suggest that the tRNA pool may be altered during HIV-1 infection, benefiting the translation of HIV-1 genes. Then, in chapter 3, we aimed to whittle away the number of false positives in tRNA datasets. We proposed and tested a phylogenetic approach to eliminate false positives from tRNA datasets, taking advantage of the different evolutionary rates between freely evolving false positives and highly conserved functional tRNA genes. Lastly, in chapter 4, we reaffirm a pattern of anticodon diversity across 21 eukaryotic species, employing our phylogenetics approach to remove falsely predicted tRNA genes that deviated from the pattern.

2 How maladapted really is the codon usage of HIV-1?

2.1 Abstract

Viruses must use the tRNA pool of their host, and therefore their codon usage must adhere to their host cell's tRNA pool for efficient production of viral proteins. However, the codon usage of HIV-1 is A-rich, despite infecting a GC-rich host, as a result of a high A-biased mutation rate. Nevertheless, it was recently found that numerous rare tRNAs are enriched into HIV-1 particles. Save for $tRNA^{Lys}$, which is known to be packaged to prime reverse-transcription of the HIV-1 genome, the reason for packaging these rare tRNAs is unknown. By examining more closely the levels of enrichment of tRNAs and adaptation of HIV-1 codons, we demonstrate the degree of enrichment to be higher for tRNAs that read poorly adapted codons. Thus, tRNA packaging is in the position to benefit the translation of HIV-1 genes. With no biochemical support for HIV-1 actively enriching these rare tRNAs, we suggest that the tRNA pool is altered during infection and then passively packaged into HIV-1 particles. Investigating when this tRNA pool may be induced, we

found that early genes adhere much better than late genes. This suggests the tRNA pool may be modified late during HIV-1 infection. Thus, the codon usage of HIV-1 may not be as poor as traditionally thought.

2.2 Contributions

The data and interpretations in this chapter were published by *Molecular Biology and Evolution* (van Weringh et al. 2011). This work was the result of a collaborative project between myself and members of the Xia lab: Manon Ragonnet-Cronin, Erinija Pranckeviciene and Dr Xia, as well as Dr Lawrence Kleiman, a researcher at McGill University, and finally Mariana Pavon-Eternod, a graduate student in the lab of Dr Tao Pan at the University of Chicago.

This project was conceptualized by XX, who proposed the indices (equations 2.2 and 2.3) and tested for a correlation (Table 2.1). The development of the hypotheses and identification of the temporal differences (Figure 2.1, Figure 2.3, Table 2.2) resulted from discussions among XX, EP, MRC and AVW. Details regarding the mechanisms of HIV packaging arose from discussion between AVW and LK. The suggestion that the data might be heterogeneous (Figure 2.2) was suggested by MPE. The suggestion to look for differences in mutation bias between early and late genes was put forward by a reviewer of the paper upon submission to *MBE*; in response, the analysis of A_{12} was done by XX.

2.3 Introduction

Viruses rely upon the translational environment of their host cell and, consequently, there is selection on viral codon usage from the host tRNA pool. Thus, a well-adapted virus is one whose genes have a codon usage bias similar to that of its host genes, as it experiences the same selection pressures from the tRNA pool. Human Immunodeficiency Virus (HIV-1) has been previously noted for its exceptionally poor codon usage (Kypr and Mrazek 1987). Similar to many RNA viruses, HIV-

1 is an AT-biased virus infecting a GC-biased human host (Berkhout et al. 2002). This maladapted codon usage has been shown to come at a cost: codon optimization of the AT-rich codons to GC-rich codons in HIV-1 coding sequences has been shown to improve the efficiency of translation and increase the yield of protein production (Haas, Park, and Seed 1996).

Traditionally, the high mutation rates of RNA viruses have been blamed for their inability to adhere to tRNA pool (Jenkins and Holmes 2003). The mutation rate of HIV-1 is extraordinarily high, estimated at 2.5×10^{-3} substitutions/site/year (Hanada, Suzuki, and Gojobori 2004). Mutations in HIV-1 are extremely A-biased: the prevalence of A nucleotides within the genomes of some HIV-1 subtypes attains 40% (Vartanian, Henry, and Wain-Hobson 2002). The mutational A-bias of HIV-1 is in part the result of a biased nucleotide pool and a viral polymerase, reverse-transcriptase, with no proof-reading capacity that increases the likelihood of erroneously added T residues in the antisense genome (and thus A residues in the complementary sense strand) (Martinez, Vartanian, and Simon 1994; Vartanian, Henry, and Wain-Hobson 2002). Moreover, in large part the A-bias results from the action of the anti-viral host protein APOBEC3, which actively mutates the viral genome (Yu et al. 2004). In addition, it has been proposed that the A-bias beneficially reduces the secondary structure of the HIV-1 RNA genome, facilitating reverse-transcription (Keating et al. 2009). As the nucleotide bias of a genome is a strong contributor to codon usage bias (Jenkins and Holmes 2003), the codons in HIV-1 tend to be A-ending. Thus, with host codon usage being strongly GC-biased, the A-biased mutation rate of HIV-1 acts in opposition to and is consequently thought to limit codon adaptation (Berkhout et al. 2002).

Nevertheless, recent evidence suggests another explanation for the poor codon usage of HIV-1. It was recently demonstrated by tRNA microarray that HIV-1 particles contain numerous tRNA species in addition to the $tRNA^{Lys}$ and $tRNA^{Ile}$ previously known to be packaged (Jiang et al. 1993; Pavon-Eternod et al. 2010). HIV-1 packages $tRNA^{Lys}$ during viral particle assembly to be used as a

primer for reverse-transcription of its genome, upon infection of a new host cell (Marquet et al. 1995). On the other hand, the reasons for packaging $tRNA^{Ile}$ and other non-lysyl (other than $tRNA^{Lys}$) tRNAs are unknown. Surprisingly, Pavon-Eternod *et al* (Pavon-Eternod et al. 2010) demonstrated that the non-lysyl tRNAs quantified within HIV-1 particles did not proportionately reflect the host tRNA pool. For example, while in human cells, the $tRNA^{Ile}$ pool is biased towards $tRNA^{Ile/IAU}$, the $tRNA^{Ile}$ most highly packaged was not the most common $tRNA^{Ile/IAU}$, but the rare $tRNA^{Ile/UAU}$ that reads the A-ending Ile codon AUA (Pavon-Eternod et al. 2010). The authors speculated that the incorporation of rare tRNAs into viral particles may enhance translation of the poorly adapted HIV-1 proteins (Pavon-Eternod et al. 2010).

The mechanism by which these rare tRNAs are enriched and packaged into viral particles also remains unknown. The packaging of $tRNA^{Lys}$ by HIV-1 has been well characterized. Multiple molecular players are required: the viral proteins Gag and Gag-Pol, viral genomic RNA and the host enzyme lysyl-tRNA synthetase (LysRS), which loads the $tRNA^{Lys}$ with lysine residues (Javanbakht et al. 2003; Kleiman, Halwani, and Javanbakht 2004; Kleiman, Jones, and Musier-Forsyth 2010). $tRNA^{Lys}$ incorporation into viral particles requires Gag-Pol (Mak et al. 1994), yet, despite this dependence Gag-Pol does not itself possess specificity for $tRNA^{Lys}$. Rather, Gag possesses an affinity for LysRS, whose natural targeting of $tRNA^{Lys}$ is responsible for binding the tRNA, while the resulting Gag-LysRS-$tRNA^{Lys}$ complex is unstable in the absence of Gag-Pol within the viral particle (Mak et al. 1994; Khorchid et al. 2000; Javanbakht et al. 2003; Pavon-Eternod et al. 2010). There is no experimental support for specificity of HIV proteins for any amino-acyl RS (aaRS) other than LysRS (Halwani et al. 2004). Moreover, even given specificity by HIV-1 proteins for other aaRS, it is unclear how rare tRNAs would be enriched over the more common tRNAs, as seen by the enrichment of $tRNA^{Ile/UAU}$. Both forms of $tRNA^{Lys}$ are incorporated into HIV-1 as a consequence of the affinity of Gag for LysRS (Pavon-Eternod et al. 2010). Moreover, without Gag-Pol very low

concentrations of tRNAs, lysyl and non-lysyl, are packaged into HIV-1 particles (Pavon-Eternod et al. 2010), emphasizing a general role for Gag-Pol in tRNA packaging with or without packaging of an aaRS.

Here, we further explore the hypothesis that non-lysyl tRNAs packaged by HIV-1 aid the translation of HIV-1 genes, overcoming their poor codon adaptation. By examining the codon usage of HIV-1 genes and the tRNAs packaged into HIV-1 particles, we demonstrate that tRNAs enriched in HIV-1 particles are those that read unfavorable codons. We subsequently show that early and late HIV-1 genes display distinct codon usage patterns, such that early genes adhere to their host better than late genes. Multiple subtypes of HIV-1 corroborate this difference between early and late genes. We discuss hypotheses that may explain the enrichment of rare tRNAs and the temporal differences in codon usage.

2.4 Methods

2.4.1 Sequences

The reference genome sequences for HIV-1 subtype B, HXB2 (NC_001802), Simian Immunodeficiency Virus (SIV, M58410), Human T-Lymphotropic Virus 1 (HTLV-1, NC_001436) were downloaded from the National Centre for Biotechnology Information (NCBI). Reference sequences for HIV-1 subtypes A1 (AB253421), A2 (AF286237), C (U52953) and J (EF614151) were downloaded from the Los Alamos National Laboratory database (www.hiv.lanl.gov).

The genes encoded within HIV-1 are expressed in a temporal manner: early genes are *tat*, *rev* and *nef* and late genes are *gag*, *pol*, *env*, *vif*, *vpu* and *vpr* (Cullen 1991a). Similarly, in SIV early genes are *rex* and *tax*, and late genes are *gag*, *pol* and *pro* (Yoshida 2005), and in HTLV-1 early genes are *tat*, *rev* and *nef* and late genes are *gag*, *pol*, *env*, *vif* and *vpx*.

Within HIV-1, a single mRNA produces both Gag and Gag-Pol proteins; the Gag-Pol fusion is produced via a ribosome frame-shift, a change of -1 in the reading frame, as caused by the sequence UUUUUUA, located a position 1631 within the HIV-1 HXB2 reference sequence. Though the polymerase (Pol) is never translated independently, to eliminate dependencies in the data due to the common N-terminal sequences of Gag and Gag-Pol, Pol was considered as a separate protein that began at the end of the Gag sequence (starting at position 1637 within the HXB2 genome). Gag virus-like particles (GagVLP) are viral particles produced from full-length HIV-1 genomes lacking this frame-shift sequence, thereby unable to translate the Gag-Pol fusion protein.

2.4.2 tRNA microarray data

Microarray data (kindly provided by Dr Tao Pan) of tRNA incorporation into HIV-1 particles as from Pavon-Eternod *et al.* (Pavon-Eternod et al. 2010) was re-examined in this paper. Briefly summarizing the experimental design followed to collect the microarray data (Pavon-Eternod et al. 2010), HEK293T cells were transfected with full length HIV-1 or GagVLP genomes. After 48h, virus-containing supernatants were collected and purified. Total RNA was extracted from the collected HIV-1 and GagVLP particles, as well as from uninfected HEK293T cells. The tRNA content within extracted RNA from these three sources was subsequently quantified using custom designed tRNA microarrays, with probes for human tRNA genes (including nuclear and mitochondrial tRNA sequences). tRNA microarrays are limited in their ability to distinguish perfectly complementary probes from interactions with up to 8 mismatches (Dittmar, Goodenbour, and Pan 2006; Coughlin et al. 2009). Thus, it is common that probes bind and quantify heterogeneous groups of sequences rather than single tRNA species, including sequences from the same amino acid family with different anticodons.

2.4.3 Nucleotide composition

The frequencies of A residues at the first and second codon position (A_{12}) were calculated in DAMBE (Xia and Xie 2001).

2.4.4 Measures of codon usage bias

Relative synonymous codon usage (RSCU) values for all HIV-1 protein coding sequences were calculated in DAMBE (Xia 2001; Xia and Xie 2001). RSCU is a normalized index of codon usage bias within synonymous codon families, calculated by equation (2.1) (Sharp, Tuohy, and Mosurski 1986).

$$RSCU_{ij} = \frac{CodFreq_{ij}}{\frac{\left(\sum_{j=1}^{NumCodon_i} CodFreq_{ij}\right)}{NumCodon_i}} \qquad (2.1)$$

Thus, for a family containing *j* family members, the RSCU values within a synonymous family sum to *j*. RSCU values for individual codons are 0 for unused codons, 1 for unbiased codon usage, up to a maximum of *j* for codons that are used 100% of the time. As codon families with a single member (Met and Trp in the standard genetic code) always carry an RSCU value of 1, these were excluded from analyses. Stop codons were also excluded from our analysis; because of their rarity (with only 1 per sequence), the data is unreliable. To avoid noisy data as from other rare coding codons, any codons that were used less than 14 times were removed. A cut-off of 14 was chosen so as to avoid removal of the Ile family in early genes (as it is of particular interest).

The codon adaptation index (CAI) (Sharp and Li 1987) for all HIV-1 genes was calculated in DAMBE (Xia 2001; Xia and Xie 2001; Xia 2007). CAI is a measure of the codon usage of an entire gene sequence; it compares the codon usage frequencies of a gene to that of a reference set of known highly or constitutively expressed genes. For this, we used the data from E_human.cut, a file

originally from the EMBOSS package (Rice, Longden, and Bleasby 2000) and available through DAMBE. Such highly expressed genes are known to be under strong selection from the tRNA pool and are therefore likely to be composed of efficiently translated codons. The codon usage data from E_human.cut was also used to calculate RSCU values for codons from humans for comparison with viral codon usage.

2.4.5 Defining indices for tRNA enrichment into HIV-1 and viral codon adaptation

We define two indices, $I_{codon.i}$ (2.2) and $I_{tRNA.i}$ (2.3) to assess the level of adaptation of HIV-1 codon usage to its host and the level of tRNA enrichment during packaging into HIV-1 particles:

$$I_{codon.i} = \log_2\left(\frac{RSCU_{HIV.i}}{RSCU_{Human.i}}\right) \qquad (2.2)$$

$$I_{tRNA.i} = \log_2\left(\frac{tRNA_{HIV.i}}{tRNA_{GagVLP.i}}\right) \qquad (2.3)$$

Both measures share an index, *i*, which links the two indices: *i* refers to a codon in equation (2.2) and its complementary anticodon in equation (2.3). When multiple anticodons are indistinguishable by the tRNA microarray, then the corresponding $I_{codon.i}$ is the mean RSCU values for HIV and host genes of the corresponding codons.

These indices were defined such that increasing values of $I_{codon.i}$ imply a stronger preference for a codon *i* within HIV-1 genes as compared to within the reference set of human genes. Similarly, with increasing values of $I_{tRNA.i}$, a tRNA is enriched to a greater extent in HIV-1 particles relative to GagVLP particles. Here, GagVLP is considered to non-selectively package tRNAs, since GagVLP tRNA content closely reflected the tRNA content of uninfected HEK293T cells (Pavon-Eternod et al. 2010). The values $tRNA_{HIV.i}$ and $tRNA_{GagVLP.i}$ are each ratios of tRNAs packaged into each viral particle type, normalized to the levels within the cell, *i.e.* $tRNA_{HIV.i} = tRNA_{HIV\_q.i}/tRNA_{HEK293T.i,}$ where $tRNA_{HIV\_q.i}$ values are those quantified from the microarray experiments. Comparing the

packaging of tRNAs into HIV-1 to GagVLP rather than to cellular tRNA content avoids introducing bias that would arise from comparing the tRNA content from total RNA of different types of sources, from cell-free viral particles (HIV-1 or GagVLP) or cells.

2.4.6 Statistical Analysis

All statistical analysis was performed in excel or SAS (SAS Institute Inc. 1989). Non-parametric tests or ranking of data were used when the assumptions of parametric testing, those of linearity and normality, were violated. Tests of significance were two-tailed.

2.5 Results

2.5.1 Unfavorable HIV-1 codons are mostly A-ending

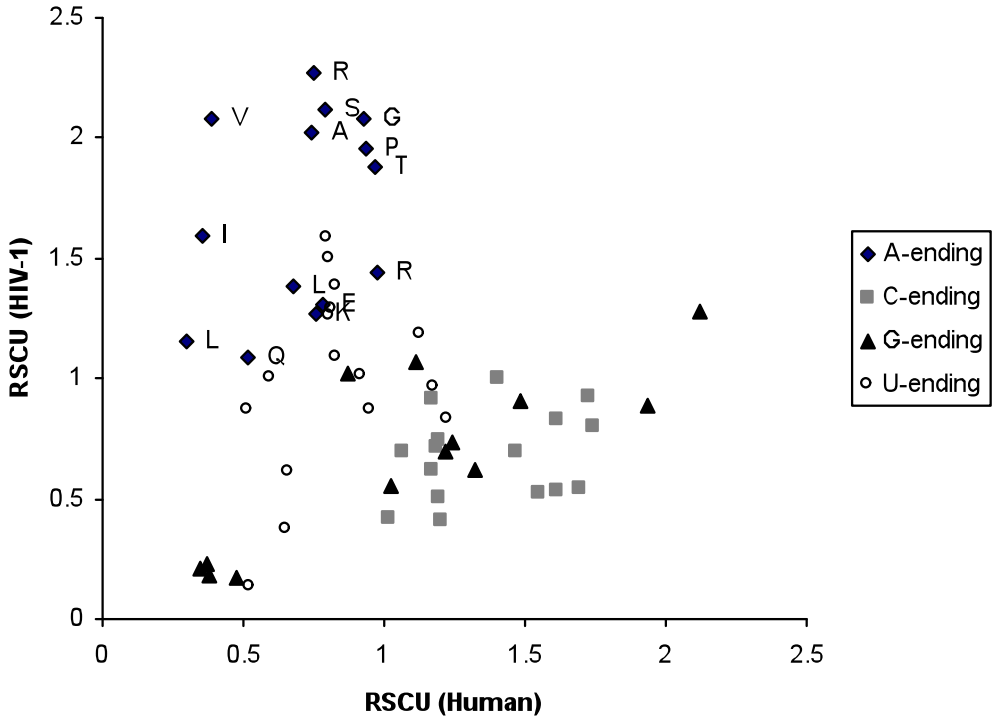
Viral codons are poorly adapted when they are found at frequencies different from those in host genes. Because codons avoided by the host implies a lack of available tRNAs, this discrepancy is unfavorable for the virus when codons are more frequent in viral genes than in host genes. To identify particularly unfavorable codons, we plotted HIV-1 RSCU values against human RSCU values (Figure 2.1). The codon usage of HIV-1 genes correlated extremely poorly with that of its host (Pearson r = -0.1470, p = 0.2665; Spearman r = 0.1829, p = 0.1657). Overall, this poor correlation can be attributed to inflated usage of A-ending codons and reduced usage of G and C ending codons in HIV-1 genes. Thus, in agreement with the A-biased mutation of this virus, it is A-ending codons that are particularly unfavorable.

2.5.2 Selectively packaged tRNAs decode the unfavorable codons of HIV-1

To investigate the hypothesis that tRNA packaging in HIV-1 benefits the translation of unfavorable codons, we first assessed agreement between the decoding capacity of the tRNAs packaged into HIV-1 particles and viral codon usage. As detailed in the methods section, indices for

packaging (I_{tRNA}) and codon usage (I_{codon}) were defined such that large values of each represent tRNAs packaged highly in HIV-1 and unfavorable codons, respectively. Thus, if tRNAs are packaged to help translation, we would expect a positive correlation between I_{tRNA} and I_{codon}.

Figure 2.1 – Relative synonymous codon usage (RSCU) of HIV-1 compared to RSCU of highly expressed human genes. Data points for codons ending with A, C, G or U are annotated with different combinations of colors and symbols. A-ending codons exhibit strong discordance in their usage between HIV-1 and human and are annotated with their coded amino acids.



By microarray, tRNA sequences with low diversity (<8 nucleotide differences) cannot be distinguished from one another (Dittmar, Goodenbour, and Pan 2006; Coughlin et al. 2009). As tRNA genes in the same family show high sequence similarity, many probes were complimentary to sequences with different anticodons; this heterogeneity limits the types of questions that may be asked from the I_{tRNA} data. With A-ending codons already identified as particularly unfavorable (Figure 2.1) and knowing already that tRNAs decoding A-ending codons are rare within host cells,

these were the particular data points of interest. As such, we chose families for which there was specific I_{tRNA} data for anticodons that only read A-ending codons. Within the available data, this was possible for 7 of the 20 amino acid families: Arg, Ile, Leu, Lys, Gly, Val and Thr. For these 7 families, pairs of I_{codon} and I_{tRNA} values were calculated (Table 2.1). For 5 of these 7 families (all but Gly and Thr), the packaging of the rare tRNA that read A-ending codons exceeded its family member(s). When ranked, I_{codon} and I_{tRNA} values were significantly and positively correlated (Pearson $r = 0.5780$, $p = 0.0304$). As both the A-ending and non-A-ending $tRNA^{Lys}$ are packaged to very high levels to serve stages of the HIV-1 lifecycle other than translation, we removed these data to avoid obscuring any signal related to translation. The correlation between ranked I_{tRNA} and I_{codon} values was improved in the absence of $tRNA^{Lys}$ data (Pearson $r = 0.6853$, $p = 0.0139$).

Table 2.1 – Relationship between codon usage measured by RSCU for human and HIV-1 ($RSCU_{Hum}$ and $RSCU_{HIV}$) and packaging of host tRNA by HIV-1. tRNA abundance data were kindly provided by Dr. Tao Pan. Rank(I_{codon}) and Rank(I_{tRNA}) are significantly and positively correlated ($r = 0.5780$, $p = 0.0304$).

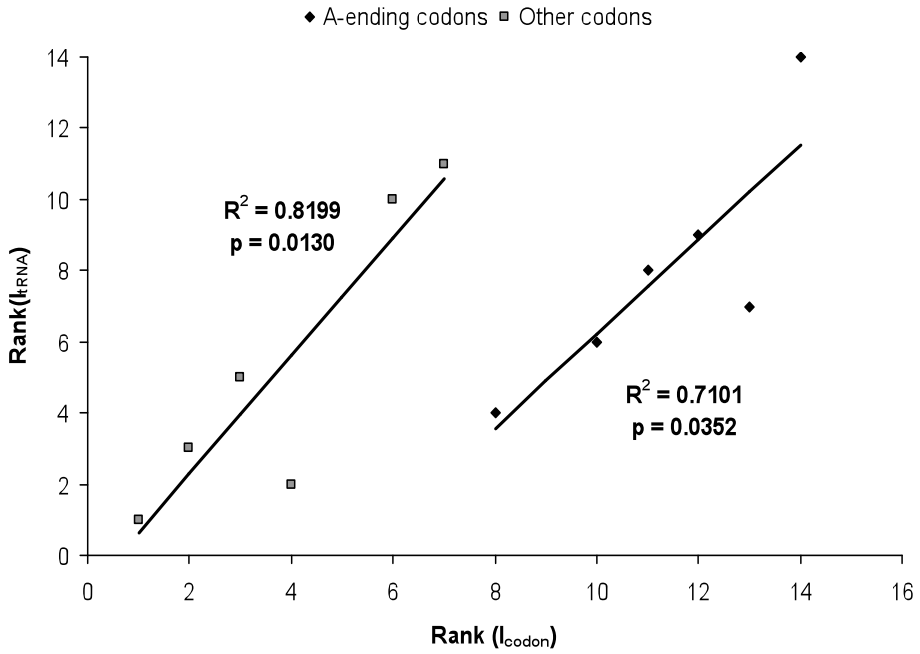
Codon	$RSCU_{Hum}$	$RSCU_{HIV}$	Rank(I_{codon})	$tRNA_{HIV}$ [1]	$tRNA_{GagVLP}$ [2]	Rank(I_{tRNA})
Arg(AGA)	0.97	1.44	8	0.049	0.028	4
Arg(AGG)	1.03	0.56	4	0.066	0.054	2
Ile(AUA)	0.24	1.59	14	1.340	0.061	14
Ile(AUY)	2.64	1.41	3	0.267	0.137	5
Leu(UUA)	0.68	1.38	11	0.090	0.037	8
Leu(UUG)	1.32	0.62	1	0.045	0.050	1
Lys(AAA)	0.76	1.27	9	0.641	0.034	13
Lys(AAG)	1.24	0.73	5	1.008	0.060	12
Gly(GGA)	0.93	2.08	12	0.071	0.029	9
Gly(GGB)	3.07	1.92	6	0.202	0.049	10
Val(GUA)	0.39	2.08	13	0.066	0.028	7
Val(GUB)	3.61	1.92	2	0.074	0.049	3
Thr(ACA)	0.97	1.94	10	0.048	0.022	6
Thr(ACB)	3.03	2.06	7	0.252	0.035	11

[1] $tRNA_{HIV}$: the relative tRNA abundance of HIV-1 virion versus human HEK293T cells

[2] $tRNA_{GagVLP}$: the relative tRNA abundance of Gag viral-like particles (GagVLP) versus human HEK293T cells.

Moreover, as A-biased mutation is dramatically altering the codon usage measure I_{codon}, the data may be heterogeneous. In order to account for the effect of the A-bias, we split the data into two groups: A-ending codons and non A-ending codons (Figure 2.2). Significant positive correlations between I_{codon} and I_{tRNA} were observed for both groups (A-ending, $p = 0.0352$; non A-ending, $p = 0.0130$). The correlations in both groups suggest that all codons, not simply A-ending codons, may be benefiting from tRNA packaging.

Figure 2.2 – Relationships between ranked I_{codon} and I_{tRNA} values for A-ending and non-A-ending codons. Correlation lines are plotted with R^2 values to indicate the goodness of fit. I_{codon} and I_{tRNA} values for lysine were not included because $tRNA^{Lys}$ is enriched for RT initiation rather than for translation.



2.5.3 Differential codon usage of early and late genes across HIV-1 subtypes and SIV

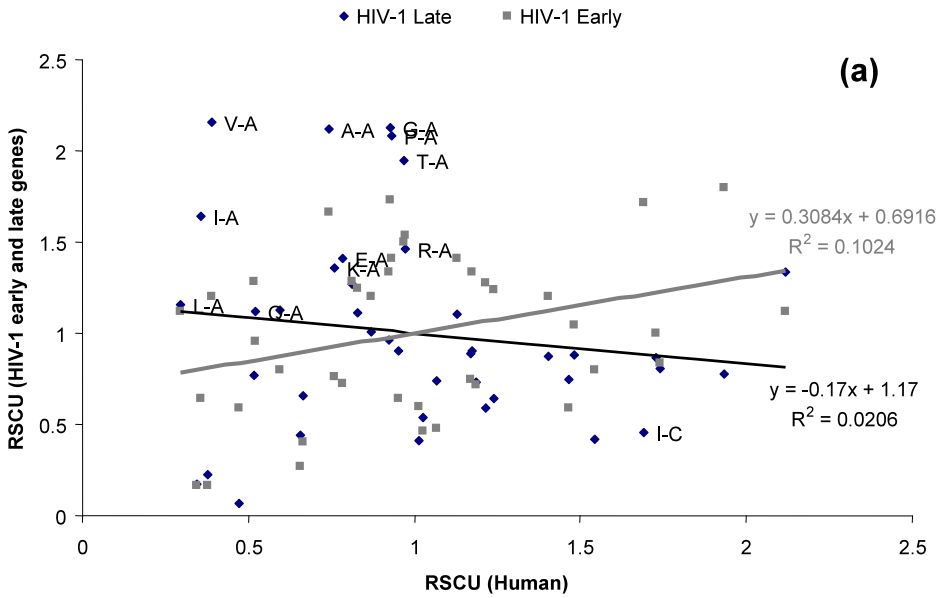
Herein, we have employed vernacular implying tRNA packaging itself to be beneficial for the codon usage of HIV-1. Clearly, translation occurs within cells, using the tRNA pool available to it, not within HIV-1 particles. For any mechanism to benefit the translation of HIV-1 proteins, rare tRNAs must then be enriched in the cytoplasm at the time of translation. We propose two possible

hypotheses. First, tRNAs may be enriched in HIV-1 particles, and delivered into the newly infected cell, to make the new cell more favorable for viral infection. Second, tRNAs in HIV-1 particles may reflect an altered tRNA pool, induced during viral infection, that is passively packaged at the time of particle assembly. As the former hypothesis is actively enriching and delivering tRNAs, while the latter is passively packaging the already altered tRNA pool, we name them H_{Active} and $H_{Passive}$.

Both hypotheses suggest alterations to the tRNA pool. Genes translated by the altered tRNA pool would then be adapting to the altered tRNA pool, rather than the typical host tRNA pool. If there are times during infection when this tRNA pool is present and times when it is not, we might expect to see differences in the codon usage between different HIV-1 genes. As by $H_{Passive}$, an altered tRNA pool must be present during the late stages of infection, since this is time of particle production. If the altered tRNA pool is not immediately induced, we may expect temporal differences in the codon usage of HIV-1 genes. By H_{Active}, tRNAs delivered at viral entry into the cell would be available during the entirety of infection or, if the delivered tRNAs had a transient effect, only at the beginning of infection. HIV-1 gene expression occurs in two major phases: early and late. Thus overall, $H_{Passive}$ would predict either all genes or only late genes to experience the altered tRNA pool, while H_{Active} would predict all genes or only early genes to experience the altered tRNA pool.

To investigate temporal differences in codon usage within HIV-1 genes, RSCU values of early and late gene codon usage were plotted against host RSCU (Figure 2.3a). Compared to the RSCU for total HIV-1 genes, the RSCU values of early genes correlated surprisingly well with host genes (Pearson $r = 0.32$, $p = 0.0413$; Spearman $r = 0.2776$, $p = 0.0789$). In contrast, HIV-1 late genes had a negative, although insignificant, correlation with the host proteins (Pearson $r = -0.01435$, $p = 0.3704$; Spearman $r = -0.1847$, $p = 0.2477$). In agreement, using CAI as a measure of codon usage, a significant difference between early and late genes was also observed (Table 2.2; two-sample t-test assuming unequal variance, $t = 2.8099$, $df = 4$, $p = 0.004832$).

Figure 2.3 – Relative synonymous codon usage (RSCU) of HIV-1 (a) and HTLV-1 (b) compared to RSCU of highly expressed human genes, with early (grey) and late (dark blue) genes shown with different colors. Linear regression lines are plotted with the fitted equations, R^2 and p values, where significant, are indicated.



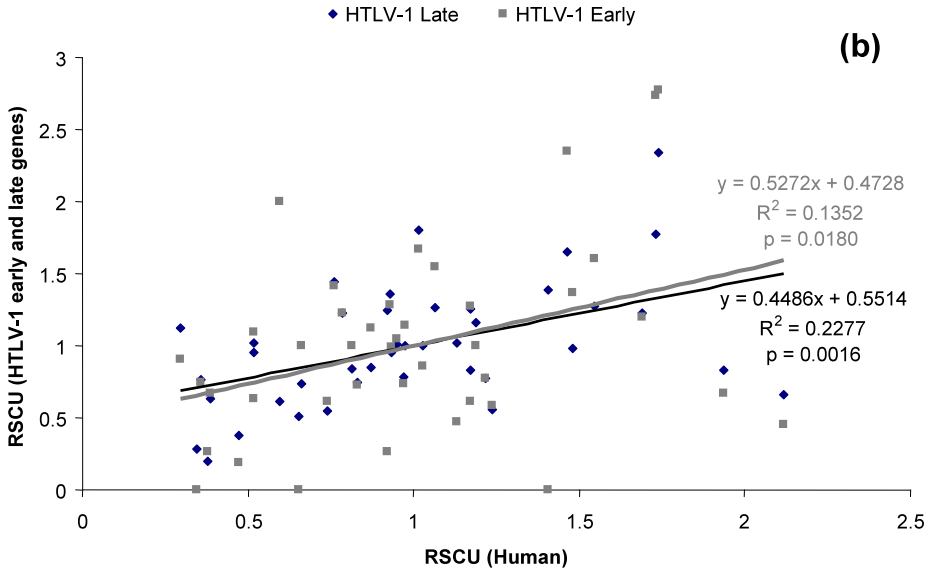


Table 2.2 – Gene length and codon adaptation index (CAI) values for the 3 early (*tat*, *rev* and *nef*) and 5 late (*gag-pol*, *vif*, *vpu*, *vpr*, and *env*) coding sequences (CDS) of HIV-1. Early genes had significantly higher CAI values than late genes (two sample t-test assuming unequal variance, t = 2.8099, df = 4, p = 0.004832).

Gene	CDS (bp)	CAI
EARLY		
tat	261	0.669
rev	351	0.662
nef	621	0.675
LATE		
gag	1503	0.628
pol	3012	0.581
vif	579	0.619
vpr	291	0.643
vpu	249	0.491
env	2571	0.619

To ensure that this was not a chance result from the HXB2 subtype, we analyzed multiple subtypes (A1, A2, C and J) of HIV-1, as well as SIV. All the subtypes of HIV-1 corroborated the result for subtype B, we observed a consistent difference between the codon usages of early and late HIV-1 genes, with early genes correlating better with host codon usage (data not shown). SIV also shared the difference between its early and late genes. Thus, these temporal differences support $H_{Passive}$.

2.5.4 The temporal differences are not explicable by differences in mutation rates

There does exist a competing, alternative hypothesis, unrelated to tRNA availability that must be considered. A higher mutation rate in late genes than early genes would similarly predict that the codon usage of late genes to then be worse than that of early. Though studies on the HIV-1 reverse-transcriptase demonstrated that mutations are induced randomly across the full genome (Vartanian, Henry, and Wain-Hobson 2002), these were *in vitro* studies. Furthermore, this does not include mutations resulting from the action of the host protein APOBEC3 or other selection

pressures, which could be non-random across HIV-1 genes. Thus, in absence of comprehensive data on the mutational rates across the different HIV-1 genes, we examined the nucleotide composition of early and late genes to distinguish translational selection from mutation rate. Translational selection has been shown to primarily influence the composition of the third codon position, while an increased mutation rate influences the composition at all codon positions (Sueoka 1961; Jenkins and Holmes 2003; Lobry 2004). We compared the prevalence of A nucleotides at the first and second position (A_{12}) between early and late genes of HIV-1. If the mutation bias were stronger in late genes, A_{12} would also be higher in late genes. However, no significant differences in A_{12} composition between early and late genes were found. The early gene *tat* (A_{12} = 0.3276) even had higher composition than the late genes *vpr* (A_{12} = 0.3144) and *vpu* (A_{12} = 0.3133), albeit only slightly. Therefore, we are lead to believe that the differences between early and late genes at the third codon position (Figure 2.3a) are not caused by an increased mutation rate in late genes.

2.5.5 HTLV-1 does not share the difference in early and late genes

Interestingly HTLV-1, a retrovirus which like HIV-1 infects CD4+ T cells (Rimsky et al. 1988), does not demonstrate a difference in codon adaptation between its early and late genes (Figure 2.3b). Moreover, HTLV-1, early and late genes combined, and human RSCU values correlate significantly and positively (Pearson r = 0.4982, $p < 0.0001$; Spearman r = 0.4688, p = 0.0002). The overall differences between HIV-1 and HTLV-1 can be explained by considering the differences in lifecycle of these two retroviruses. HTLV-1 does not have a lytic replicative lifecycle like HIV-1, but rather follows a latent lifecycle, remaining integrated in the genome of its host cell. This replicative process is called clonal expansion, as daughters of infected cells are infected, increasing the number of infected cells in the host. The majority of viral expansion is then accomplished by the host polymerase rather than the error-prone reverse transcriptase. Thus, HTLV-1 does not share the strong

A-biased mutation rate of HIV-1, but like its host, prefers C and G-ending codons (van Hemert and Berkhout 1995; Van Dooren et al. 2004). In addition, the mutation rate of HTLV-1, estimated to be $5.2*10^{-6}$ substitutions/site/year (Hanada, Suzuki, and Gojobori 2004; Van Dooren et al. 2004), is much lower than that of HIV-1, estimated to be a formidable $2.5*10^{-3}$ substitutions/site/year (Hanada, Suzuki, and Gojobori 2004).

Furthermore, when taking into account the differences between early and late genes in HIV-1, there are three classes of codon adaptation between HTLV-1 and HIV-1 genes. Ordered from best to worst: first are HTLV-1 early and late genes, then HIV-1 early genes and lastly HIV-1 late genes. The strong A-biased mutation could explain the difference between the early genes of HIV-1 and genes from HTLV-1. Whereas, the even worse codon usage of HIV-1 late genes could be explained as a result of A-biased mutation along with a more favorable, altered tRNA pool, which does not as strongly discourage the use of A-ending codons.

2.6 Discussion

tRNAs that read unfavorable codons were found to be enriched in the tRNA population packaged into HIV-1 particles. This relationship suggests that the translation of unfavorable codons might be helped in some way by alterations to the tRNA pool. We subsequently found evidence for temporal differences between codon usage of early and late genes, such that early genes correlate better with host codon usage than late genes. The differences between early and late codon usage are shared across different subtypes of HIV-1 and its near relative SIV, but not by HTLV-1, another retrovirus that does not share the A-biased mutation of HIV-1.

Unfortunately, in this study we could only examine the data for 7 amino acid families including Lysine, less than half of the total 20 families. The available tRNA microarray data did not allow for more families to be included, as different anticodons were indistinguishable due to high

sequence similarity between tRNA family members. When methods with higher sensitivity are available, these results should be corroborated in the remaining 13 families. An additional simplification made to complete this study was to only account for Watson-Crick pairing as codon-anticodons interactions, no wobble pairing was considered. Nevertheless, for the purpose of this study, we were primarily looking at A-ending versus non A-ending codons. Within the majority of families, wU anticodons read A-ending codons with very little wobble capacity for other codons while the anticodons that wobble more freely (wG or wI) were investigated as a group.

It is particularly surprising that late structural proteins have such poor codon usage, worse than early genes. The opposite trend was found examining the codon usage of bacteriophages (Lucks et al. 2008) and among adenoviruses (Das, Paul, and Dutta 2006): higher translational selection was observed in the genes encoding structural proteins. This is likely due to their high expression levels, as demanded for particle formation and key to viral success. This conflicting result however, would be assuaged by the presence of an altered tRNA pool, more favorable for translation of these late genes.

We have previously proposed two hypotheses to explain the correlation between tRNA packaging and viral codon usage. With H_{Active}, we proposed the packaging of non-lysyl tRNA by HIV-1 to be selective, actively enriching rare tRNAs to deliver tRNAs into newly infected cells. With $H_{Passive}$, we proposed that tRNA packaging was simply a reflection of an altered tRNA pool at the time of packaging. As we previously argued, the temporal differences we observed between early and late genes support $H_{Passive}$, since the Delivery hypothesis would not make rare tRNAs available to only late genes. In addition, there are multiple arguments that can be made against H_{Active}. First, as discussed in the introduction, enrichment during packaging would require HIV-1 proteins to have specific affinities for tRNAs other than $tRNA^{Lys}$, for which there is no supporting biochemical data (Halwani et al. 2004). Moreover, it is unknown how rare tRNAs would be enriched seeing as an

affinity for the aaRS would result in the packaging of all tRNAs in a family, as observed for $tRNA^{Lys}$. Second, so few tRNAs are packaged into viral particles that they are likely insufficient to alter the tRNA pool in any meaningful way (Huang et al. 1994; Halwani et al. 2004). Lastly, though tRNA packaging is well suited for the delivery of $tRNA^{Lys}$ for reverse-transcription, being the first stage of the viral lifecycle, translation is a late event in viral infection. With the delay between cell entry and translation, any tRNAs brought into the host cell may no longer be present by the time of translation, particularly if the virus enters latency (Saksela et al. 1993). Thus, we reject H_{Active}. Contrastingly, we find few arguments against the altered tRNA pool induced by $H_{Passive}$. Moreover, $H_{Passive}$ satisfactorily addresses the question of how non-lysyl tRNAs could be enriched in the absence of specific affinities for the tRNAs or their respective aaRS. Yet, how would an altered pool prior to packaging be induced?

It is unknown whether changes to the tRNA pool proposed by $H_{Passive}$ are actively induced by viral proteins or by the host in the response to viral infection. There are no available examples of viral proteins having such a function. Nevertheless, viral infection induces multiple changes within the host cell. Many of these pathways, including the innate anti-viral response and the unfolded protein response (UPR), cause shutdown of cap-dependent translation by phosphorylation of the eukaryotic initiation factor eIF-2α (DuRose et al. 2009; Goujon and Malim 2010). Interestingly, HIV-1 mRNAs carry internal ribosome entry sites (IRES), thus their translation is uninhibited by shutdown of cap-dependent translation via eIF-2α phosphorylation (Yilmaz, Bolinger, and Boris-Lawrie 2006). The presence of IRES elements in HIV-1 supports that host-shutdown occurs at some point during infection. Though HIV-1 has been shown to block the innate anti-viral response (Goujon and Malim 2010), it is unknown whether a UPR is induced during HIV-1 infection. Nevertheless, UPR is known to be activated by many viruses, including Hepatitis C (Chan and Egan 2009), SARS (Minakshi et al. 2009), Japanese Encephalitis Virus (Su, Liao, and Lin 2002) and Coxsackie B2

(Zhang et al. 2010). Drastic reductions in translation of host proteins would likely affect the tRNA pool, as demand upon tRNAs is drastically altered, possibly resulting in the altered tRNA pool packaged into HIV-1 particles. Moreover, UPR is activated in response to the production of highly expressed viral proteins, especially structural glycoproteins (Kaufman 2002). Hence, this response would likely be activated late in infection, in agreement with the temporal differences between early and late genes. Moreover, if it is the UPR pathways that are causing the altered tRNA pool, the same pool may be available to other viruses as well. A-biased codon usage is common to many RNA viruses, which have high mutation rates. Perhaps the altered tRNA pool explains why the codon usage of RNA viruses may be so poorly adapted with little consequence to completing their lifecycle.

Suggesting an altered tRNA pool draws into question previous experimental evidence demonstrating the codon usage of HIV-1 to be poor. Such experiments compared protein production from transient transfection of wild type and codon optimized messages (Haas, Park, and Seed 1996). Consequently, the cell in which these proteins are being translated is an uninfected cell, which would be in absence of any alterations to the tRNA pool during infection. These estimates of translational efficiency may then be under-estimates, as the altered tRNA pool we are suggesting would be more favorable for HIV-1 translation.

Overall, our results highlight that the general perception of the tRNA pool as a stagnant entity is perhaps too simplistic. It is possible that changes in the tRNA pool are not infrequent, but have yet to be fully investigated. The current availability of large-scale methods, such as the tRNA microarray, are well suited for the preliminary study of tRNA pools across various conditions and shed light on the dynamics of the tRNA pool. Although these data would still be limited by the inability to distinguish certain tRNA species by tRNA microarray; new, more sensitive methods are needed to further investigate the dynamics of tRNA pools.

In summary, we have formulated that HIV-1 infection causes alterations in the tRNA pool during infection, potentially via induction of the host UPR pathway. Along with further corroboration of our results with a comprehensive dataset of tRNA packaging, we suggest two follow-up experiments to directly test this hypothesis. First, we suggest to test whether there are detectable difference between the tRNA pools of infected and uninfected cells and second, to test whether this, different tRNA pool is shared when the UPR response alone is induced.

Currently, no anti-HIV therapies target translation (de Clercq 2007). A reduced ability to produce structural genes could rein in particle production during HIV-1 infection. Lower viral titres could have the result of not only slowing the progress of infection, but, as smaller viral populations within a host could also reduce the degree of total genetic variation, this might curtail the development of drug resistance or immune evasion. Moreover, high viral titres increase the risk of transmitting HIV-1 to a new host. With HIV-1 ever in need of new drug therapies, a better understanding of the translation of HIV-1 genes might explore new avenues to control this proficient pathogen.

3 Cleaning up the clutter: removal of falsely predicted tRNA genes with phylogenetics

3.1 Abstract

Although tRNA prediction algorithms quite successfully identify true tRNA genes, they are nevertheless prone to falsely identifying tRNA-derived repetitive sequences as true tRNA genes. As tRNA-derived repetitive sequences are freely accruing mutations, they may clutter numerous tRNA families by mutating within the anticodon. Moreover, repetitive sequences are found in very high numbers in the larger eukaryotic genomes. Therefore, they significantly increase the number of tRNA genes predicted, creating noisy datasets. We aimed to address the problem of false positives in tRNA datasets, and proposed that phylogenetics may discern the false positives by their inflated evolutionary rates and heterogeneous clustering patterns. Putting our method to the test, we demonstrate that known repetitive elements are distinguishable by phylogenetics that employed methods appropriate for structured, non-coding RNA sequences. Repetitive sequences either clustered together, forming clusters where all members had longer branch lengths compared to their functional counterparts, or had exceptionally long branches. Thus, phylogenetics is capable of

highlighting repetitive sequences in tRNA datasets and may be applied to filter out these troublesome falsities.

3.2 Introduction

Although the algorithm employed by the gtRNAdb (Chan and Lowe 2009), tRNAscan-SE, has a very low false positive rate against random sequences (<0.00007/Mbp), tRNA prediction is known to be sensitive to the non-random background of tRNA-derived repetitive sequences such as Short INterspersed Elements (SINEs) (Lowe and Eddy 1997). Repetitive sequences are found in particularly high numbers across the genomes of multicellular eukaryotes (Kramerov and Vasetskii 2009), retaining sufficient similarity to their progenitor tRNA to be erroneously predicted as tRNA genes. For example, in *Canis familiaris*, 440 of 898 predicted tRNA genes are derived from tRNA Lys, and are acknowledged to be caused by repetitive sequences (Tang et al. 2009). Furthermore, being non-functional, these repetitive sequences may accumulate mutations at the anticodon position and clutter other families. These noisy tRNA datasets are problematic as they reduce the ability to make clear conclusions.

How then may false positives be distinguished from true tRNA genes? Clearly, experimental validation of tRNA gene sets would be ideal. One commonly used large-scale method is the tRNA microarray, though this method is limited by its inability to discriminate between complete and partial hybridization, until sequences have more than 8 nucleotide differences to a probe (Dittmar, Goodenbour, and Pan 2006; Coughlin et al. 2009). Thus, tRNA microarrays are only capable of characterizing the expression levels of groups of related tRNA genes rather than single genes. To distinguish single genes, 'small-scale' experimentation methods must be employed, such as Northern blotting (Coughlin et al. 2009). But, as eukaryotic species are predicted to have hundreds, sometimes even thousands of tRNAs, only large-scale methods are feasible. Moreover, with the potential of

differential expression of tRNAs between tissues (Dittmar, Goodenbour, and Pan 2006), the number of samples requiring experimental validation is dramatically increased. Currently, tRNA microarrays have only been performed for a few species, thus there is no experimentally validated, gold standard for datasets of multiple species.

Several computational approaches have been developed to eliminate false positives from predicted tRNA datasets. The program RepeatMasker (Smit, Hubley, and Green 1996-2010) is a standard computational method to minimize the number of repetitive sequences in predicted tRNA datasets (Coughlin et al. 2009; Tang et al. 2009; Rogers, Bergman, and Griffiths-Jones 2010). RepeatMasker compares input sequences to a compilation of known repetitive sequences from the database RepBase (Jurka et al. 2005); this method is therefore limited by the need for repetitive sequences in each species to have been previously annotated. An evolutionary approach, using comparative genomics to identify tRNA genes in *Bos taurus* has been previously used, as the *B. taurus* genome has a particularly high number of repetitive sequences with very limited annotation in RepBase (Tang et al. 2009). Tang *et al.* expected that functional tRNAs would share 95% identity with their orthologs, although this cut-off is extremely stringent as in reality this represents only 2 nucleotide changes in these short sequences. Furthermore, their alignment and phylogenetic methods were not appropriate for tRNA sequences for two reasons. First, it is crucial to incorporate secondary structure into RNA alignments to optimize the identification of homology (Kjer 1995; Hickson, Simon, and Perrey 2000; Xia 2000; Xia, Xie, and Kjer 2003; Hochsmann, Voss, and Giegerich 2004; Siebert and Backofen 2005). Second, compensatory mutations occur frequently in these sequences (Higgs 1998) and thus it is important to employ evolutionary models that can accommodate such substitution patterns (Savill, Hoyle, and Higgs 2001; Knies et al. 2008).

Here, we developed a novel systematic approach to remove false positives in eukaryotic tRNA datasets. Functional tRNA genes evolve at an extremely low rate, under functional constraint

to maintain both the cloverleaf structure and sequence elements. In contrast, the evolution of repetitive sequences is unhindered. Therefore, we proposed phylogenetic analysis as a method to distinguish tRNA genes from their freely evolving, non-functional counterparts based on their evolutionary rate. To better model the compensatory evolution of these tRNA genes, we employed both multiple sequence alignment and phylogenetic methods that consider secondary structure. Using RepeatMasker identified sequences as known repetitive sequences, we demonstrate that phylogenetics is indeed capable of distinguishing false positives.

3.3 Methods

3.3.1 Sequence datasets

tRNA sequences and predicted structures from *Bos taurus*, *Homo sapiens* and *Mus musculus* were downloaded from the Genomic tRNA Database (gtRNAdb) (Chan and Lowe 2009). As sequences and structures are output separately from the gtRNAdb, sequence and structure were combined with scripts written in MATLAB.

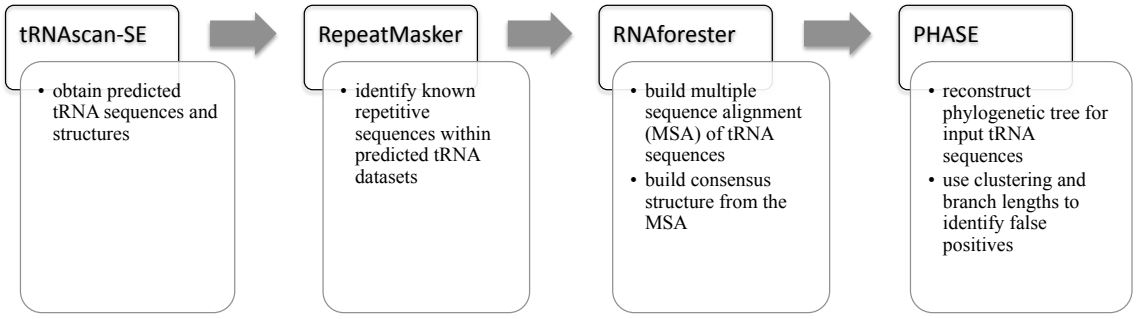
3.3.2 RepeatMasker

Although the genomes used by the gtRNAdb to run tRNAscan-SE are pre-screened by RepeatMasker, there are on-going updates to RepBase. Therefore, to remove any repetitive elements annotated since tRNAscan-SE was run, all downloaded tRNA genes were screened again by RepeatMasker. RepeatMasker (version 3.2.9) was run from the online server from the Institute for Systems Biology (http://www.repeatmasker.org/), using the settings *cross-match* and *slow* for high sensitivity and specificity, as per user instructions. For each run, species-specific options were chosen (Smit, Hubley, and Green 1996-2010). Sequences identified by RepeatMasker will serve as our true positives when testing whether phylogenetics may indeed identify repetitive sequences. Thus, sequences identified by RepeatMasker were not eliminated, but were highlighted for further analysis.

3.3.3 Phylogenetics

As structure is central to tRNA function, it is advised to analyze tRNA genes using methods that incorporate structural information. Thus, the following steps were taken to perform phylogenetics using multiple sequence alignment (MSA) and phylogenetics appropriate for RNA sequences (Figure 3.1).

Figure 3.1 – Flow diagram of steps required for the analysis of tRNA genes with our proposed phylogenetics approach.



RNAforester, downloaded as part of the RNA Vienna Package version 1.8.4, was used to build MSA and consensus structures. tRNA introns are found within the anticodon loop in some tRNA families, and are identified by tRNAscan-SE. Introns were cut out and pasted at the 5' end of the sequence, prior to the first stem, ensuring that the intron and anticodon loop were properly and separately aligned relying on the partitioning of paired and unpaired sites by RNAforester. Perl scripts were written to facilitate the building of sequence subsets for analysis and to change file formatting, linking different programs.

Bayesian phylogenetic analysis was performed using the program PHASE (PHylogenetics And Sequence Evolution); (Jow et al. 2002). In PHASE, we compared two different methods: paired and unpaired evolutionary models used together, and unpaired evolutionary models alone. When both paired and unpaired evolutionary models were employed, sites were partitioned, as per the generated consensus structures, and the parameters for two substitution models were simultaneously estimated

during MCMC (Markov Chain Monte Carlo): the GTR (general time reversible) + γ (gamma, estimated to assess rate heterogeneity across sites) 4x4 matrix was chosen for unpaired sites, and a 7x7 matrix + γ (7D in PHASE) for paired sites (Savill, Hoyle, and Higgs 2001). For the unpaired evolutionary model trial, GTR + γ was used across all sites. Due to the short length of tRNA sequences, we chose the simplest of the 7x7 evolutionary models (7D in PHASE) to avoid over-parameterization. Meanwhile, for the unpaired sequences, we chose the GTR model, which considers each nucleotide exchange independently. We wished to avoid models with simplifications designed for the evolution of protein coding sequences.

Phylogenetic analyses were run to completion in duplicate, each initiated randomly and each with different non-informative priors for rate ratios. The functions *mcmcphase* and *mcmcsummarize* in PHASE were used to complete the MCMC runs and reconstruct the consensus tree. To assess convergence, likelihood values sampled from the posterior distribution during MCMC and tree topologies of the run pairs were compared. The branch lengths of the consensus tree are the average of branch lengths sampled during the MCMC, therefore branch lengths were further optimized in the final tree topology by maximum likelihood, using *optimizer* in the PHASE package.

3.4 Results

3.4.1 RNA phylogenetics can distinguish known repetitive elements

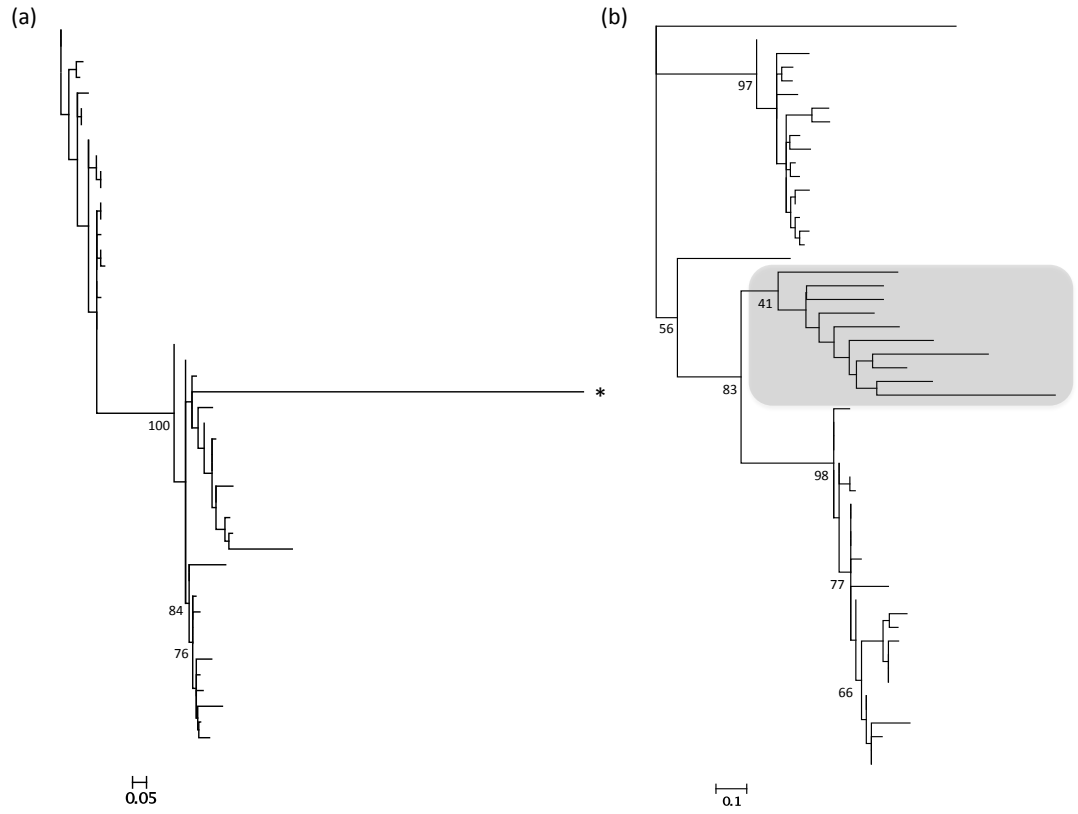
True tRNA genes are under strong constraint to maintain sequence elements and the conserved cloverleaf structure, both required for proper function. This includes carrying the tRNA identity sites, unique to each amino acid family, which dictate the aaRS and consequently the amino acid that tRNA will carry. Proper loading of tRNAs is crucial to maintain the specificity of protein synthesis. As tRNA identity includes numerous sites across the tRNA, not only the anticodon, tRNAs that encode the same amino acid are likely to be evolutionarily related. Conversely, non-functional

sequences, in particular falsely predicted from tRNA-derived repetitive sequences, are freely evolving, including within the anticodon. Thus, we proposed that phylogenetics might distinguish false positives from true tRNA genes. In particular, we expected true tRNA genes to cluster with their family members, while false positives would cluster with sequences from any amino acid family. Additionally, we predicted false positives would have a higher evolutionary rate, as determined by increased branch lengths.

To evaluate whether phylogenetics may distinguish known repetitive sequences from their family members, two preliminary test sets were examined: first a set of human $tRNA^{Ala}$ genes with one sequence identified by RepeatMasker (Figure 3.2a), and next a set of human $tRNA^{Gln}$ genes with ten masked sequences (Figure 3.2b). Sequences were aligned with RNAforester and examined by RNA phylogenetics in PHASE, as detailed above (Figure 3.1). Visually, the single $tRNA^{Ala}$ repetitive sequence was easily identified by its exceptionally long branch length, as compared to the other sequences (Figure 3.2a).

The RepeatMasked $tRNA^{Gln}$ sequences were also easily identified by phylogenetics: all ten sequences clustered together, with branch lengths within that cluster longer than those of the unmasked sequences (Figure 3.2b). The topology of these clusters matched our expectation of an increased evolutionary rate among freely evolving repetitive sequences as compared to their presumably functional counterparts. We call such long branched clusters *repetitive-like clusters*. It must be noted that while support values for the separation of the repetitive-like cluster were high (Figure 3.2b), posterior probability values for the interior branches of clusters were quite low. This is likely due to the short sequence length and low variability of these sequences, and thus an overall low phylogenetic signal. Despite this, independent runs converged. Thus, within these preliminary test sets, RepeatMasked sequences were identifiable by RNA phylogenetics.

Figure 3.2 – Testing the ability of phylogenetics to distinguish known repetitive sequences, as identified by RepeatMasker. In panel (a), test set with human $tRNA^{Ala}$ with only one known repetitive sequence, highlighted by a *; in panel (b), test set with human $tRNA^{Gln}$ genes including 10 repetitive sequences and a $tRNA^{Ala}$ outgroup. The cluster formed by the repetitive sequences is shaded in (b). Posterior probabilities are shown for major clusters.

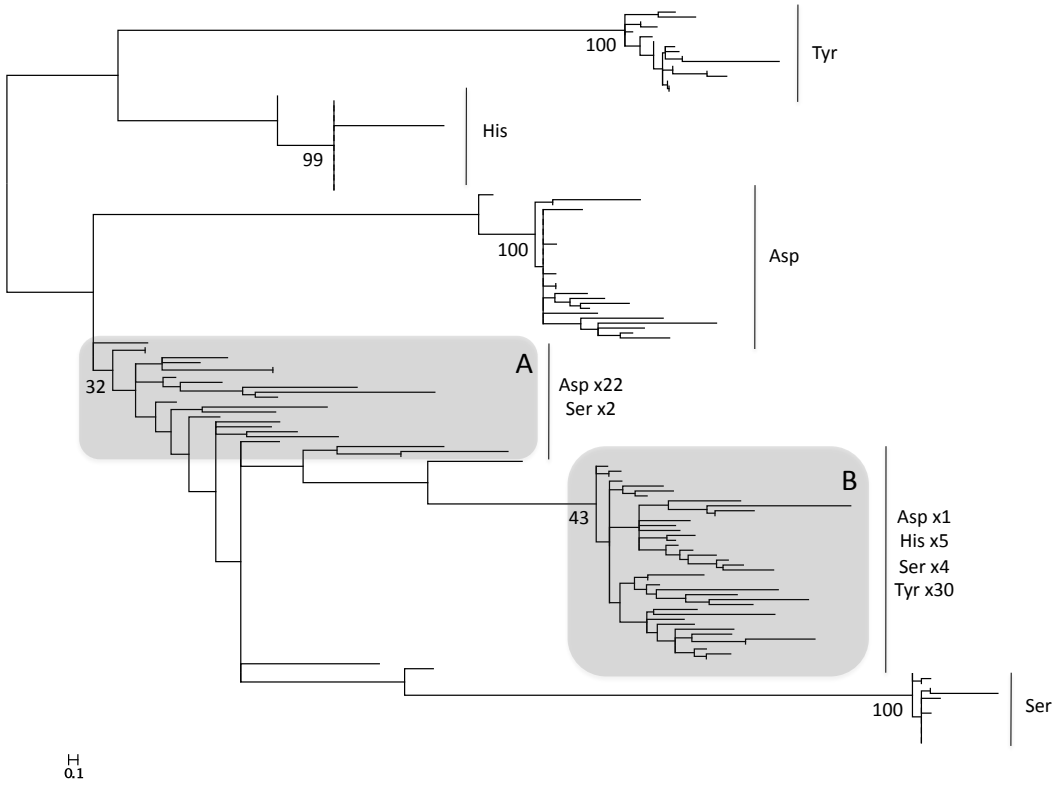


3.4.2 Analyzing multiple amino acids together identifies novel repetitive sequences and produces heterogeneous clusters

We then turned to a larger dataset from *Bos taurus*, composed of tRNAs encoding 4 amino acid families: serine, aspartic acid, tyrosine and histidine. In the entire *Bos taurus* dataset, numerous repetitive sequences were identified by RepeatMasker yet many more appeared to have been missed. *B. taurus* has an exceptionally high number of tRNA genes predicted, 4161 (Chan and Lowe 2009), which was only reduced to 3523 by RepeatMasker. The reconstructed tree (Figure 3.3) was composed of four homogeneous clusters, one for each of the amino acid families, and then two

heterogeneous repetitive-like clusters that included the RepeatMasked sequences. This is exactly in line with our expectation that repetitive sequences would not be constrained to cluster with their family members.

Figure 3.3 – Analysis of $tRNA^{Tyr}$, $tRNA^{Ser}$, $tRNA^{His}$ and $tRNA^{Asp}$ genes from *B. taurus*. Highlighted clusters (A and B) were made up mostly by RepeatMasked sequences. Posterior probabilities are shown for major clusters.



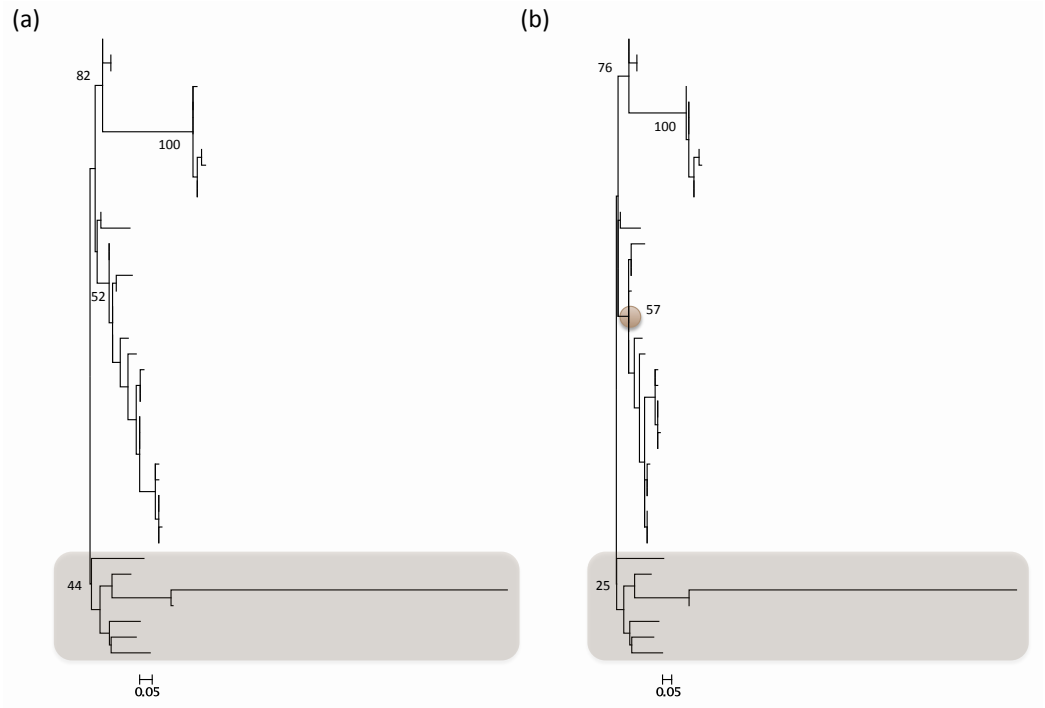
Most of the sequences in these highlighted clusters were identified by RepeatMasker, but not all (unmarked on Figure 3.3). There were 4 unmasked sequences within repetitive-like cluster *A* (with Asp and Ser only) and 3 within repetitive-like cluster *B*. Thus, this method is also capable of identifying erroneous genes unmasked by RepeatMasker. As these sequences are all from the same species, they have not been missed due to lack of annotation. Rather, the sensitivity of RepeatMasker appears insufficient. Thus, the consensus sequence of a repetitive sequence from RepeatMasker may not capture the full sequence diversity of its members.

Moreover, some sequences within the Asp cluster showed longer branches. With the dramatic inflation of tRNA genes in *B. taurus*, these may very well also be falsely predicted.

3.4.3 RNA guided alignment is necessary for repetitive sequence identification while RNA evolutionary models are not

After demonstrating that phylogenetics correctly identifies known repetitive sequences in our preliminary test sets, we wanted to investigate whether its aptitude depended on the use of RNA evolutionary models. Thus, we performed two separate phylogenetic analyses using the same RNA-appropriate MSA built with RNAforester, but modifying the evolutionary models. First, we proceeded as above using both paired and unpaired evolutionary models. Second, we simply ran the unpaired evolutionary model across all sites of the tRNA. Yet again, when alanine genes from *Mus musculus* were analyzed with paired and unpaired models, a single cluster of repetitive sequences was easily identified (Figure 3.4a). Using only unpaired models, the same cluster was also identified (Figure 3.4b).

Figure 3.4 – Identification of RepeatMasked $tRNA^{Ala}$ genes from *M. musculus* employing (a) two separate evolutionary models: 7D for paired sites and GTR for unpaired sites (b) the unpaired GTR evolutionary model for all sites. The cluster of repetitive sequences in both panels is shaded. Posterior probabilities for major clusters are shown; the posterior probability of 57 in panel (b) corresponds to the highlighted node.



Overall, both trees showed near identical clustering and branch lengths. Within the clusters sequence order varied, yet these interior branches had low posterior values, suggesting little support for these differences. Moreover, the posterior probability values between the trees were similar. Thus, it appears that the evolutionary models may not be as important for the identification of repetitive sequences, as long as a high quality structure-guided MSA has been built.

3.5 Discussion

Preliminary results suggest it is indeed possible to distinguish repetitive sequences from functional tRNAs based on differences in evolutionary rates and phylogenetic history. Strongly matching our expectations, known repetitive sequences were identified by their co-clustering in long-branched clusters or by their exceptionally high evolutionary rates. Moreover, when numerous amino acid families were analyzed together, repetitive sequences formed heterogeneous clusters, substantiating their rejection. Every one of the RepeatMasked sequences included in our preliminary test sets was recognized in constructed trees, along with some that were previously unmasked. In its ability to detect repetitive sequences, this method may overcome the lack of annotation or sensitivity of RepeatMasker, to help clean-up tRNA datasets. A clear picture of the tRNA datasets in eukaryotes is fundamental for further studies investigating tRNA evolution and the extent of codon-anticodon adaptation.

Our approach differs from common approaches to clean-up tRNA datasets. Typically, datasets of previously known true tRNAs are required: tRNA prediction methods use statistical descriptions of tRNAs, while approaches to clean datasets often employ comparative genomics. Conversely, our phylogenetic approach takes advantage of the high tRNA gene numbers in eukaryotic genomes, by using the information within the tRNA dataset. There will always be true tRNA genes amid the group, and we have demonstrated that these have a clear signature that is different from repetitive sequences. Thus, it is possible to identify falsely predicted tRNAs using only the tRNAs within the dataset in question. Nevertheless, with known tRNA datasets unavailable

for most species we are unable to determine the accuracy of our method. It is possible that within repetitive-like clusters there are true tRNA genes that are being falsely rejected. In absence of this data, for future research the choice of using phylogenetics or not to clean up datasets should depend on whether the results would be more skewed by the absence of true genes or the presence of false positives.

The power of our method relies strongly on a good quality alignment, since the use of RNA evolutionary models made little difference to the results. Nevertheless, we examined only a single dataset; more extensive comparisons are required to confirm that unpaired evolutionary models are unnecessary. Use of unpaired models could overestimate the distance between true tRNA genes, reducing the contrast between true and false genes. Thus, until this claim is further substantiated, it would be advised to still use paired models, as they more accurately depict tRNA evolution.

One limitation of our phylogenetic method is that Bayesian phylogenetics rapidly becomes computationally expensive with increasing numbers of sequences, particularly when complicated evolutionary models are being employed. Despite this, we hope that this method can be adapted and used to clean-up entire tRNA datasets for various species. With preliminary results demonstrating the effectiveness of phylogenetics, we believe that the development and use of this method for large datasets is warranted. Perhaps the use of simpler phylogenetic methods could reduce the computational time. If unpaired evolutionary models truly detect repetitive sequences effectively, then it may be possible to use simpler phylogenetic methods, since RNA models are not widely implemented. Moreover, during the last stage of our analysis, repetitive-like cluster identification was completed manually from visual analysis of trees. To tackle larger datasets, we then suggest that cluster identification be done automatically. Due to their long branches, it may be possible to automatically calculate the intra-cluster distances of repetitive-like clusters, which would be much higher than those of tight clusters of functional genes. As such, a stringent intra-cluster distance cutoff may be chosen and repetitive-like clusters may be classified as such if their intra-cluster

distance exceeds the cutoff. Moreover, for analysis of entire tRNA datasets, it should be ensured that misidentified genes are not excluded, but are rather placed back within their correct family.

During the next chapter, we continue to test the reliability of our method against additional RepeatMasked sequences. Moreover, we will apply this method to address a biological question, to see if our phylogenetic approach may redeem an anticodon pattern, by removing deviating genes that RepeatMasker could not.

4 Reaffirming a parsimonious tRNA usage pattern across eukaryotes

4.1 Abstract

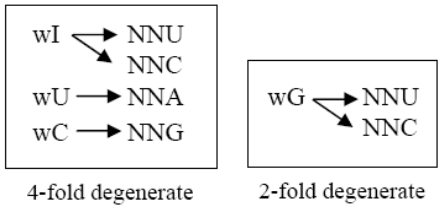
The study of the tRNA gene content of simple eukaryotes has yielded a general anticodon pattern: U and C-ending codons appear to be decoded by tRNAs with either a wobble G (wG) or a wobble I (wI), never both. This mutually exclusive tRNA pattern, however, is called into question by extensive datasets of predicted tRNA genes. In numerous species, there are genes that violate this pattern. We here examine the possibility that tRNAs violating the usage pattern are false positives, implementing our phylogenetic method. Specifically, of the tRNAs that decode C and U-ending codons from 21 representative eukaryotic species, 89 wobble A (wA) and 134 wG deviating tRNA genes are left unmasked by RepeatMasker. Yet, 91% and 72% of wA and wG genes were identified to be false positives by phylogenetics, as determined by their exceptionally high evolutionary rates. After phylogenetic analysis, rice was the only species that retained clusters of deviating $tRNA^{Val}$ and $tRNA^{Thr}$ genes. Upon further inspection, these genes appear to have been acquired from the chloroplast. We therefore conclude that the wG or wI pattern generally holds true in eukaryotes and

subsequently discuss the origin and maintenance of this usage pattern, particularly in light of gene conversion.

4.2 Introduction

A general tRNA usage pattern has been described from a recent compilation of tRNA genes, by which all tRNA families dichotomously carry a wobble inosine (wI) or a wobble G (wG) to decode Y-ending codons (where Y is either C or U), never both (Marck and Grosjean 2002; Grosjean, Marck, and de Crecy-Lagard 2007). This structure of tRNA usage, hereafter referred to as the Marck-Grosjean pattern, appears to be conserved across all three kingdoms of life (Marck and Grosjean 2002). Specifically, except for glycine in all eukaryotes and leucine in a few fungal species, the Y-ending codons from 4-fold and 3-fold degenerate families are decoded by wI anticodons, and not by wG, whereas 2-fold degenerate (NNY) families are decoded by wG anticodons, and not by wI or wA (Figure 4.1) (Marck and Grosjean 2002; Grosjean, Marck, and de Crecy-Lagard 2007). The wI is a modified wobble A (wA) (Auxilien et al. 1996), which expands its decoding capacity to not only read U-ending codons, but also C-ending and, rarely, A-ending codons (Crick 1966; Curran 1995; Johansson et al. 2008). Because wG and wI nucleotides share the ability to decode the Y-ending codons, there is no loss of decoding ability associated with carrying only one of the two anticodons (Agris 2004).

Figure 4.1 – Schematic of the decoding of typical 4-fold and 2-fold degenerate families, as described by the Marck-Grosjean pattern.



In eukaryotes, a single enzyme catalyzes the conversion of wA to wI of all anticodons that carry it. This enzyme has a preference for anticodons with a purine at position 35 of the anticodon (the wobble nucleotide being position 34) with the only exception being $tRNA^{Arg/ICG}$ (Auxilien et al.

1996). Thus, anticodons with a wI would read codons with a pyrimidine at the second position, while those with a purine do not. In agreement, the 2-fold degenerate codon families, as well as the exceptional glycine, all carry a purine at the 2nd codon position. Conversely, in prokaryotes, $tRNA^{Arg/ICG}$ is the only tRNA that undergoes conversion to wI (Haumont et al. 1984). It is then clear that eukaryotes have evolved to expand the use of wI anticodons. However the reason for choosing wI over wG residues and why neither anticodon are found concurrently are unknown.

Meanwhile, a cost has been proposed for the unmodified wA nucleotide: a modeling study predicted that when wA anticodons are in the P-site of the ribosome, they disrupt codon-anticodon interactions in the A-site (Lim 1994; Lim 1995), *i.e.* wA anticodons disrupt the reading of the next codon. This hypothesis is often called upon to explain the avoidance of unmodified wA_d anticodons (Agris, Vendeix, and Graham 2007; Johansson et al. 2008), although other studies suggest no such cost (Watanabe et al. 1997; Chen et al. 2002).

A recent large scale compilation of tRNA genes, the Genomic tRNA Database (gtRNAdb), however, casts doubt on the validity of the Marck-Grosjean pattern (Chan and Lowe 2009). Scattered across eukaryotic species are numerous wG tRNAs in 4-fold and 3-fold families, which by the Marck-Grosjean pattern should only carry a wI. Similarly, there are numerous wA tRNAs in 2-fold families, which should only carry a wG. The wG and wA tRNAs that deviate from the Marck-Grosjean pattern will be hereafter referred to as wG_d and wA_d genes, with the subscript 'd' an abbreviation for 'deviating'.

The presence of wG_d and wA_d genes in higher eukaryotes poses a serious challenge to the Marck-Grosjean pattern, particularly as this rule has never been truly tested across a representative set of eukaryotic species; the study by Marck *et al.* (Marck and Grosjean 2002) included only 7 eukaryotic species, with simpler eukaryotes over-represented in their dataset. For the two higher eukaryotes studied, only the scaffold genome sequence of *Drosophila melanogaster* was available at the time, and the early draft of the *Homo sapiens* genome allowed for the inclusion of fewer than 200 predicted tRNAs (Marck and Grosjean 2002), while current predictions approach 500 (Chan and

Lowe 2009). Do the deviating predicted anticodons then demand a revision of the Marck-Grosjean pattern?

One possible explanation that would redeem the Marck-Grosjean pattern would be that these wG_d and wA_d genes are non-functional, falsely predicted genes. Thus, we undertook analysis of the rogue wG_d and wA_d genes to test whether there truly was a conflict between tRNA gene data and the Marck-Grosjean pattern. We inspected the tRNA families from 21 representative eukaryotic species in which wG_d and wA_d were predicted, using first RepeatMasker and then RNA phylogenetics. RepeatMasker identified some wG_d and wA_d genes within the datasets as false positives, but many wG_d and wA_d genes remained unfiltered. In contrast, our phylogenetic approach removed nearly all wG_d genes predicted in 3-fold and 4-fold degenerate families and wA_d genes predicted in 2-fold degenerate NNY families. These results reaffirm the Marck-Grosjean pattern.

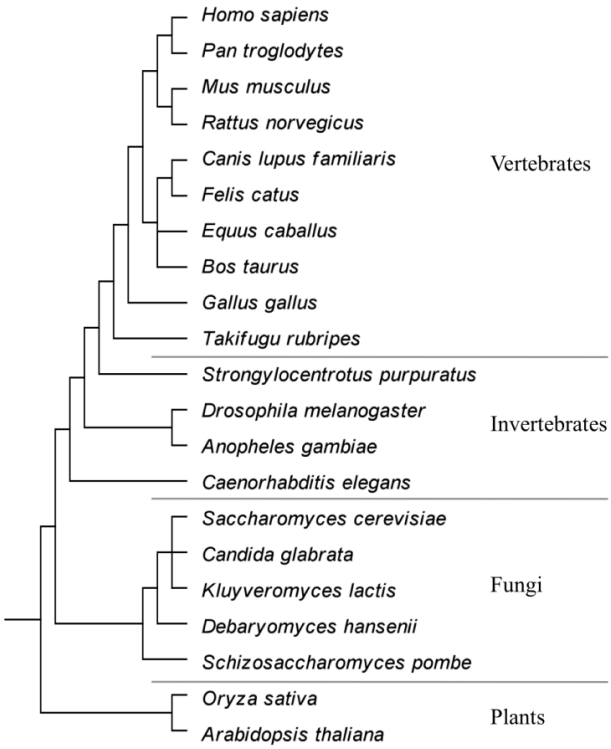
4.3 Methods

4.3.1 Sequence datasets

tRNA sequences and predicted structures were downloaded from the Genomic tRNA Database (gtRNAdb) (Chan and Lowe 2009) for 21 eukaryotic species: the fungi: *Candida glabrata, Debaryomyces hansenii, Kluyveromyces lactis, Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (baker's yeast); the invertebrates: *Anopheles gambiae* (mosquito), *Drosophila melanogaster* (fruit fly), *Stongylocentrotus purpuratus* (sea urchin) and *Caenorhabditis elegans* (nematode); the plants: *Arabidopsis thaliana* (arabidopsis) and *Oryza sativa* (rice); and the vertebrates: *Bos taurus* (cow), *Canis familiaris* (dog), *Equus caballus* (horse), *Felis catus* (cat), *Gallus gallus* (chicken), *Homo sapiens (hg19 - NCBI Build 37.1 Feb 2009)* (human), *Mus musculus* (mouse), *Pan troglodytes* (chimpanzee), *Rattus norvegicus* (rat) and *Takifugu rubripes* (pufferfish). The phylogenetic relationship between these species (Figure 4.2) was taken from the Taxonomy Common Tree tool available from NCBI (National Centre for Biotechnology Information). As

sequences and structures are output separately from the gtRNAdb, sequence and structure were combined with scripts written in MATLAB.

Figure 4.2 – Phylogenetic relationship between the 21 species included in this study, as per the NCBI taxonomy builder. Branch lengths do not reflect the true distances between species.



4.3.2 BLAST

To identify homologous sequences for tRNA genes of interest from *O. sativa*, nucleotide BLAST (BLASTn) was run against databases on the NCBI website as well as specifically against the complete genome sequence of the *O. sativa* chloroplast (NC_008155).

4.3.3 RepeatMasker

RepeatMasker (version 3.2.9) was run from the online server from the Institute for Systems Biology (http://www.repeatmasker.org/), using the settings *cross-match* and *slow* for high sensitivity and specificity, as per user instructions. For each run, species-specific options were chosen (Smit, Hubley, and Green 1996-2010).

4.3.4 Multiple Sequence Alignment, Consensus Structure & Phylogenetics

Following the methodology put forth in the previous chapter, RNAforester was used to build multiple sequence alignments and consensus structures. tRNA introns identified by tRNAscan-SE are found within the anticodon loop in some tRNA families. Introns were cut out and pasted at the 5' end of the sequence, prior to the first stem, ensuring that the intron and anticodon loop were properly, separately aligned relying on the partitioning of paired and unpaired sites by RNAforester. Perl scripts were written to facilitate the building of sequence subsets for analysis and to change file formatting, linking different programs.

Bayesian phylogenetic analysis was performed using the program PHASE (PHylogenetics And Sequence Evolution); (Jow et al. 2002). In PHASE, paired and unpaired sites were partitioned, as per the generated consensus structures, and the parameters for two substitution models were simultaneously estimated during MCMC (Markov Chain Monte Carlo): the GTR + γ (general time reversible) 4x4 matrix was chosen for unpaired sites, and a 7x7 matrix + γ (7D in PHASE) for paired sites (Savill, Hoyle, and Higgs 2001). Phylogenetic analyses were run to completion in duplicate, each initiated randomly and each with different non-informative priors for rate ratios. To assess convergence, likelihood values sampled from the posterior distribution during MCMC and tree topologies of the run pairs were compared. The branch lengths of the consensus tree are the average of branch lengths sampled during the MCMC, therefore branch lengths were further optimized to the final tree topology by maximum likelihood, using Optimizer in the PHASE package. The posterior probabilities were then parsed into the optimized tree file using a script written in MATLAB.

4.3.5 Counting tRNA anticodons

To help tally anticodons after phylogenetic analysis, trees were viewed using the Phylogenetic Tree Tool graphic user interface in MATLAB. In this milieu, sequences in repetitive-like clusters may be deleted, and the names of remaining sequences output. Anticodon counts were then obtained

using a Perl script from the sequence names, which included unique anticodon, amino acid and species identifiers.

4.4 Results

4.4.1 RepeatMasker is insufficient to remove predicted deviating wG_d and wA_d genes

The Marck-Grosjean pattern describes the choice of decoding Y-ending codons with either a wI or a wG, stating that there are no modified (wI) or unmodified wA anticodons when there is a wG, and that there are no wG anticodons when there is a wI. To determine the frequency of wA and wG genes deviating from the Marck-Grosjean pattern, tRNA genes for 21 eukaryotic species were downloaded from the genomic tRNA database (gtRNAdb). To remove any remaining repetitive elements all tRNA genes were screened by RepeatMasker: this eliminated some, but not all, wA_d and wG_d genes (Table 4.1).

Table 4.1 – Number of total predicted tRNA genes, deviating wA (wA_d) genes and deviating wG (wG_d) genes by species before and after RepeatMasker. Genes left unmasked by RepeatMasker are denoted by RM+.

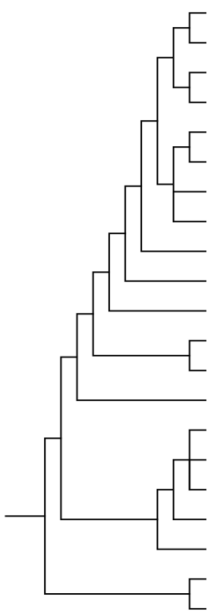
	All	**wA_d**	**wG_d**	**RM+**	**wA_d RM+**	**wG_d RM+**
A. gambiae	422	0	0	422	0	0
A. thaliana	639	1	2	610	1	2
B. taurus	4161	180	42	3523	8	15
C. elegans	820	36	37	773	32	27
C. familiaris	906	3	0	906	3	0
C. glabrata	227	0	1	227	0	1
D. hansenii	226	0	1	226	0	1
D. melanogaster	304	0	0	304	0	0
E. caballus	503	4	4	503	4	4
F. catus	2393	12	9	2393	12	9
G. gallus	242	2	2	242	2	2
H. sapiens	625	5	4	602	4	3
K. lactis	183	0	1	183	0	1
M. musculus	433	2	3	423	2	2
O. sativa	764	8	60	764	8	60
P. troglodytes	463	5	2	463	5	2
R. norvegicus	441	3	2	379	1	1
S. cerevisiae	295	0	1	289	0	1
S. pombe	188	0	1	187	0	1
S. purpuratus	1067	3	2	811	3	0
T. rubripes	722	4	2	709	4	2
Total		268	176		89	134

Of a total of 268 and 176 predicted wA_d and wG_d genes, 89 and 134 were left unmasked across the 21 species. The success of RepeatMasker in eliminating these deviating anticodons varied by species, the largest decrease in number of wA genes was for *B. taurus*, where 180 wA_d genes were predicted, all but 8 of which were masked; while all but 15 of the 42 predicted wG_d genes of *B. taurus* were removed. *C. elegans* still retained 32 of its 36 wA_d genes and 27 of its 37 wG_d genes. In the unmentioned species, few genes were filtered by screening with RepeatMasker.

4.4.2 Unmasked wG_d and wA_d are sporadically distributed across tRNA families

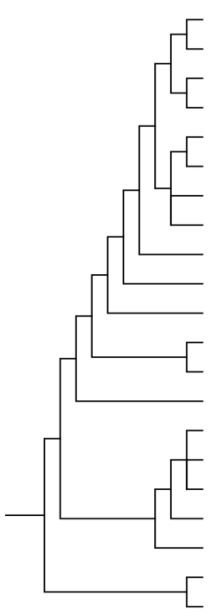
If deviating tRNA genes were functional, once acquired, they would be expected to persist and be conserved across species. To examine the distribution of deviating tRNA genes from an evolutionary perspective, the phylogenetic relationship between these species (Figure 4.2) was used to guide the analysis of wA_d and wG_d genes, separated by anticodon (Figures 4.3 and 4.4, respectively).

Figure 4.3 – Distribution of unmasked predicted wA_d genes sorted by species and by anticodon. The phylogenetic relationship shows only the topology of the relationship between species, branch lengths are uninformative.



	Phe AAA	Asn ATT	Asp ATC	His ATG	Ser ACT	Tyr ATA	Cys ACA	Gly ACC	**wA_d total**
H. sapiens		2			1	1			4
P. troglodytes		1			2	1	1		5
M. musculus				1				1	2
R. norvegicus					1				1
C. familiaris		1	2						3
F. catus	2	5	1		1	1	1	1	12
E. caballus		1	1	1		1			4
B. taurus	1		6		1				8
G. gallus	1			1					2
T. rubripes			1	2		1			4
S. purpuratus	2				1				3
D. melanogaster									0
A. gambiae									0
C. elegans			4	25		1		2	32
S. cerevisiae									0
C. glabrata									0
K. lactis									0
D. hansenii									0
S. pombe									0
O. sativa			2			4	2		8
A. thaliana								1	1
Total wA_d	6	10	17	30	7	10	4	5	**89**
wA_d families	4	5	7	5	6	7	3	4	**41**

Figure 4.4 – Distribution of unmasked predicted wG_d genes sorted by species and by anticodon. The phylogenetic relationship shows only the topology of the relationship between species, branch lengths are uninformative.



	Ala GGC	Pro GGG	Thr GGT	Val GAC	Ser GGA	Arg GCG	Leu GAG	Ile GAT	**wG_d total**
H. sapiens								3	3
P. troglodytes						1		1	2
M. musculus					1			1	2
R. norvegicus	1								1
C. familiaris									0
F. catus	1	1			1	6			9
E. caballus				3		1			4
B. taurus	2		1		3	9			15
G. gallus	1			1					2
T. rubripes						1	1		2
S. purpuratus									0
D. melanogaster									0
A. gambiae									0
C. elegans	1	6	10	1		2	7		27
S. cerevisiae							1[a]	1	1
C. glabrata							1[a]	1	1
K. lactis							1[a]	1	1
D. hansenii								1	1
S. pombe								1	1
O. sativa	2		16	34	8				60
A. thaliana					1		1		2
Total wG_d	8	7	27	39	14	20	9	10	**134**
wG_d families	6	2	3	4	5	6	3	8	**37**

[a] In Leu families of fungal species, wG is the expected wobble nucleotide, therefore they are not counted as deviating.

Across all families of the 21 species, only 24.4% of the wG families have wA_d genes predicted, and 22% of the wI families have wG_d genes. There is no amino acid family that consistently has predicted wA_d or wG_d anticodons, nor is there a species which consistently has predicted deviating tRNAs for all amino acid families. Examining the figures, these anticodons appear to be gained and lost sporadically across various species. For example, $tRNA^{Phe/AAA}$ sequences form a paraphyletic group, being predicted in *F. catus* and *B. taurus*, but not in their relatives *C. familiaris* and *E. caballus* (Figure 4.3). Therefore, the $tRNA^{Phe/AAA}$ anticodon would have had to be gained or lost multiple times between these species. Overall, the random distribution and low numbers of these deviating genes suggest that they might be false positives, though RepeatMasker failed to remove them.

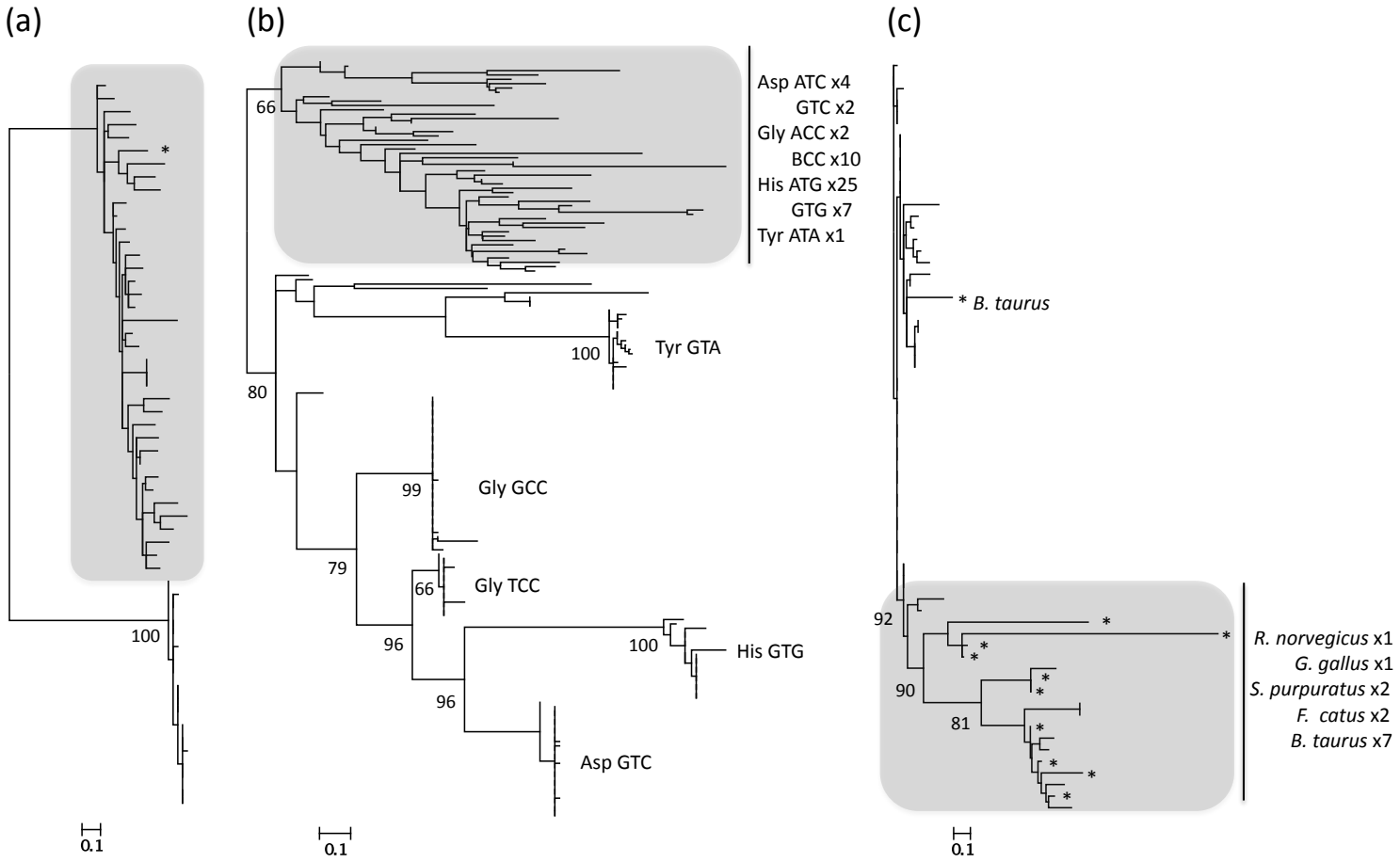
4.4.3 The majority of deviating genes are filtered by phylogenetics

After initial validation that phylogenetics can identify known repetitive sequences, all wA_d and wG_d genes were analyzed to determine whether they might be unfiltered repetitive sequences. In all the datasets containing wA_d or wG_d genes, masked tRNAs were included to further substantiate the ability of phylogenetics to identify known false positives. As this is a comparative approach, wA_d and wG_d sequences were sorted into datasets, which minimally included all the family members of the same species. As exemplified by our analysis of the three trees in Figure 4.5, we used the criteria determined from our test sets to identify novel repetitive sequences.

Analysis of the $tRNA^{Pro}$ genes from *F. catus* alone revealed two clusters: one with very short branches, as expected for true tRNA genes, and one repetitive-like cluster, with each member of the cluster having a long branch (Figure 4.5a), similar to the repetitive-like cluster identified in the glutamine test set (Figure 3.1a). The repetitive-like cluster, as highlighted in the figure, included the predicted wG_d gene (Figure 4.5a). All members of the repetitive-like cluster are considered to be repetitive sequences; hence the wG_d gene was filtered out by phylogenetics.

Four amino acid families of *C. elegans* were analyzed together: $tRNA^{Asp}$, $tRNA^{Gly}$, $tRNA^{His}$, and $tRNA^{Tyr}$. The constructed tree consisted of 5 tight clusters of homogeneous anticodons, one for each of the anticodons expected by the Marck-Grosjean pattern (glycine being a 4-fold degenerate family must have wU in addition to wG anticodons) and one large repetitive-like cluster, which was composed of genes from all families, including all the wA_d genes (Figure 4.5b). This is therefore a clear example where all deviating genes were filtered by phylogenetics. Similarly, a repetitive-like cluster included all but one of the phenylalanine wA_d genes of five species (Figure 4.5c). Since repetitive sequences would be free to accumulate mutations at any position, including the anticodon, the formation of such heterogeneous clusters further supporting that, though unmasked, these sequences truly did arise from repetitive sequences.

Figure 4.5 – Output trees examining tRNA genes for (a) $tRNA^{Pro}$ from *F. catus* with a single wG_d gene, (b) $tRNA^{Asp}$, $tRNA^{Gly}$, $tRNA^{His}$, and $tRNA^{Tyr}$ from *C. elegans* and (c) $tRNA^{Phe}$ genes from the species *F. catus, B. taurus, G. gallus, S. purpuratus* and *R. norvegicus*. The shaded clusters were those identified to be repetitive-like due to their long branches. wA_d genes in (c) and wG_d genes in (a) were highlighted by a *, the counts by the repetitive cluster in (c) only include wA_d genes. For major clusters, posterior probabilities over 50 are shown.



Moreover, an additional notable characteristic was observed in the repetitive-like cluster of *C. elegans* genes: one of the Asp wA_d genes spontaneously gained an intron. Thus, in addition to rapidly accumulating mutations, sequences in repetitive-like clusters are experiencing indel (insertion/deletion) events. Spontaneous gains and losses of introns were also observed in numerous datasets, other than the examples presented here. Contrastingly, within the tight clusters of seemingly true tRNA genes, either all possessed an intron or none did. This further supports that sequences in repetitive-like clusters are evolving rapidly, and are therefore likely to be non-functional.

Lastly, after eliminating those sequences belonging to repetitive-like clusters, genes with extremely long branches were singled out for removal, as in the test set (Fig 3.2a). For example, a long branch was observed for the $tRNA^{Phe/AAA}$ in *B. taurus*; although the wA_d gene clustered with its family members, it had a particularly long branch (Figure 4.5c). In such cases, we did not attempt to determine whether the long branch might be due to repetitive sequences, single genes which have undergone pseudogenization or even genes whose anticodons have been misidentified. Both tRNA-derived repetitive sequences and pseudogenes are non-functional, which would warrant their exclusion. Meanwhile, genes that are misidentified might not warrant their exclusion from the tRNA gene set, but they would no longer be a wA_d or wG_d. Thus regardless of these possible types of genes, for our purposes these genes should be eliminated.

Summarizing the analysis of all generated trees: 73 wA_d and 93 wG_d genes were found to cluster in repetitive-like clusters, with a further 8 wA_d and 3 wG_d genes removed for having exceptionally long branches (Tables 4.2 and 4.3). Together, tRNAs were removed from almost all families, eliminating 91.0% of the unmasked wA_d genes and 71.6% of the unmasked wG_d genes, suggesting the majority of these deviating genes arose from repetitive sequences. The remaining genes occurred across few species: all wA_d or wG_d anticodons were removed for the fungal species as well as for *S. purpuratus, R. norvegicus, C. familiaris, E. caballus*, *G. gallus* and *C. elegans*, increasing the total number of species with no predicted deviating anticodons to 11. Families that retained their wA_d or wG_d anticodons tended to have very small numbers of deviating genes. Many

Table 4.2 – Number of wA_d genes remaining after removal of those found in repetitive clusters or with long branches after phylogenetic analysis. Zeros indicate families that had wA_d genes, which were unfiltered by RepeatMasker, but that were removed by phylogenetic analysis.

	Phe AAA	Asn ATT	Asp ATC	His ATG	Ser ACT	Tyr ATA	Cys ACA	Gly ACC	**wA_d total**	**% filtered**
H. sapiens		2			0	1	0		3	25.0%
P. troglodytes		1			0	1	0		2	60.0%
M. musculus				0				0	0	100.0%
R. norvegicus					0				0	100.0%
C. familiaris		0	0						0	100.0%
F. catus	0	0	0		0	1	0	0	1	91.7%
E. caballus		0	0	0		0			0	100.0%
B. taurus	0		0		0				0	100.0%
G. gallus	0			0					0	100.0%
T. rubripes			1	0		0			1	75.0%
S. purpuratus	0				0				0	100.0%
D. melanogaster									0	
A. gambiae									0	
C. elegans			0	0		0		0	0	100.0%
S. cerevisiae									0	
C. glabrata									0	
K. lactis									0	
D. hansenii									0	
S. pombe									0	
O. sativa			1			0	0		1	87.5%
A. thaliana								0	0	100.0%
Total wA_d	0	3	2	0	0	3	0	0	**8**	**91.0%**
wA_d families	0	2	2	0	0	3	0	0	**7**	**82.9%**

Table 4.3 – Number of wG_d genes remaining after removal of those found in repetitive clusters or with long branches after phylogenetic analysis. Zeros indicate families that had wG_d genes, which were unfiltered by RepeatMasker, but that were removed by phylogenetic analysis.

	Ala GGC	Pro GGG	Thr GGT	Val GAC	Ser GGA	Arg GCG	Leu GAG	Ile GAT	**wG_d total**	**% filtered**
H. sapiens								3	3	0.0%
P. troglodytes						0		1	1	50.0%
M. musculus				0	0			1	1	50.0%
R. norvegicus	0								0	100.0%
C. familiaris									0	
F. catus	0	0			0	0			0	100.0%
E. caballus				0		0			0	100.0%
B. taurus	0		0		3	0			3	80.0%
G. gallus	0			0					0	100.0%
T. rubripes						0	0		0	100.0%
S. purpuratus									0	
D. melanogaster									0	
A. gambiae									0	
C. elegans	0	0	0	0		0	0		0	100.0%
S. cerevisiae							1[a]	0	0	100.0%
C. glabrata							1[a]	0	0	100.0%
K. lactis							1[a]	0	0	100.0%
D. hansenii								0	0	100.0%
S. pombe								0	0	100.0%
O. sativa	0		8	17	4				29	51.7%
A. thaliana					1		0		1	50.0%
Total wG_d	0	0	8	17	8	0	0	5	**38**	**71.6%**
wG_d families	0	0	1	1	3	0	0	3	**8**	**78.3%**

[a] In Leu families of fungal species, wG is the expected wobble nucleotide, therefore they are not counted as deviating.

had only 1, a few up to 4, besides the notable exception of $tRNA^{Val/GAC}$ and $tRNA^{Thr/GGT}$ genes in *O. sativa*, which retained 17 and 8 deviating genes, respectively.

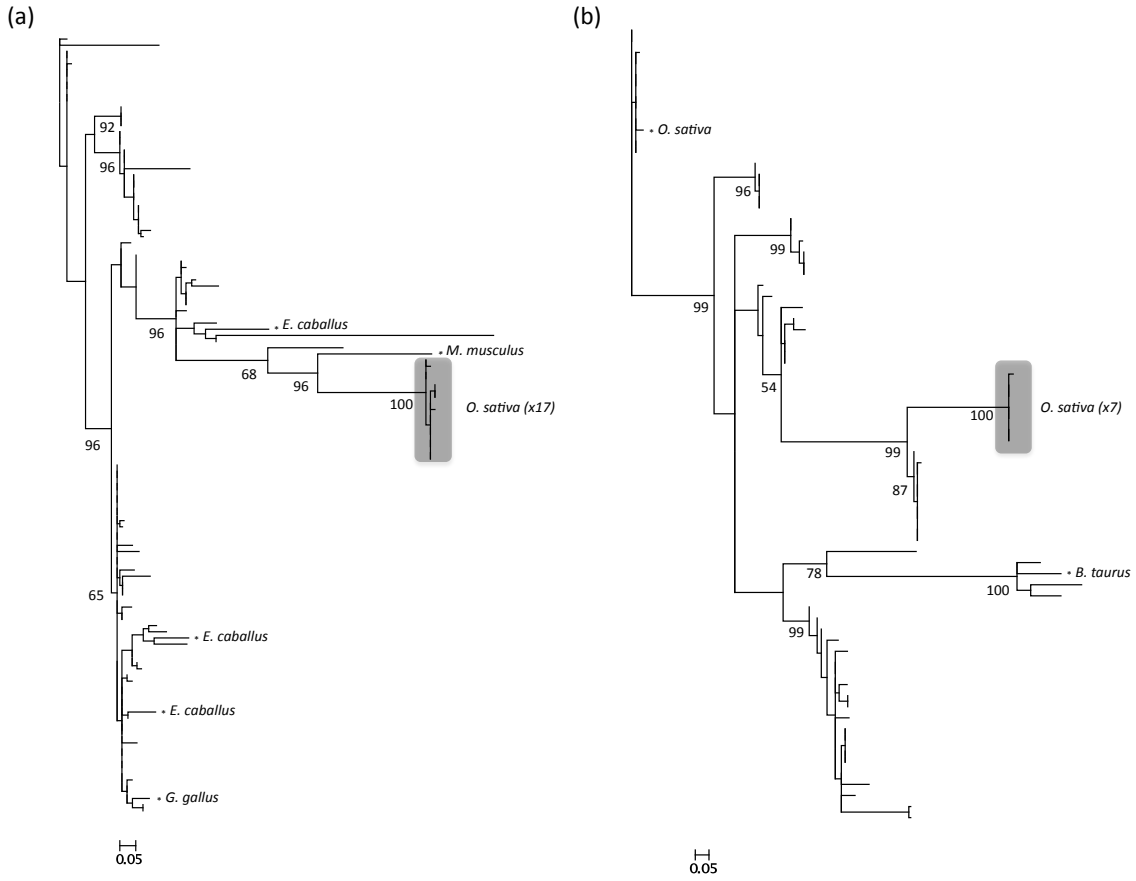
4.4.4 Deviating wG anticodons resemble chloroplast tRNAs in O. sativa

Being the only families with un-filtered deviating genes in sufficient numbers, we further analyzed the 17 $tRNA^{Val/GAC}$ and the 8 $tRNA^{Thr/GGT}$ genes of *O. sativa*. All the $tRNA^{Val/GAC}$ and 7 of the 8 $tRNA^{Thr/GGT}$ clustered together with extremely short branches (Figure 4.6), similar to our expectation for functional genes. Nevertheless, these clusters were rather distant from other genes of the same family. When single, randomly chosen sequences from each cluster were analyzed by BLAST, matches with the highest similarity were tRNA genes from chloroplast genomes of various plants. Both valine and threonine sequences showed 99% identity to sections of the *O. sativa* chloroplast genome, suggesting that wG_d genes in *O. sativa* arose by transfer from the chloroplast to the nucleus.

The insertion of chloroplast sequences into the rice genome has been previously noted, including the tRNA genes included in this sequence (RC10SC 2003; IRGSP 2005); (RC10SC: Rice Chromosome 10 Sequencing Consortium, IRGSP: International Rice Genome Sequencing Project). Being of bacterial origin, chloroplasts have no wI anticodons in families other than arginine (Karcher and Bock 2009), thus the presence of valine and threonine wG are not surprising from this source. The presence of chloroplast sequences within the nuclear genome could be explained by contamination during the sequencing process. However, homologous wG_d genes are also predicted in other plants species, including *Glycine max* (soybean), *Medicago trunculata* (clover), and *Zea mays* (corn) (Chan and Lowe 2009). It is therefore more likely that a true acquisition of chloroplast sequences occurred during the evolution of these plant genomes. From sequences alone, it cannot be determined whether these genes are expressed and, if expressed, whether they are functional.

These genes represent 24 of the 43 remaining wG_d genes. Therefore, 17.9% of the unmasked wG_d genes were derived from chloroplast genes and 71.6% were filtered by phylogenetics, leaving only 10.5% of the predicted wG_d genes clustering with short branches with their family members.

Figure 4.6 – Generated trees for the (a) $tRNA^{Val}$ genes from *E. caballus, G. gallus, M. musculus* and *O. sativa* and (b) $tRNA^{Thr}$ genes from *B. taurus* and *O. sativa*. In both panels, the clusters of wG_d genes with short branch lengths in *O. sativa* are shaded. wG_d genes, outside of the highlighted *O. sativa* cluster, are highlighted by a *. For major clusters, posterior probabilities over 50 are shown.



4.5 Discussion

Phylogenetic analysis of tRNA genes that violated the Marck-Grosjean pattern revealed 90.7% and 71.6% of unmasked wA_d and wG_d genes to be false positives. The vast majority of these sequences were rejected for belonging to long-branched, repetitive-like clusters, and a minority for having exceptionally long branches. The few wA_d and wG_d genes remaining that clustered tightly with their family members would have arisen by mutation at the anticodon site. Determining whether

these genes are functional is beyond the scope of this thesis; they may be true deviations from the Marck-Grosjean pattern or they may be recent pseudogenes that have not yet acquired sufficient mutations to be identified as such. We are therefore unable to conclude that the avoidance of wA_d or wG_d genes is absolute. Still, with deviating genes so rarely and sporadically distributed across the 21 eukaryotic species included in this study, the Marck-Grosjean pattern appears to be broadly conserved. Moreover, every one of the RepeatMasked sequences included in our datasets was recognized in constructed trees, along with the many previously unmasked sequences, in further support of the phylogenetic method's ability to identify repetitive sequences.

Our analysis also identified outlying clusters of valine and threonine wG_d genes in *O. sativa,* which topologically seemed to be functional; they made up a further 17.9% of the total wG_d genes. Displaying high similarity to sequences from the chloroplast, these genes appear to have arisen via transfer from this organelle. If organelle-derived tRNA genes are expressed and properly undergo maturation, they might yield novel sequence diversity and even, as seen in this case, novel anticodon diversity. Nevertheless, it seems questionable whether these tRNA genes, being so distant to their family members, would be properly recognized by the correct aaRS. Moreover, their prokaryotic origin makes it unlikely that these genes would be expressed, as the transcription machinery, and consequently promoter sequences, between prokaryotes and eukaryotes are different (RC10SC 2003). With no known cases of functional tRNA genes acquired from organelles, it cannot be deduced whether these genes are functional without further investigation.

Though the avoidance of wA_d and wG_d genes is expected from the Marck-Grosjean pattern, it has yet to be determined whether the Marck-Grosjean pattern itself is expected. No evolutionary mechanism has been proposed for the creation and maintenance of this pattern. How then do the available models of tRNA evolution fare in explaining these extremely biased, if not absolute, anticodon choices?

For the removal of deviating anticodons, a selective or mutative force would need to be acting against them. Such a rationale was employed in the Wobble Cost Hypothesis, which states that

tRNA choice acts to minimize the occurrence of poor codon-anticodon interactions (Xia 2008). Although, such a cost has been suggested for wA anticodons, this alone is insufficient to yield the Marck-Grosjean pattern. There would need to be a similar cost for wG nucleotides in wI families, while none has been suggested. Moreover, given the prevalence of the wG and that evolutionarily speaking the wG was historically used in these families such a cost seems unlikely. It would be inappropriate to rely on a cost-based hypothesis for the avoidance of wA, ignoring that an equivalent phenomenon is observed for the avoidance of the wG in absence of a cost. Thus, the Marck-Grosjean pattern must be driven by another mechanism.

Following the documentation of agreement between tRNA anticodons and codon usage in *Escherichia coli* (Ikemura 1981) and *S. cerevisiae* (Ikemura 1982), it was proposed that codons and tRNAs co-evolve (Bulmer 1987). According to a recent modeling study of this co-evolution, the optimal anticodon content was heterogeneous not homogeneous, with the ratio of anticodons within families weighed by their decoding capacity of each codon and by codon frequencies (Higgs and Ran 2008; Ran and Higgs 2010). Thus, an absolute avoidance of wA would be unexpected, even with its reduced decoding capacity. wA anticodons would even be advantageous in AT-rich genomes. However, this is not what is observed, for example *S. cerevisiae* has an AT-rich codon usage bias (as from the gtRNAdb), yet avoids wA anticodons. Moreover, if the wI and wG are considered to have the same decoding capacity, within this framework there is no reason to choose one over the other in families that carry a wI. Thus, the Marck-Grosjean pattern cannot be predicted by codon-anticodon adaptation alone.

This pattern of anticodon choice has also been described as ‘parsimonious’, as most codons are read by a single anticodon, thus reducing redundancy in decoding (Percudani 2001). Thus, without calling upon costs or co-adaptation, a lack of redundancy alone is called upon as support (Percudani 2001). In other words, within this framework, wG anticodons would be considered to ‘complicate’ the tRNA content if wI anticodons are predominantly used. If there were single copies of these anticodons, perhaps this might be true, but as wI anticodons are found in multiple copies (for

example Ala^{AGC} is found in 11 copies in *S. cerevisiae* and 29 in *H.sapiens* (gtRNAdb)), there is redundancy, simply between the multiple copies of these genes. Moreover, from the perspective of a codon, it is the concentration of cognate tRNAs and not the composition that influences translational efficiency (Fluitt, Pienaar, and Viljoen 2007).

An important assumption made so far in our discussion is that wG and wI anticodons are equivalent. But, these nucleotides are not equal, in that they form different numbers of hydrogen bonds: with C nucleotides wG residues form 3 hydrogen bonds, while wI residues form only 2 (Percudani and Ottonello 1999). It has been proposed that there is an avoidance of 6 or 9 hydrogen bonds between the complete codon-anticodon interaction, being too weak and too strong. This may also be guiding the choice of wI or wG nucleotides (Percudani and Ottonello 1999). But, this hypothesis can only be applied to a subset of families: only for AT-rich codons does this model suggest a wG, to increase stability, and for GC-rich codons, this model suggests a wI to decrease stability. Hence, taking as an example the 3 and 4-fold degenerate families, only 5 families fall into the appropriate categories, 2 of which, isoleucine and glycine, break this rule. More importantly, this hypothesis would suggest that for the remaining families that form 7 or 8 hydrogen bonds (threonine, valine, serine and leucine), we would expect to see both wG and wI anticodons as neither are considered to be detrimental, a pattern different from the Marck-Grosjean pattern.

Finally, another possible explanation is that the aaRS would have adapted to the dominant anticodon, wI or wG. Such a preference has been demonstrated for the isoleucine family, where *in vitro* $aaRS^{Ile}$ was shown to have a higher affinity for the wI and pseudouridine anticodons actually used in the isoleucine family over the other possible anticodons: the loading of the $tRNA^{Ile/GAU}$ was 3 fold less than $tRNA^{Ile/IAU}$ (Senger et al. 1997). Interestingly, the *E. coli* $tRNA^{Ile/GAU}$ is recognized more strongly than the *S. cerevisiae* $tRNA^{Ile/IAU}$, thus there are sites other than the anticodon influencing this specificity, which must have changed during the development of the Marck-Grosjean pattern (Senger et al. 1997). Thus, a preference by the aaRS could potentially explain the choice between wG and wI.

Thus overall, although many of the existing hypotheses were unable to explain the Marck-Grosjean pattern, we found differential affinities for tRNA anticodons by the aaRS may be a feasible explanation for the preferences in anticodons. However, many of the hypotheses that describe codon-anticodon evolution were formulated from analysis of prokaryotes, where codon-anticodon selection is much stronger than in eukaryotes. We consequently found ourselves doubting whether the same intensity of selection could be extended to the higher tRNA gene numbers of eukaryotes. For example, using the preference by $aaRS^{Ile}$ as an example, there are 13 copies of the $tRNA^{Ile/IAU}$ in *S. cerevisiae*. Thus, employing a highly simplified scheme, each tRNA is responsible for 7.7% of the decoding responsibility. If one of the wI anticodons were to mutate to a wG, then there would be a 3-fold loss in the decoding ability due to slower loading. Thus, only 2.6% of the 7.7% responsibility would be met by the wG anticodon. However, this is a drop of only 5.1% of the decoding ability overall, with the 13 tRNA gene copies. We are thus skeptical whether the strength of selection is sufficient to cause reversions, or rather whether a constant, though small, number of the deviating genes would be expected. Particularly, these concerns are intensified for species where the numbers of tRNA genes are further increased, *e.g.* the aforementioned 29 copies of $tRNA^{Ala/IGC}$ in *H. sapiens* where each tRNA holds only 3.4% of the decoding responsibility.

We hereby propose an alternative hypothesis, which might alleviate the need for such strong selection pressures. Gene families are unavoidably subjected to gene conversion, a process that results in concerted evolution between homologous genes (Ohta 1983). Shown to be acting among tRNAs in fungal species, gene conversion has been blamed for the extremely low evolutionary rate of tRNA genes (Amstutz et al. 1985). Concerted evolution would also homogenize the anticodon content of tRNA families, in all likelihood in the direction that would increase over-represented anticodons. Therefore, anticodons deviating from the Marck-Grosjean pattern may well be functional, and lost not because they are detrimental, but because they are homogenized by their family members. The dominant anticodon would therefore be found in an evolutionarily stable state, with anticodon content maintained by gene conversion. Thus, the influence of gene conversion might

sufficiently enhance a preference in anticodon to an absolute choice. According to this model of evolution, we would expect anticodons to cluster together rather than interspersed with their family members. This phenotype is exemplified by the independent clustering of wU and wG anticodons from glycine (Figure 4.5b). As both wU and wG genes are necessary to decode all genes effectively, there would be pressure for these anticodons to become distinct enough from one another to escape homogenization.

If gene conversion is producing concerted evolution in tRNA gene families, then mutations at the wobble site would not be truly free. This is distinct from what is assumed in many of the previously discussed evolutionary theories. This difference may be exploited in future studies to test whether or not gene conversion is occurring. We did not attempt to do this with our results for two reasons. First, we did not examine all families, only those with deviating wA and wG genes. Second, we did not account for gene duplication and gene loss events, of which gene duplication could also predict co-clustering of identical anticodons.

We hope in future research to expand the use of this RNA phylogenetic method to entire tRNA datasets for various species and further explore how codon-anticodon adaptation and gene conversion shape the anticodon content of eukaryotes.

5 Concluding statements

Throughout this thesis, we have presented three different projects, beginning from the perspective of codon usage and ending from the perspective of tRNA genes. We first identified that tRNAs packaged by HIV-1 may reflect an altered tRNA pool. We propose this tRNA pool to be induced late during infection, and available to benefit the exceptionally poor codon usage of late genes. We then turned to the removal of false positives from predicted tRNA datasets with phylogenetics. tRNA gene prediction is prone to falsely predicting tRNA-derived repetitive sequences as tRNA genes. Indeed, our proposed phylogenetic method was capable of distinguishing known repetitive sequences. Finally, we examined the anticodon content reading NNY families in eukaryotes, where a reduced pattern was expected, yet there were deviations found within predicted tRNA datasets. Across 21 representative eukaryotic species, our phylogenetic method was able to remove the vast majority of these deviating genes, reaffirming the tRNA pattern in eukaryotes. Each of these projects suggests future studies and highlight novel aspects about translation in eukaryotes.

The story of HIV codon usage would suggest that further studies be completed to examine the dynamics of the tRNA pool during HIV infection. Moreover, the hypothesis of an altered tRNA pool calls into question the traditional view of the tRNA pool as stagnant within a cell. Having such a

widespread influence, were the tRNA pool altered under stressful or other conditions in a cell, overall shifts in the state of that cell may be induced. Moreover, if an altered tRNA pool is induced in response to HIV-1, a similar phenomenon could also be experienced with other viruses. This would particularly be the case, were it anti-viral stress responses that induce changes in tRNA concentrations. Future studies should then not only examine the tRNA pool in response to infection by HIV-1 but also numerous different viruses. Nevertheless, for such detailed study, more sensitive methods need to also be developed, which overcome the limitations of tRNA microarrays.

While the altered tRNA pool may then further complicate the story of codon-anticodon adaptation, our phylogenetics approach to removing falsely predicted tRNA genes takes one step in the direction of resolving codon-anticodon adaptation in eukaryotes. The method shows promise, quite clearly being able to distinguish false positives derived from repetitive sequences. We hope this method will be used in the future to clean up entire tRNA datasets, which will reduce the noisy false positives and may allow more accurate conclusions to be made. Moreover, this method requires no prior information about the repetitive sequence content of a genome, meaning that it may be conducted just after new genomes are sequenced. Additionally, with the Marck-Grosjean pattern conserved across eukaryotes, the presence of wA_d and wG_d genes may serve as a marker of repetitive sequences and false predictions. Seemingly, in genomes with an increased frequency of repetitive sequences, there was also an increased frequency of wA_d and wG_d genes.

After reaffirming the anticodon saving pattern, we paid particular attention to the lack of a hypothesis to explain the creation and maintenance of these strict anticodon choices. In particular, we drew attention to the influence that gene conversion may be playing in the tRNA gene family, a factor typically ignored. Another major concern that arose during our discussion was whether evolutionary hypotheses, which had been modeled after organisms where there is very strong codon-anticodon adaptation, could be extended to higher eukaryotes where co-adaptation may be a weaker force, minimally due to the increase in tRNA gene numbers. These concerns remain unaddressed and should be the topic of future studies.

6 References

Adams, M. J., and J. F. Antoniw. 2004. Codon usage bias amongst plant viruses. Arch Virol **149**:113-135.

Agris, P. F. 2008. Bringing order to translation: the contributions of transfer RNA anticodon-domain modifications. EMBO Rep **9**:629-635.

Agris, P. F. 2004. Decoding the genome: a modified view. Nucleic Acids Res **32**:223-238.

Agris, P. F. 1991. Wobble position modified nucleosides evolved to select transfer RNA codon recognition: a modified-wobble hypothesis. Biochimie **73**:1345-1349.

Agris, P. F., F. A. Vendeix, and W. D. Graham. 2007. tRNA's wobble decoding of the genome: 40 years of modification. J Mol Biol **366**:1-13.

Akashi, H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics **136**:927-935.

Amstutz, H., P. Munz, W. D. Heyer, U. Leupoid, and J. Kohli. 1985. Concerted evolution of tRNA genes: intergenic conversion among three unlinked serine tRNA genes in S. pombe. Cell **40**:879-886.

Ashe, M. P., P. Griffin, W. James, and N. J. Proudfoot. 1995. Poly(A) site selection in the HIV-1 provirus: inhibition of promoter-proximal polyadenylation by the downstream major splice donor site. Genes Dev **9**:3008-3025.

Auxilien, S., P. F. Crain, R. W. Trewyn, and H. Grosjean. 1996. Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of inosine 34 in the anticodon of transfer RNA. J Mol Biol **262**:437-458.

Bauer, M., G. W. Klau, and K. Reinert. 2007. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. BMC Bioinformatics **8**:271.

Bennetzen, J. L., and B. D. Hall. 1982. Codon selection in yeast. J Biol Chem **257**:3026-3031.

Berg, J., J. Tymoczko, and L. Stryer. 2002. Biochemistry. WH Freeman, New York.

Berkhout, B., A. Grigoriev, M. Bakker, and V. V. Lukashov. 2002. Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. AIDS Res Hum Retroviruses **18**:133-141.

Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature **325**:728-730.

Cannarozzi, G., N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral. 2010. A role for codon order in translation dynamics. Cell **141**:355-367.

CDC. 2010. Basic information about HIV and AIDS. Centres for Disease Control and Prevention; http://www.cdc.gov/hiv/topics/basic/.

Chan, P. P., and T. M. Lowe. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res **37**:D93-97.

Chan, S.-W., and P. Egan. 2009. Effects of hepatitis C virus envelope glycoprotein unfolded protein response activation on translation and transcription. Archives of Virology **154**:1631.

Chen, P., Q. Qian, S. Zhang, L. A. Isaksson, and G. R. Bjork. 2002. A cytosolic tRNA with an unmodified adenosine in the wobble position reads a codon ending with the non-complementary nucleoside cytidine. J Mol Biol **317**:481-492.

Chen, Y., D. B. Carlini, J. F. Baines, J. Parsch, J. M. Braverman, S. Tanda, and W. Stephan. 1999. RNA secondary structure and compensatory evolution. Genes Genet Syst **74**:271-286.

Coleman, J. R., D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, and S. Mueller. 2008. Virus attenuation by genome-scale changes in codon pair bias. Science **320**:1784-1787.

Coughlin, D. J., T. Babak, C. Nihranz, T. R. Hughes, and D. R. Engelke. 2009. Prediction and verification of mouse tRNA gene families. RNA Biol **6**:195-202.

Crick, F. H. 1966. Codon-anticodon pairing: the wobble hypothesis. J Mol Biol **19**:548-555.

Cullen, B. R. 1991a. Regulation of HIV-1 gene expression. Faseb J **5**:2361-2368.

Cullen, B. R. 1991b. Human immunodeficiency virus as a prototypic complex retrovirus. J Virol **65**:1053-1056.

Curran, J. F. 1995. Decoding with the A:I wobble pair is inefficient. Nucleic Acids Res **23**:683-688.

Das, S., S. Paul, and C. Dutta. 2006. Synonymous codon usage in adenoviruses: influence of mutation, selection and protein hydropathy. Virus Res **117**:227-236.

de Clercq, E. 2007. The design of drugs for HIV and HCV. Nat Rev Drug Discov **6**:1001.

Dittmar, K. A., J. M. Goodenbour, and T. Pan. 2006. Tissue-specific differences in human transfer RNA expression. PLoS Genet **2**:e221.

Duret, L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet **16**:287-289.

DuRose, J. B., D. Scheuner, R. J. Kaufman, L. I. Rothblum, and M. Niwa. 2009. Phosphorylation of eukaryotic translation initiation factor 2alpha coordinates rRNA transcription and translation inhibition during endoplasmic reticulum stress. Mol Cell Biol **29**:4295-4307.

Eddy, S. R., and R. Durbin. 1994. RNA sequence analysis using covariance models. Nucleic Acids Res **22**:2079-2088.

Emerman, M., and M. H. Malim. 1998. HIV-1 regulatory/accessory genes: keys to unraveling viral and host cell biology. Science **280**:1880-1884.

Fichant, G. A., and C. Burks. 1991. Identifying potential tRNA genes in genomic DNA sequences. J Mol Biol **220**:659-671.

Fluitt, A., E. Pienaar, and H. Viljoen. 2007. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. Comput Biol Chem **31**:335-346.

Garrett, R. H., and C. M. Grisham. 2005. Biochemistry. Thomson - Brooks/Cole, Belmont, CA.

Gayle, H. D., and G. L. Hill. 2001. Global impact of human immunodeficiency virus and AIDS. Clin Microbiol Rev **14**:327-335.

Goodenbour, J. M., and T. Pan. 2006. Diversity of tRNA genes in eukaryotes. Nucleic Acids Research **34**:6137-6146.

Goujon, C., and M. H. Malim. 2010. Characterization of the alpha interferon-induced postentry block to HIV-1 infection in primary human macrophages and T cells. J Virol **84**:9254-9266.

Gouy, M., and C. Gautier. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10**: 7055-7064.

Grantham, R., C. Gautier, M. Gouy, R. Mercier, and A. Pave. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res **8**:r49-r62.

Grosjean, H., C. Marck, and V. de Crecy-Lagard. 2007. The various strategies of codon decoding in organisms of the three domains of life: evolutionary implications. Nucleic Acids Symp Ser (Oxf):15-16.

Gu, W., T. Zhou, J. Ma, X. Sun, and Z. Lu. 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res **101**:155-161.

Gultyaev, A. P., T. Franch, and K. Gerdes. 2000. Coupled nucleotide covariations reveal dynamic RNA interaction patterns. Rna **6**:1483-1491.

Haas, J., E.-C. Park, and B. Seed. 1996. Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Current Biology **6**:315.

Halwani, R., S. Cen, H. Javanbakht, J. Saadatmand, S. Kim, K. Shiba, and L. Kleiman. 2004. Cellular distribution of Lysyl-tRNA synthetase and its interaction with Gag during human immunodeficiency virus type 1 assembly. J Virol **78**:7553-7564.

Hanada, K., Y. Suzuki, and T. Gojobori. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Molecular Biology and Evolution **21**:1074-1080.

Haumont, E., M. Fournier, S. de Henau, and H. Grosjean. 1984. Enzymatic conversion of adenosine to inosine in the wobble position of yeast tRNAAsp: the dependence on the anticodon sequence. Nucleic Acids Res **12**:2705-2715.

Hickson, R. E., C. Simon, and S. W. Perrey. 2000. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. Molecular Biology and Evolution **17**:530-539.

Higgs, P. G. 2000. RNA secondary structure: physical and computational aspects. Q Rev Biophys **33**:199-253.

Higgs, P. G. 1998. Compensatory neutral mutations and the evolution of RNA. Genetica **102-103**:91-101.

Higgs, P. G., and W. Ran. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol **25**:2279-2291.

Hochsmann, M., B. Voss, and R. Giegerich. 2004. Pure multiple RNA secondary structure alignments: a progressive profile approach. IEEE/ACM Trans Comput Biol Bioinform **1**:53-62.

Huang, Y., J. Mak, Q. Cao, Z. Li, M. A. Wainberg, and L. Kleiman. 1994. Incorporation of excess wild-type and mutant tRNA₃/(Lys) into human immunodeficiency virus type 1. Journal of Virology **68**:7676.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310-2314.

Iida, K., H. Jin, and J. K. Zhu. 2009. Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in Arabidopsis thaliana. BMC Genomics **10**:155.

Ikemura, T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol **158**:573-597.

Ikemura, T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol **151**:389-409.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol **2**:13-34.

IRGSP, I. R. G. S. P. 2005. The map-based sequence of the rice genome. Nature **436**:793-800.

Javanbakht, H., R. Halwani, S. Cen, J. Saadatmand, K. Musier-Forsyth, H. Gottlinger, and L. Kleiman. 2003. The Interaction between HIV-1 Gag and Human Lysyl-tRNA Synthetase during Viral Assembly. Journal of Biological Chemistry **278**:27644-27651.

Jenkins, G. M., and E. C. Holmes. 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res **92**:1-7.

Jiang, M., J. Mak, A. Ladha, E. Cohen, M. Klein, B. Rovinski, and L. Kleiman. 1993. Identification of tRNAs incorporated into wild-type and mutant human immunodeficiency virus type 1. J. Virol. **67**:3246-3253.

Johansson, M. J., A. Esberg, B. Huang, G. R. Bjork, and A. S. Bystrom. 2008. Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. Mol Cell Biol **28**:3301-3312.

Jow, H., C. Hudelot, M. Rattray, and P. G. Higgs. 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol Biol Evol **19**:1591-1601.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110**:462-467.

Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol **53**:290-298.

Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**:143-155.

Karcher, D., and R. Bock. 2009. Identification of the chloroplast adenosine-to-inosine tRNA editing enzyme. Rna **15**:1251-1257.

Kaufman, R. J. 2002. Orchestrating the unfolded protein response in health and disease. J Clin Invest **110**:1389-1398.

Keating, C. P., M. K. Hill, D. J. Hawkes, R. P. Smyth, C. Isel, S. Y. Le, A. C. Palmenberg, J. A. Marshall, R. Marquet, G. J. Nabel, and J. Mak. 2009. The A-rich RNA sequences of HIV-1 pol are important for the synthesis of viral cDNA. Nucleic Acids Res **37**:945-956.

Khorchid, A., H. Javanbakht, S. Wise, R. Halwani, M. A. Parniak, M. A. Wainberg, and L. Kleiman. 2000. Sequences within Pr160gag-pol affecting the selective packaging of primer tRNA(Lys3) into HIV-1. J Mol Biol **299**:17-26.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol **16**:111-120.

Kimura, M. 1985. The role of compensatory neutral mutations in molecular evolution. Journal of Genetics **64**:7-19.

Kirby, D. A., S. V. Muse, and W. Stephan. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. Proc Natl Acad Sci U S A **92**:9047-9051.

Kjer, K. M. 1995. Use of Ribosomal-RNA Secondary Structure in Phylogenetic Studies to Identify Homologous Positions - an Example of Alignment and Data Presentation from the Frogs. Molecular Phylogenetics and Evolution **4**:314-330.

Kleiman, L., R. Halwani, and H. Javanbakht. 2004. The selective packaging and annealing of primer tRNALys3 in HIV-1. Curr HIV Res **2**:163-175.

Kleiman, L., C. P. Jones, and K. Musier-Forsyth. 2010. Formation of the tRNALys packaging complex in HIV-1. FEBS Letters **584**:359-365.

Knies, J. L., K. K. Dang, T. J. Vision, N. G. Hoffman, R. Swanstrom, and C. L. Burch. 2008. Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. Mol Biol Evol **25**:1778-1787.

Kramerov, D., and N. Vasetskii. 2009. Short interspersed repetitive sequences as a phylogenetic tool. Mol Biol **43**:735-746.

Kudla, G., A. W. Murray, D. Tollervey, and J. B. Plotkin. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science **324**:255-258.
Kypr, J., and J. A. N. Mrazek. 1987. Unusual codon usage of HIV. Nature **327**:20.
Larget, B., and D. L. Simon. 1999. Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. Mol Biol Evol **16**:750-759.
Lengauer, T., and T. Sing. 2006. Bioinformatics-assisted anti-HIV therapy. Nat Rev Microbiol **4**:790-797.
Lim, V. I. 1994. Analysis of action of wobble nucleoside modifications on codon-anticodon pairing within the ribosome. J Mol Biol **240**:8-19.
Lim, V. I. 1995. Analysis of action of the wobble adenine on codon reading within the ribosome. J Mol Biol **252**:277-282.
Limor-Waisberg, K., A. Carmi, A. Scherz, Y. Pilpel, and I. Furman. 2011. Specialization versus adaptation: two strategies employed by cyanophages to enhance their translation efficiencies. Nucleic Acids Res.
Lobry, J. R. 2004. Life history traits and genome structure: aerobiosis and G+C content in bacteria. Lecture Notes in Computer Science **3039**:679-686.
Lowe, T. M., and S. R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25**:955-964.
Lucks, J. B., D. R. Nelson, G. R. Kudla, and J. B. Plotkin. 2008. Genome landscapes and bacteriophage codon usage. PLoS Comput Biol **4**:e1000001.
Mak, J., M. Jiang, M. A. Wainberg, M. L. Hammarskjold, D. Rekosh, and L. Kleiman. 1994. Role of Pr160gag-pol in mediating the selective incorporation of tRNA(Lys) into human immunodeficiency virus type 1 particles. J Virol **68**:2065-2072.
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**:1189-1232.
Marquet, R., C. Isel, C. Ehresmann, and B. Ehresmann. 1995. tRNAs as primer of reverse transcriptases. Biochimie **77**:113.
Martinez, M. A., J.-P. Vartanian, and W.-H. Simon. 1994. Hypermutagenesis of RNA Using Human Immunodeficiency Virus Type 1 Reverse Transcriptase and Biased dNTP Concentrations. Proceedings of the National Academy of Sciences of the United States of America **91**:11787.
Mau, B., M. A. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55**:1-12.
Minakshi, R., K. Padhan, M. Rani, N. Khan, F. Ahmad, and S. Jameel. 2009. The SARS Coronavirus 3a Protein Causes Endoplasmic Reticulum Stress and Induces Ligand-Independent Downregulation of the Type 1 Interferon Receptor. PLoS One **4**:e8342.
Moriyama, E. N., and J. R. Powell. 1997. Codon usage bias and tRNA abundance in Drosophila. J Mol Evol **45**:514-523.
Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics **139**:1429-1439.
Nishitsuji, H., Y. Tamura, T. Fuse, Y. Habu, N. Miyano-Kurosaki, and H. Takaku. 2001. Inhibition of HIV-1 replication by 5'LTR decoy RNA. Nucleic Acids Res Suppl:141-142.
Ogle, J. M., D. E. Brodersen, W. M. Clemons, Jr., M. J. Tarry, A. P. Carter, and V. Ramakrishnan. 2001. Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science **292**:897-902.
Ogle, J. M., F. V. Murphy, M. J. Tarry, and V. Ramakrishnan. 2002. Selection of tRNA by the ribosome requires a transition from an open to a closed form. Cell **111**:721-732.
Ohta, T. 1983. On the evolution of multigene families. Theor Popul Biol **23**:216-240.
Osawa, S., A. Muto, T. H. Jukes, and T. Ohama. 1990. Evolutionary changes in the genetic code. Proc Biol Sci **241**:19-28.

Pavesi, A., F. Conterio, A. Bolchi, G. Dieci, and S. Ottonello. 1994. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. Nucleic Acids Res **22**:1247-1256.

Pavon-Eternod, M., M. Wei, T. Pan, and L. Kleiman. 2010. Profiling non-lysyl tRNAs in HIV-1. Rna **16**:267-273.

Percudani, R. 2001. Restricted wobble rules for eukaryotic genomes. Trends Genet **17**:133-135.

Percudani, R., and S. Ottonello. 1999. Selection at the wobble position of codons read by the same tRNA in Saccharomyces cerevisiae. Mol Biol Evol **16**:1752-1762.

Plotkin, J. B., H. Robins, and A. J. Levine. 2004. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A **101**:12588-12591.

Ran, W., and P. G. Higgs. 2010. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol Biol Evol **27**:2129-2140.

RC10SC, R. C. S. C. 2003. In-depth view of structure, activity, and evolution of rice chromosome 10. Science **300**:1566-1569.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16**:276-277.

Rimsky, L., J. Hauber, M. Dukovich, M. H. Malim, A. Langlois, B. R. Cullen, and W. C. Greene. 1988. Functional replacement of the HIV-1 rev protein by the HTLV-1 rex protein. Nature **335**:738-740.

Robinson, M., R. Lilley, S. Little, J. S. Emtage, G. Yarranton, P. Stephens, A. Millican, M. Eaton, and G. Humphreys. 1984. Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res **12**:6663-6671.

Rocha, E. P. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res **14**:2279-2286.

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

Rousset, F., M. Pelandakis, and M. Solignac. 1991. Evolution of compensatory substitutions through G.U intermediate state in Drosophila rRNA. Proc Natl Acad Sci U S A **88**:10032-10036.

Saksela, K., E. Muchmore, M. Girard, P. Fultz, and D. Baltimore. 1993. High viral load in lymph nodes and latent human immunodeficiency virus (HIV) in peripheral blood cells of HIV-1-infected chimpanzees. J Virol **67**:7423-7427.

SAS Institute Inc. 1989. SAS/STAT User's guide. Version 6. SAS Institute Inc., Cary, NC.

Sau, K., S. K. Gupta, S. Sau, S. C. Mandal, and T. C. Ghosh. 2006. Factors influencing synonymous codon and amino acid usage biases in Mimivirus. Biosystems **85**:107-113.

Savill, N. J., D. C. Hoyle, and P. G. Higgs. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics **157**:399-411.

Schoniger, M., and A. von Haeseler. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol Phylogenet Evol **3**:240-247.

Schoniger, M., and A. von Haeseler. 1999. Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. J Mol Evol **49**:691-698.

Scorer, C. A., M. J. Carrier, and R. F. Rosenberger. 1991. Amino acid misincorporation during high-level expression of mouse epidermal growth factor in Escherichia coli. Nucleic Acids Res **19**:3511-3516.

Senger, B., S. Auxilien, U. Englisch, F. Cramer, and F. Fasiolo. 1997. The modified wobble base inosine in yeast tRNAIle is a positive determinant for aminoacylation by isoleucyl-tRNA synthetase. Biochemistry **36**:8269-8275.

Sharp, P. M., and K. M. Devine. 1989. Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res **17**:5029-5039.

Sharp, P. M., L. R. Emery, and K. Zeng. 2010. Forces that influence the evolution of codon bias. Philos Trans R Soc Lond B Biol Sci **365**:1203-1212.

Sharp, P. M., and W. H. Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res **15**:1281-1295.

Sharp, P. M., M. S. Rogers, and D. J. McConnell. 1984. Selection pressures on codon usage in the complete genome of bacteriophage T7. J Mol Evol **21**:150-160.

Sharp, P. M., T. M. Tuohy, and K. R. Mosurski. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res **14**:5125-5143.

Shields, D. C. 1990. Switches in species-specific codon preferences: the influence of mutation biases. J Mol Evol **31**:71-80.

Siebert, S., and R. Backofen. 2005. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics **21**:3352-3359.

Sierra, S., B. Kupfer, and R. Kaiser. 2005. Basics of the virology of HIV-1 and its replication. J Clin Virol **34**:233-244.

Sipley, J., and E. Goldman. 1993. Increased ribosomal accuracy increases a programmed translational frameshift in Escherichia coli. Proc Natl Acad Sci U S A **90**:2315-2319.

Smit, A. F. A., R. Hubley, and P. Green. 1996-2010. RepeatMasker Open-3.2.9.

Sorensen, M. A., and S. Pedersen. 1991. Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. J Mol Biol **222**:265-280.

Sprinzl, M., T. Hartmann, F. Meissner, J. Moll, and T. Vorderwulbecke. 1987. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res **15 Suppl**:r53-188.

Stephan, W. 1996. The rate of compensatory evolution. Genetics **144**:419-426.

Stevenson, M. 2003. HIV-1 pathogenesis. Nat Med **9**:853-860.

Su, H. L., C. L. Liao, and Y. L. Lin. 2002. Japanese encephalitis virus infection initiates endoplasmic reticulum stress and an unfolded protein response. J Virol **76**:4162-4171.

Sueoka, N. 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of proteins. Proceedings of the National Academy of Sciences, USA **47**:1141-1149.

Sueoka, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A **48**:582-592.

Sylvers, L. A., K. C. Rogers, M. Shimizu, E. Ohtsuka, and D. Soll. 1993. A 2-thiouridine derivative in tRNAGlu is a positive determinant for aminoacylation by Escherichia coli glutamyl-tRNA synthetase. Biochemistry **32**:3836-3841.

Tang, D. T., E. A. Glazov, S. M. McWilliam, W. C. Barris, and B. P. Dalrymple. 2009. Analysis of the complement and molecular evolution of tRNA genes in cow. BMC Genomics **10**:188.

Telford, M. J., M. J. Wise, and V. Gowri-Shankar. 2005. Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the bilateria. Mol Biol Evol **22**:1129-1136.

Tillier, E. R., and R. A. Collins. 1995. Neighbor Joining and Maximum Likelihood with RNA Sequences: Addressing the Interdependence of Sites. Molecular Biology and Evolution **12**:7-15.

Tillier, E. R., and R. A. Collins. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics **148**:1993-2002.

Van Dooren, S., O. G. Pybus, M. Salemi, H. F. Liu, P. Goubau, C. Remondegui, A. Talarmin, E. Gotuzzo, L. C. Alcantara, B. Galvao-Castro, and A. M. Vandamme. 2004. The low evolutionary rate of human T-cell lymphotropic virus type-1 confirmed by analysis of vertical transmission chains. Mol Biol Evol **21**:603-611.

van Hemert, F. J., and B. Berkhout. 1995. The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol **41**:132-140.

van Weringh, A., M. Ragonnet-Cronin, E. Pranckeviciene, M. Pavon-Eternod, L. Kleiman, and X. Xia. 2011. HIV-1 modulates the tRNA pool to improve translation efficiency. Mol Biol Evol.

Vartanian, J.-P., M. Henry, and S. Wain-Hobson. 2002. Sustained G->A hypermutation during reverse transcription of an entire human immunodeficiency virus type 1 strain Vau group O genome. J Gen Virol **83**:801-805.

Watanabe, Y., H. Tsurui, T. Ueda, R. Furusihima-Shimogawara, S. Takamiya, K. Kita, K. Nishikawa, and K. Watanabe. 1997. Primary sequence of mitochondrial tRNA(Arg) of a nematode Ascaris suum: occurrence of unmodified adenosine at the first position of the anticodon. Biochim Biophys Acta **1350**:119-122.

Westhof, E., and P. Auffinger. 2000. RNA Tertiary Structure. Pp. 5222-5232 *in* R. A. Meyers, ed. Encyclopedia of Analytical Chemistry. John Wiley & Sons Ltd, Chichester.

Wong, E. H., D. K. Smith, R. Rabadan, M. Peiris, and L. L. Poon. 2010. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. BMC Evol Biol **10**:253.

Xia, X. 2001. Data analysis in molecular biology and evolution. Kluwer Academic Publishers, Boston.

Xia, X. 2000. Phylogenetic Relationship among Horseshoe Crab Species: The Effect of Substitution Models on Phylogenetic Analyses. Systematic Biology **49**:87-100.

Xia, X. 2008. The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. BMC Evol Biol **8**:211.

Xia, X. 2005. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene **345**:13-20.

Xia, X., and Z. Xie. 2001. DAMBE: Software package for data analysis in molecular biology and evolution. Journal of Heredity **92**:371-373.

Xia, X. H., Z. Xie, and K. M. Kjer. 2003. 18S ribosomal RNA and tetrapod phylogeny. Systematic Biology **52**:283-295.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol **14**:717-724.

Yilmaz, A., C. Bolinger, and K. Boris-Lawrie. 2006. Retrovirus translation initiation: Issues and hypotheses derived from study of HIV-1. Curr HIV Res **4**:131-139.

Yokoyama, S., T. Watanabe, K. Murao, H. Ishikura, Z. Yamaizumi, S. Nishimura, and T. Miyazawa. 1985. Molecular mechanism of codon recognition by tRNA species with modified uridine in the first position of the anticodon. Proc Natl Acad Sci U S A **82**:4905-4909.

Yoshida, M. 2005. Discovery of HTLV-1, the first human retrovirus, its unique regulatory mechanisms, and insights into pathogenesis. Oncogene **24**:5931-5937.

Yu, Q., D. Chen, R. König, R. Mariani, D. Unutmaz, and N. R. Landau. 2004. APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. Journal of Biological Chemistry **279**:53379-53386.

Zhang, H. M., X. Ye, Y. Su, J. Yuan, Z. Liu, D. A. Stein, and D. Yang. 2010. Coxsackievirus B3 infection activates the unfolded protein response and induces apoptosis through downregulation of p58IPK and activation of CHOP and SREBP1. J Virol **84**:8446-8459.