BEYOND THE LIAISON: EXPLORING NOVEL INSIGHTS INTO THE GENOMIC DISTRIBUTION OF tRNA AND tRNA-MEDIATED TRANSCRIPTIONAL AND TRANSLATIONAL REGULATION

A Thesis

by

BRIAN ANDREW WHITE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Rodolfo Aramayo
Committee Members,	Luís René Garcia
	James Hu
	Steve Lockless
Interdepartmental Program Chair,	David Threadgill

August 2019

Major Subject: Genetics

ABSTRACT

Transfer RNAs (tRNAs) are primeval molecules ubiquitous to all domains of life. The interactions between aminoacyl-tRNAs (aa-tRNAs) and actively translating ribosomes are critical components to the central dogma of biology as they are directly involved in the transformation of genetic code into protein. It is generally believed that this limited interaction is the extent of cooperation between tRNAs and protein-coding transcripts, however, recent findings suggest this relationship is much more complex. Using robust computational methods, we identify intact tRNA genes that intersect 79 protein-coding genes, 30 long intergenic non-coding RNA genes (lincRNA), and 11 antisense genes, among other gene types. A tRNA sequence that overlaps the interval of another gene is likely to solicit fundamental aspects of tRNA biology to the overlapped gene where they are otherwise not expected. Here, we present the hypothesis that when the interval of a tRNA gene is found to overlap the interval of another gene, the tRNA gene will introduce regulatory mechanisms that affect both the transcription and translation of the overlapped gene by various processes normally associated with tRNA biology. Furthermore, we describe an uneven distribution of tRNA genes in the human genome that reveals an acute concentration of tRNA genes that cluster with regions related to nucleosome assembly and the major histocompatibility complex (MHC). Our findings highlight the possibility that overlapping tRNA genes play a role in the transcriptional and post-transcriptional regulation of overlapped genes and these overlaps affect previously undescribed mechanisms of transcriptional and translational regulation.

Moreover, the identified clustering of tRNA genes with regions associated with nucleosome assembly and the MHC suggests tRNA biology may facilitate necessary processes to histone organization and adaptive immunology.

DEDICATION

This work is dedicated to my wife Jessica for all of her love and support, and to our son Oliver for being the best part of our lives.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Aramayo, for giving me the opportunity to work on this project, and for his patience as I adapted to the computational nature of this research. I would also like to thank my committee members, Dr. Garcia, Dr. Lockless, and Dr. Hu, for their guidance, wisdom, and support throughout this process.

Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee chaired by Dr. Rodolfo Aramayo and including members Dr. Luís René Garcia, and Dr. Steve Lockless of the Department of Biology and Dr. James Hu of the Department of Biochemistry and Biophysics.

Funding Sources

Graduate study was supported by the Texas A&M University Department of Biology Teaching Assistant awards.

NOMENCLATURE

aa-tRNA	*Aminoacyl-tRNA*
asRNA	*Antisense RNA*
ASL	*Anti-codon Stem Loop*
dsRNA	*Double Stranded RNA*
FDR	*False Discovery Rate*
FE	*Fold Enrichment*
GO	*Gene Ontology*
GTH	*Genomic Tag Hypothesis*
MHC	*Major Histocompatibility Complex*
mRNA	*Messenger RNA*
ORF	*Open Reading Frame*
SAS	*Sense AntiSense*
SNP	*Single Nucleotide Polymorphism*
sRNA	*Soluble RNA*
siRNA	*Small Interfering RNA*
TEC	*To be Experimentally Confirmed*
tRFs	*tRNA-derived Fragments*
tRNA	*Transfer RNA*
TSS	*Transcription Start Site*

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Page

1. INTRODUCTION: A BRIEF HISTORY OF tRNAs

Transfer RNAs (tRNAs) are primeval molecules ubiquitous to all domains of life. Francis Crick first hypothesized their existence in the 1950s when he postulated the need for an intermediary in the transformation of nucleic acid sequence to amino acid sequence (Crick, F. HC., 1958). Concurrent work in cell-free protein synthesis utilized immobilized enzymes that were found to attach radiolabeled amino acids to unknown RNA molecules. These labeled amino acids were later shown to have been incorporated into the polypeptide sequence of proteins (Hoagland, M. B., et al. 1956, 1957, 1958). Initially, these unknown RNA acceptor molecules were termed soluble RNA (sRNA) by Hoagland and colleagues because they were observed to be independent of the insoluble enzyme fraction. It was sRNA, which is now known as tRNA, that satisfied Crick's intermediary hypothesis just a year prior.

tRNA was subsequently confirmed as the adapter between DNA and protein and was independently discovered by several different groups within approximately two years of each other (Hoagland, M. B., et al. 1956, 1957, 1958; Ogata, K., et al. 1957; Holley, R. W. 1957). A period of robust research throughout the 1960-70s contributed to the illumination of tRNA's distinct molecular structure that, in turn, allowed researchers to ascribe functional regions, the most consequential of which was perhaps the anticodon. The nucleotide sequence contained therein enabled researchers to capitalize on the adapter quality of tRNAs in techniques used to unlock the genetic code. Today, ongoing research continues to connect tRNAs to a diverse array of unorthodox processes that often reveal characteristics of tRNA biology that are both surprising and unexpected.

Canonically, tRNAs act as the obligatory liaison between an amino acid residue and an elongating polypeptide. Under normal conditions, aminoacyl-tRNA (aa-tRNA) synthetase will ‘charge’ cytosolic tRNA molecules by attaching the cognate amino acid onto the 3’-CCA tail of the respective tRNA molecule. aa-tRNAs will then deliver the coupled amino acid to an actively translating ribosome where it will be integrated into a growing polypeptide; however, research continues to implicate tRNAs in a wide range of non-canonical biological functions. For example, aa-tRNA transferases can use a tRNA charged with a primary destabilizing amino acid (pro-N degrons) to target specific peptides by transferring the destabilizing amino acid moiety to the N-terminus of an aberrant peptide (Mogk, A., et al. 2007). This helps to eradicate particular proteins in a cellular environment by circumventing the typical function of lysosomes and vacuoles that work to degrade proteins in a non-specific manner. Additionally, tRNAs were discovered to be principal components in the lateral transmission of epigenetic information (Chen, et al. 2016). Small pieces of tRNA molecules, termed tRNA-derived fragments (tRFs; discussed below), were found to be transmitted to offspring in the head of sperm cells of mice. These offspring were shown to exhibit physiological characteristics reflective of the paternal experimental environment in lieu of their own controlled environment (Chen, et al. 2016). This study implicates tRNAs, the parental molecules required in the production of tRFs, as an intergenerational signaling mechanism that can affect developmental processes. Modern molecular techniques and increasing computational power have allowed researchers to generate a broadening catalog of disparate functionality connecting tRNA to mechanisms that far surpass the dutiful service of amino acid delivery. Here, we aim to add to this growing repertoire by uncovering additional properties of tRNA biology through the implementation of computational methods.

Our analysis is focused on four key aspects of tRNA biology; tRNA population dynamics, the genomic organization of tRNA genes, the transcriptional processes of tRNA, and the post-transcriptional modifications and structure of tRNA molecules. Using the most recent sequencing data, we provide evidence to suggest a subset of these key aspects are implicated in an otherwise undescribed approach to the transcriptional and translational regulation of certain protein coding genes. Furthermore, our findings demonstrate an acute clustering of tRNA genes within separate regions of the human genome associated with nucleosome assembly and adaptive immunology. A clear indication of conservation amongst a wide primate clade indicates a strong selective pressure to maintain this distinct distributive property.

tRNA molecules, as we understand them, are likely the direct descendants of pre-biotic chemistry that preceded life as we know it. There are certain characteristics of tRNA biology that are better understood within the context of this ancient nature. Accordingly, this work begins with a short review of the tRNA origin story, and how life has evolved to be completely dependent on their existence.

1.1. RNA World First

The quantitative characterization of life is an impossibly complex interface between physics, chemistry, and biology. Abiogenesis necessarily requires a physical and chemical description, but our current understanding of protein-based life lacks a comprehensive and cohesive quantitative explanation. The pervasive complex chemical interactions between DNA, RNA, and protein we observe in even the simplest of organisms begs the question, which one of these molecules came first? The intricacies of protein-based life coded in DNA are generally believed to be far too complex to have arisen without a preceding period of simpler, precursor molecules. This preclusion illuminate's RNA as the likely progenitor, notwithstanding some unknown or otherwise undiscovered precursor.

The RNA world first hypothesis is an elegant, albeit imperfect, estimation of how life as we know it may have arisen. The main ideas presented therein forwards the argument that it was RNA that preceded DNA and protein. Although several critical gaps in our scientific understanding of the pre-biotic chemistry responsible for the origin of life remain, current research in this field indicates the emergence of RNA, not DNA or protein, from the primordial soup (Sutherland, J. D., 2016). The myriad of naturalistic processes and chemical modifications necessary to produce the monomers required for an RNA polymer are beyond the scope of this thesis but remains an active area of research. There are several excellent articles that summarize the progress and identify the weaknesses of this leading hypothesis and our current understanding (see Joyce, G. F., 2002; Bernhardt, H. S., et al. 2012; Kua, J., et al. 2011; Orgel, L. E., 2004). Moreover, credible and justifiable objections to an RNA dominated world have been raised and previously addressed, so we will not dwell on the points here (for a review of the main criticisms and rebuttals, see Bernhardt, H. S., 2012). For the purpose of this thesis, we necessarily assume that environmental conditions are compatible with the processes required for the generation of an initial RNA polymer. To better understand the origin of tRNA, and subsequently interpret our findings within this context, we must first peer into the pre-biotic RNA world to gain some insight into the ancestral deeds of the earliest RNA molecules and how they have maintained biological relevance throughout billions of years of evolution.

1.1.1. Self-replicating RNA

A fledgling planet Earth sustained chemical and physical processes that primed the early environment for an initial RNA oligonucleotide. To be sure, this process was remarkable, but how does an unlikely molecule initiate the transformation from an RNA dominated world to a DNA/protein dominated world? In a presumed hostile environment devoid of either DNA or protein, the most probable answer requires an RNA molecule with the inherent ability to self-replicate. We know that certain RNA molecules are capable of protein independent self-replication and they have been observed to assemble themselves within thermostable protocells in the presence of single-strand amphiphiles (Lincoln, T. A., et al. 2009; Mansy, S. S., et al. 2008). Indeed, the surest way for any molecule to persist is by sustaining a rate of replication that is greater than that of degradation and by maintaining the replicative process at a high degree of accuracy. Interestingly, due to the intrinsic differential fidelity of molecular replication, a primordial RNA molecular playground ripe with self-replicating RNAs was quite possibly the womb from which Darwinian evolution was born. The imposition of selective pressures would facilitate the optimization of replication and allow populations of these RNAs to adapt to changing environmental conditions.

In an RNA dominated world, changing environmental conditions like pH, temperature, and cation concentrations (i.e., Mg^{2+}) can act to degrade single-stranded RNA molecules (Larralde, R., et al. 1995; Szostak, J. W., et al. 2012). Modern-day mRNAs overcome degradative forces for a finite period of time through the enzymatic modifications that add a 5’ cap and 3’ poly-A tail to the molecule. Methylation patterns and secondary structures also help to ensure the longevity of mRNAs in a cellular environment. In a pre-protein world, the former modifications were not available, although, single-stranded RNA molecules can assume a stable secondary structure in the absence of proteins when their sequence allows for Watson-Crick complementary base pairing. Not only can a secondary structure increase the resilience of an RNA molecule, but it is necessary to facilitate a molecular conformation that engenders the ability to accurately self-replicate (Johnston, W. K., et al. 2001). Characteristics like secondary structure and copying fidelity are excellent fodder for the omnipotent surveillance of natural selection. As such, selective pressures would likely lead to a population of molecules with increased structural integrity and accurate replicative capabilities. Molecules undergoing selection for these types of properties would indubitably persist and their proliferation would be certain among other RNAs in Earth's early molecular laboratory. This is compatible with the belief that RNAs flourished in the budding global ecosystem and it provides some insight into the ancient nature of tRNA.

1.1.2. The Genomic Tag Hypothesis

Deliberate and consistent self-replication is dependent on a stable initiation signal. The genomic tag hypothesis (GTH) describes certain characteristics that were likely critical factors in the persistence and proliferation of ancient RNA genomes. According to the GTH, a stem-loop structure on the 3'-terminus of RNA molecules not only acted as an initiation site for replication, but it also helped protect the molecules from degradation. In fact, a similar stem-loop structure was naturally developed by RNA molecules undergoing an in vitro evolution experiment to optimize self-replication (Lincoln, T. A., et al. 2009). In addition to a stem-loop structure, an important component of the GTH states that a CCA sequence added to the 3'-terminus by a catalytic RNA (ribozyme) facilitates the initiation of replication and acts as a rudimentary telomere (Weiner, A. M., 1987; Maizels, N., et al. 1999). Characteristics like these (3' stem-loop secondary structure and a CCA terminal sequence) not only provide an initiation site for replication but help the molecule evade deterioration as well. A sub-population of RNA molecules that can assume a stable secondary confirmation and contain a replication enhancing genomic tag will have a selective advantage to those that do not and will thus move towards the maintenance and optimization of these properties.

RNA replicase requires the evolutionarily optimized stem-loop 3' terminal structure as a guide for RNA synthesis. Models proposed by Weiner suggest an adventitious affinity between the active site of RNA replicases and certain amino acids (Weiner, A. M., 1987). This association makes it likely that amino acids were added to the 3' terminus of early RNA genomes that had a genomic tag and thus endowed them with a replicative advantage (Maziels, N., et al. 1999). This ensures the persistence of 3' aminoacylation and marks the likely origin of an RNA intermediate. The subsequent co-evolution of decoding DNA into protein during the transition from RNA to DNA based genomes would develop this intermediary into the tRNA molecules we are familiar with today. The ancient nature of tRNA is reflected in nuances of these primordial characteristics and may continue to play a currently undescribed role in many vital biological processes.

1.2. The Origin of tRNA

Self-replication is a major tenet of the early RNA world. Models proposed by Di Giulio, and Dick and Schamel both indicate RNA precursor polymers with the intrinsic ability to copy itself as the originators of modern tRNAs (Appendix A Figure 1A-E; Di Giulio, M. G., 1992; Dick, T. P., et al. 1995). According to this model, a precursor molecule with the attributes required for self-replication will not only generate copies of itself, but in the process will sometimes generate erroneous copies that contain an additional 3'-run-off sequence (Appendix A Figures 1A and 1B). The subsequent duplex formation of these complimentary stem-loop structured molecules results in a coaxial double stem-loop structure (Appendix A Figure 1C). As a side note, it is also possible for two independent stem-loop molecules to duplex as well (i.e., those that have complementary base pairs but have not been copied from the other). In either case, when a duplex is formed with one of the molecules that contains a run-off sequence, the model predicts the 3' end of the run-off will be ligated to the 5' end of the original stem-loop structure and then self-excised (Appendix A Figure 1D). This is a similar process observed in modern tRNA molecules that contain an intron and it is consistent with some Archaean organisms that have been observed to ligate two tRNA halves that are transcribed from various genomic loci (van Tol, H., et al. 1989; Weber, U., et al. 1996; Riepe, A., et al. 1999; Randau, L., et al. 2005A and 2005B; Fujishima, K., et al. 2009). Following the ligation of complex stem-loop structures and the excision of an extraneous run-off sequence, the resultant molecule displays the basic secondary structure of modern tRNA and consists of the antecedent functional regions (Appendix A Figure 1E).

1.3. The Anatomy of tRNA

The anatomical features of tRNA provide compelling evidence to suggest the earliest self-replicating RNAs are indeed the progenitors of tRNA. The crystal structure of tRNA indicated two perpendicular coaxial stacks that have been subsequently termed the "top half" and "bottom half" (Appendix A Figure 2A; Quigley, G. J., et al. 1976; Maizels, N., et al. 1999). The top half includes the 3'-amino acid attachment site and the T-arm, and the bottom half includes the anti-codon and D-arms. Conspicuous similarities emerge when comparing the top half of modern tRNAs to characteristics of early RNA genomes as described by the GTH. For example, the top half of tRNAs retain both a 3'-terminal stem-loop (the TψC loop of modern tRNA) and are enzymatically modified with the addition of a non-templated 3'-CCA sequence. These characteristics are not shared with the bottom half and therefore provide some evidence to suggest the top half is likely the more ancestral portion. Additionally, the top half of tRNA interacts almost exclusively with the large ribosomal subunit during protein elongation (Samaha, R. R., 1995; Green, R., 1998; Thompson, J., et al. 2001; Green, R., et al. 1997). This is the site of peptide synthesis which is generally understood to be more ancestral than decoding; a process that occurs in the small ribosomal subunit (Bokov, K., 2009). These conclusions are consistent with phylogenetic analysis based on molecular structure, as well as thermodynamic and mechanical features that also suggest the top half of tRNA is the most ancient (Sun, F., et al. 2008A and 2009A). Taken together, several independent lines of evidence suggest the top half of tRNA is, in the least, more ancestral-like with respect to the bottom half. Thus, the top half of tRNA is the likely descendent of an initial self-replicating RNA molecule.

The potential of the top half of tRNA to retain the ancestral enzymatic activity is what we are most interested in as it may help to explain observations presented later in this work. For example, tRNA molecules undergo several post-transcriptional modifications (discussed in detail later), but a potentially consequential modification worth mentioning here is the splicing of a tRNA molecule at the anticodon loop by the ribonuclease angiogenin. This results in two tRNA halves; the 5'-half and the 3'-half (distinct from the top and bottom halves). The 3'-half is of particular interest because it retains both a stem-loop structure and the 3'-CCA terminal sequence similar to ancient self-replicating RNA molecules. The 3'-CCA terminal sequence is not present in the 5'-half. Given the proposed ancestry of tRNA, if 3'-halves retain enzymatic activity, we can reasonably implicate them in critical biological processes that have, to our knowledge, not been previously described.

Broadly defined, genes are sequences of nucleotides that code for some function and are under some form of regulation. For the purpose of this thesis, we will be using the term 'gene' to describe any principal feature defined as such by the respective database from which the feature is included. For example, Gencode.v28 defines protein coding genes based on annotations from Ensembl or Havana, or both (Appendix B Table 1). A genomic region annotated as a protein coding gene can also code for a variety of transcripts known as alternative transcripts or isoforms. The number of isoforms transcribed from the genic region can vary widely and do not necessarily have the same function as the primary gene (i.e., isoforms may code for a different functioning protein or not code for protein at all). Throughout this thesis, when we refer to a 'gene,' we are referring to the genomic interval in which the primary sequence of the feature exists. Writ large, we will not consider isoforms that derive from these defined regions as separate from the primary gene interval unless otherwise noted. Furthermore, gene types are highly variant and are primarily defined by their observed, or putative function (Appendix B Table 1). For an exhaustive list of gene types in the human genome (GRCh38.p12), as defined by Ensembl and Havana, see Supplemental 1 Table 1.[1]

[1] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

2. TRNA BIOLOGY

tRNA biology is an incredibly broad and encompassing subject that aims to describe the complexities that manifest at the interface of tRNA mechanisms and genetic code. Given the enormous scope of the subject matter contained therein, we will necessarily narrow our focus to just a few of the most fundamental characteristics of tRNA biology and how they are both biologically relevant and support the hypotheses forwarded later in this thesis.

This chapter begins with a qualitative characterization of the dynamic nature of tRNA populations as a response element to environmental stimuli and a driver of variant cell state conditions (i.e., proliferation or disease state). We then evaluate the genomic organization of tRNA genes in humans and reveal an evident uneven distribution. Later in this work we highlight some of the functional consequences of this observed genomic distribution and expand our analysis to discuss an evident conservation. Finally, we review the transcriptional processes of tRNA genes along with the many post-transcriptional modifications tRNA transcripts are subject to. Many of these modifications are implicated in the structural integrity of mature tRNA molecules and are critically important to translational efficiency. Juxtaposed to this, other modifications deliberately degenerate tRNA molecules into small fragments that can act to stifle translation.

2.1. tRNA Population Dynamics

Under normal conditions, the translational needs of a given cell type are accommodated, in part, by the sufficient availability of tRNAs. However, the regulation of tRNA molecules in a cellular environment is dynamic. tRNA populations can respond to environmental stimuli, act as a marker for certain diseases, and affect the fate of a given cell type.

Prokaryotes and eukaryotes often encounter common environmental stressors like nutrient starvation or hypoxia. Bacterial cells experiencing amino acid starvation can use uncharged tRNAs as effector molecules in a pathway that impedes translation while simultaneously promoting the transcription of genes related to amino acid synthesis (Haseltine, W. A., et al. 1973; Sy, J., et al. 1973; Ross, W., et al. 2013). In contrast, yeast cells under nutrient starvation conditions will limit cytosolic translation by enacting mechanisms that facilitate the shuttling of tRNA molecules into the nucleus where they are unable to deliver amino acids to the translational machinery (Whitney, M. L., et al. 2007). Given the divergent evolutionary trajectory of prokaryotes and eukaryotes, it is not surprising that they have evolved diverse strategies to overcome similar environmental stressors, but it is, at least interesting that each have developed coping mechanisms that implement the manipulation of tRNA populations.

Multicellular organisms have variant tissue types that require a distinctive protein constituency. Evidence of modulating tRNA expression to accommodate the amino acid needs of a given cell type is presented by a strong correlation between the expression of tRNA genes and the translational needs of a given cell type or cell fate (i.e., differentiation or proliferation) (Dittmar, K. A., et al. 2006; Gingold, H., et al. 2014). When the regulatory mechanisms that control tRNA populations break down, the consequences can be deadly. For example, a positive association between excessive tRNA expression and codon usage was demonstrated when looking at genes involved in the development and growth of tumors in human breast cancer cell lines (Pavon-Eternod, M., et al. 2009). This suggests cellular tRNA abundance can both be a driver and a marker of certain breast cancers (Pavon-Eternod, M., et al. 2009). Adversely, when a tRNA species is erroneously down regulated, the translation of transcripts coding for the amino acid specific to that tRNA will be delayed. If these transcripts code for proteins that are necessary to critical cellular processes, the consequences of delayed translation can be detrimental to a cell.

Translational efficiency can be modulated by tRNAs through the regulation of cognate aminoacyl tRNA (aa-tRNA) synthetase transcripts. aa-tRNA synthetases are the enzymes responsible for charging tRNA molecules with their cognate amino acid. Gram-positive bacteria can use uncharged tRNAs to regulate the expression of aa-tRNA synthetase genes through interactions between the uncharged tRNA and the 5' untranslated region (5' UTR) of the aa-tRNA synthetase transcript (Nelson, A. R., et al.

2006). This interaction acts as a control on the expression of specific aa-tRNA synthetases due to the anticodon specificity of this interaction. In a situation in which there is an overabundance of a certain tRNA species, the uncharged tRNAs of this species will bind the aa-tRNA synthetase transcript that codes for the enzyme responsible for charging it. Conversely, when the abundance of a certain tRNA species is low, there will be less uncharged tRNAs available to bind to its respective aa-tRNA synthetase transcript. As a result, the translation of this transcript is more likely to occur and the subsequent charging of the low abundance tRNA molecules will commence. To our knowledge, this type of translational control is not observed in eukaryotes, although in humans, a component of a multi aa-tRNA synthetase complex has been shown to associate with other proteins to silence the translation of ceruloplasmin transcripts by associating with the 3' UTR (Sampath, P., et al. 2004). Regardless, the regulation and manipulation of tRNA populations in both prokaryotes and eukaryotes is not only a critical environmental response mechanism, but it also plays an important role in cell state, cell differentiation, and translational efficiency.

2.2. Genomic Organization of tRNAs

A fundamental characteristic in genetics, with respect to genomic structure and organization, is the distribution of genes in a given genome. Prokaryotes often cluster co-expressed genes into operons that usually produce interacting proteins (Dandekar, T., et al. 1998). Linking genes that code for interacting proteins will likely facilitate the simultaneous expression of each gene and help to ensure the physical interactions of the resultant proteins. Thus, there is a selective advantage in the maintenance of linkage disequilibrium when the interaction of the encoded proteins is necessary. In Eukaryotes, however, operons are not widely utilized, and the null expectation is that genes are evenly distributed throughout the genome. With the exception of housekeeping genes (e.g., typical constitutive genes related to expression, metabolism, cell surface, etc.), little clustering is observed in the human genome, although this does not preclude tissue-specific clustering (Lercher, M. J., et al. 2002). It is probable then, that any apparent clustering of genes in the human genome indicates fundamental cellular processes and evolutionarily conservation.

Karyology is the study of whole sets of chromosomes and works to organize them by pairs (in diploid organisms) and length, or by the location of the centromere if more than one pair of chromosomes are the same length. The somatic chromosomes in humans are numbered following this convention. For example, the longest somatic chromosome in humans is named chromosome 1. The remaining somatic chromosomes are numbered sequentially by decreasing length, with a couple modest exceptions. Chromosome 11 is about 1.3 million nucleotides longer than chromosome 10, chromosome 20 is about 5.8 million nucleotides longer than chromosome 19, and chromosome 22 is about 4.1 million nucleotides longer than chromosome 21 (Appendix B Table 2). The sex chromosomes are not numbered in the same manner as somatic chromosomes. Instead, the sex chromosomes are differentiated from the somatic chromosomes by lettering rather than numbering. In humans, the sex chromosomes are named X and Y.

In the haplotype of eukaryotes where n > 2, we expect to see a greater abundance of genes located in the longer chromosomes rather than the shorter chromosomes because of the greater potential to code for genes in the longer chromosomes. In general, there does tend to be more total genes in the longer chromosomes compared to the shorter chromosomes, however, this pattern is not observed with tRNA genes (Appendix B Table 2). As expected, chromosome 1 has the most tRNA genes (149), but chromosomes 2, 3, 4, and 5 all have less tRNA genes than other, much shorter chromosomes (Appendix B Table 2). For example, chromosome 4 has only 2 tRNA genes whereas chromosome 17 has 41 despite being 107 million nucleotides shorter (Appendix B Table 2).

2.3. tRNA Transcription

In Eukaryotes, RNA polymerase III (RNA-pol III) transcribes various types of small RNAs including tRNAs. There are three types of promoters recognized by RNA-pol III. Types-1 and -2 have intragenic promoter elements and do not contain a TATA box whereas type-3 promoters can have distal and proximal sequence elements upstream of the transcription start site (TSS) and do contain a TATA box. RNA-pol III genes have a terminal poly-T sequence that facilitates the termination of transcription.

tRNA genes have a type-2 promoter. There are two intragenic promoter elements called the A- and B-box (Appendix A Figure 2B; Schramm, L., et al. 2002; Galli, G., et al. 1981; Hofstetter, H., et al. 1981; Sharp, S., et al. 1981; Allison, D. S., et al. 1983). Sequence conservation of the A- and B-box elements amongst tRNA genes has been repeatedly observed, although the spacing between them can be variant. The conservation of these regions is most likely because they form the functional D- and TψC-loops of mature tRNAs (Appendix A Figure 2A; Schramm, L., et al. 2002). In humans, the transcription of tRNA genes begins when the six-subunit transcription factor TFIIIC binds to the A- and B-box intragenic promoter region (Dumay-Odelot, H., et al. 2007). Once bound, the three-subunit TFIIIB is subsequently recruited (Lassar, A. B., et al. 1983; Bieker, J. J., et al. 1985; Setzer, D. R., et al. 1985). When TFIIIC and TFIIIB are complexed, the 17-subunit RNA-pol III is enlisted, mainly through protein-protein interactions with TFIIIB and possibly TFIIIC. (Dumay-Odelot, H., et al. 2007; Schramm, L., et al. 2002). After RNA-pol III is bound, the assembly of the elongation complex is complete and transcription of the tRNA gene will proceed.

Two independent processes are required to complete the transcription of tRNA genes. The first step occurs when the RNA-pol III complex stalls on the poly-T termination sequence. The elongation complex becomes enzymatically inoperative and begins to backtrack. Embedded secondary structures in the body of the tRNA transcript act to dissociate the elongation complex from the template (Nielsen, S., et al. 2013). There are, however, exceptions to this process.

RNA-pol III has been shown to read-through many tRNA poly-T terminators (Turowski, T., et al. 2013). Rather than stalling and backtracking, RNA-pol III will sometimes continue transcribing through the poly-T sequence generating long 3' extended transcripts. Various mechanisms have been implicated in the generation of RNA-pol III read-through transcripts, for example, mutations that disrupt the poly-T sequence (Schramm, L., et al. 2002). Moreover, NF1 polypeptides are a family of proteins that can associate with the TFIIIC1 fraction of the RNA-pol III elongation complex and can play a role in the termination of transcription by binding specificity in a region downstream of the poly-T sequence (Schramm, L., et al. 2002). Mutations in either the NF1 polypeptide or the sequence recognized by them, can also result in read-through transcripts (Schramm, L., et al. 2002).

2.4. tRNA Structure and Modifications

The structure of tRNA is essential for proper function during the canonical process of translation. The primary structure of a processed tRNA transcript is a relatively short length of approximately 76-90 nucleotides (Appendix A Figure 2B; Sharp, S. J., et al. 1985). Differences in lengths are due to a variable region between the anticodon and TψC loops, while some tRNAs, like tRNA^{Ser}, tRNA^{Leu}, and tRNA^{Sel}, which is the longest, have an extra arm between these loops (Appendix A Figure 2A; Itoh, Y., et al. 2013). tRNAs have a distinct cloverleaf secondary structure with discrete functional regions that include the 5’ phosphate group, D-loop, anticodon loop, TψC loop, and the acceptor stem (Appendix A Figure 2A). Conserved and semi-conserved residues in the D- and TψC-loops facilitate the functional tertiary L-shape structure necessary for the integration and conformational plasticity when interacting with the ribosomal A, P, and E sites during the elongation process of translation (Giegé, R., 2008). tRNA molecules undergo an extensive post-transcriptional modification regime to ensure proper function during the many molecular interactions encountered during aminoacylation and translation (Agris, P. F., et al. 2007; Helm, M., 2006). These modifications to are not only necessary to ensure a stable molecular conformation, but they also expand the cognate amino acid repertoire and help maintain the structural integrity of the molecule as well.

At the time of this writing, 111 post-transcriptional RNA modifications have been identified (Agris, P., et al. 2019). Of these, at least 92 modifications (~83%) have been shown to occur in tRNAs and are initiated soon after transcription. (Agris, P., et al. 2019). The modification processes transform the transcript from a precursor tRNA (pre-tRNA) to a mature tRNA beginning with the removal of the 5' leader and 3' trailer sequences by RNase P and RNase Z respectively (Phizicky, E. M., et al. 2010; Frank, D. N., et al. 1998; Maraia, R. J., et al. 2011). An untemplated CCA sequence is then enzymatically added to the 3' end by a nucleotidyl transferase protein. This conserved sequence addition is required for the aminoacylation of tRNA by aa-tRNA synthetase and is the final step of the maturation process. The matured tRNA molecule is escorted out of the nucleus and into the cytoplasm where it will undergo, on average, 13 additional modifications (Maraia, R. J., et al. 2011). Amongst tRNA molecules, position 34 of the anticodon stem loop (ASL) is the most frequently modified base. This is known as the wobble position because the modifications here allow for differing species of tRNA molecules to deliver the same amino acid residue to alternate codons. These modified tRNAs are known as isoacceptors and account for the degeneracy of the genetic code by enabling different codon-anticodon specificity for the same amino acid. Furthermore, the modification of anticodon nucleosides can induce codon bias by altering the affinities to cognate-codons. Thus, mRNAs enriched with favored codons are preferentially expressed. This implicates tRNA modifications in the regulation of gene expression (Duechler, M., et al. 2016). Nearly all other modifications to tRNA molecules are structural in nature, and along with secondary and tertiary contacts, engender tRNA with a robust stability seldom observed in any other RNA molecule (Gebetsberger, J., et al. 2013). Paradoxically, there is an additional set of modifications that have an entirely opposite effect.

tRNA derived fragments (tRFs) are small pieces of tRNA molecules that are known to be consistently and deliberately produced. Advancements in extraction and high-throughput sequencing technologies lead to the discovery of short RNAs (< 40 nt) that fueled a wave of interest focused on characterizing these populations of short, non-coding sequences. For a long time, tRFs were regarded as random products of degradation and were literally washed away, however, an increasing body of research implicates them in specific biological processes and demonstrates they are as ubiquitous as their tRNA progenitors and their biogenesis is separate from miRNA (Kumar, P., et al. 2014). The modification of tRNA molecules into tRFs is a relatively recent area of study but has already amassed a profusion of literature. There are several properties of tRFs that are relevant to this thesis, but the topic is much too broad to cover with any due justice here (for excellent reviews on tRFs see; Kumar, P., et al. 2014; Fu Y., et al. 2015; Keam. S., et al, 2015).

tRNA biology is an immense topic that covers a wide breadth of relevancy. The scope of our discussion is limited to four fundamental aspects that include tRNA population dynamics, genomic organization, transcription and translation, and the structure and modification of tRNA. The regulatory pathways of tRNA populations continue to illuminate the role tRNAs have in critical non-canonical functions that are continuously being discovered with unprecedented resolution. For example, specialized tRNAs can act as primers during reverse transcription and, specific to prokaryotes, aa-tRNAs capable of ribosome independent peptide formation were found to be involved in the biosynthesis of peptidoglycan, as well as antibiotics and resistance pathways (Marquet, R., et al. 1995; Mak, J., et al. 1997; Sheppard, J., et al. 2013). Here, we provide evidence to suggest the distribution of tRNA genes in the human genome is non-random and implicated in transcriptional and post-transcriptional processes that affect the expression of certain genes. Furthermore, we hypothesize that the uneven distribution of tRNA genes in the human genome is related to genomic structure and adaptive immunology.

The experimental validation of each tRNA gene in a given genome is often beyond the practical limitations of research laboratories. An easier way to characterize genomic tRNAs is to implement well established software programs designed explicitly for this purpose. tRNAscan-SE is putatively the most popular tRNA prediction program and is both efficient and accurate (Pavesi, A., et al. 1994). It identifies putative tRNA genes by searching a sequence query with a tRNA model that has been trained on known tRNAs specific to phylogenetic groupings (i.e., mammals, Archaea, or Bacteria). It also allows the user to customize output options tailored for specific purposes. For instance, tRNAscan-SE can be configured to generate output files in BED format, allowing the user to visualize the predicted tRNAs in a genome browser. tRNAscan-SE can also output FASTA files that can be used to align and analyze the sequences of predicted tRNAs. General output files summarize the scan and include key information and statistics on each predicted tRNA sequence.

Following a relatively permissive first-pass scan, tRNAscan-SE performs a more stringent second-pass that predicts the secondary structure of tRNA by the implementation of Infernal v1.1; a covariance model search engine that will score DNA sequence based on the consensus of sequence alignment and secondary structure (Nawrocki, E. P., et al. 2013). Pragmatically, Infernal scores > 50 indicate robust tRNA genes that are likely to assume the canonical cloverleaf secondary structure and thus are assumed to participate in translation. Moreover, tRNAscan-SE will typically define some tRNA predictions as pseudogenes. The program considers these sequences atypical variants in which the secondary structure lacks the usual conserved features found in typical tRNAs. These variants are not known to function in translation, however, their participation in non-canonical functions have been observed (Rogers, T. E., et al. 2012).

tRNAscan-SE has been in use for over two decades and has built a reputation for accuracy and reliability, however, like any algorithm, it has its limits. tRNAscan-SE is unable to identify tRNA sequences that are split within a genome. The discovery of archaeon *Nanoarchaeum equitans* in 2005 and *Caldivirga maquilingensis* in 2009 highlight this challenge as they each contain tRNA isoacceptors that are products of two independently transcribed sequences that are subsequently ligated (Randau, L., et al. 2005A and 2005B; Fujishima, K., et al. 2009). As such, any such occurrences of split tRNAs have not been identified in our analysis and will not be considered in our conclusions.

The pervasiveness of tRNAs amongst all living things should not be underappreciated. They are the likely derivatives of Earth's earliest molecules and preceded life as we know it. Whether tRNAs are destined to deliver an amino acid to a ribosome or participate in a non-canonical pathway, we are persistently reminded that our comprehensive understanding of this ancient and dynamic molecule is incomplete and there is likely much more to uncover.

3. METHODS

tRNA genes were predicted using tRNAscanSE-2.0 on the FASTA file of each respective taxa (Pavesi, A. et al. 1994). The FASTA file corresponding to the human genome was downloaded from the Gencode database (GRCh38.p12; Gencode.v28). The FASTA files corresponding to the five other primates analyzed here were downloaded from the Ensembl database Release 95 via FTP and correspond to the following assemblies; bonobo (panpan.1), chimp (Pan tro 3.0), gorilla (gorGor4), orangutan (PPYG2), and macaque (Mmul 8.0.1). The FASTA files that correspond to the model organisms analyzed here were downloaded from the Ensembl database Release 95 via FTP and correspond to the following assemblies; mouse (GRCm38), fruit fly (BDGP6), and nematode (WBcel235). For each implementation of tRNAscan-SE, any tRNA genes that were called from the sequences of contigs or scaffolding included in the FASTA files were not included in our analysis. The output tRNA gene intervals do not include the 5' leader or 3' trailing sequences. The intervals used in our analysis range from the 5' phosphorus to the 3' terminus of the processed tRNA transcript. For those tRNAs that have a retained intron, the intronic sequence is included in the interval.

To determine whether or not tRNA genes intersect the interval of features annotated as genes, all of the entries defined as 'gene' in the third column of the GFF3 file were extracted and converted to a BED file (Appendix B, Supplemental 1 s.1 and s.2).[2] This eliminated redundant intersect counts that would result in overlapping transcripts of some genes. Two BED files, one containing all features annotated as a gene and the other is the output BED file from tRNAscan-SE-2.0, were loaded to Reveille; Texas A&M's implementation of the Galaxy software framework. In Reveille, dataset 1 was the BED file with all the features annotated as 'gene,' and dataset 2 was the BED file with all of the tRNA genes predicted by tRNAscan-SE. To identify any possible intersections of predicted tRNA genes with regions annotated as genes, the Join tool (v.1.0.0) was used to return only the overlapping intervals (inner join). The output of this operation was grouped by name (column 4) using the Group tool (v.2.1.0) and a count function was added. This step allows us to consolidate into a single line instances in which more than one tRNA gene intersect a unique feature annotated as a gene. It also provides a summation of these intersects to indicate possible redundancy. This file was then joined side by side with the dataset 1 BED file using the name column from each (column 4 from dataset 1 and column 1 from dataset 4). This was performed using the Join Two Datasets tool (v.2.0.1). The columns from this output were then reordered using the Cut (reorder) tool (v.1.0.2) to conform to the format specifications of a BED file. The resultant BED file was then loaded into IGV (2.3.82) for visual inspection of the reported intersections. Intersect analysis was independently repeated with equivalent

[2] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

genomic files from the RefSeq and Ensembl databases in an effort to eliminate false positives.

In order to visualize the linear distribution of tRNA genes in the human genome, the sequence length of each chromosome was calculated from the FASTA file using the Biostrings package (v.2.48.0) in RStudio (v.1.1.383; R v.3.5.1; Appendix B, Supplemental 1 s.3).[3] To generate an ordinal vector of tRNA gene positions for the human genome, 1 was subtracted from the start position of each tRNA gene in chromosome 1 and divided by the length of the genome. For chromosome 2, the length of chromosome 1 was added to the start position of each tRNA gene, 1 was subtracted from the start position of each tRNA gene and divided by the length of the genome. This was repeated for the remaining chromosomes such that the sum of the preceding chromosome lengths was added to the start position of each tRNA gene in the respective chromosome, 1 was subtracted from the start position of each tRNA gene in a respective chromosome and the length of the genome was divided out. A histogram of the resultant ordinal vector was plotted. The same logic was followed when plotting the tRNA gene distribution for chromosomes 1 and 6 in humans and primates and when plotting tRNA loci in different MHC assemblies.

[3] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

Protein coding genes within the regions of human chromosomes 1 and 6 in which the density of tRNA genes was highest (chr1:143,486,629-150,098,821 and chr6:26,240,093-29,022,932 respectively) were extracted using the BioMart server (Appendix B, Supplemental 1 s.4).[4] The parameters used identified 171 unique UniProtKB IDs within the region of chromosome 1 and 153 IDs within the region of chromosome 6. These UniProtKB IDs were compiled into a list and submitted to the Panther Classification System web server for a statistical overrepresentation test (Appendix B, Supplemental 1 s.5; Mi H., et al. 2016).[5] All parameters were set to default settings and the Annotation Data Set was set to 'PANTHER GO-Slim Biological Process.' To control for gene density and statistical enrichment, chromosomes 1 and 6 were split into intervals containing 171 and 153 protein coding genes respectively. These two values correspond to the number of protein coding genes within the intervals of chromosomes 1 and 6 that have the densest tRNA gene clusters (see above). These intervals were sorted randomly and the first three were selected and piped through the analysis workflow described above (Appendix B, Supplemental 1 s.6).[6]

[4] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

[5] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

[6] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

4. COMPUTATIONAL CHARACTERIZATION OF tRNA

Our implementation of tRNAscan-SE 2.0.0 has predicted 636 total tRNA genes in the human genome (HGRCh38.p12; Gencode.v28). Of these, 107 are classified by the program as pseudogenes and 189 have an Infernal v1.1 score of < 50. Just less than 3% of the predicted pseudogenes have an Infernal score >50, indicating the likelihood that greater than 97% of pseudogenes do not function in translation or assume the distinct cloverleaf secondary structure, although, they may participate in non-canonical pathways as previously observed. Eliminating pseudogenes and those predictions which have an Infernal scores < 50 results in the most conservative estimate of predictions totaling 445 tRNA genes in the human genome. Taken together, when the secondary structure of tRNA is implicated in proposed functionality, pseudogenes will not be included in our analysis, however, we do include them in our overall analysis with respect to genomic distribution and abundance due to the implication of their involvement in non-canonical biological pathways.

We compared the predictions made by tRNAscan-SE to data mined from several databases that identify tRNA genes in the human genome. Those reported here indicate a similar amount of tRNA genes with the exception of tRNAdb (Appendix B Table 3). UCSC, tRNAscan-SE (the database), and tRFdb report an average of 625 tRNA genes while tRNAdb reports well below this average at 359 tRNA genes. It is unclear why tRNAdb reports 266 less tRNA genes than the average of the other three databases, although the server hosting the data for tRNAdb does not indicate the date of the latest update and the most recent article describing tRNAdb was published over ten years ago (Jühling, Frank, et al., 2008).

In addition to comparing tRNAscan-SE results to the tRNA genes reported in these databases, we utilized another popular tRNA search program called Aragorn (Laslett, D., et al., 2004). Our implementation of Aragorn predicts just over 30% more tRNA genes than tRNAscan-SE (Appendix B Table 3). A paired t-test indicates a significant difference between the number of tRNA sequences predicted by tRNAscan-SE and Aragorn (*p* = 2.1e-9), although the per chromosome abundance patterns are very similar, (Supplemental 1 Figure 1).[7] Aragorn employs a heuristic algorithm to predict tRNA secondary structure which is more efficient and runs faster than the more stringent covariance modeler used by tRNAscan-SE. As a result, Aragorn is less constrained than tRNAscan-SE and is likely to predict more tRNA sequences than tRNAscan-SE. For our purposes, using the more conservative set of tRNA gene predictions as output from tRNAscan-SE increases our confidence that the inferences we make are less likely to be based on false positives.

[7] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

4.1. Genomic Distribution of tRNA Genes

When considering the manner in which genes are distributed throughout a given genome, the null expectation is an even distribution. Accordingly, longer chromosomes are predicted to contain more genes than shorter chromosomes. We tested this prediction by first plotting the correlation between all genes and chromosome length in the human genome. We observe a moderate positive correlation (r^2=0.57) indicating the longer chromosomes generally contain more genes than the shorter chromosomes (Appendix A Figure 3A).

Next, we plotted the correlation between the number of tRNA genes per chromosome and chromosome length. We found this correlation to be very weak in comparison (r^2=0.19). This suggests, with respect to tRNA genes, the longer chromosomes do not necessarily contain more tRNA genes than the shorter chromosomes (Appendix A Figure 3B). Our linear model indicates chromosomes 1 and 6 as the two statistical outliers driving the correlation coefficient down with respect to the distribution of tRNA genes in the human genome. Clustering of tRNA genes in these chromosomes has been previously described and is further analyzed in Chapter 4.2 (Mungall, A. J., et al. 2003).

In order to visualize the genomic distribution of tRNA genes in the human genome, we generated an ordinal vector of start positions for each predicted tRNA gene and plotted a histogram in which each chromosome is a unique bin color (Appendix A Figure 4).

Confirming the indications of the linear model, we can see a clear enrichment of tRNA genes in chromosomes 1 and 6 with the most noticeable concentration occurring in chromosome 6.

There are 149 tRNA genes in chromosome 1 and 188 tRNA genes in chromosome 6. To put these counts into perspective, chromosome 17 has the next most abundant tRNA gene count of 41. A closer look at the distribution of tRNA genes in chromosome 1 reveals a dense cluster within a region of about 6,600kb (chr1:143,486,629-150,098,821). There are 66 tRNA genes within this region. Comparatively, there is a cluster of 165 tRNA genes in chromosome 6 within a region spanning just over 2,700kb (chr6:26,240,093-29,022,932). There are nearly 3 times the number of tRNA genes in chromosome 6 that are grouped within a region that is about 2.5 times smaller than chromosome 1. These dense clusters indicate an uneven distribution of tRNA genes in the human genome.

In an effort to determine if the observed clustering of tRNA genes is evolutionarily conserved, we broadened our analysis to include three popular and well annotated model organisms; mouse (*Mus musculus*), fruit fly (*Drosophila melanogaster*), and nematode (*Caenorhabditis elegans*). Amongst these genomes, we did not find an acute concentration of tRNA genes in a particular chromosome that approaches the magnitude observed in the humans (Appendix A Figure 5). The dense concentrations of tRNA genes appears to be unique to humans. Below, we perform the same distributive analysis on a select primate clade, but first, we perform a more resolute examination of human chromosomes 1 and 6 within the region of this apparent tRNA gene enrichment.

4.1.1. Clustering of tRNA Genes and Genes Associated with Nucleosome Assembly and Adaptive Immunology

The eukaryotic nucleosome is a stretch of DNA (~147 bp) that is coiled around a spool-like protein octamer called histone. The assembly of nucleosomes is essential to the overall stability of the genome and is intimately involved in the regulation of gene expression; however, the regulatory pathways remain unresolved (Ransom, M., et al. 2010; Groth, A., et al. 2007). The dynamic nature of nucleosomes is revealed by the transient associative fluctuations of DNA from the nucleosome core that frequently shift between loose and tight associations (Polach, K. J., et al. 1995; Anderson, J. D., et al. 2000). These oscillations allow just enough time for high affinity DNA binding factors to bind and impedes those factors with a lower affinity (Polach, K. J., et al. 1995; Anderson, J. D., et al. 2000). The histone components of nucleosomes provide a substantial framework for epigenetic markers as they are subjected to several types of modifications (i.e., acetylation, methylation, phosphorylation) that carry an enormous regulatory potential. For example, the acetylation of a histone tail will alter the affinity of DNA to the nucleosome core such that the bonds between the two are relaxed and transcription factors can bind.

The region of human chromosome 1 in which the tRNA gene density is highest also harbors a statistical overrepresentation of gene ontology terms related to the biological processes of nucleosome assembly, protein folding, and peptidyl-amino acid modification (Appendix B Table 4). The overrepresentation of these gene ontology terms does not necessarily mean there are interactions between the genes associated with the ontology terms and the proximate tRNA genes, although, based on previous observations, gene clustering can facilitate the coordinated transcription and interaction of gene products (Thompson, M., et al. 2003). For example, we know that aa-tRNA transferases can transfer amino acids from charged tRNAs to the N-terminus of a peptide, but to our knowledge, the addition of an amino acid onto histone tails mediated by charged tRNAs is not known (Mogk, A., et al. 2007). The tRNA-mediated addition of amino acids onto histone tails would engender the nucleosome core with additional material that can be further modified, thus imposing an additional, and yet undescribed, regulatory mechanism. We expect to see this kind of tRNA-mediated histone modulation throughout the genome, however, the clustering of tRNA genes and histone genes observed here may facilitate this type of interaction.

The region of chromosome 6 in which we observe a dense cluster of tRNA genes also overlaps with statistically overrepresented gene ontology terms associated with nucleosome assembly as well as adaptive immunology (Appendix B Table 4). A sub-cluster of tRNA genes in this region overlaps with a relatively small portion of the major histocompatibility complex (MHC; Supplemental 1 Figure 2).[8] The MHC is a collection of genes that code for proteins responsible for binding and presenting epitopes on the cell surface for T-cell recognition. The presentation of the epitope is necessary for lymphocytes to differentiate between self and non-self. The clustering of tRNA genes and MHC genes observed here was recently and independently corroborated by Tao Pan (Pan, T., 2018).

MHC genes are considered to be the most polymorphic of all genes and are only found in the jawed vertebrates. There are three gene classes associated with the MHC. Class I molecules present peptide fragments that come from either the nucleus or from the cytoplasm and are present on all nucleated cells and platelets. Class II molecules present peptide fragments from vesicles within the cell and are derived from cytosolic or extracellular proteins. Class III MHC molecules do not present epitopes; however, they are involved with facilitating the efficiency of immune response as well as cellular stress response. The area of the MHC in which we observe a cluster of tRNA genes is populated exclusively by Class I genes (Vandiedonck, C., 2009).

A high level of allelic polymorphism is repeatedly observed in the MHC region and indicates the struggle for pathogens to evade detection and the adaptive immune system’s ability to surveil and identify those threats (Beck, S., et al. 2000; Trowsdale, J.,

[8] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

et al. 2013; Bernatchez, L., et al. 2003; Spurgin, L. G., et al. 2010). Interestingly, in an apparent attempt to ensure heterozygosity at the MHC locus, research has shown a correlation with odor preference and mate choice amongst human individuals with MHC loci dissimilar to one another (Wedekind, C., et al. 1995; Yamazaki, K., et al. 1976 and 1979; Ober, C., et al. 1997). We are not aware of any described mechanism that is responsible for the maintenance of allelic polymorphism at this locus that implicates the proximity to tRNA genes, although, the sexual selective pressure to maintain heterozygosity at the MHC region highlights the selective proclivity to make certain this area remains inordinately variant.

The genomic intervals that contain the tRNA gene clusters for chromosomes 1 and 6 contain 171 and 153 protein coding genes respectively. In an effort to determine if the proximate localization of the previously described tRNA gene clusters with genes associated with nucleosome assembly and adaptive immunology are unique to these intervals or are found throughout each respective chromosome, three random intervals were generated and analyzed for gene ontology enrichment (see Methods). All but one of these random intervals did not show a statistical overrepresentation of gene ontology terms related to nucleosome assembly or adaptive immunology for either chromosome. The single exception was a randomly generated interval (chr6:29,555,515-31,446,973) that happens to map to a region within the MHC. In this case, the statistical overrepresentation of gene ontology terms related to adaptive immunology is expected. Interestingly, this random interval is just over 500,000 nucleotides to the 3' boundary of the tRNA gene cluster within the MHC. This region has statistically overrepresented gene ontology terms related to the positive regulation of immune response (88.9-fold enrichment; p=2.2e-7; FDR=4.0e-4) whereas the interval that contains the tRNA gene cluster 5' to this random interval has a 99.9-fold enrichment for T cell receptor signaling pathway gene ontology terms (Appendix B Table 4). It is unclear whether or not there is a functional relationship between tRNA biology and T cell receptor signaling pathways that is not utilized for the positive regulation of immune response. What is clear, however, is the observation that a significant enrichment of gene ontology terms related to nucleosome assembly and adaptive immunology share a proximate distribution with dense tRNA gene clusters in chromosomes 1 (i.e., nucleosome assembly) and 6 (i.e., nucleosome assembly and adaptive immunology).

The polymorphic nature of the MHC has made the assembly of this region very difficult. Both NCBI (release 109) and Ensembl (release 95) highlight this region in their respective genome browser application with assembly exceptions. A tRNAscan-SE run on each exception revealed a cluster of tRNA genes immediately downstream of the MHC 5' boundary. There is a remarkable conservation of tRNA gene order and species type between each assembly exception and the reference despite the high polymorphism of the MHC region (Appendix A Figure 6). The assembly exceptions vary in length with SSTO being the longest (4,929,268 nt) and DBB being the shortest (4,604,810 nt), however, the area in which we observe this dense cluster of tRNA genes is consistently assembled between exceptions. It is possible that this cluster of tRNA genes is part of a linkage group that remains intact regardless of the tendency to induce variation within this region. The coevolution of this tRNA gene cluster and class I MHC genes would facilitate linkage and imply co-dependency or interaction with the gene products therein. We have been unable to identify any reports that describe an interaction between tRNA genes and MHC genes or any proposed functionality of maintaining a dense cluster of tRNA genes within this region. This will be the focus of future work.

4.2. Alignment of Genomic Blocks

The Ensembl synteny analysis tool (release 95) was implemented to systematically test chromosomes 1 and 6 (chromosome 4 in the macaque) of a select primate clade against human chromosomes 1 and 6 (Zerbino, D. R., et al. 2018). In all comparisons, the aligned genomic blocks indicate a high degree of shared synteny, although chromosome 6 (chromosome 4 in macaque) consistently displays a more uniform alignment than chromosome 1 (Supplemental 1 Figures 3A-J).[9] This provides evidence to suggest that amongst these primates, chromosome 6 may be under stronger selection than chromosome 1. Synteny amongst a relatively recent diverged monophyletic group is not surprising and tells us little about the larger evolutionary history specific to tRNA gene distribution. Therefore, a slightly deeper phylogenetic analysis was performed in an effort to provide further insight into this distribution.

[9] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

We repeated our analysis using the genomes of the cat (*Felis catus*), dog (*Canis familiaris*), mouse (*Mus musculus*), and zebra fish (*Danio rerio*; Zerbino, D. R., et al. 2018). The nematode and fruit fly genomes were unavailable for this release of the Ensembl synteny tool. The tRNA gene cluster in human chromosome 1 appears to share aligned genomic blocks with the genomes of cats and dogs, although there appear to be no blocks shared with the mouse or zebra fish genomes in this region (Supplemental 1 Figures 4A-D).[10] On the other hand, the tRNA gene cluster in human chromosome 6 appears to share aligned genomic blocks with the cat, dog, and mouse genomes, however, there does not appear to be shared blocks with the zebra fish genome in this region (Supplemental 1 Figures 4E-H).[11] Although, Sültmann and colleagues did identify a region of conserved synteny between human chromosome 6 and a linkage group that includes 27 loci associated with MHC genes of the zebra fish (Sültmann, H., et al. 2000). At the time, this was the largest conserved synteny between mammals and fishes.

Taken together, the dense clusters of tRNA genes in human chromosomes 1 and 6 have a higher degree of synteny with chromosome 6 than there is with chromosome 1. Both chromosome 1 and 6 are enriched with GO terms associated with nucleosome assembly (FDR = 4.8e-2 and 1.2e-15 respectively), but chromosome 6 is also enriched with GO terms associated with adaptive immunology (Appendix B Table 4; FDR = 3.4e-8). The

[10] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1
[11] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

FDR for each set of GO terms associated with chromosome 6 is orders of magnitude lower than chromosome 1, and we observe a striking level of similarity with respect to tRNA gene distribution and species type within the boundary of the MHC in chromosome 6 (Appendix A Figure 6). These independent lines of evidence suggest human chromosome 6 is experiencing a more stringent evolutionary constraint compared to chromosome 1. It is unclear exactly where in the ancestry of jawed vertebrates the clustering of nucleosome assembly and adaptive immunology genes within the immediate proximity to tRNA genes occurred. The degree of synteny we observe in these regions makes it more probable than not that this association was beneficial and thus conferred some selective advantage. However, there is always the possibility that this association is by chance, or selectively neutral, however, if there was no functional relationship between these gene types, we would not expect to see the degree of synteny or conservation we have demonstrated.

The utilization of computational methods to characterize and compare the genomic distribution of tRNA genes amongst taxa is portable, quick, robust, and cost effective. Our implementation of tRNAscan-SE allowed us to predict and plot the distribution of tRNA genes in a wide variety of taxa that would have been well beyond the limitations of experimental validation. Furthermore, our comparative approach revealed a distributive pattern of tRNA genes that appears to be both highly conserved and biologically relevant. Specifically, the regions of dense tRNA gene clustering we identified in human chromosomes 1 and 6 overlap with genes associated with nucleosome assembly and the MHC respectively. These overlapping regions share aligned genomic blocks most notably amongst mammals and to a lesser degree in fish suggesting an evolutionarily conserved condition. This implies functionality, and as we discussed previously, gene clustering often reflects gene interaction. We hypothesize that there are some aspects of tRNA biology that are being exploited in the mechanisms of nucleosome assembly and adaptive immunology.

4.2.1. Conservation of tRNA Genomic Distribution

Our genome wide visualization of tRNA gene distribution was repeated to include the genomes of a select group of our closest ancestors. The genomes of bonobo (*Pan paniscus*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo abelii*), and macaque (*Macaca mulatta*) were scanned for the presence of tRNA genes and totaled per chromosome. The macaque is an Old World monkey and serves as the outgroup for this primate clade. Among the apes, tRNA genes share a near identical distribution pattern in an apparent evolutionarily conserved condition (Appendix A Figure 7). We find the same tRNA gene enrichment in chromosomes 1 and 6 among the apes, although the bonobo is slightly different in that there does not appear to be a significant enrichment in chromosome 1. Unlike the apes, the macaque exhibits an enrichment of tRNA genes on chromosomes 1 and 4 instead of chromosomes 1 and 6. The conservation of tRNA gene enrichment is especially poignant in chromosome 6 of the apes as they all exhibit an abundance of tRNA genes orders of magnitude higher than the rest of the chromosomes in their respective genomes. Furthermore, it appears that chromosome 4 in the macaque is more similar to chromosome 6 in the apes, with respect to tRNA gene abundance. To facilitate a more resolute comparison in an attempt to better understand this evident distributive conservation, we narrowed our comparative approach to include only chromosomes 1 and 6 in the apes, and chromosomes 1 and 4 in the macaque.

There is a general similarity of tRNA gene distribution with respect to species type in chromosome 1 of the primates (Appendix A Figure 8). The orangutan seems to be the exception and appears to have a more unique distribution amongst the others, especially around the 0.6 region. The clustering of tRNA genes in human chromosome 1 discussed previously is clearly visible around the 0.6 region. Dense clustering of tRNA genes in chromosome 1 is evident and shared with the gorilla and macaque but is located just upstream of the human cluster. However, the distribution of tRNA genes in chromosome 6 (chromosome 4 in macaque) amongst the primates demonstrates a remarkable similarity throughout. Just like chromosome 1, the clustering of tRNA genes is clear, but occurs around the 0.16 region. Unlike chromosome 1, the apparent correlation of the cluster amongst these primates is much more stringent in chromosome 6.

The synteny observed earlier among these primates is reflected by the similarity of tRNA gene distribution in chromosomes 1 and 6 (chromosome 4 in macaque) and implies strong evolutionary conservation. Interestingly, primates share an almost duplicate MHC class 1 architecture with humans suggesting the observed colocalization of tRNA genes and class I MHC associated genes is adaptive (Kelley, J., et al. 2005). Accordingly, this region is likely under strong selective pressure. This implies a selective advantage in having a colocalization of tRNA genes and class I MHC genes. Whatever the fitness effect this association has will be an additional area of future research.

5. THE INTERSECTION OF tRNA GENES AND VARIOUS OTHER GENES

For the most part, genes fall into two large categories; coding or non-coding. Coding genes are translated into protein whereas non-coding genes are not. Biologists often use gene models as an abstraction to graphically illustrate the boundaries and functional regions that define coding or non-coding genes. A typical gene model indicates the directionality of a given gene and includes features like the transcription start site (TSS) as well as exonic and intronic regions (Appendix A Figure 9). They are typically used in genome browsers because they make it easy to visualize and interpret the genetic structure of a given genomic region. Additionally, gene models are used to demonstrate the overlapping features of a given region, for example, transcript isoforms, sense antisense (SAS) gene pairs, and sense overlapping genes.

Due to the double helix structure of DNA, both the sense and antisense strands have the potential for harboring sequences that code for genes. Moreover, the anti-parallel nature of these strands introduces the probability that the intervals of genes on one strand (i.e., the sense strand) overlap with the intervals of genes on the other strand (i.e., the antisense strand). In fact, overlapping features in the human genome are more widespread than previously thought and are estimated to account for about 25% of all known transcripts (Yelin, R., et al. 2003; Wood, E. J., et al. 2013). Based on September 2004 Ensembl data, Makalowska and colleagues analyzed the human genome for overlapping features and found about 13% of genes occur in 1766 overlaps

(Makalowska, I., et al. 2005). The manner in which genes overlap often provide some insight into whether or not there is a relationship with the gene products.

In general, SAS loci can produce transcripts that remain independent of one another, or conversely, they can produce two transcripts that have some interaction with one another. Protein coding antisense transcripts have a wide range of biological functions; however, non-coding antisense transcripts often have a more regulatory role (Kelley, R. L., et al. 2000). For example, an antisense noncoding transcript will have a complementary sequence to the sense coding transcript. The annealing of these two transcripts is an effective recruitment signal to the ribonuclease Dicer which will excise small interfering RNAs (siRNAs) from the double stranded RNA (dsRNA) inducing gene silencing by RNA interference (Bass, B. L., 2000; Zamore, P. D., 2002). The results of our intersect analysis (discussed below) provide evidence to suggest this mechanism of gene regulation may be elicited by certain genes.

Orthologous SAS loci are not well conserved. There are genetic structural differences in these regions that may have played a role in phenotypic differentiation between humans and mice (Wood, E. J., 2013). The non-conserved regions where SAS pairs exist could also impose a differential regulatory regime on overlapping protein coding genes that can facilitate the establishment of variant evolutionary trajectories (Wood, E. J., 2013). Thus, the regulatory implications of overlapping genes can have serious downstream consequences and should not be underestimated.

5.1. The Intersect Hypothesis

An intersect analysis was performed on the human genome to determine whether there are any loci in which the intervals of tRNA genes overlap with the intervals of other gene types (see Methods). We have identified four possible ways tRNA genes can intersect the intervals of other genes, in either orientation, for a possibility of eight configurations (Appendix A Figure 10): (i) sense or antisense 5' UTR, (ii) sense or antisense 3' UTR, (iii) sense or antisense coding exon, and (iv) sense or antisense intron. Of course, the boundaries of these features are completely arbitrary with respect to how a tRNA gene interval may overlap a given feature. For example, the interval of a tRNA gene may overlap the terminal boundary of a 3' UTR and extend into intergenic space. The possible biological implications of tRNA genes overlapping these regions depends on the region, and the orientation of the intersecting tRNA gene and will be discussed below.

A tRNA gene that overlaps the 5' UTR of a protein coding gene in the sense orientation with respect to the protein coding gene has the potential to affect the protein coding gene at the DNA and RNA level (Appendix A Figure 10 1A). For example, certain mechanisms of transcriptional silencing do not necessarily preclude the transcription of an overlapping tRNA gene. If the RNA-pol III complex is still able to bind and transcribe the overlapping tRNA gene, read-through transcription of the tRNA gene could generate transcripts from within the interval of the silenced protein coding gene. If the protein coding gene is not transcriptionally silenced and the RNA-pol III complex is stalled on the overlapping tRNA sequence, the RNA-pol II machinery will be mechanically prevented from assembling on the 5’ UTR. The precedent for this type of promoter competition has been established in prokaryotes and has been observed in eukaryotes as well (Wang, P., et al. 1998; Hirschman, J. E., et al. 1988). If the RNA-pol II machinery is not impeded and transcription of the protein coding gene proceeds normally, a tRNA-like structure embedded in the 5’ UTR of the protein coding transcript is likely to recruit modification enzymes normally associated with tRNA molecules that can act to splice or otherwise reinforce the embedded secondary structure. These processes will affect translational efficiency by either truncating the 5’ UTR or by reinforcing a secondary structure within the 5’ UTR that prevents or otherwise disturbs the assembly of the translational machinery. Furthermore, the secondary structure of an embedded tRNA sequence within a transcript wields an exposed anticodon sequence that could bind with a cognate codon within the body of the transcript. This complementary pairing would cause the transcript to fold in on itself in such a manner that could facilitate the formation of an additional secondary structure that will likely affect the translational efficiency of the transcript.

A tRNA gene that overlaps the 5’ UTR of a protein coding gene in the antisense orientation with respect to the protein coding gene could still affect the protein coding gene at the level of DNA and RNA, although the reverse complement of a tRNA sequence is unlikely to assume the same cloverleaf structure we expect to see in sense overlapping tRNAs (Appendix A Figure 10 1B; data not shown). Therefore, in this scenario, we do not anticipate the recruitment of tRNA modifying enzymes to the overlapping region of the protein coding transcript. If there is transcriptional silencing of the protein coding gene and there is nothing preventing the assembly of the RNA-pol III complex, the transcription of an antisense overlapping tRNA is possible and could be a source of transcripts that are complementary to the 5' UTR of the protein coding transcript. If silencing of the protein coding gene is reversed, a population of RNAs that are complementary to the 5' UTR of the protein coding transcript could impose translational regulation. Furthermore, any RNA-pol III read-through product would generate a transcript that is mostly upstream from the sense TSS and outside of the defined genic interval. However, it will still have a portion of sequence that is complementary to the 5' UTR of the sense transcript and could also impose regulatory processes on the protein coding gene transcript. If the protein coding gene is not silenced, the RNA-pol II transcriptional complex may preclude the assembly or transcriptional processes of the RNA-pol III complex in the manner just previously described.

A tRNA gene that overlaps the 3' UTR of a protein coding gene in the sense orientation with respect to the protein coding gene has the potential to affect the protein coding gene at the level of DNA and RNA similar to the mechanisms proposed above (Appendix A Figure 10 2A). If the protein coding gene is transcriptionally silenced in such a way that does not prevent the assembly of the RNA-pol III complex, then the overlapping tRNA gene can be transcribed normally. RNA-pol III could also generate a read-through transcript that would extend beyond the 3’ terminus of the protein coding gene region. If the protein coding gene is not transcriptionally silenced, transcriptional interference is unlikely unless the RNA-pol III complex is stalled on the tRNA sequence. This could prematurely disassociate the RNA-pol II complex resulting in a truncated protein coding transcript. A tRNA sequence embedded in the 3’ UTR can affect translational efficiency by the same mechanisms proposed above.

A tRNA gene that overlaps the 3’ UTR of a protein coding gene in the antisense orientation with respect to the protein coding gene has the potential to affect the intersected gene by similar mechanisms described above (Appendix A Figure 10 2B). If the protein coding gene is transcriptionally silenced and does not prevent the RNA-pol III complex from assembling, the tRNA gene can be transcribed normally. Read-through transcripts would also complement the 3’ UTR and any transcribed exons of the silenced protein coding gene and may be available to bind to the protein coding transcript if it is unsilenced therefore affecting translational efficiency. As mentioned above, we do not expect an overlapping tRNA sequence that is antisense with respect to the protein coding gene to assume a tRNA-like secondary structure within the protein coding transcript. Accordingly, the modification enzymes associated with tRNA molecules are not expected to be recruited. If the protein coding gene is not transcriptionally silent, we do not expect the respective RNA-pol complexes to preclude the assemblies of one another because of the distance separating them. However, there is still a possibility that the two complexes interfere with one another as they could both be convergently and simultaneously transcribing.

A tRNA gene that overlaps the intronic region of a protein coding gene in either the sense or antisense orientation can potentially affect the protein coding gene by similar mechanisms described above (Appendix A Figure 10 4A and 4B). If the protein coding gene is transcriptionally silenced and the assembly of the RNA-pol III complex is not impeded, the overlapping tRNA gene can be transcribed. When the intersecting tRNA gene is in the sense orientation with respect to the protein coding gene, read-through transcription could generate alternative transcripts from the intronic genic region. These transcripts may be a novel source of RNA-pol II gene transcript variants that are typically produced by alternative splicing. If the tRNA is in the antisense orientation, any read-through transcripts will be complementary to the intronic region of the protein coding transcript and may interfere with the processes of splicing. If the protein coding gene is not transcriptionally silenced, overlapping tRNA genes that are both sense and antisense could cause transcriptional interference with either the assembly or active transcription of the respective polymerase complexes as described above. Furthermore, tRNA genes that overlap intronic regions in the sense orientation will likely form a tRNA-like structure. If the recruitment and subsequent modifications, in this case splicing, occur prior to the excision of the intron, the translation of the protein will not occur.

As described earlier, tRNA transcripts undergo several post-transcriptional modifications. When considering the overlap of tRNA genes and other gene types, there are two modifications in particular that are most consequential to gene regulation; those that reinforce the distinctive clover-leaf secondary structure of a tRNA transcript and those that splice a tRNA transcript into tRFs. The former is most relevant when overlapping regions are exonic (Appendix A Figure 10 3A and 3B). The normal process of splicing an mRNA will eliminate any intronic region from the primary transcript regardless if a tRNA gene has intersected it or not. We are unaware of any described mechanism that implicates a tRNA or a tRNA-like structure within an intronic region that is responsible for, or otherwise related to, the facilitation of splicing, although this does not preclude the possibility. Furthermore, an intersecting tRNA sequence in the sense orientation with respect to the excised intronic sequence is likely to engender the intronic sequence with a tRNA-like secondary structure that could help avoid a hasty degradation and may be involved in some other yet discovered biological function.

The recruitment of modification enzymes to regions in which tRNA genes overlap protein coding genes introduces fundamental aspects of tRNA biology to mRNA biology. For example, a tRNA-like structure in a protein coding transcript, either intronic or exonic, could be subject to splicing by angiogenin or RNase P. This would effectively cut the protein coding transcript short, thus inhibiting complete translation (Appendix A Figure 11A). Of course, in the intronic case, the splicing would have to occur before the excision of the intron.

The recruitment of modification enzymes to a tRNA-like structure within the 5' UTR of a protein coding transcript may act to inhibit translation by also being spliced, or by the fortification of the structure which could cause physical obstruction of the elongation complex (Appendix A Figure 11B). To our knowledge, the interaction between enzymes known to modify tRNA and RNA-pol II transcripts has not been explored and may be a novel mechanism of translational regulation.

In general, when tRNA genes overlap protein coding genes, there are two fundamental implications we are interested in exploring: (i) the regulation of transcription and translation of the protein coding gene, and (ii) read-through transcription that generates RNA polymers that are complementary to protein coding transcripts. Transcriptomic data was not analyzed as part of this thesis, so we have yet to validate the regulatory implications of tRNA genes that may overlap protein coding genes, however, with the implementation of the IGV genome browser (Version 2.3.82 (130)), we are able to visually validate overlapping regions. Regardless, experimental validation is preferred, however it is beyond the scope of this thesis but will be the focus of future work.

5.2. Intersect Analysis

When analyzing the human genome for the intersection of protein coding and tRNA genes, we considered the overlap between a tRNA gene and the entire protein coding gene interval. The complete interval of a protein coding gene contains non-coding exons,

exons, and introns (Appendix A Figure 12). At the time of writing, Gencode.v28 has identified 19,901 features in the human genome annotated as protein coding genes (GRCh.38.p12). tRNAscan-SE predicted a total of 636 tRNA genes which accounts for just over 3% of the genes between these two gene types (i.e., protein coding and tRNA). tRNA genes have an average sequence length of 77 nucleotides, and protein coding genes have an average length of 66,577 nucleotides (Piovesan, A., et al. 2016). The lengths of tRNAs used in this analysis include the distance between the 5' phosphorus group and the 3' terminus of a processed tRNA transcript (i.e., this does not include the 5' leader or 3' trailing sequences). We calculated the haplotype sequence of the human genome to be 3,031,042,417 nucleotides in length. If there are 19,901 protein coding genes with an average sequence length of 66,577 nucleotides each, then, on average, the total length of protein coding gene intervals in the human genome is (19,901*66,577) = 1.3e9 nucleotides. This represents about 44% of the length of the genome. If there are 636 tRNA genes with an average length of 77 nucleotides, then, on average, the total length of tRNA genes is (636*77) = 4.9e4. This represents about 0.0016% of the length of the human genome. The probability then that the sequences of a tRNA gene and protein coding gene of average length overlap at any given locus in the human genome, assuming both gene types are evenly distributed, is 0.0007%. Despite this low probability, we have identified intact tRNA genes that overlap 79 protein-coding genes, as well as 30 long-intergenic non-coding RNAs (lincRNAs) and 11 antisense genes amongst others (Appendix B Table 5). These overlaps are not mutually exclusive. For example, in the situation in which the interval of a tRNA gene overlaps the interval of an antisense gene, by definition, the same tRNA gene interval is simultaneously overlapping the sense gene as well. If this sense gene happens to be a protein coding gene, then the tRNA gene interval is found to overlap both the protein coding and antisense genes. The values reported in Table 5 does not make this distinction and only reports individual counts for each gene type overlap. For a comprehensive summary of simultaneous overlaps, see Supplemental 2.[12]

Protein coding genes are sequences of DNA that contain all of the structural units of a gene (i.e., non-coding exons, exon, introns, promoter, enhancer, and terminator) and has an open reading frame (ORF). The transcript of a protein coding gene is post-transcriptionally modified by the addition of a 5' cap, 3' poly-A sequence, and the removal of intronic regions. The resultant mature mRNA contains two non-coding exons (5' and 3' UTRs) and a series of triplet DNA sequences called codons that code for specific amino acids. We identified tRNA genes that overlap 79 protein coding genes (Appendix B Table 5).

Processed transcripts do not contain an ORF and are divided into three main categories; long non-coding RNAs (lncRNAs), pseudogenes, and genes designated to be experimentally confirmed (TEC). In addition to the 79 protein coding genes, our analysis

[12] https://etd.tamu.edu/submit/22476/file/171940/ancillary_table.xlsx

has identified 59 other occurrences in which a tRNA gene was found to overlap defined genomic intervals and they all fall within these three classes of processed transcripts. lncRNAs are processed transcripts that exceed 200 nucleotides in length and are not known to be translated into protein. Many sub-categories of lncRNAs have been defined, but we will limit our discussion to those indicated in our analysis. For example, sense overlapping, sense intronic, antisense, and lincRNAs are all types of lncRNAs. Sense overlapping genes can generate a long non-coding transcript that contain coding genes within its intron while sense intronic genes can generate long non-coding transcripts from an intron of coding genes but does not overlap an exon. We have identified tRNA genes that intersect 2 sense overlapping genes and 2 sense intronic genes (Appendix B Table 5). Antisense genes produce processed transcripts that overlap the genomic region of a protein coding gene on the opposite strand. lincRNAs (lincRNAs) are defined the same way as lncRNAs except they do not overlap the intervals of protein coding genes (Ransohoff, J. D., et al. 2018). We have identified tRNA genes that overlap the regions of 30 lincRNA genes and 11 antisense genes (Appendix B Table 5). Bi-directional promoters are regions within the promoter of protein coding genes but facilitate the transcription lncRNAs from the opposite strand. We have identified tRNA genes that overlap 2 bi-directional promoters.

Pseudogenes are similar to protein coding genes, but they contain a frameshift or aberrant stop codon that disrupts the ORF. There are two types of pseudogenes indicated in our analysis; unprocessed and polymorphic. Unprocessed pseudogenes are typically produced by gene duplication but the transcripts and not completely processed and still contain intronic regions. Polymorphic pseudogenes arise by SNP or indels and the gene is usually translated among the individuals in a population that do not have these mutations. We have identified tRNA genes that overlap the regions of 8 unprocessed pseudogenes and 1 polymorphic pseudogene (Appendix B Table 5).

Lastly, TEC is a designation for transcripts that appear to be protein coding but need experimental validation. Our analysis has identified a tRNA gene that overlaps a single region identified as TEC. We have also identified tRNA genes overlapping the regions of 2 unclassified processed transcripts (Appendix B Table 5). These are transcripts that cannot be placed into existing designations.

According to our predictions, overlapping tRNA genes have the potential to introduce key aspects of tRNA biology to the genes and transcripts they overlap. This can fundamentally alter the function of these genes and transcripts through processes like molecular interactions and modifications. Our analysis indicates protein coding genes, lncRNAs, and antisense genes as the most abundant classes of tRNA intersects (Appendix B Table 5). Accordingly, we will narrow our analysis specific to these three classes.

5.2.1. The Intersection of tRNA and Protein Coding Genes

The transcription of protein coding genes results in a pre-mRNA that contains non-coding exons, exons, and introns. Introns are typically spliced out of the pre-mRNA and are not part of the sequence that gets translated to protein. Exons are the segments of genes that are retained in the mRNA and can either be coding or non-coding (Appendix A Figure 9). For example, UTRs are non-coding exons because they are a part of the mRNA but do not encode a sequence that will be translated to protein. Coding exons on the other hand are part of the mRNA and encode the sequence that will be translated to protein. Because the coding regions of protein coding genes dictate the ultimate protein product, we expect to see more evolutionary constraint amongst the coding regions of genes as opposed to a more relaxed constraint amongst the non-coding regions like UTRs and introns. Accordingly, we predict that the 79 protein coding genes indicated in our intersect analysis are most likely to contain tRNA genes within the non-coding regions and we do not expect to find tRNA sequences intersecting the coding regions. In line with our prediction, 67 protein-coding genes have tRNA sequences within intronic regions, 11 have tRNA sequences in non-coding exons, and, surprisingly, 1 has a tRNA sequence intersecting the interval of a coding exon.

The only protein coding gene to have a coding exonic overlapping tRNA is the pleckstrin homology domain interacting protein (PHIP; Supplemental 1 Figure 5A).[13] PHIP is in human chromosome 6 and is associated with glucose regulation and melanoma metastasis. The overlapping tRNA sequence identified here is in the same orientation as the protein coding gene and occurs in a coding exon of a protein coding isoform of this gene, although, the primary transcript is intronic at this interval. According to our predictions, a tRNA sequence that overlaps the sequence of a gene in the same orientation can act as an independent promoter unit and induce RNA-pol III transcription at this locus. In this particular gene (PHIP), the overlapping tRNA sequence occurs on the 10th exon of a 17-exon model. There are no isoforms indicated that begin near this region, so the transcription of isoforms by means of RNA-pol III for this gene is unlikely. However, post-transcriptional modifications that could splice or strengthen tRNA-like secondary structures found within transcripts introduces the potential to regulate the expression of the PHIP (or an isoform thereof) by the premature termination of translation or the physical inhibition of translation. In either situation, the overlapping tRNA sequence in PHIP may impose a regulatory mechanism that has yet to be described and to our knowledge has not been experimentally validated.

Unlike coding exons, non-coding exons are not translated to protein. Regardless, these regions (3’ and 5’ UTRs) are implicated in the regulation of gene expression and are thus

[13] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

likely to be under some level of evolutionary constraint (Jackson, R. J., et al. 1990; Conne, B., et al. 2000; Hughes, T. A., 2006; Van Der Velden, A. W., et al. 1999). The 5' UTR of protein coding transcripts needs to accommodate the translational machinery and is likely under a slightly more stringent selective pressure than the 3' UTR (Conne, B., et al. 2000). According to our predictions, a tRNA gene that overlaps the 5' UTR of a protein coding gene in either orientation can act as an independent promoter unit to recruit the RNA-pol III transcription complex to the type-2 intragenic promoter of the tRNA gene. If the RNA-pol II transcription complex assembles within the same temporal framework as the RNA-pol III complex, it is possible that this mutual assembly can interfere with each other and inhibit the assembly of both complexes. Alternatively, if the overlapping tRNA sequence is in an opposing orientation with respect to the protein coding gene and there is no interference with the assembly of each respective transcription complex, the actively transcribing complexes on opposite strands moving towards each other are likely to interfere with each other upon contact. It is unclear whether or not this interaction would interrupt transcription. On the other hand, the RNA-pol III transcription complex could simply assemble on the type-2 promoter of the tRNA gene and transcribe the tRNA gene, or perhaps generate an alternative transcript by reading through the termination sequence of the tRNA gene.

There are eleven protein coding genes indicated by our analysis in which a non-coding exon overlaps with at least one tRNA gene. Seven of these protein coding genes have more than one overlapping tRNA. For example, VAC14 and CTC1 have 4 and 3 tRNA genes that overlap the 3’ UTRs respectively. SHF and ZBED9 have 3 tRNA genes that overlap each of the respective 5’ UTRs (Appendix B Table 6). CTC1 is of particular interest, not only because it is a component of the CST complex that protects telomeres from degradation, and is therefore of great research value, but each of the three overlapping

tRNA sequences are in the same orientation as the protein coding gene and are different species of tRNA from one another. This suggests they did not arise by duplication and have been independently recruited to this region. Furthermore, the spacing between these three overlapping tRNAs appears to be periodic as well. There are 301 nucleotides that separate the first and second tRNA gene and 300 nucleotides separating the second and third tRNA gene suggesting the spatial distribution of these tRNA genes is non-random. We are unaware of any literature that has implicated overlapping tRNA genes in the 3’ UTR of the CTC1 gene that affects transcriptional or translational regulation or other functional processes. Experimental validation is required to determine whether or not these overlapping tRNA genes are implicated in the expression or otherwise general function of CTC1.

Our predictions indicate eight different ways a tRNA gene can overlap a region of a protein coding gene. Our analysis has identified seven of these eight possible overlaps in the human genome which occur in eleven protein coding genes. We did not find an example of a tRNA gene that overlaps a coding exon of a protein coding gene in the antisense orientation. Moreover, we visually validated the indicated intersects and found some evidence to suggest the generation of alternative transcript predicted by our model are credible. For example, $tRNA^{Val}$ intersects the 5’ UTR of DPP9 in the same orientation (they are both in the Crick orientation). There is a protein coding isoform annotated by Havana as having an alternative 5’ UTR that, in the genome browser, appears to start at the same locus as the overlapping tRNA gene (Supplemental 1 Figure 5B).[14] This is what we would expect to see when a tRNA gene acts as an independent promoter region that can facilitate read-through transcription.

5.2.2. The Intersection of tRNA and Long Non-coding RNA Genes

Advancements in molecular techniques and computational power in the last fifteen years have helped erode a long-standing dogma that supposed most of the human genome was transcriptionally inactive and that the bulk of the transcriptome consisted of protein-coding exons. Our current understanding is that a majority of the genome is in fact transcriptionally active and protein-coding exons make up a small fraction of the transcriptome. Throughout much of this period of discovery, the rate at which novel transcripts were identified outpaced the rate at which they are functionally annotated, although there have been considerable efforts in recent years to attribute function to a growing catalog of non-translated transcripts.

[14] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

lncRNAs are typically more concentrated in the nucleus and are maintained at lower levels of expression compared to protein-coding transcripts (Derrien, T., et al. 2012). Maintaining an assemblage of lncRNA transcripts within the nuclear envelope may be a function of interactions with neighboring genes. For example, *Malat1* (metastasis associated lung adenocarcinoma transcript 1) is a highly conserved lncRNA that has been implicated in *cis*-acting regulatory pathways (Zhang, B., et al. 2012). Interestingly, *Malat1* has a 3’ terminal tRNA-like secondary structure similar to that described by the GTH (Weiner, A. M., 1987). This structure is cleaved by the same ribonuclease RNase P that cleaves the 5’ end of pre-tRNA. This cleavage results in two distinct molecules; a mature lncRNA transcript with a stabilizing 3’ triple helix structure and a tRNA-like *Malat1*-associated small cytoplasmic RNA (mascRNA; Wilusz, J. E., et al. 2008). The matured *Malat1* remains in the nucleus where it functions in the regulation of alternative splicing and the cleaved tRNA-like mascRNA is exported from the nucleus where it undergoes a similar modification regime to that of canonical tRNA (Wilusz, J. E., et al. 2012; Tripathi, V., et al. 2010; Brown, J. A., et al. 2012; Wilusz, J. E., et al. 2008). The function of tRNA-like mascRNA remains unknown, but it is unlikely it participates in translation because it does not have a conserved anticodon sequence and it is not aminoacylated (Wilusz, J. E., et al. 2008). Because the lncRNA remains in the nucleus and the mascRNA is exported into the cytoplasm, it is possible the mascRNA could act as a signaling molecule to inform some cytoplasmic process that the parental lncRNA has been transcribed and matured (Wilusz, J. E., et al. 2008). Regardless, a tRNA-like structure in the body of a lncRNA transcript recruits enzymes known to modify tRNA transcripts. This provides experimental data that supports a fundamental prediction of our hypothesis.

There are currently 7,490 lincRNA genes and 48 lncRNA genes annotated in the human genome (Gencode.v28; GRCh38.p12). We have identified tRNA genes overlapping the intervals of 30 lincRNA genes and 2 lncRNA genes (Appendix B Table 5). Additionally, there are 118 transcripts that are mostly derived from the 30 lincRNA genes that retain the overlapping tRNA sequence. This averages out to be about 4 transcripts per lincRNA gene that have an overlapping tRNA gene suggesting a functional parameter that is conserving this condition within the lincRNAs indicated in our analysis.

As with the *Malat1* example, and in-line with our predictions, tRNA structures within the transcripts of the lincRNA and lncRNA genes are likely recruiting enzymes known to splice or otherwise modify the embedded tRNA structure. tRNA$^{\text{Ala}}$ is the most abundant species of tRNAs in the human genome. Despite this, there appears to be a preference for asparagine when it comes to the species of tRNA intersecting lincRNA (Supplemental 1 Figure 6).[15] The apparent bias for tRNA$^{\text{Asn}}$ intersecting lncRNAs indicates a property of the asparagine anticodon that is not present in the rest of the tRNA anticodon population. To our knowledge, this characteristic (the apparent bias for specific anticodons) has not been experimentally explored.

[15] https://etd.tamu.edu/submit/22476/file/172859/Supplemental+1

5.2.3. The Intersection of tRNA and Antisense Genes

We have identified tRNA genes that overlap the intervals of 11 antisense genes (Appendix B Table 5). A tRNA gene that overlaps an antisense gene is effectively intersecting two genes simultaneously; the sense gene and the antisense gene. Thus, transcriptional interference by the physical contact of transcriptional complexes can occur with the complexes from either the sense or the antisense gene. Furthermore, it is possible that an antisense gene responsible for downregulating the expression of a sense gene can itself become downregulated upon the recruitment of an overlapping tRNA gene. This would effectively rescue the expression of the otherwise suppressed sense gene. We have identified this type of overlap (i.e., a tRNA gene that simultaneously intersects a sense and antisense gene) amongst 3 of the 11 antisense genes indicated in our intersect analysis. Interestingly, 2 of these 3 trisects occur within the 5' UTR of the sense protein coding gene. Of particular interest is the SHF gene. According to GeneCards, SHF has been implicated in the regulation of apoptosis in response to a growth factor that regulates cell growth and division (Stelzer, G., et al. 2016). The tRNA gene overlapping this antisense gene is in the opposite orientation with respect to the antisense gene. Thus, the overlapping tRNA gene could act as an independent promoter to transcribe a sequence that is complementary to the antisense gene therefore preventing the antisense gene to otherwise downregulate the expression of the sense SHF gene. To our knowledge, this mechanism of gene regulation has not been experimentally explored.

Antisense genes are a class of lncRNAs that are, by definition, complementary to a sense gene. The sense genes in these respective pairings are not necessarily protein coding. For many years, antisense transcripts were considered to be little more than transcriptional noise, however, relatively recent and consistent observations of pervasive transcription and biological relevancy has challenged this long-held idea. For example, antisense transcripts have been shown to induce DNA methylation and histone modification patterns that can affect the initiation of transcription and the subsequent expression of their paired sense genes (Tufarelli, C., et al. 2003; Yu, W., et al. 2008). Moreover, antisense transcripts can also work synergistically with their sense transcripts to enhance the translational efficiency of the sense transcript (Carrieri, C., et al. 2012). Adversely, the transcript of an antisense gene, also known as antisense RNA (asRNA), can hybridize with the sense transcript which not only prevents the translation of the sense gene but can also recruit endonucleases that have an affinity for double-stranded RNA (dsRNA). Antisense genes have also been implicated as an underlying cause in disease state expression of an otherwise apparently normal gene and are increasingly being recognized as critical regulators of both the transcription and translation of their sense genes (Tufarelli, C., et al. 2003). SAS gene pairs can also impose regulatory processes based on their spatial and temporal characteristics.

The respective orientation of SAS gene pairs introduces the possibility that the transcriptional machinery of two overlapping genes will interfere with each other. When the promoter regions of genes on opposing strands overlap, or are otherwise close enough to each other, competition between transcription factors assembling on or near one of the promoters can preclude the transcription factors from assembling on or near the other. This was shown to occur in prokaryotes and at enhancer sites in eukaryotes (Wang, P., et al. 1998; Hirschman, J. E., et al. 1988; Conte, C., 2002). If the promoter regions of SAS gene pairs are distal enough from each other such that the assembly of transcription complexes is not impeded, interference can still occur by the physical interaction of the respective transcriptional complexes if they are each actively transcribing. This establishes a mechanism in which the respective transcriptional complexes are on a collision course with one another and has been shown to occur in *S. cerevisiae* (Prescott, E. M., et al. 2002). This type of transcriptional interference is believed to be rare in nature, although most genomic searches for convergent promoters has been limited to RNA-pol II genes. We propose the presence of an RNA-pol III type-2 intragenic promoter in addition to an RNA-pol II promoter can cause this type of interference. It is likely that the type-2 promoter of tRNA genes have evaded detection of previous work describing this mechanism of transcriptional interference and is therefore, to our knowledge, not well explored.

6. DISCUSSION AND CONCLUSION

The origin story of tRNA is relevant to most research in tRNA biology but it is often ignored or underestimated by researchers. An understanding of this dynamic and ancient history helps to put a unique perspective on current findings. It is easy to lose sight of the probability that precursor tRNA-like molecules preceded the origin of life and were likely instrumental through the processes of abiogenesis. Today, most research centered on tRNA biology is focused on pathway dependent interactions between primary tRNA transcripts, highly modified mature tRNA, and more recently, tRFs. Most findings are interpreted with an emphasis on the effects these molecules have on downstream transcriptional and translational efficiency and homeostatic processes. Attention is also given to the clinical role tRNA molecules and their derivatives have on disease state tissues. What is ominously missing from much of the contemporary literature focused on some aspect of tRNA biology is the possibility of retained ancient catalytic functionality.

Modern molecular techniques and increasing computational power have allowed us to untangle much of the complexity of tRNA biology broadening a catalog of disparate functionality connecting tRNAs to mechanisms that far surpass the dutiful service of amino acid delivery. Many of these newfound discoveries have been characterized as having unique, or alternative functionality with respect to dogmatic translational activities, but when considering the tRNA origin story, these alternatives are likely reflective of the original functions that dominated an ancient pre-DNA/protein world.

Ironically, our definition of tRNA canonical function is likely the alternative. Despite this, tRNAs remain inextricably linked to the central dogma of biology on account of the integral role they play in the translation of genetic code into protein.

In our most conservative estimation, we have identified 445 cytosolic predicted tRNA gene copies in the human genome. This amount redundancy suggests peak tRNA transcription rates cannot be facilitated by a single template and tRNA genes can tolerate various types of mutations (e.g., point mutations or insertions and deletions) with little to negligible deleterious functional consequence (Sharp, S. J., et al. 1985). Alternatively, the generation of mascRNAs, or other tRNA-like mimics may have elicited evolutionary mechanisms that work to degrade and eliminate these types of molecules in an effort to inhibit any deleterious interactions. These destructive mechanisms would likely act on canonical tRNA molecules as well (e.g., angiogenin, RNase P, RNase Z, etc.). Thus, the proliferation of tRNA genes would result in the observed redundancy in copy number and would be required to maintain tRNA populations at a sustainable level. This process could also explain the presence and pervasiveness of tRFs.

The human mitochondrial genome (ignored in our analysis) encodes only 20 tRNA genes. This demonstrates an alarming vulnerability to dysfunction and disease not observed in cytosolic tRNA populations. Indeed, greater than 50% of mutations that occur in mitochondria are located within tRNA genes (Lott, M. T., et al. 2013). Although cytosolic heteroplasmy allows for some level of cellular redundancy, mutations in mt-tRNA genes result in a wide breadth of syndromes. For example, MERFF (myoclonic epilepsy and ragged red fibers) syndrome results from a point mutation in the TΨC-loop of tRNA^{Lys} and Mitochondrial Encephalopathy with Lactic Acidosis and Stroke-like episodes (MELAS) is an extreme example with no known treatment (Tryoen-Tóth, Petra, et al. 2003; Perli, E., et al. 2014). Several mechanisms in which cytosolic tRNAs can be imported into the mitochondria are known, and it may be possible to rescue these disease states by the manipulation of such a mechanism (Rubio, M. A. T., et al. 2008). Moreover, it would be an enormous waste of metabolic resources if each tRNA gene copy were to be expressed simultaneously (Sharp, S. J., et al. 1985). Accordingly, processes involved in the controlled expression of tRNA genes are critically important. Research focused on the mechanisms that coordinate and regulate transcription have been crucial in expanding our understanding of tRNA biology (for reviews, see Willis, I. M., et al. 2007; Cieśla, M., et al. 2008).

Regulating the expression of tRNA genes is essential to the maintenance of cellular and organismal health. In eukaryotes, there are three classes of RNA polymerase (I, II, and III) which are regulated by the conserved protein Maf1 (Pluta, K., et al. 2001; Reina, J. H., et al. 2006; Johnson, S. S., et al. 2007). Under normal conditions, Maf1 is phosphorylated and unable to negatively regulate the activity of RNA-pol III. However, when cellular conditions deteriorate (e.g., stress or disease), Maf1 becomes unphosphorylated and actively begins to negatively regulate RNA-pol III through interactions with RNA-pol III subunits and associated transcription factors (Pluta, K., et al. 2001; Gavin, A-C., et al. 2006; Oficjalska-Pham, D., et al. 2006; Reina, J. H., et al. 2006; Upadhya, R., et al. 2002; Desai, N., et al. 2005; Rollins, J., et al. 2007). Despite this downregulation during cellular stress, there is a subset of tRNA genes in which transcription appears to be impervious to negative Maf1 regulation (Turowski, T. W., et al. 2016; Orioli, A., et al. 2016). Of the many mechanisms proposed that can maintain transcription of these so-called 'housekeeping' tRNA genes, the one most favorable to our findings suggests the proximity of RNA-pol III genes to actively transcribing RNA-pol II protein-coding genes (Turowski, T. W., et al. 2016). Within the regions of chromosomes 1 and 6 in which we found tRNA gene density to be punctuated, we also observe a statistical overrepresentation of genes implicated in nucleosome assembly and adaptive immunity. It is highly probable these regions exhibit attributes similar to euchromatin and may facilitate the ongoing transcription of tRNA genes, even in episodes of cellular stress. Regulating the transcription of tRNA molecules is a critical process as the consequences of dysregulation can be devastating. Regardless of the cellular attempt to dynamically regulate the expression of tRNA, we suggest the clustering of tRNA genes amongst other critical cellular protein-coding genes on chromosomes 1 and 6 may impart a mechanism to safeguard the continuing transcription of tRNA genes when cellular signals mandate otherwise.

We have identified two regions of the human genome that exhibit a sharp increase in the frequency of tRNA genes and are populated by genes implicated in nucleosome assembly and the MHC (chromosomes 1 and 6 respectively). Perhaps the most pressing question is what genetic aspect(s) of tRNA biology are necessary, or otherwise related, to sets of genes implicated in nucleosome assembly and the adaptive immune system, or vice versa? Given the critical nature of these genes, it is possible the regulatory aspect of these regions may play a role. For example, the clustering of and colocalization of these genes may facilitate mutual transcription. Moreover, the polymorphic nature of the MHC region suggests the clustering of tRNA and MHC genes could be a source of generating and maintaining variation, although this type of heterozygosity is not observed in the nucleosome assembly genes on chromosome 1 that also display clustering with tRNA genes. It is possible there are separate regulatory mechanisms in place to ensure these disparate properties of gene clustering do not overlap. Regardless, maintaining dense clusters of tRNA genes in close proximity to genes associated with nucleosome assembly and adaptive immunity suggests tRNA genes play a role in the maintenance of genomic stability, immunology, and sexual selection further broadening its non-canonical repertoire.

Early work in tRNA biology seldom examined processes beyond the understood function of tRNA acting as adapter molecules required for the translation of mRNA into proteins. More recently, however, a more dynamic landscape has emerged that highlight tRNAs as principal components involved in an array of biological processes that range from homeostatic to disease state. Our analysis essentially forwards two main hypotheses; H_{A1}: The genomic organization of tRNA genes in the human genome is non-random and plays a key role in nucleosome assembly and adaptive immunology, and

H_{A2}: The intersection of tRNA genes with various genomic features provides critical regulatory mechanisms to that be both *cis*- and *trans*-acting. Here, we have provided evidence that identified a spatial relationship between tRNA genes and genes associated with nucleosome assembly and adaptive immunology. Whether or not this spatial relationship will translate into a functional one will be determined by future work. We have also forwarded evidence to suggest certain coding and non-coding genes may be recruiting enzymes normally associated with tRNA modification. We have predicted regulatory implications of these associations that will need to be validated in future work.

The origin story of tRNA and subsequent evolutionary optimization imply this molecule, in one form or another, has been present from a time in which life did not exist. When recognizing the likely origin of tRNA from a population of self-replicating RNA molecules, the ubiquity of tRNA across domains should not be surprising and the pervasiveness of tRNA in a myriad of biological processes should be expected. It is within this framework that we have interpreted our findings. Furthermore, it is our conviction that future research will be served well to, in the least, apply the context of a foregone origin when interpreting data that involves any aspect of tRNA biology.

REFERENCES

Agris, P., et al. “The RNA Modification Database.” The RNA Institute, University At Albany, State University of New York, 2019, https://mods.rna.albany.edu/home

Agris, Paul F., Franck AP Vendeix, and William D. Graham. "tRNA’s wobble decoding of the genome: 40 years of modification." Journal of molecular biology 366.1 (2007): 1-13.

Allison, D. S., Goh, S. H., & Hall, B. D. (1983). The promoter sequence of a yeast tRNAtyr gene. Cell, 34(2), 655-663.

Anderson, J. D., and J. Widom. "Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites." Journal of molecular biology 296.4 (2000): 979-987.

Bachman, Nancy J., et al. "The 5′ region of the COX4 gene contains a novel overlapping gene, NOC4." Mammalian genome 10.5 (1999): 506-512.

Bass, Brenda L. "Double-stranded RNA as a template for gene silencing." Cell 101.3 (2000): 235-238.

Beck, Stephan, and John Trowsdale. "The human major histocompatibility complex: lessons from the DNA sequence." Annual review of genomics and human genetics 1.1 (2000): 117-137.

Bernatchez, L., and C. Landry. "MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years?." Journal of evolutionary biology 16.3 (2003): 363-377.

Bernhardt, Harold S. "The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others) a." Biology direct 7.1 (2012): 23.

Bernhardt, Harold S., and Warren P. Tate. "Primordial soup or vinaigrette: did the RNA world evolve at acidic pH?." Biology direct 7.1 (2012): 4.

Bieker, J. J., Martin, P. L., & Roeder, R. G. (1985). Formation of a rate-limiting intermediate in 5S RNA gene transcription. Cell, 40(1), 119-127.

Bokov, Konstantin, and Sergey V. Steinberg. "A hierarchical model for evolution of 23S ribosomal RNA." Nature 457.7232 (2009): 977.

Brown, Jessica A., et al. "Formation of triple-helical structures by the 3′-end sequences of MALAT1 and MENβ noncoding RNAs." Proceedings of the National Academy of Sciences 109.47 (2012): 19202-19207.

Carrieri, Claudia, et al. "Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat." Nature 491.7424 (2012): 454.

Chen, Qi, et al. "Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder." Science 351.6271 (2016): 397-400.

Cieśla, Małgorzata, and Magdalena Boguta. "Regulation of RNA polymerase III transcription by Maf1 protein." Acta Biochimica Polonica 55.2 (2008): 215-225.

Conne, Béatrice, André Stutz, and Jean-Dominique Vassalli. "The 3′ untranslated region of messenger RNA: a molecular 'hotspot'for pathology?." Nature medicine 6.6 (2000): 637.

Conte, Caroline, Bernard Dastugue, and Chantal Vaury. "Promoter competition as a mechanism of transcriptional interference mediated by retrotransposons." The EMBO journal 21.14 (2002): 3908-3916.

Crick, Francis HC. "On protein synthesis." Symp Soc Exp Biol. Vol. 12. No. 138-63. 1958.

Dandekar, Thomas, et al. "Conservation of gene order: a fingerprint of proteins that physically interact." Trends in biochemical sciences 23.9 (1998): 324-328.

Derrien, Thomas, et al. "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." Genome research 22.9 (2012): 1775-1789.

Desai, Neelam, et al. "Two steps in Maf1-dependent repression of transcription by RNA polymerase III." Journal of Biological Chemistry 280.8 (2005): 6455-6462.

Di, M. Giulio. "On the origin of the transfer RNA molecule." Journal of theoretical biology 159.2 (1992): 199-214.

Dick, Tobias P., and Wolfgang WA Schamel. "Molecular evolution of transfer RNA from two precursor hairpins: implications for the origin of protein synthesis." Journal of molecular evolution 41.1 (1995): 1-9.

Dittmar, Kimberly A., Jeffrey M. Goodenbour, and Tao Pan. "Tissue-specific differences in human transfer RNA expression." PLoS genetics 2.12 (2006): e221.

Dumay-Odelot, H., Marck, C., Durrieu-Gaillard, S., Lefebvre, O., Jourdain, S., Prochazkova, M., ... & Teichmann, M. (2007). Identification, molecular cloning, and characterization of the sixth subunit of human transcription factor TFIIIC. Journal of Biological Chemistry, 282(23), 17179-17189.

Frank, Daniel N., and Norman R. Pace. "Ribonuclease P: unity and diversity in a tRNA processing ribozyme." (1998): 153-180.

Fu, Yu, et al. "Small non-coding transfer RNA-derived RNA fragments (tRFs): their biogenesis, function and implication in human diseases." Genomics & informatics 13.4 (2015): 94.

Fujishima, Kosuke, et al. "Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea." Proceedings of the National Academy of Sciences 106.8 (2009): 2683-2687.

Galli, G., Hofstetter, H., and Birnstiel, M.L. 1981. Two con-served sequence blocks within eukaryotic tRNA genes are major promoter elements. Nature 294: 626–631.

Gangopadhyay, Samudra S., et al. "Unusual genome organisation in Entamoeba histolytica leads to two overlapping transcripts." Molecular and biochemical parasitology 89.1 (1997): 73-83.

Gavin, Anne-Claude, et al. "Proteome survey reveals modularity of the yeast cell machinery." Nature 440.7084 (2006): 631.

Gebetsberger, Jennifer, and Norbert Polacek. "Slicing tRNAs to boost functional ncRNA diversity." RNA biology 10.12 (2013): 1798-1806.

Giegé, Richard. "Toward a more complete view of tRNA biology." Nature structural & molecular biology 15.10 (2008): 1007.

Gingold, Hila, et al. "A dual program for translation regulation in cellular proliferation and differentiation." Cell 158.6 (2014): 1281-1292.

Green, Rachel, Christopher Switzer, and Harry F. Noller. "Ribosome-catalyzed peptide-bond formation with an A-site substrate covalently linked to 23S ribosomal RNA." Science 280.5361 (1998): 286-289.

Green, Rachel, and Harry F. Noller. "Ribosomes and translation." Annual review of biochemistry 66.1 (1997): 679-716.

Groth, Anja, et al. "Chromatin challenges during DNA replication and repair." Cell 128.4 (2007): 721-733.

Haseltine, William A., and Ricardo Block. "Synthesis of guanosine tetra-and pentaphosphate requires the presence of a codon-specific, uncharged transfer ribonucleic acid in the acceptor site of ribosomes." Proceedings of the National Academy of Sciences 70.5 (1973): 1564-1568.

Heikkilä, P., Raija Soininen, and K. Tryggvason. "Directional regulatory activity of cis-acting elements in the bidirectional alpha 1 (IV) and alpha 2 (IV) collagen gene promoter." Journal of Biological Chemistry 268.33 (1993): 24677-24682.

Helm, Mark. "Post-transcriptional nucleotide modification and alternative folding of RNA." Nucleic acids research 34.2 (2006): 721-733.

Hirschman, J. E., K. J. Durbin, and F. Winston. "Genetic evidence for promoter competition in Saccharomyces cerevisiae." Molecular and cellular biology 8.11 (1988): 4608-4615.

Hoagland, M. B., et al. "Enzymatic carboxyl activation of amino acids." J Biol Chem 218.1 (1956): 345-358.

Hoagland, M. B., et al. "A soluble ribonucleic acid intermediate in protein synthesis." Journal of Biological Chemistry 231.1 (1958): 241-257.

Hoagland, M. B., et al. "Intermediate reactions in protein biosynthesis." Biochimica et biophysica acta 24.1 (1957): 215.

Hofstetter, H., Kressmann, A., & Birnstiel, M. L. (1981). A split promoter for a eucaryotic tRNA gene. Cell, 24(2), 573-585.

Holley, R. W. "An alanine-dependent, ribonuclease-inhibited conversion of AMP to ATP, and its possible relationship to protein synthesis." Journal of the American Chemical Society 79.3 (1957): 658-662.

Hughes, Thomas A. "Regulation of gene expression by alternative untranslated regions." Trends in Genetics 22.3 (2006): 119-122.

Ito, Emi, et al. "A core-promoter region functions bi-directionally for human opioid-receptor-like gene ORL1 and its 5′-adjacent gene GAIP." Journal of molecular biology 304.3 (2000): 259-270.

Itoh, Yuzuru, et al. "Tertiary structure of bacterial selenocysteine tRNA." Nucleic acids research 41.13 (2013): 6729-6738.

Jackson, Richard J., and Nancy Standart. "Do the poly (A) tail and 3′ untranslated region control mRNA translation?." Cell 62.1 (1990): 15-24.

Jankowski, Jacek M., et al. "In vitro expression of two proteins from overlapping reading frames in a eukaryotic DNA sequence." Journal of molecular evolution 24.1-2 (1986): 61-71.

Johnson, S. S., et al. "Mammalian Maf1 is a negative regulator of transcription by all three nuclear RNA polymerases." Molecular cell 26.3 (2007): 367-379.

Johnston, Wendy K., et al. "RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension." Science 292.5520 (2001): 1319-1325.

Joseph, David R. "The rat androgen-binding protein (ABP/SHBG) gene contains triplet repeats similar to unstable triplets: evidence that the ABP/SHBG and the fragile X-related 2 genes overlap." Steroids 63.1 (1998): 2-4.

Joyce, Gerald F. "The antiquity of RNA-based evolution." Nature 418.6894 (2002): 214.

Jühling, Frank, et al. "tRNAdb 2009: compilation of tRNA sequences and tRNA genes." Nucleic acids research 37.suppl_1 (2008): D159-D162.

Kanai, A. "Molecular evolution of disrupted transfer RNA genes and their introns in archaea." Evolutionary Biology: Exobiology and Evolutionary Mechanisms. Springer, Berlin, Heidelberg, 2013. 181-193.

Keam, Simon, and Gyorgy Hutvagner. "tRNA-derived fragments (tRFs): emerging new roles for an ancient RNA in the regulation of gene expression." Life 5.4 (2015): 1638-1651.

Kelley, James, Lutz Walter, and John Trowsdale. "Comparative genomics of major histocompatibility complexes." Immunogenetics 56.10 (2005): 683-695.

Kelley, Richard L., and Mitzi I. Kuroda. "Noncoding RNA genes in dosage compensation and imprinting." Cell 103.1 (2000): 9-12.

Kua, Jeremy, and Jeffrey L. Bada. "Primordial ocean chemistry and its compatibility with the RNA world." Origins of Life and Evolution of Biospheres 41.6 (2011): 553-558.

Kumar, P., et al. “tRFdb: a database for transfer RNA fragments.” Nucleic Acids Research (Database Issue) doi:10.1093/nar/gku1138 (2014): Pubmed Link.

Larralde, Rosa, Michael P. Robertson, and Stanley L. Miller. "Rates of decomposition of ribose and other sugars: implications for chemical evolution." Proceedings of the National Academy of Sciences 92.18 (1995): 8158-8160.

Laslett, D., et al. "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences." Nucleic acids research 32.1 (2004): 11-16.

Lassar, A. B., Mrtin, P. L., & Roeder, R. G. (1983). Transcription of class III genes: formation of preinitiation complexes. Science, 222, 740-748.

Lercher, Martin J., Araxi O. Urrutia, and Laurence D. Hurst. "Clustering of housekeeping genes provides a unified model of gene order in the human genome." Nature genetics 31.2 (2002): 180.

Lincoln, T. A., et al. "Self-sustained replication of an RNA enzyme." Science 323.5918 (2009): 1229-1232.

Lott, M.T., et al. (2013). “mtDNA variation and analysis using MITOMAP and MITOMASTER. Current Protocols in Bioinformatics.” 1(123):1.23.1-26. PMID: 25489354 URL: http://www.mitomap.org

Maizels, N., et al. "The genomic tag hypothesis: what molecular fossils tell us about the evolution of tRNA." COLD SPRING HARBOR MONOGRAPH SERIES 37 (1999): 79-112.

Mak, Johnson, and Lawrence Kleiman. "Primer tRNAs for reverse transcription." Journal of Virology 71.11 (1997): 8087.

Makalowska, Izabela, Chiao-Feng Lin, and Wojciech Makalowski. "Overlapping genes in vertebrate genomes." Computational biology and chemistry 29.1 (2005): 1-12.

Mansy, S. S., et al. "Thermostability of model protocell membranes." Proceedings of the National Academy of Sciences 105.36 (2008): 13351-13355.

Maraia, Richard J., and Tek N. Lamichhane. "3′ processing of eukaryotic precursor tRNAs." Wiley Interdisciplinary Reviews: RNA 2.3 (2011): 362-375.

Marquet, R., et al. "tRNAs as primer of reverse transcriptases." Biochimie 77.1-2 (1995): 113-124.

Mi H., et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements.
Nucl. Acids Res. (2016) doi: 10.1093/nar/gkw1138

Mogk, Axel, Ronny Schmidt, and Bernd Bukau. "The N-end rule pathway for regulated proteolysis: prokaryotic and eukaryotic strategies." Trends in cell biology 17.4 (2007): 165-172.

Mungall, A. J., et al. "The DNA sequence and analysis of human chromosome 6." Nature 425.6960 (2003): 805.

Nawrocki, E.P. and Eddy, S.R. (2013) "Infernal 1.1: 100-fold Faster RNA Homology Searches", Bioinformatics, 29, 2933-2935.

Nelson, Audrey R., Tina M. Henkin, and Paul F. Agris. "tRNA regulation of gene expression: interactions of an mRNA 5′-UTR with a regulatory tRNA." Rna 12.7 (2006): 1254-1261.

Nielsen, Soren, Yulia Yuzenkova, and Nikolay Zenkin. "Mechanism of eukaryotic RNA polymerase III transcription termination." Science 340.6140 (2013): 1577-1580.

Ober, Carole, et al. "HLA and mate choice in humans." The American Journal of Human Genetics 61.3 (1997): 497-504.

Oficjalska-Pham, Danuta, et al. "General repression of RNA polymerase III transcription is triggered by protein phosphatase type 2A-mediated dephosphorylation of Maf1." Molecular cell 22.5 (2006): 623-632.

Ogata, K., et al. "The possible role of the ribonucleic acid (RNA) of the pH 5 enzyme in amino acid activation." Biochimica et biophysica acta 25.3 (1957): 659-660.

Orgel, L. E., "Prebiotic chemistry and the origin of the RNA world." Critical reviews in biochemistry and molecular biology 39.2 (2004): 99-123.

Orioli, Andrea, et al. "Human MAF1 targets and represses active RNA polymerase III genes by preventing recruitment rather than inducing long-term transcriptional arrest." Genome research (2016).

Pan, Tao. "Modifications and functional genomics of human transfer RNA." Cell research (2018): 1.

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

Pavon-Eternod, M., Gomes, S., Geslain, R., Dai, Q., Rosner, M. R., & Pan, T. (2009). tRNA over-expression in breast cancer and functional consequences. Nucleic acids research, 37(21), 7268-7280.

Perli, Elena, et al. "The isolated carboxy-terminal domain of human mitochondrial leucyl-tRNA synthetase rescues the pathological phenotype of mitochondrial tRNA mutations in human cells." EMBO molecular medicine (2014): e201303198.

Phizicky, Eric M., and Anita K. Hopper. "tRNA biology charges to the front." Genes & development 24.17 (2010): 1832-1860.

Piovesan, Allison, et al. "GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics." Database 2016 (2016).

Pluta, Krzysztof, et al. "Maf1p, a Negative Effector of RNA Polymerase III inSaccharomyces cerevisiae." Molecular and cellular biology 21.15 (2001): 5031-5040.

Polach, K. J., and J. Widom. "Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation." Journal of molecular biology 254.2 (1995): 130-149.

Prescott, Elizabeth M., and Nick J. Proudfoot. "Transcriptional collision between convergent genes in budding yeast." Proceedings of the National Academy of Sciences 99.13 (2002): 8796-8801.

Quigley, Gary J., and Alexander Rich. "Structural domains of transfer RNA molecules." Science 194.4267 (1976): 796-806.

Randau, Lennart, et al. "Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5′-and 3′-halves." Nature 433.7025 (2005A): 537.

Randau, Lennart, Michael Pearson, and Dieter Söll. "The complete set of tRNA species in Nanoarchaeum equitans." FEBS letters 579.13 (2005B): 2945-2947.

Ransohoff, Julia D., Yuning Wei, and Paul A. Khavari. "The functions and unique features of long intergenic non-coding RNA." Nature reviews Molecular cell biology 19.3 (2018): 143.

Ransom, Monica, Briana K. Dennehey, and Jessica K. Tyler. "Chaperoning histones during DNA replication and repair." Cell 140.2 (2010): 183-195.

Reina, Jaime H., Teldja N. Azzouz, and Nouria Hernandez. "Maf1, a new player in the regulation of human RNA polymerase III transcription." PloS one 1.1 (2006): e134.

Riepe, Andrea, Hildburg Beier, and Hans J. Gross. "Enhancement of RNA self-cleavage by micellar catalysis." FEBS letters 457.2 (1999): 193-199.

Rogers, T. E., et al. "A pseudo-tRNA modulates antibiotic resistance in Bacillus cereus." PLoS One 7.7 (2012): e41248.

Rollins, Janet, et al. "Human Maf1 negatively regulates RNA polymerase III transcription via the TFIIB family members Brf1 and Brf2." International journal of biological sciences 3.5 (2007): 292.

Ross, Wilma, et al. "The magic spot: a ppGpp binding site on E. coli RNA polymerase responsible for regulation of transcription initiation." Molecular cell 50.3 (2013): 420-429.

Rubio, M. A. T., et al. "Mammalian mitochondria have the innate ability to import tRNAs by a mechanism distinct from protein import." Proceedings of the National Academy of Sciences 105.27 (2008): 9186-9191.

Samaha, Raymond R., Rachel Green, and Harry F. Noller. "A base pair between tRNA and 23S rRNA in the peptidyl transferase centre of the ribosome." Nature 377.6547 (1995): 309.

Sampath, Prabha, et al. "Noncanonical function of glutamyl-prolyl-tRNA synthetase: gene-specific silencing of translation." Cell 119.2 (2004): 195-208.

Schramm, L., & Hernandez, N. (2002). Recruitment of RNA polymerase III to its target promoters. Genes & development, 16(20), 2593-2620.

Setzer, D. R., & Brown, D. D. (1985). Formation and stability of the 5 S RNA transcription complex. Journal of Biological Chemistry, 260(4), 2483-2492.

Sharp, S., et al. "Internal control regions for transcription of eukaryotic tRNA genes." Proceedings of the National Academy of Sciences 78.11 (1981): 6657-6661.

Sharp, S. J., et al. "Structure and transcription of eukaryotic tRNA gene." Critical Reviews in Biochemistry 19.2 (1985): 107-144.

Sloan, Joan, James R. Kinghorn, and Shiela E. Unkles. "The two subunits of human molybdopterin synthase: evidence for a bicistronic messenger RNA with overlapping reading frames." Nucleic acids research 27.3 (1999): 854-858.

Spurgin, Lewis G., and David S. Richardson. "How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings." Proceedings of the Royal Society B: Biological Sciences 277.1684 (2010): 979-988.

Stallmeyer, B., et al. "Human molybdopterin synthase gene: identification of a bicistronic transcript with overlapping reading frames." The American Journal of Human Genetics 64.3 (1999): 698-705.

Stelzer, G., Rosen, R., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Iny Stein, T., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., and Lancet, D. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis , Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5. [PDF]

Sültmann, Holger, et al. "Conservation of Mhc class III region synteny between zebrafish and human as determined by radiation hybrid mapping." The Journal of Immunology 165.12 (2000): 6984-6993.

Sun, Feng-Jie, and Gustavo Caetano-Anolles. "Transfer RNA and the origins of diversified life." Science Progress 91.3 (2008A): 265.

Sun, Feng-Jie, and Gustavo Caetano-Anollés. "The origin and evolution of tRNA inferred from phylogenetic analysis of structure." Journal of molecular evolution 66.1 (2008B): 21-35.

Sutherland, John D. "The origin of life—out of the blue." Angewandte Chemie International Edition 55.1 (2016): 104-121.

Sy, Jose, and Fritz Lipmann. "Identification of the synthesis of guanosine tetraphosphate (MS I) as insertion of a pyrophosphoryl group into the 3′-position in guanosine 5′-diphosphate." Proceedings of the National Academy of Sciences 70.2 (1973): 306-309.

Szostak, Jack W. "The eightfold path to non-enzymatic RNA replication." Journal of Systems Chemistry 3.1 (2012): 2.

Thompson, Jill, et al. "Analysis of mutations at residues A2451 and G2447 of 23S rRNA in the peptidyltransferase active site of the 50S ribosomal subunit." Proceedings of the National Academy of Sciences 98.16 (2001): 9002-9007.

Thompson, Martin, et al. "Nucleolar clustering of dispersed tRNA genes." Science 302.5649 (2003): 1399-1401.

Tripathi, Vidisha, et al. "The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation." Molecular cell 39.6 (2010): 925-938.

Trowsdale, John, and Julian C. Knight. "Major histocompatibility complex genomics and human disease." Annual review of genomics and human genetics 14 (2013): 301-323.

Tufarelli, Cristina, et al. "Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease." Nature genetics 34.2 (2003): 157.

Turowski, Tomasz W., et al. "Global analysis of transcriptionally engaged yeast RNA polymerase III reveals extended tRNA transcripts." Genome research (2016).

Turowski, Tomasz W., et al. "Transcription by RNA polymerase III: insights into mechanism and regulation." Biochemical Society Transactions 44.5 (2016): 1367-1375.
Tryoen-Tóth, Petra, et al. "Proteomic consequences of a human mitochondrial tRNA mutation beyond the frame of mitochondrial translation." Journal of Biological Chemistry 278.27 (2003): 24314-24323.

Upadhya, Rajendra, JaeHoon Lee, and Ian M. Willis. "Maf1 is an essential mediator of diverse signals that repress RNA polymerase III transcription." Molecular cell 10.6 (2002): 1489-1494.

Van Der Velden, Alike W., and Adri AM Thomas. "The role of the 5′ untranslated region of an mRNA in translation regulation during development." The international journal of biochemistry & cell biology 31.1 (1999): 87-106.

van Tol, Hans, Hans J. Gross, and Hildburg Beier. "Non-enzymatic excision of pre-tRNA introns?." The EMBO journal 8.1 (1989): 293-300.

Vandiedonck, Claire, and Julian C. Knight. "The human Major Histocompatibility Complex as a paradigm in genomics research." Briefings in functional genomics and proteomics 8.5 (2009): 379-394.

Wang, Peixiang, et al. "Demonstration that the TyrR Protein and RNA Polymerase Complex Formed at the Divergent P3 Promoter Inhibits Binding of RNA Polymerase to the Major Promoter, P1, of the aroP Gene ofEscherichia coli." Journal of bacteriology 180.20 (1998): 5466-5472.

Weber, Ute, Hildburg Beier, and Hans J. Gross. "Another heritage from the RNA world: self-excision of intron sequences from nuclear pre-tRNAs." Nucleic acids research 24.12 (1996): 2212-2219.

Wedekind, Claus, et al. "MHC-dependent mate preferences in humans." Proc. R. Soc. Lond. B 260.1359 (1995): 245-249.

Weiner, Alan M., and Nancy Maizels. "tRNA-like structures tag the 3'ends of genomic RNA molecules for replication: implications for the origin of protein synthesis." Proceedings of the National Academy of Sciences 84.21 (1987): 7383-7387.

West, Andrew B., et al. "Identification of a novel gene linked to parkin via a bi-directional promoter." Journal of molecular biology 326.1 (2003): 11-19.

Willis, Ian M., and Robyn D. Moir. "Integration of nutritional and stress signaling pathways by Maf1." Trends in biochemical sciences 32.2 (2007): 51-53.

Wilusz, Jeremy E., Susan M. Freier, and David L. Spector. "3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA." Cell 135.5 (2008): 919-932.

Wilusz, Jeremy E., et al. "A triple helix stabilizes the 3′ ends of long noncoding RNAs that lack poly (A) tails." Genes & development 26.21 (2012): 2392-2407.

Wood, Emily Jane, et al. "Sense-antisense gene pairs: sequence, transcription, and structure are not conserved between human and mouse." Frontiers in genetics 4 (2013): 183.

Yamazaki, K., et al. "Control of mating preferences in mice by genes in the major histocompatibility complex." Journal of Experimental Medicine 144.5 (1976): 1324-1335.

Yamazaki, K., et al. "Recognition among mice. Evidence from the use of a Y-maze differentially scented by congenic mice of different major histocompatibility types." Journal of Experimental Medicine 150.4 (1979): 755-760.

Yelin, Rodrigo, et al. "Widespread occurrence of antisense transcription in the human genome." Nature biotechnology 21.4 (2003): 379.

Yu, Wenqiang, et al. "Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA." Nature 451.7175 (2008): 202.

Zamore, Phillip D. "Ancient pathways programmed by small RNAs." Science 296.5571 (2002): 1265-1269.

Zerbino, Daniel R., et al. Ensembl 2018. PubMed PMID: 29155950. doi:10.1093/nar/gkx1098

Zhang, Bin, et al. "The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult." Cell reports 2.1 (2012): 111-123.

APPENDIX A

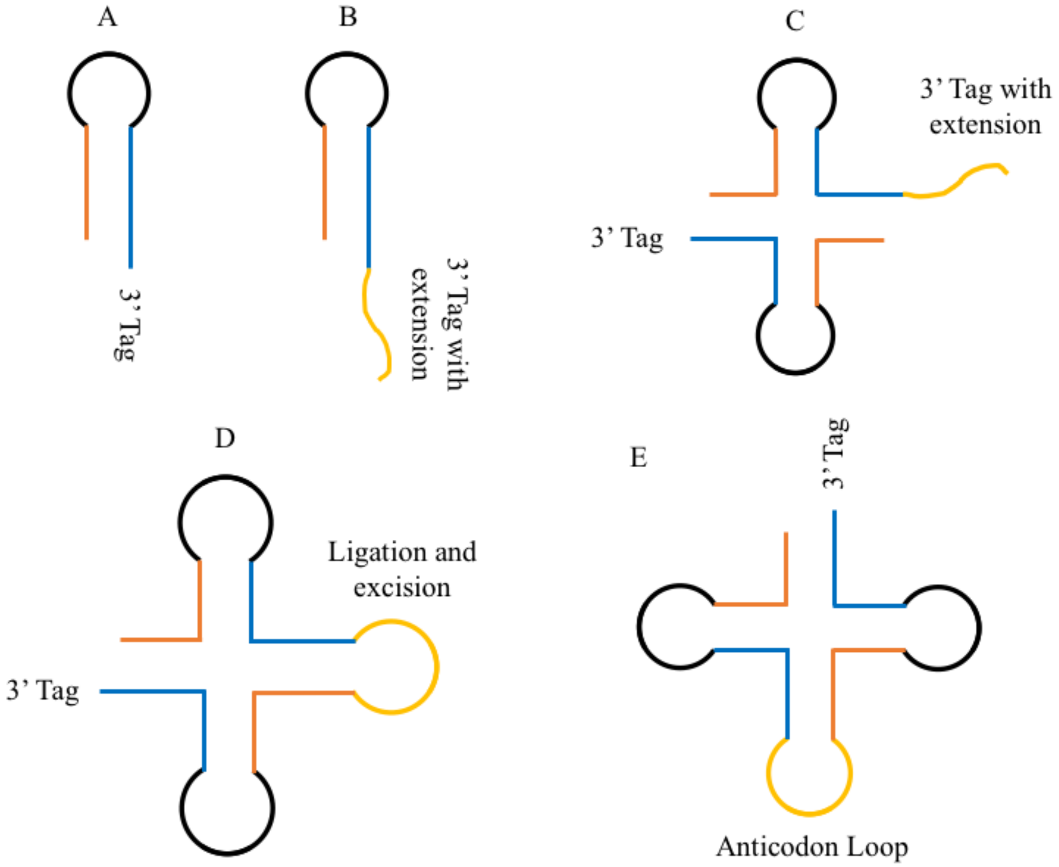

Figure 1. The origin of tRNA adapted from models presented by Di Giulio, and Dick and Schamel (Di Giulio, M. G., 1992; Dick, T. P., et al. 1995). A. An original RNA oligonucleotide with a stem and loop secondary structure. B. A replicated RNA oligonucleotide with an aberrant run-off sequence. C. Complexed molecules based on Watson and Crick pairing rules. D. The run-off sequence is ligated and the intronic region is self-spliced. E. A final tRNA-like molecule.

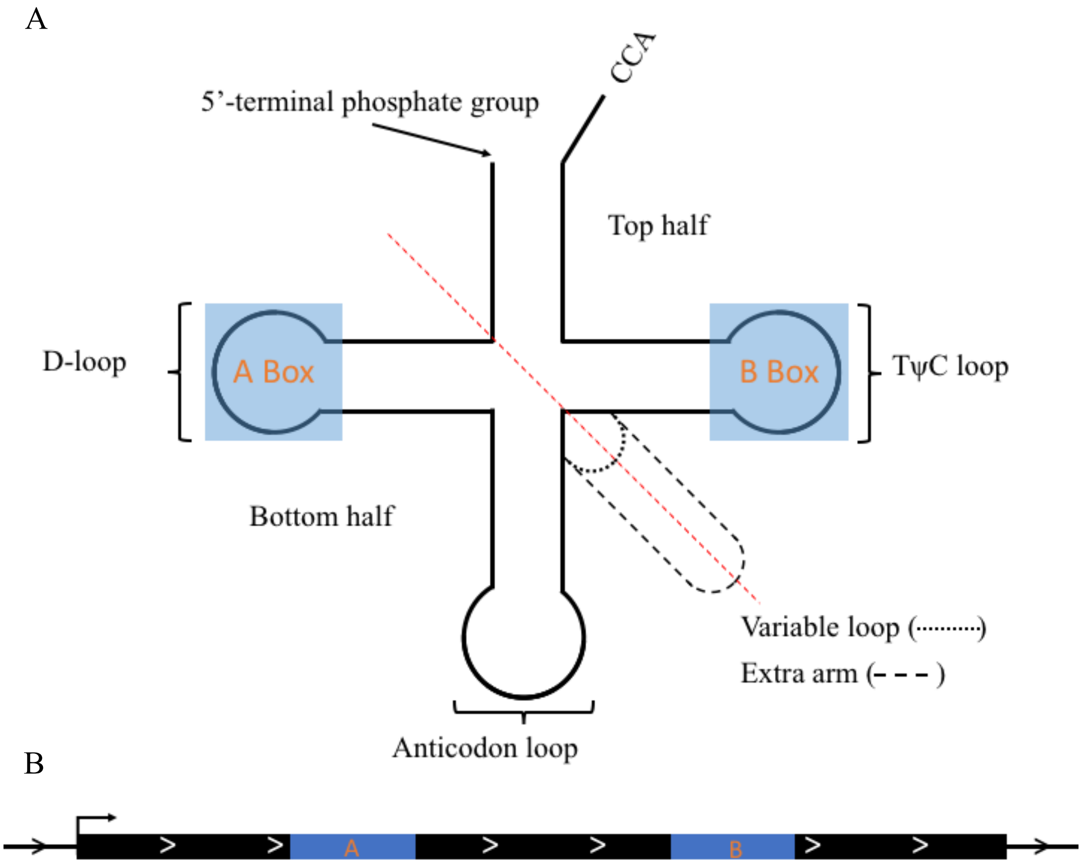

Figure 2. A typical tRNA gene model and distinct cloverleaf secondary structure. A. The distinct cloverleaf secondary structure of a typical tRNA transcript. The conserved A- and B-box promoter regions (light blue boxes) form the D-loop and TψC-loop functional regions. B. A typical tRNA gene model highlighting the type-2 A- and B-box intragenic promoters (blue boxes). The black arrow represents the transcription start site (TSS) of the gene.

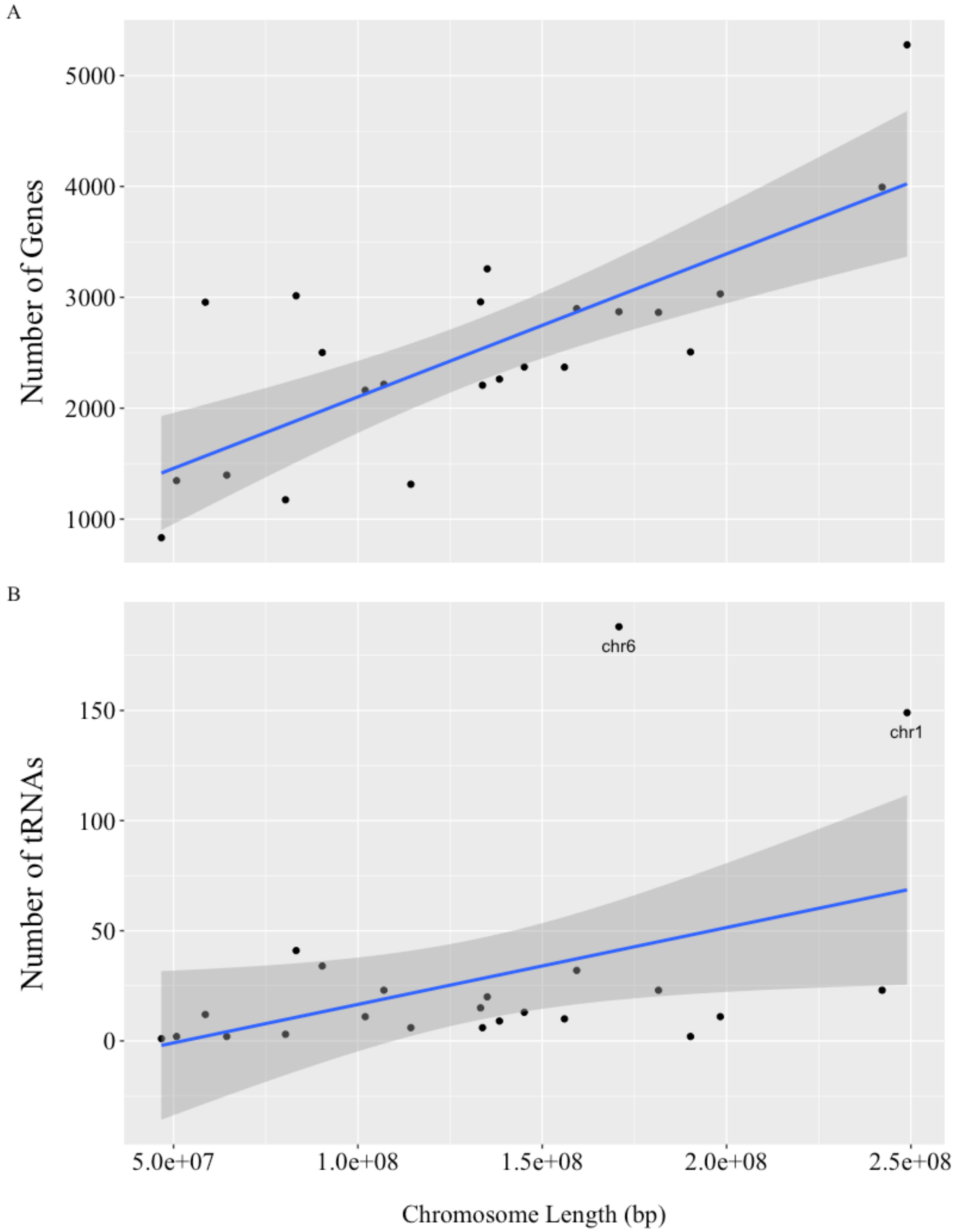

Figure 3. The correlation of all genes, tRNA genes, and chromosome length. A. The correlation of the total number of genes per chromosome and the chromosome length in the human genome (r^2=0.57). B. The correlation of the total number of tRNA genes per chromosome and the chromosome length in the human genome (r^2=0.19).

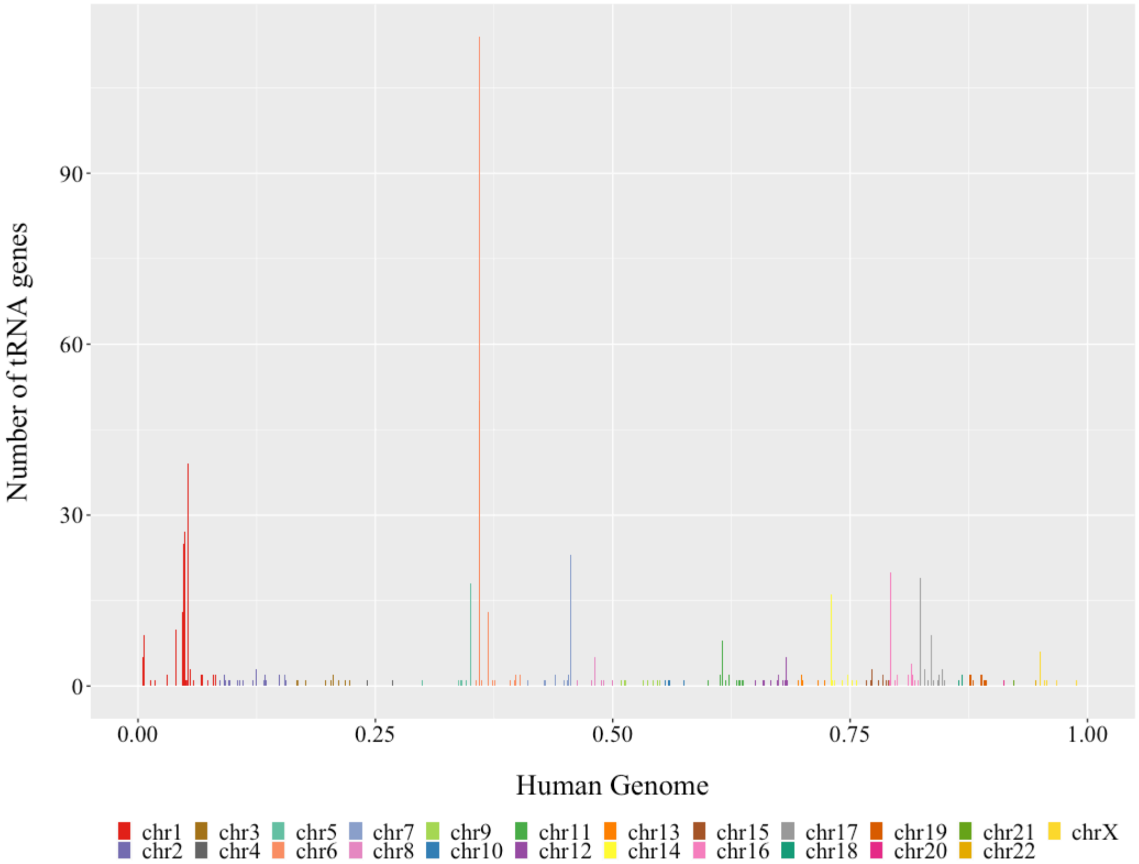

Figure 4. A histogram showing the ordinal distribution of tRNA genes in the human genome (GRCh38 Gencode.v28). Human chromosomes are ordered such that the first base of chromosome 1 corresponds to position '0' on the x-axis, and the last base of chromosome X corresponds to position '1' on the x-axis. The start positions of all tRNA genes were calculated relative to their position in the genome. There is a clear enrichment of tRNA genes in chromosomes 1 (red) and 6 (orange) with a distinct concentration in chromosome 6. Binwidth = 3Mbp.

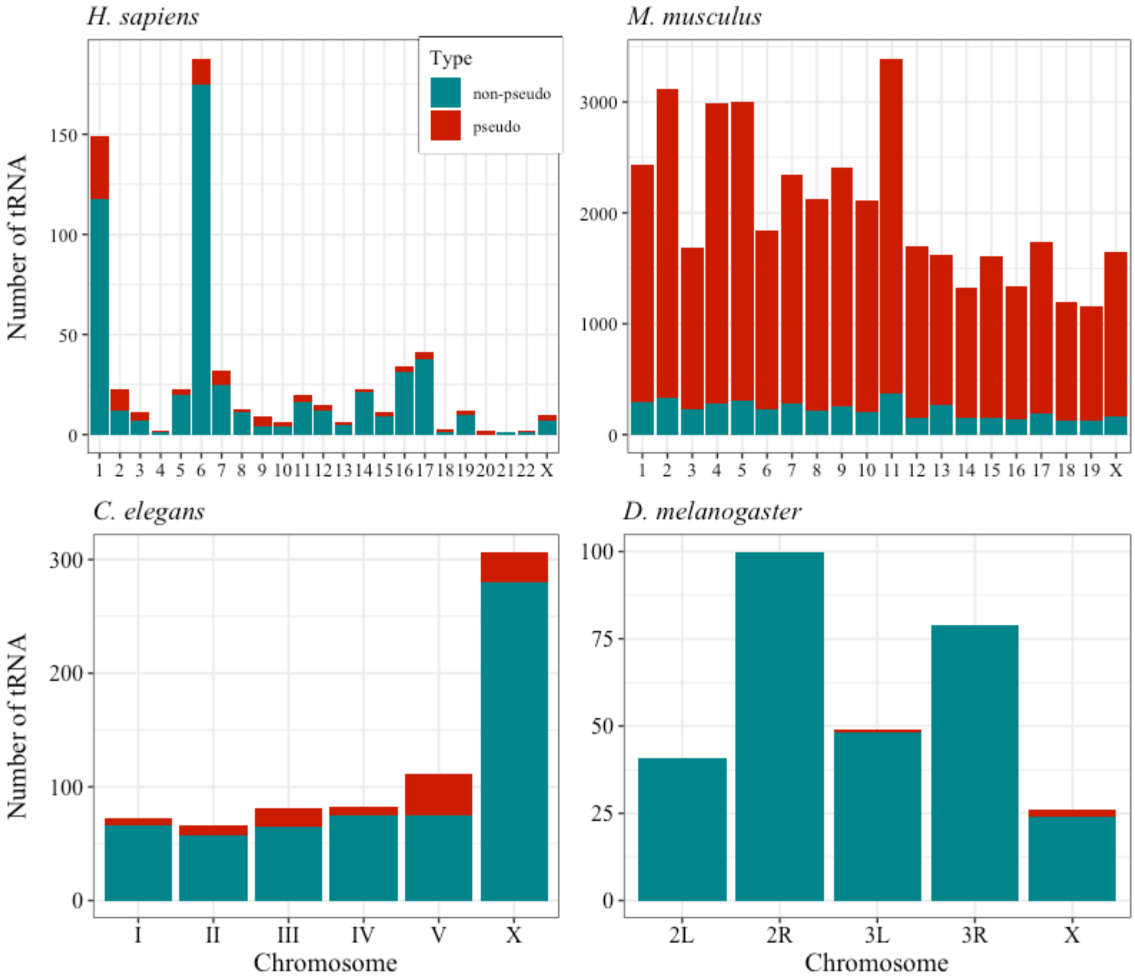

Figure 5. The total number of tRNA genes predicted by tRNAscan-SE in humans and three popular model organisms; *Mus musculus* (mouse), *Caenorhabditis elegans* (nematode), and *Drosophila melanogaster* (fruit fly). Pseudo genes are shown in red and non-pseudo genes are shown in turquoise. Comparatively, humans are the only species shown here that exhibit a clear and distinct tRNA gene enrichment.

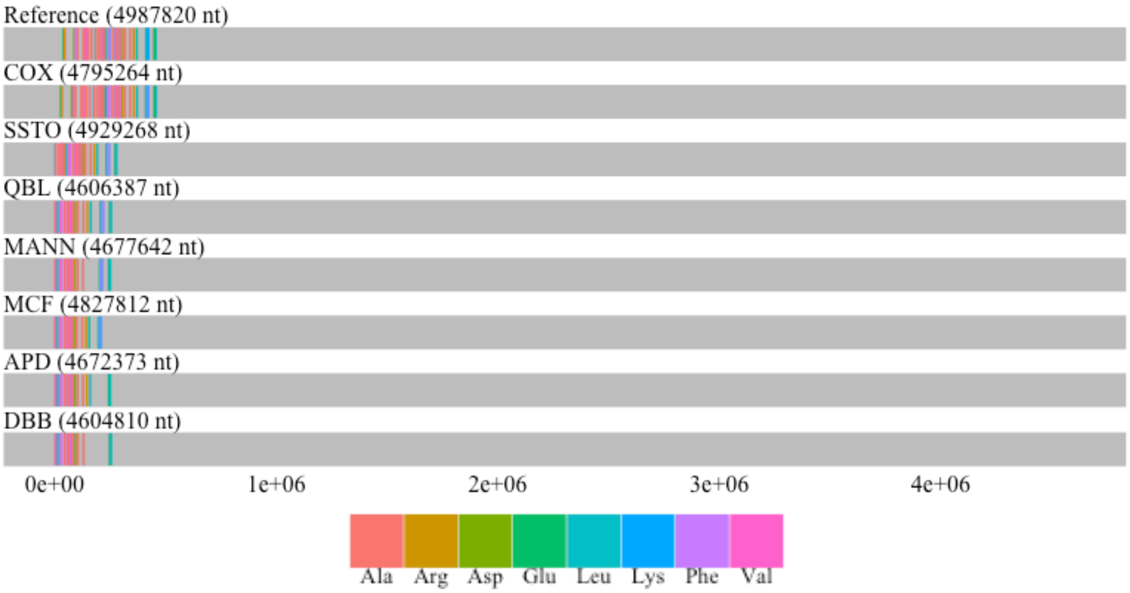

Figure 6. Assembly exceptions at the MHC region of human chromosome 6. Ensembl (release 95) has indicated seven assembly exceptions named COX, SSTO, QBL, MANN, MCF, APD, and DBB. tRNAscan-SE identifies a dense cluster of tRNA genes immediately downstream of the 5' boundary of each exception with no other tRNAs predicted within the region. The vertical lines represent the absolute start position of each predicted tRNA gene colored by species type. The length of each assembly, including the reference sequence, is indicated in parentheses.

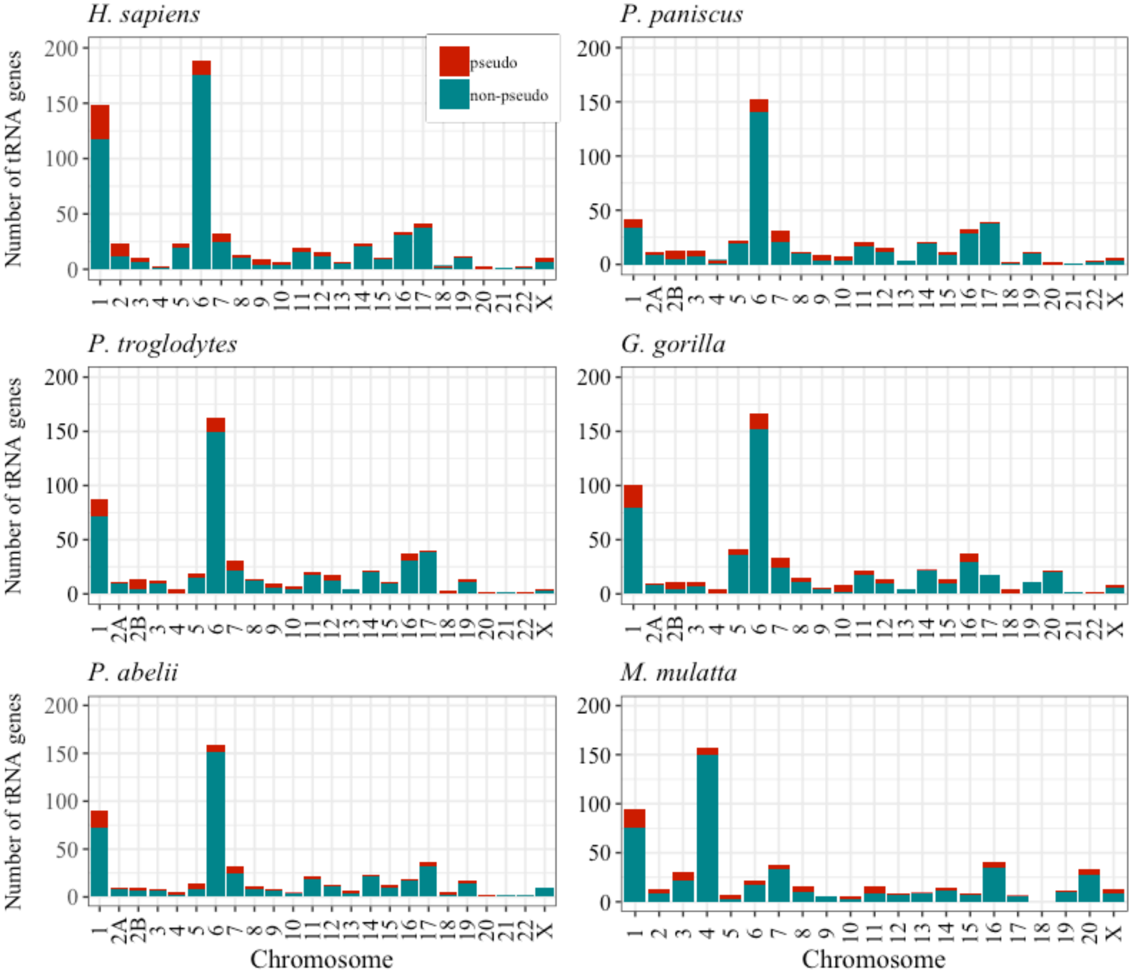

Figure 7. tRNA gene distribution in five apes (*Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla*, and *Pongo abelii*), and one Old World monkey (*Macaca mulatta*). The apes exhibit a similar distribution with an apparent tRNA gene enrichment in chromosomes 1 and 6, whereas the Old World monkey appears to have an enrichment in chromosomes 1 and 4. Pseudo genes are shown in red and non-pseudo genes are shown in turquoise.

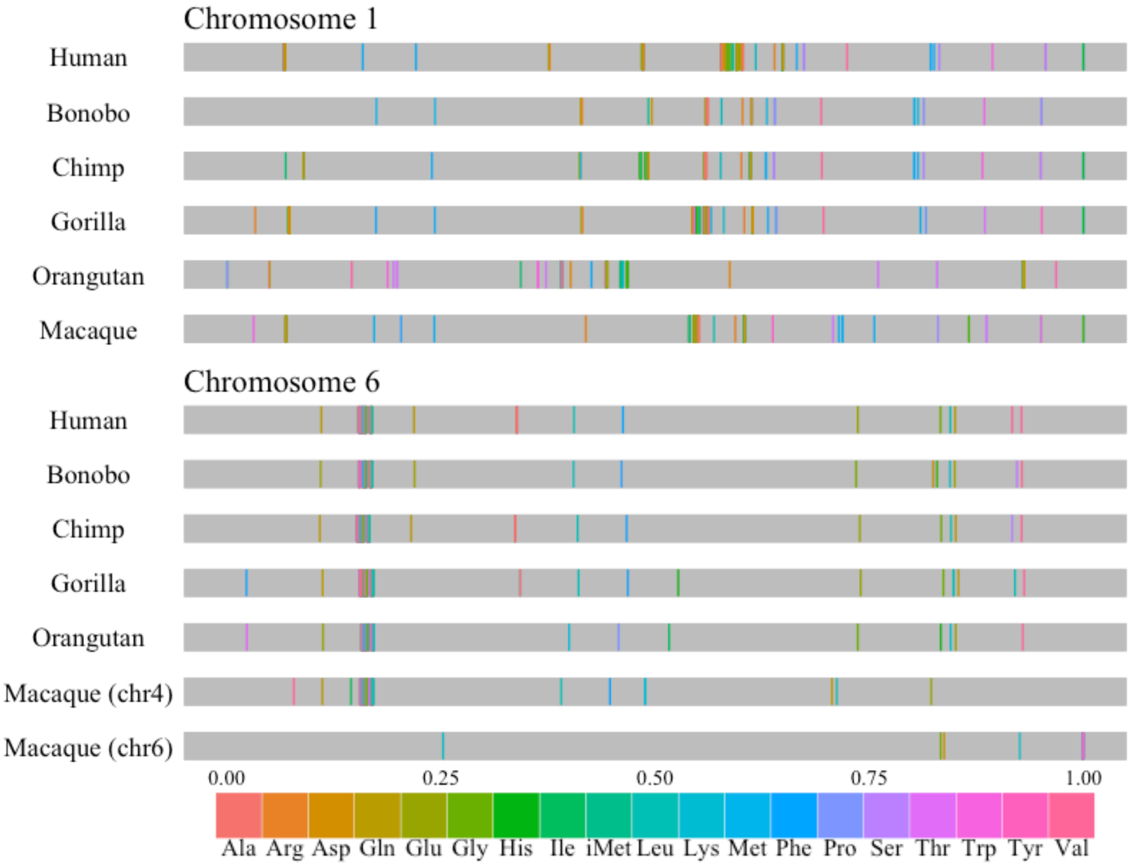

Figure 8. The distribution of tRNA genes with respect to species type in chromosomes 1 and 6 of human, bonobo, chimp, gorilla, orangutan, and macaque. For the macaque, chromosomes 6 and 4 are shown as chromosome 4 more closely resembles the order and species of tRNA genes with the apes rather than chromosome 6. Colored vertical lines indicate the loci of tRNA genes as predicted by tRNAscan-SE. Chromosome length was normalized and ranges in value from 0-1.

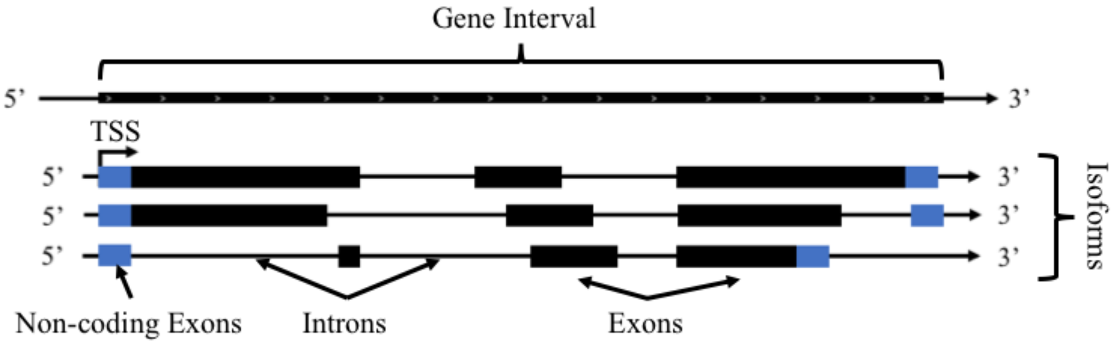

Figure 9. A typical gene model. The top bar represents the entire interval of the gene. Blue boxes represent non-coding exons (e.g., UTRs), black boxes represent exons, thin black lines represent the intronic regions, and the bent black arrow represents the transcription start site (TSS). This example illustrates three possible isoforms transcribed from the same genic interval.

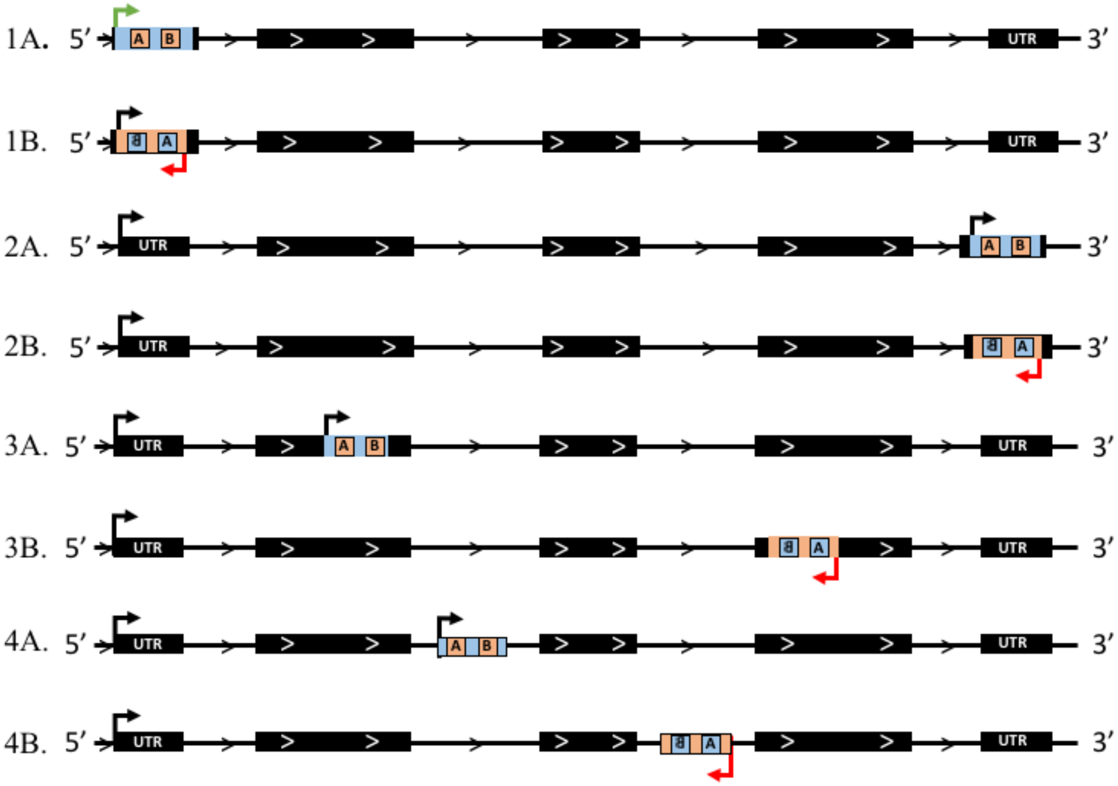

Figure 10. Eight possible ways a tRNA gene can intersect a protein coding gene. (1A and 1B) A tRNA gene intersecting a 5' UTR in the sense and antisense orientation respectively. (2A and 2B) A tRNA gene intersecting a 3' UTR in the sense and antisense orientation respectively. (3A and 3B) A tRNA gene intersecting a coding exon in the sense and antisense orientation respectively. (4A and 4B) A tRNA gene intersecting an intronic region in the sense and antisense orientation respectively. Light blue boxes with orange A and B boxes represent a tRNA gene in the sense orientation with respect to the protein coding gene. Orange boxes with backwards light blue A and B boxes represent a tRNA gene in the antisense orientation with repsect to the protein coding gene. Green arrows represent a possible TSS overlap with RNA-pol II and RNA-pol III. Black arrows represent a typical TSS for RNA-pol II. Red arrows represent possible transcriptional interference between RNA-pol II and RNA-pol III.

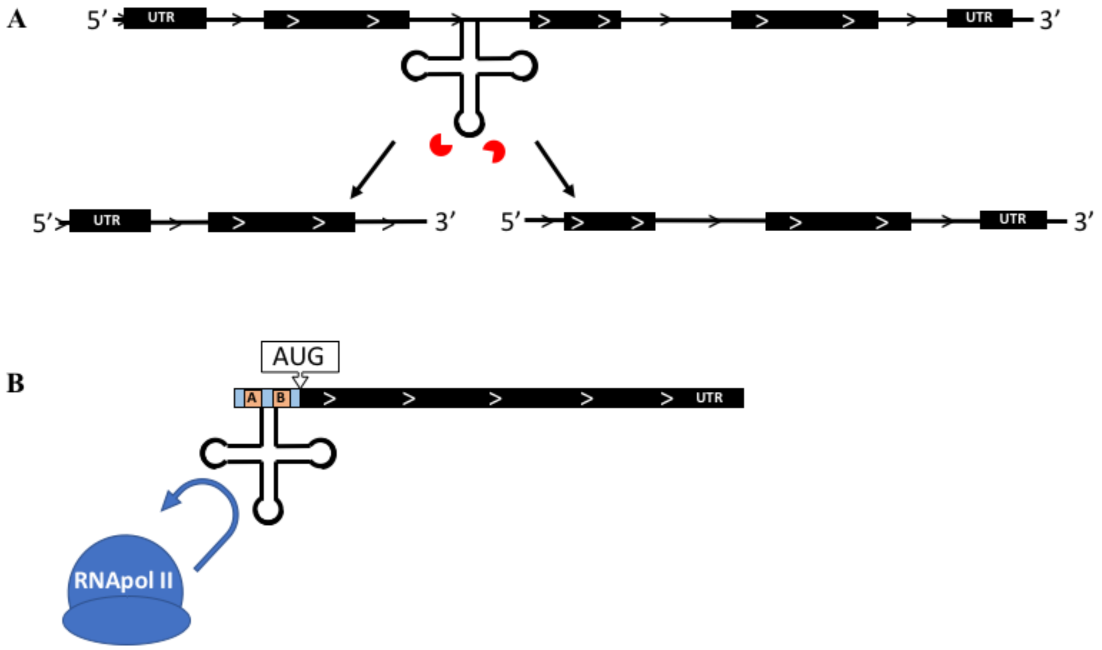

Figure 11. Possible modifications to tRNA genes intersecting protein coding genes. A. A tRNA-like structure in an intronic region of a primary transcript. Endonucleases may be recruited to the structure and splice the transcript. B. A tRNA-like structure in the 5’ UTR of a mature transcript. Modifications that stabilize the structure may block the translational machinery from binding.

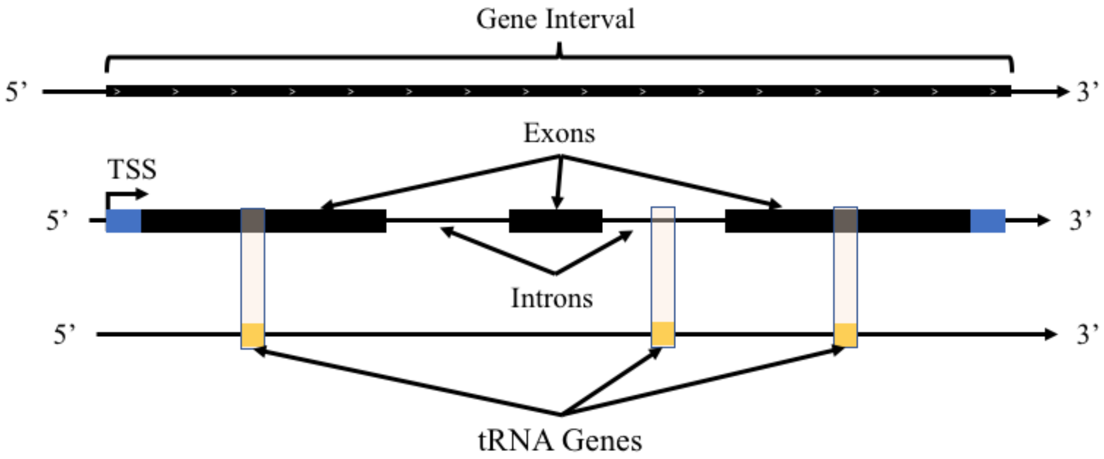

Figure 12. A graphical representation of the intersect analysis. The top model is the gene interval which contains non-coding (blue rectangles) and coding (black rectangles) exons, and introns (thin black lines). The bottom model indicates the location of tRNA genes (yellow squares). The vertical transparent yellow rectangles represent the overlap of tRNA genes and certain features within the gene interval (i.e., coding exon or intron).

APPENDIX B

Gene Type	Count
Protein Coding Isoforms	149592
Protein Coding Genes	19901
Processed Pseudogene	10219
lincRNA	7490
Antisense	5501
Unprocessed Pseudogene	2664
Miscellaneous RNA	2213
snRNA	1900
miRNA	1881
TEC	1067
snoRNA	943

Table 1. The most abundant gene types as defined by Gencode.v28 (GRCh38.p12).

Chromosome	Length (bp)	Total Number of Genes	Total Number of tRNA
1	248956422	5277	149
2	242193529	3993	23
3	198295559	3031	11
4	190214555	2507	2
5	181538259	2864	23
6	170805979	2870	188
7	159345973	2898	32
8	145138636	2372	13
9	138394717	2262	9
10	133797422	2207	6
11	135086622	3257	20
12	133275309	2959	15
13	114364328	1314	6
14	107043718	2216	23
15	101991189	2162	11
16	90338345	2502	34
17	83257441	3014	41
18	80373285	1174	3
19	58617616	2956	12
20	64444167	1397	2
21	46709983	832	1
22	50818468	1347	2
X	156040895	2370	10

Table 2. The length of each chromosome and the number of total genes and tRNA genes in the human genome (GRCh38.p12; Gencode.v28).

Database	Total tRNA Genes
UCSC	631
tRNAscan-SE	619
tRNAdb	359
tRFdb	625
Program	**-**
tRNAscan-SE	636
Aragorn	916

Table 3. The total number of tRNA genes reported by four databases and two programs.

PANTHER GO-Slim Biological Process Chr1	FE	P-value	FDR
Nucleosome Assembly (GO:0006334)	39.49	7.93E-05	4.75E-02
Protein Folding (GO:0006457)	19.58	8.07E-07	1.45E-03
Peptidyl-amino Acid Modification (GO:0018193)	12.45	1.00E-05	9.00E-03
PANTHER GO-Slim Biological Process Chr6	**FE**	**P-value**	**FDR**
Nucleosome Assembly (GO:0006334)	> 100	6.71E-19	1.21E-15
T cell Receptor Signaling Pathway (GO:0050852)	99.27	9.43E-11	3.39E-08
Cellular Component Assembly (GO:0022607)	71.22	1.83E-17	1.64E-14
Cellular Component Organization or Biogenesis (GO:0071840)	24.48	8.56E-14	5.13E-11
Cellular Component Biogenesis (GO:0044085)	24.48	8.56E-14	3.84E-11
Antigen Receptor-mediated Signaling Pathway (GO:0050851)	21.44	4.52E-07	1.35E-04
Immune Response-activating Cell Surface Receptor Signaling Pathway (GO:0002429)	19.71	7.26E-07	1.86E-04
Immune Response-regulating Cell Surface Receptor Signaling Pathway (GO:0002768)	19.71	7.26E-07	1.63E-04
Immune Response (GO:0006955)	7.66	1.34E-04	2.68E-02

Table 4. Regions of human chromosomes 1 and 6 with dense tRNA gene clusters also enriched with gene ontology (GO) terms related to nucleosome assembly and adaptive immunology. There is a nearly 40-fold and 100-fold enrichment (FE) of GO terms related to nucleosome assembly in chromosomes 1 and 6 respectively. The false discovery rate (FDR) is much lower in chromosome 6 (1.2e-15) than chromosome 1 (4.8e-2). Additionally, chromosome 6 has a nearly 100-fold enrichment (FE) of GO terms associated with T cell receptor signaling pathways with an FDR of 3.4e-8.

Gene Type	Count
Protein Coding	79
lincRNA	30
Antisense	11
Transcribed Unprocessed Pseudogene	8
Bidirectional Promoter (lncRNA)	2
Processed Transcript	2
Sense Intronic	2
Sense Overlapping	2
TEC	1
Polymorphic Pseudogene	1

Table 5. The gene type and total count of genes that have an intersecting tRNA gene.

	3UTR Convergent Between Ensembl, RefSeq, and Gencode			
Gene	FCGR2A (+)	HES7 (-)	CTC1 (-)	ZNF136 (+)
Human	2 trnas predicted; AspGTC (+), GlyGCC (-)	1 tRNA predicted; ArgTCT (+)	3 tRNAs predicted; IleAAT (-), SerAGA (-), ThrAGT (-)	1 tRNA predicted; AlaGGC (+)
Mouse	no tRNA predicted	1 tRNA predicted; ArgTCT	1 tRNA predicted; SerGCT [pseudo]	1 tRNA predicted; SerGCT
Worm	no orthologs	no orthologs	no orthologs	no orthologs
Fly	no orthologs	no orthologs	no orthologs	no orthologs
	5UTR Convergent Between Ensembl, RefSeq, and Gencode			
Gene	SHF (-)	NDUFS7 (+)	DPP9 (-)	
Human	3 tRNAs predicted; 3xHisGTG (-) (-) (+)	2 tRNAs predicted; AsnGTT (+), PheGAA (-)	2 tRNAs predicted; GlyTCC, ValCAC (-)	
Mouse	no tRNA predicted	2 tRNAs predicted; AsnGTT, PheGAA	1 tRNA predicted; SerGCT	
Worm	no orthologs	no tRNA predicted	no tRNA predicted	
Fly	no tRNA predicted	ND20: no tRNA predicted	no tRNA predicted	
		ND20L: no tRNA predicted		
	3UTR Unique to Gencode			
Gene	SACM1L (+)	TEC (?)	VAC14 (-)	
Human	1 tRNA predicted; ArgACG (-)	no tRNA predicted	4 tRNAs predicted; 4xGlyGCC 2x(-) 2x(+)	
Mouse	2 tRNAs predicted; SerGCT, ArgACG	2 tRNAs predicted; 2xSerGCT [pseudo]	4 tRNAs predicted; 2xGlyGCC, SerGCT, AlaGGC	
Worm	no tRNA predicted	no orthologs		
Fly	no tRNA predicted	1 tRNA predicted; LysTTT		
	5UTR Unique to Gencode			
Gene	ZBED9 (-)			
Human	3 tRNAs predicted; GlnTTG* (-), AlaAGC (+), SerGCT (-)			
Mouse	no orthologs			
Worm	no orthologs			
Fly	no orthologs			
	Exon (Gencode)			
Gene	PHIP (-)	MAP1LC3B (+)		
Human	1 tRNA predicted; PheGAA (-)	1 tRNA predicted; MetCAT (-)		
Mouse	1 tRNA predicted; SerGCT	no tRNA predicted		
Worm	no orthologs	no orthologs		
Fly	no orthologs	no orthologs		

Table 6. tRNA genes found to intersect coding and non-coding exons. FCGR2A, HES7, CTC1, SHF, NDUFS7, and DPP9 were identified as having tRNA genes intersecting either the 5' or 3' UTRs and were verified by running the analysis using data from Ensembl, RefSeq, and Gencode databases. SACM1L, VAC12, ZBED9, MAPILC3B, and PHIP were unique to Gencode. (+) and (-) indicate strict (not relative) directionality. 'OL' indicates whether there is an overlapping feature at the site of an intersect. The sequence of all indicated genes was extracted and independently analyzed for tRNA sequences by tRNAscan-SE.