Title

Transfer RNA Dynamics in Animal Neurogenesis and Fungal Evolution

Author

Wint,, Rhondene

Publication Date

2022

UNIVERSITY OF CALIFORNIA, MERCED

TRANSFER RNA DYNAMICS IN ANIMAL NEUROGENESIS AND FUNGAL EVOLUTION

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

in

Quantitative and Systems Biology

by

Rhondene Wint

Dissertation Committee:
Professor Ramendra Saha, Chair
Professor Suzanne Sindi
Professor Juris Grasis
Professor David Ardell
Professor Michael Cleary
2022

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Professor Michael Cleary (Principal Advisor)

Professor David Ardell (Co-Advisor)

Professor Ramendra Saha (Chair)

Professor Suzanne Sindi (Committee Member)

Professor Juris Grasis (Committee Member)

University of California, Merced 2022

Dedication

To my late father, Lloyd L. Wint, who fostered my love for reading and, unlike me, did not need permission from a committee of learned scholars to periodically refer to himself as ‘Dr. Wint’. And to my mother, Judith Hall-Wint, for her continuous support and unrelenting belief in education as a vehicle for socioeconomic mobility.

Table of Contents

Chapter 1: Literature Review

Chapter 2: Dynamic changes in tRNA expression establish proliferation- and differentiation-specific codon optimality in neurogenesis

Chapter 3: Parametric and Non-Parametric Estimation of Differential tRNA expression analysis using DESeq2 and NOISeqBio

Chapter 4: Kingdom-wide Analysis of Fungal Transcriptomes and tRNAs Reveals Conserved Patterns of Adaptive Evolution

Lists of Figures

Chapter 2: Dynamic changes in tRNA expression establish proliferation- and differentiation-specific codon optimality in neurogenesis

Chapter 3 Figures: Parametric and Non-Parametric Estimation of Differential tRNA expression analysis using DESeq2 and NOISeqBio

Lists of Figures

Chapter 4 Figures: Kingdom-wide Analysis of Fungal Transcriptomes and tRNAs Reveals Conserved Patterns of Adaptive Evolution

Acknowledgments

This research was supported by funding from the National Institutes of Health, the University of California Cancer Research Coordinating Committee (CRCC), and Cal-BRAIN (the California Blueprint for Research to Advance Innovations in Neuroscience).

Chapter 4 of this dissertation is a reprint of the material as it appears in "*Kingdom-wide Analysis of Fungal Protein-coding Genes and tRNAs Reveals Conserved Patterns of Adaptive Evolution*." I thank Igor Grigoriev (Lawrence Berkeley National Laboratory/ Joint Genome Institute), who is listed as the corresponding author in this publication, for supervising this research which is the basis of Chapter 4." The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

My deepest gratitude to Dr. Michael Cleary for his unwavering support and mentorship, especially for providing me with the opportunity to be trained as a scientist in developmental and molecular biology.

I would like to express my deepest appreciation to my co-advisor, Dr. David Ardell, for his constant guidance and encouragement. I have benefitted tremendously from his mentorship in computational genomics and genome evolution.

Finally, I would like to thank the current and former members of my labs for their support and camaraderie. From the Cleary Lab: Naoki Hida, Mohamed Aboukilila, Jade Fee Quail, and Josephine Sami. From the Ardell Lab: Fateme Hadi Nezhad and Andrea Collins-Hed.

Curriculum Vitae

Rhondene J. Wint
5200 N Lake Road Merced, CA 95343| rwint@ucmerced.edu
https://github.com/rhondene

CAREER OBJECTIVE

Obtain a full-time position in computational biology, ideally at the intersection of genomics and cellular development/human diseases.

EDUCATION

08/2017 to 07/2022 University of California, Merced
Doctor of Philosophy in Molecular and Cell Biology
Ph.D. Advisors: Dr. Michael Cleary and Dr. David Ardell

2010-2014 Northern Caribbean University
B.Sc. in Biological Sciences (*magna cum laude)*

Selected online certifications: Machine Learning (Stanford/Coursera), DeepLearning.ai (Coursera)

COMPUTATIONAL SKILLS

- Programming Languages: Python, R, UNIX/Bash, git for version control
- Analysis of next-generation sequence (NGS) data
- Statistical modeling and hypothesis testing
- Predictive modeling using machine learning and deep learning
- Phylogenetic comparative methods
- Experience working in a high-performance computing environment (SLURM)

MOLECULAR BIOLOGY “Wet Lab” SKILLS

- Standard molecular biology assays: RT-qPCR, gel electrophoresis, plasmid purification
- RNA metabolic labeling for estimating mRNA decay rates
- cDNA library preparation for Illumina sequencing (especially small RNA/ tRNA sequencing)

WORK EXPERIENCE

2019 to Present *Certified Software & Data Carpentry Instructor* Data Carpentry
Organize and teach workshops for introductory programming and data science

June-Aug 2018 *Bioinformatics Inter*n LBNL/Joint Genome Institute

- Performed large-scale genomics analysis of over 400 newly sequenced fungal species
- Delivered weekly research presentations at the intra- and inter-departmental JGI lab meetings

TEACHING EXPERIENCE

08/2020 – 05/2021 *UC Merced Living Learning Community Teaching Fellowship*

- Served as the instructor of record for *BeyondMD LLC* and designed a curriculum aimed at fostering a sense of belonging among 1st year STEM undergrads through mentorship and professional development activities

01/2018 – 06/2020 *Teaching Assistant* UC Merced School of Natural Sciences

- Taught biology and bioinformatics courses to undergraduates

09/2014 – 07/2017 *High School Teacher* Westwood High School (Jamaica)

- Taught biology, chemistry, and physics at CAPE level (equivalent to AP in the U.S.)

SELECTED AWARDS & FELLOWSHIPS

02/2022 *Diverse: Issues in Higher Education's Rising Graduate Scholar*
06/2019 *UC Merced Quantitative and Systems Biology Summer Research Fellowship*
06/2018 *UC Merced Computational Biology Fellowship*
08/2017 *UC Merced Graduate Division Recruitment Fellowship*

2012&2013 Jamaica National Bank Tertiary Full Scholarship

2010-2014 Jamaica Union Undergraduate Scholarship (50% Tuition waiver)

PROFESSIONAL DEVELOPMENT

06/2020- 08/2020 *Facebook Above & Beyond Bootcamp for Non-CS STEM Ph.D. students*

- 8-week invitation-only training workshop on data structures & algorithms and best practices for technical coding interviews led by Facebook (Meta) software engineers.

02/2019 *Full-Stack Deep Learning bootcamp Organizers: UC Berkeley, OpenAI and Turnit-In*

- The 2-day workshop focussed on deploying deep learning models in production

01/2018 - 05/2018 *NSF-NRT Interdisciplinary Computational Graduate Education (ICGE)*

- A project-oriented program that trains 1st-year graduate students in computational methods for research. As a capstone project, we collaborated with UC Merced's IT department to test our machine learning phishing detector that outperformed college-enrolled humans by 1.4%.

OUTREACH & EXTRACURRICULAR ACTIVITIES

2018 & 2019 *Co-organizer of Northern California Computational Biology Symposium*

- Annual student-led conference (>100 attendees) organized by graduate students from UC Berkeley, UC Santa Cruz, UCSF, UC-Davis, UC Merced, and Stanford (nccb.io).

Abstract

Transfer RNA Dynamics in Animal Neurogenesis and Fungal Evolution

By

Rhondene Wint

Doctor of Philosophy in Quantitative and Systems Biology
University of California, Merced
Professor Michael Cleary, Dissertation Advisor
Professor David Ardell, Dissertation Co-Advisor

Systems biology takes an integrative approach to connecting molecular-level mechanisms to cellular physiology and developmental outcomes. In metazoans, cellular development converges onto either a proliferative or differentiated state. Precise spatiotemporal regulation of gene expression is paramount for defining cellular proliferation and differentiation programs. During gene expression, the genetic information stored in a segment of DNA is copied into a messenger RNA (mRNA). This “message” within the mRNA is organized as a sequence of discrete units called codons. Each mRNA codon is then translated by its complementary transfer RNA (tRNA) molecule into a protein, the molecules that ultimately shape the cellular identity and function. Recently, codon identity emerged as a key regulatory grammar for post-transcriptional control of animal gene expression. Prior work from my laboratory in *Drosophila* demonstrated that certain “optimal” codons enhanced the stability of mRNAs in embryonic tissue, but the stabilizing effect of these codons was attenuated in the differentiated embryonic neural tissue. Current models suggest that optimal codons – codons that are enriched in high expressed and/or stable mRNAs – are preferentially decoded by abundant tRNAs. tRNAs are the adaptor molecules of mRNA translation. Mounting evidence from genomics experiments now overrides the longstanding view of tRNAs as uniformly expressed housekeeping molecules. However, genome-wide tRNA measurements, especially *in vivo* data, are usually missing from codon optimality studies in metazoans. Hence, our understanding of tRNA regulation in normal tissue development remains fragmentary. The *Drosophila* larval central nervous system offers a tractable model for studying the genetic control of cell proliferation and differentiation because of the suite of available genetic tools that enables cell type-specific assays. Here I demonstrate how altered tRNA expression establishes cell-type specific codon optimality that dynamic programs of mRNA decay and translation efficiency in neurogenesis, using the *Drosophila melanogaster* central nervous system.

By quantifying the tRNA transcriptome, mRNA transcriptome, and mRNA degradome in neural progenitor and post-mitotic neurons, for the first time, I present evidence that supports the dynamic regulation of the tRNA levels and post-transcriptional modification across neural differentiation serves as a mechanism for coordinating the translation and stability of functionally related mRNAs in a codon-dependent manner. Collectively, these findings support a mechanistic link between translational control by tRNA and neural differentiation.

Thus, my work lays the foundation for future investigations of dynamic tRNA regulation in cell-fate determination in other tissue types as well as in other complex animal models.

Chapter 1: Literature Review

1. Evolution of Codon Usage Bias in Simple Organisms and Animals

“Nothing in Biology Makes Sense Except in the Light of Evolution” **-** evolutionary biologist Theodosius Dobzhansky (1973).

Codon optimality – the principle that specific mRNA codons confer a fitness advantage – is a product of evolution. Emerging evidence supports codon optimality as a key determinant of post-transcriptional regulation in animal cell development. But this was not always the case. Long-held assumptions about the influence of selection on animal codon usage patterns have only recently been challenged due to the greater availability of tissue-specific RNAseq and bulk tRNA measurements. Here, I review primary research that has both shaped and challenged our understanding of the evolutionary basis of codon usage bias (CUB).

1.1 “Secondary genetic code”: Synonymous Codon Usage Bias

All living cells operate under the central dogma of molecular biology (Crick 1970), the universal algorithm that specifies the flow of genetic information from nucleic acids to proteins. Gene expression describes the mechanisms responsible for transducing the biological instructions stored in DNA to direct the synthesis of proteins, the molecules that ultimately establish cellular identity and function. During gene expression, a segment of DNA – a gene – is first copied (transcription) into a messenger RNA (mRNA) molecule. The coding region of the mRNA specifies the sequence of the protein product. mRNA coding regions are organized as a sequence of nucleotide triplets, known as codons. Each mRNA codon is then translated by its complementary transfer RNA (tRNA) molecule into a specific amino acid, the building block of proteins.

The degenerate genetic code maps 61 codons to the 20 amino acids, such that 18 of the 20 amino acids are encoded by two to six different *synonymous* codons. The advent of first-generation DNA sequencing 40 years ago (Sanger et al., 1977; Maxam and Gilbert, 1977) led to the discovery of non-uniform usage of synonymous codons across genes in select bacteria and phages studied at the time. Post et al. provided one of the earliest discoveries of intragenomic synonymous codon usage variation based on their DNA sequencing of the *E.coli* 88-min operon consisting of four ribosomal protein genes (Post et al., 1979). Although their primary goal was to uncover the regulatory elements that control the transcription of this essential ribosomal operon (they later concluded that their sequencing efforts did not answer this study's aim), they discovered that codon usage of those ribosomal protein genes was highly non-random and different from the previously sequenced β-lactamase gene(Post et al., 1979).

Since then, decades of genome sequencing have identified synonymous codon usage variation between and within genomes of most species from all kingdoms of life (Grantham et al., 1980; Novoa et al., 2019). Grantham et al. coined *codon usage bias* to describe the unequal distribution of synonymous codons within and between genomes (Grantham et al., 1980).
Remarkably – and, at first, unexpectedly – despite not altering the protein sequence, numerous biochemical and genomics studies have elucidated how the choice of synonymous codons regulates different stages of gene expression, from the rate of transcription (Zhou et al., 2016), exon splicing (Chamary and Hurst., 2005), mRNA stability (Presnyak et al., 2015; Burow et al., 2018), mRNA secondary structure (Shabalina et al., 2006; Wan et al., 2014), translation elongation rate via altered ribosome processivity (Tuller et al., 2010), protein levels (Carlini and Stephan, 2003; Spencer et al., 2012) and co-translational protein folding (Pechmann and Frydman, 2011; Yu et al., 2015).

Mutation-Selection-Drift Theory of the Evolution of Codon Usage Bias

What explains the ubiquity of synonymous codon usage preferences across different species? The prevailing *mutation-selection-drift* theory explains codon usage bias as an interplay between adaptive (selection for optimizing gene expression) and non-adaptive evolutionary processes (genetic drift and mutation bias) (Grantham et al.,1980; Bulmer, 1991). The *mutation-selection-drift* theory does not mean that the influences of adaptive and non-adaptive processes are not mutually exclusive. Rather the extent of their contribution varies between species. Therefore, one of the key challenges of evolutionary biology is to disentangle the role of various processes in shaping genomic codon usage. In the subsequent sections, I review the literature on evolutionary hypotheses surrounding neutral directional mutation and translation selection on shaping codon usage bias.

1.2 Directional mutation establishes neutral codon usage bias

The mutational bias model predicts that the synonymous codon usage variation is explained by neutral events that result in biased patterns of nucleotide conversion, which in turn gives rise to genome-wide GC-compositional bias (Smith & Eyre-Walker, 2001). As a result, the GC-content at the silent third codon position (GC3) is expected to correlate with the GC-composition of non-coding regions (Grantham et al., 1980). GC-compositional bias is thought to arise by GC-biased gene conversion (gBGC), an event that increases the fixation of G and C alleles during meiotic recombination or biased base incorporation by DNA repair enzymes (Duret & Galtier, 2009; de Boer et al., 2015). One extreme example of GC-compositional bias driving codon usage patterns is observed in the A+T-rich *Mycoplasma capricolum* that has genomic GC-content of 25% and, unsurprisingly, 90% of its codons are A/U-ending (Sharp et al., 1993). However, it may be misleading to assume strong selection (or weak directional mutation) if the base composition of synonymous sites deviates from the genome-wide composition since mutational dynamics vary within the genome due to context-dependent mutation probabilities (Morton, 2003; Zhu et al., 2017; Morton, 2022). For example, in bacteria, mutation biases vary between leading and lagging DNA strands. Thus codon usage was dependent upon the strand of a coding sequence (Romero et al., 2000). Therefore, a correlation between CUB and high expressed genes may reflect transcription-associated mutagenesis rather than selection for gene expression optimization. In mammals, the genome is partitioned into long variegated regions (>300kb) of distinct GC-content (isochores), where codon usage bias of the genes was observed to covary with the GC-content of the host isochore (Mouchiroud et al.,1988; Pouyet et al., 2017).

1.3 Natural Selection Shapes Codon Usage Bias to Optimize Gene Expression

Evidence for selection on codon usage is primarily gleaned from biochemical and bioinformatic analyses in unicellular organisms. Because of the high energetic and fitness cost of protein synthesis, the selection model predicts that the codon composition of highly expressed genes is under stronger selection. It was observed in *E.coli* and *S.cerevisiae* that high expressed and low expressed genes used distinct codon sets (Ikemura, 1982; Sharp, 1987). Ribosomal protein genes, the core component of every cell’s protein synthesis machinery, are often constitutively and highly expressed and exhibit greater codon bias than the rest of the genome (Sharp and Li, 1987; Rochoa, 2003). Thus, codons enriched in highly expressed genes are commonly referred to as “optimal codons.” These observations led Sharp and Li to propose the Codon Adaptation Index (CAI), which measures the similarity of a gene’s codon usage bias to that of a reference set of ribosomal proteins. Why are ribosomal proteins (RPs) used as a reference set for codon optimality? This is based on the observations that: 1) ribosomes are the most abundant molecules in fast-growing cells (Nomura et al., 1984), 2)RPs are more restricted in their codon bias (i.e., low sequence entropy) compared to the rest of the genome (Post et al., 1979), and 3) not only are RPs distinctly codon biased but their coding sequences exhibit a bias for codons matching abundant tRNAs (Ikemura 1983a). These observations collectively suggest that growth is a selection pressure shaping RP codon usage. Thus, by quantifying similarity to ribosomal codon usage, CAI estimates the expressivity of a gene (Sharp and Li, 1987). Since then, CAI has been used extensively in microbial systems to predict mRNA levels (Wu et al., 2005), protein expression levels (Lu et al., 2007), and in synthetic biology for optimization of DNA sequences for heterologous gene expression in microbial (Al-Hawash 2017) and mammalian systems (Inouye et al., 2015).

1.4 Co-evolution between tRNA availability and codon usage to optimize translation efficiency

Gene expression is metabolically expensive as this process consumes ATP and GTP for energy and nucleotides and amino acids as building blocks in transcription and protein synthesis. Additionally, the bioenergetic cost of protein synthesis was estimated to be at least 100 times greater than transcription (Wagner A, 2005; Lynch et al., 2015). Thus, the selection model further posits that tRNA availability is the major selection pressure that shapes the codon bias of highly expressed genes. tRNAs decode the genetic information in mRNA codons to amino acids of protein sequences. Ideally, for each of the 61 sense codons, there should be 61 distinct tRNA anticodon types. However, nearly all sequenced genomes lack the full complement of 61 tRNA types because tRNAs often engage in wobble-decoding at the 3rd codon position. Moreover, the genomic dosage of different tRNAs takes on a dynamic range, from single-copy to even hundreds of identical or near-identical copies within the same genome (Marck and Grosjean, 2002; Goodenbour and Pan, 2006). As such, the prevailing theory is that the imbalance in the supply decoding tRNAs confers distinct elongation rates to codons that underlie the genome-wide distribution of synonymous codons [Ikemura, 1985]. Thus, optimal codons are expected to be cognate to abundant tRNAs, while rare codons should match low abundance tRNAs. Translation efficiency optimizes for two but not mutually exclusive parameters: translation elongation rates (speed) and translation accuracy. Translational efficiency is defined as the rate of protein production from mRNA. Translation accuracy is the rate of amino acid misincorporation (Akashi, 1994). In *E.coli,* selection for translation accuracy was inferred by substituting favored codons with non-favored codons, which resulted in a 10-fold increase in amino acid misincorporation errors (Precup et al.,1989). But most studies on translation efficiency focused on the rate of translation elongation than translation accuracy.

Toshimichi Ikemura first provided experimental evidence for codon-anticodon co-adaptation in shaping gene expression when he quantified relative tRNA levels in *E.coli* using two-dimensional gel electrophoresis (Ikemura 1981). This paradigm-shifting work first elucidated the positive correlation between codon bias in high expressed genes and tRNA concentration, suggesting that codons have distinct translation elongation rates, further buttressing the selectionist model. Even so, translation rates were not directly measured, so it was still unknown if mRNAs were translated unevenly at different sites. However, a year later, Randall and colleagues first elucidated that elongation rates are non-uniform along the mRNA, based on the identification of two distinct nascent premature polypeptides (differed by a 1000 kDa) of maltose-binding membrane protein in *E.coli*. Here they went on to show that the presence of distinct precursor intermediates was not due to aberrant premature termination but was indicative of ribosomal pausing at different sites on the mRNA during elongation [Randall et al., 1982]. Later, Vavrenne and colleagues showed these paused sites correlated with a short stretch of rare codons that matched rare tRNAs [Vavrenne et al., 1984; Vavrenne et al., 1986]. Collectively, these early studies demonstrated that codon-dependent elongation rates are modulated by their tRNA concentration. Since then, multiple studies that span the tree of life have identified competition for the limited supply of tRNAs as an evolutionary determinant of synonymous codon usage variation (Ikemura, 1982; Duret, 2000; Novoa et al., 2012; Wint et al., 2022). To quantify translation selection for tRNA availability, dos Reis et al. formulated the tRNA adaptive index (tAI) based on the tRNA gene copy number and anticodon-codon binding efficiencies of Watson-Crick and wobble interactions (dos Reis et al., 2004). The recurring correlation between transcriptome-wide codon usage and tRNA gene copy number across many species led to the 'tRNA-codon coevolution" hypothesis (Bulmer,1987; Higgs and Ran, 2008).

1.5 Energy Constraints on Gene Expression and Translation Selection

Exponential cellular growth is linearly correlated with global protein synthesis rates (Scott et al., 2010). Indeed, data from growth assays revealed that codon usage correlates better with tRNA gene copy in fast-growing bacteria than in slow-growing species (Rochoa, 2003; Wei et al., 2019). From a resource allocation perspective, it was proposed that translation selection on highly expressed genes evolved as an adaption to optimize global protein synthesis rates by coordinating the efficient use of ribosomes [Andersson and Kurland, 1990] which are known to be the primary limiting factor under high growth conditions (Warner, 2005). Based on this model, codon optimization reduces the dwell time of actively elongating ribosomes, which in turn increases the turnover of free ribosomes to initiate translation on other transcripts. This claim was later validated by genome-engineering experiments in *E.coli*, where Frumkin and colleagues replaced abundant codons in all highly expressed genes with rare synonymous codons. They went on to measure proteome-wide changes using mass spectrometry which revealed a decrease in the abundance of proteins encoded by not only the recoded mRNAs but also the non-recoded mRNAs that use codons that overlap with substituted codons of the re-coded highly expressed genes (Frumkin et al., 2018)

1.6 Evolution of Codon Usage in Animal Genomes

While simpler organisms exhibit clear signatures of adaptive codon usage bias, the influence of translation selection in animal genomes was historically disputatious (Duret, 2002; Galtier et al., 2018). A major reason was the lack of correlation between tRNA gene copy frequency and codon bias in animals (Kanaya et al., 2001). Although genomic tRNA copy number scales with cytosolic tRNA levels in simpler organisms (Dong et al., 1996; Harismendy, 2003), recent developments in chromatin profiling and bulk tRNA sequencing now support tissue-specific and cell-type patterns of tRNA expression sequencing (Ditmar 2009; Gingold et al., 2014; Gogakos et al., 2017). Recent studies have uncovered tissue-specific patterns of codon usage in *Drosophila* (Allen et al., 2022; Payne and Ponce, 2018) and humans (Kames et al., 2020). Because this represents a fairly recent shift in expectations, and the model organism of focus in my dissertation is *Drosophila melanogaster*, I will summarize key studies that supported and contradicted translation selection in drosophilids and mammals.

Deep population genomics suggest that selection on codon usage bias was underestimated in D. melanogaster and Human Genome

The advent of DNA microarrays in the 1990s led to an increase in the comparative studies of gene expression differences between species. Kanaya and colleagues investigated how species-specific patterns of codon usage correlated with tRNA gene copy number in the model animals *Caenorhabditis elegans* (nematode worm), *Xenopus laevis (*frog), and *Drosophila melanogaster* (fruit fly) in addition to *Homo sapiens (Kanaya et al., 2001)*. Although they found that codon bias modestly correlated with levels of expression (based on expressed sequence tags) in both *C. elegans* and *D. melanogaster*, optimal codons and tRNA gene content were only positively correlated in *C. elegans* (also in Duret, 2000) but not in *D.melanogaster*, opposing the selectionist model of codon optimality. Similar analyses in the higher vertebrates*, H.sapiens,* and *X.laevis*, found no correlation between gene expression and tRNA gene copy number (based on comparing ribosomal genes with other genes) and concluded that translation efficiency could not explain codon usage bias in these animals (Kanaya et al., 2001). Later population genetics studies in *Drosophila* also supported weak or near absent selection on synonymous sites and, by extension, codon usage bias (Zeng and Charlesworth, 2009). However, Machado et al. reasoned that these previous population studies may have failed to detect strong selection in *D.melanogaster* because strong purifying selection at synonymous sites would produce low-frequency allele variants. So, by utilizing deep genomic population sequencing of two *D.melanogaster* populations, thus affording greater statistical power, they were able to identify signals ranging from weak to strong selection on codon usage bias (Machado et al., 2020).

In mammals, notably humans, GC-compositional bias was proposed to be the major determinant of codon usage bias as it was observed that the GC-content at the wobble codon position (GC3%) correlated with the isochoric GC-content (Mouchiroud et al., 1988; Clay and Bernardi, 2011). However, ground-breaking work by Gingold et al. provided experimental evidence of translation selection in humans by directly quantifying the tRNA expression using microarrays across hundreds of proliferative and differentiated human cell lines. Here, they uncovered distinct proliferation-specific and differentiation-specific patterns of codon-tRNA co-adaptation (Gingold et al., 2014). Later, Gingold's findings were later challenged by Pouyet et al., who analyzed the proliferation and differentiation gene sets used in Gingold et al. and found that the GC3% of those genes strongly correlated with the GC-content of their isochore and their meiotic expression level, features that are suggestive of neutral mutation bias. They then went on to rule out the influence of translation selection (and all other selection pressures) on human genes based on the expectation that amino acids decoded by a single tRNA type (mono-isoacceptor amino acids), those tRNAs should be uniform across cell types. They reasoned that adaptation to tRNA abundance could not explain the observed variation of synonymous codons that encode these mono-isoacceptor amino acids in the cell-type specific genes (Pouyet et al., 2017). However, this conclusion is rather short-sighted because directional mutation and selection for matching codon demand with tRNA supply are not mutually exclusive. Nor did they consider that tRNA expression may be under selection to match the tissue-specific codon demand, and despite the lack of *in vivo* tRNA expression at the time, the authors argue against tRNA isoacceptor pools changing to match the cell-type specific codon demand. Finally a recent deep population genomics analysis of over 60,000 human genomes (similar to the approach used by Machado et al., 2020) opposes the lack of selection on human codon usage (Dhindsa et al., 2020). This study identified selection against variants that reduced codon optimality, particularly in DNA-damage repair genes, and synonymous mutations that reduce codon optimality occur at lower allele frequency than neutral variants. Additionally, they uncovered that selection on codon optimality was stronger in dosage-sensitive genes that are implicated in Mendelian diseases (Dhindsa et al., 2020).

1.7 Summary of Evolution of Codon Usage Bias

Since the discovery of codon usage bias over 40 years ago, decades of research reveal that, in addition to neutral mutational pressures, natural selection has evolved genome-wide codon usage patterns and reflects optimization of protein synthesis under energy constraints. Signatures of codon optimality are prevalent in microbial systems but historically dubious in more complex multicellular organisms. Because of this longstanding belief, the role of codon optimality as a potential regulator of animal gene expression was largely overlooked until the last seven years [Gingold et al., 2014]. However, the increased availability of tissue/cell type-specific genomic assays and deep population genomics sequencing now lend evidence in favor of stronger selection on animal codon usage than previously believed. In the next section, I will review biochemical studies that have advanced our view of how adaptive codon usage, i.e., codon optimality, influences animal gene expression and physiology.

2. Codon optimality as a genetic determinant of animal development

Here, I review how genome-wide studies demonstrate the importance of post-transcriptional control, mRNA decay, and translation, in animal development and the contribution of codon optimality to these processes, with a focus on D. melanogaster (fruit fly) and mammals (humans and mice).

How multicellular organisms develop from a single-celled fertilized egg is the *raison d'etre* of developmental biology. Multicellularity generates two fundamental and distinct cellular states: proliferative and differentiated. Healthy tissue development requires a balance between cell proliferation and differentiation. Central to the process of cell lineage specification is the precise and dynamic remodeling of gene expression that is required to generate the distinct proteomic changes that, in turn, shape cellular identity and function. A hallmark of cell proliferation is elevated global protein synthesis as a form of growth adaptation compared to more emphasis on local translation in differentiated cells which are no longer actively dividing.

Protein steady-state levels reflect a balance between mRNA synthesis and mRNA degradation. Regulation of gene expression is controlled at multiple levels. Processes that regulate mRNA transcription, such as transcription factors, chromatin dynamics, and RNA polymerase activity, represent the better-studied half of this equation. However, discordant changes between mRNA levels and protein abundance have been demonstrated across multiple systems from yeast (Liu et al., 2016) to mammals, including humans (Vogel et al., 2010; Swindell et al., 2015; Wang et al., 2019). For example, Schwanhausser measured global steady-state levels of mRNA and protein using RNA sequencing and mass spectrometry of mouse fibroblasts and found that mRNA levels explain only 40% of protein levels, leading the authors to conclude that mRNA decay is as important as mRNA synthesis in controlling cellular protein concentrations. On the other hand, coordinated changes in genome-wide decay rates of functionally related mRNAs have been observed across mammalian (Raghavan et al., 2002) and insect development (Thomsen et al., 2010). For example, the induction of quiescence in human fibroblasts resulted in a global destabilization of mRNAs involved in ribosome biogenesis (Johnson et al., 2017).

2.1 RNA decay is a major contributor to post-transcriptional control of animal development

RNA degradation plays a crucial role in modulating mRNA steady-state levels and coordinating quality control by removing defective transcripts. Genome-wide measurements of mRNA decay rate are typically obtained by serially quantifying the mRNA abundance over time using one of two strategies: 1) pharmacological inhibition of transcription using Actinomycin D or α-Amanitin (Friedel and Dolken, 2009; Lugowski et al., 2018) or 2) pulse-chase methods that incorporate nucleoside analogs, (e.g., 4-thiouracil, 5-ethynyl uridine, 5-bromo-uridine) to metabolically label nascent RNA transcripts (Tani et al., 2012; Burow et al., 2015). The maternal-to-zygotic transition (MZT), a conserved event in early animal embryogenesis, represents a well-studied developmental event wherein differential mRNA degradation dramatically alters the transcriptome via the synchronized decay of maternally derived transcripts (Vastenhouw et al., 2019) and the activation of the zygotic genome (Sha et al., 2020).

Most mRNAs in eukaryotes are degraded by the conserved exonuclease-mediated 5’-3’ decay pathway (Mugridge et al., 2018). This multi-step decay process begins with the shortening of the 3’poly-A tail (de-adenylation) by the CCR4-NOT complex, followed by the removal of the 5’ 7-methylguanosine cap by the decapping enzyme complex (Mugridge et al.,2016) which then exposes the 5’ phosphate group that acts as the signal for the binding of the conserved exoribonuclease, *Xrn1. Xrn1* processively degrades the mRNA in the 5’-3’ direction. Despite sharing a general mechanism for decay, eukaryotic mRNA half-lives vary widely between transcripts, as short as a few minutes to hours, and may extend to even more than a day (Tani et al., 2012; Burow et al., 2018). mRNA stability was estimated to control up to 10% of all human genes (Bolognani and Perrone-Bizzozero, 2008). In addition to differential decay rates, mRNA stability correlates with functional category. For example, mRNAs with housekeeping functions tend to have long half-lives (>4 hours), whereas mRNAs with regulatory and cell-type specific functions exhibit shorter half-lives (<4 hours) (Tani et al., 2012; Schwanhäusser et al., 2013). Collectively, these genome-wide studies support mRNA degradation as an essential mechanism for controlling the persistence of genetic information, necessitating a clearer understanding of how mRNA stability is developmentally regulated.

2.2 Codon Optimality and post-transcriptional regulation

Codon optimality broadly refers to how codon identity within the mRNA coding sequence (CDS) regulates the stability and elongation rates of mRNAs. Ground-breaking work in *S.cerevisiae* revealed that codon optimality is a major genetic determinant of mRNA half-lives, wherein certain codons either stabilized ‘optimal codons’ or destabilized ‘non-optimal’ their mRNAs. Furthermore, causality between codon identity and mRNA half-lives was established by monitoring changes in decay rates after re-coding with synonymous codons. To quantify this association, the same study proposed the Codon Stabilization Coefficient (CSCs), which is the Pearson's R correlation coefficient between codon frequency and mRNA half-lives. They also reported that CSCs correlated with genomic tRNA abundance in the yeast, suggesting crosstalk between mRNA translation efficiency and degradation (Presynyak et al., 2015). Since then, the influence of codon optimality on mRNA degradation rates has been characterized in several animal species under various developmental states.

Maternal-to-zygotic transition represents a major developmental phase in early embryogenesis that is defined by the elimination of maternally deposited mRNAs and proteins and the *de novo* expression of the zygotic genome. Therefore, the genetic mechanisms controlling maternal RNA degradation remain an active area in developmental and reproductive biology. Two groups independently elucidated a causal link between non-optimal codon enrichment and destabilization of a subset of maternal mRNAs after zygote genome activation in *Xenopus laevis* (frog) and Danio *rerio* (zebrafish) (Bazzini et al., 2016; Mishima and Tomari, 2016). Bazzini et al. also uncovered a similar pattern in mouse and *D.melanogaster* embryos. CSCs also explained mRNA stability in human HeLa and Chinese hamster ovary cell lines (Forrest et al., 2020). Our lab elucidated that codon-mediated mRNA decay is a major determinant of differential mRNA decay in the *Drosophila* embryos but – for reasons unknown - not in the embryonic nervous tissues, suggestive that codon optimality is context-dependent (Burow et al., 2018). Fragile-X syndrome (FXS) is the most common hereditary neurodevelopmental disorder and is caused by defects in the fragile X mental retardation protein, *FMRP*. *FRMP* is a neuronally enriched RNA-binding protein, with *dFMR1* homolog in *Drosophila*, that regulates different aspects of mRNA metabolism (reviewed in Richter et al., 2021). Knockdown of *FRMP* in the mouse brain led to destabilization and reduced levels of *FRMP*- mRNA targets, and it was found that the target mRNAs fastest degradation rates were enriched with optimal codons. This led to authors to conclude that *FMRP* couples mRNA stability to codon optimality in neurons, possibly through direct interactions with ribosomal machinery (Shu et al., 2020). An emergent theme from these animal studies is that mRNAs encoding functionally related proteins that have similar mRNA half-lives also tend to share similar codon usage profiles. Collectively these findings position codon optimality as an important regulatory code governing the stability of functionally related genes.

Codon optimality and differential mRNA translation

Translation efficiency determines the rate and quality of protein synthesis. Translation efficiency itself is parametrized by initiation and elongation rates. Compared to translation initiation - the rate-limiting step of protein synthesis - control of elongation dynamics is less studied. Current models suggest that synonymous codons have distinct elongation rates, in part due to the differences in the supply of cognate tRNAs. Microarray profiling in hundreds of human cell lines revealed that tissue-specific patterns of tRNA expression matched the codon demand of fate-determining mRNAs resulting in proliferation-specific and differentiation-specific codon usage signatures (Gingold et al., 2014). Gingold et al. speculated that this tissue-specific adaptation between the mRNA and tRNA pool served to modulate context-specific protein levels. However, direct evidence for codon-dependent changes in translation efficiency leading to changes in protein abundance during development is based primarily on cancer studies (Goodarzi et al., 2016; Benisty, 2020; Passarelli et al., 2022). A striking example is the oncogene *KRAS*, which shares 85% protein sequence similarity with the non-oncogenic RAS proteins, but only 15% codon similarity because its codon usage matches that of the proliferation-related genes. Consequently, the synonymous recoding of the *KRAS* reduced its protein abundance when expressed in quiescent fibroblasts compared to proliferative fibroblasts. This study raises the possibility that the codon usage adaptation of some oncogenes allows cancers to hijack the translation program of proliferative cells (Benisty et al., 2020)

Optimal codons are expected to be rapidly decoded by the ribosomes compared to non-optimal codons. Toward this end, ribosomal profiling experiments have yielded valuable insights into how the interplay between codon identity and ribosome kinetics impacts protein expression. Ribosome profiling involves the deep-sequencing of ribosome-protected mRNA fragments and enables the precise monitoring of global translation *in vivo* at the nucleotide resolution (Ignolia et al., 2009). In breast cancer cell lines, the overexpression of tRNA-Arg-CCG and
tRNA-Glu-UCC led to an increase in the ribosome occupancy on mRNAs enriched with cognate GAA and GAG codons, indicative of elevated translation efficiency. This was further validated by direct proteomic quantification using mass spectrometry that also reported elevated levels of proteins enriched with GAA and GAG codons (Goodarzi et al., 2016). This study demonstrates that tRNA measurements can be informative about the tissue-specific proteome. However, a caveat with these studies is that gene expression programs in cancer/disease states do not always mirror normal development.

Codon optimality couples mRNA decay and translation efficiency

In eukaryotic cells, the efficiency with which an mRNA is translated and the rate of its decay are intimately coupled (Wu et al., 2009;) in a complex and context-dependent fashion (Roy and Jacobson, 2013). Translation-coupled decay is utilized for both mRNA quality via ribosome surveillance pathways such as No-Go decay (NGD) and nonsense-mediated decay (NMD) that target transcripts with stalled ribosomes (Veltri et al., 2020) and regulation of mRNA abundance (Bazzini et al., 2016), with the former role being better characterized than the latter. Whilst I do not directly investigate translation-coupled mRNA decay, this

topic is relevant for explaining our results in Chapter 2 of this manuscript, so I will discuss the molecular insights we have gleaned so far regarding how codon optimality – the differential elongation rates of synonymous codons - is transduced into a signal for mRNA degradation.
Immunoprecipitation studies in yeast revealed that the eukaryotic conserved ribosome recycling protein, *Dhh1p* (*DDX6* in humans) preferentially associates with ribosomes on mRNAs with high non-optimal codon content, which suggest that *Dhh1p* senses slow-moving ribosomes and somehow directed those transcripts for degradation (Radhakrishnan et al., 2016). Another major mechanistic insight came from Buschauer and colleagues, who combined cryo-electron microscopy and mRNA sequencing in S.cerevisiae to show that the Not5 subunit of the *CCR4-NOT* complex directly binds the empty E-site of ribosomes that have a vacant A-site, a conformation state adopted by a ribosome that suggests pausing such as waiting on a low abundant charged tRNA to bind the A-site codon (Buschauer et al., 2020). Moreover, *DDX6* is known to interact with the *CCR4-NOT* complex, so possibly both proteins act in concert at linking codon optimality to mRNA stability. Collectively, these two key studies establish a mechanistic link between COMD and the canonical 5'-3' co-translational decay pathway.

2.3 Untranslated regions and post-transcriptional control in animal development

Much of the current knowledge on post-transcription control is based on UTR-mediated *cis-trans* interactions and local secondary structure effects, so I will briefly highlight their roles. Sequence elements within the untranslated regions (UTRs) and coding regions define the post-transcriptional regulatory grammar that modulates mRNA stability and translation efficiency. Alternative polyadenylation and 3'UTR extension have been observed in both *Drosophila* and mammalian neural differentiation (Hilgers et al., 2012; Blair et al., 2016). Many studies show that 3'UTRs contain *cis*-elements that are targets for RNA binding proteins (RBP) and microRNAs (miRNAs) that dynamically regulate mRNA degradation (Zaid 1994; Dini Modigliani et al., 2014; Pereira et al., 2017). The 5'UTR serves as the entry point for ribosomes during translation initiation (Hinnebusch et al., 2016); thus, 5'UTR isoforms may confer differential translation efficiency. Additionally, the 5'UTRs of genes encoding proteins associated with the translation machinery and ribosome biogenesis contain regulatory motifs consisting of a short tract of 4-15 pyrimidine bases known as 5' Terminal OligoPyrimidine (5' TOP). 5'TOP mRNAs are targets for the nutrient-sensing mammalian target of rapamycin complex 1 (*mTORC1*) signaling pathway to selectively modulate their translation rate under stress and diseased conditions such as cancers (Abraham 1998; Holland 2008). Like the 3'UTR, the 5'UTRs also modulate mRNA trafficking in the nervous system (Merianda et al., 2013). For example, alternative 5'UTR splicing modulates the axonal and dendritic trafficking of the conserved brain-derived neurotrophic factor (*BDNF*)

(Collivia et al., 2021) and neuronal nitric-oxide synthase (*nNOS*) mRNAs (Newton et al., 2003).

2.4 Summary of Codon Optimality mediated mRNA regulation in metazoans

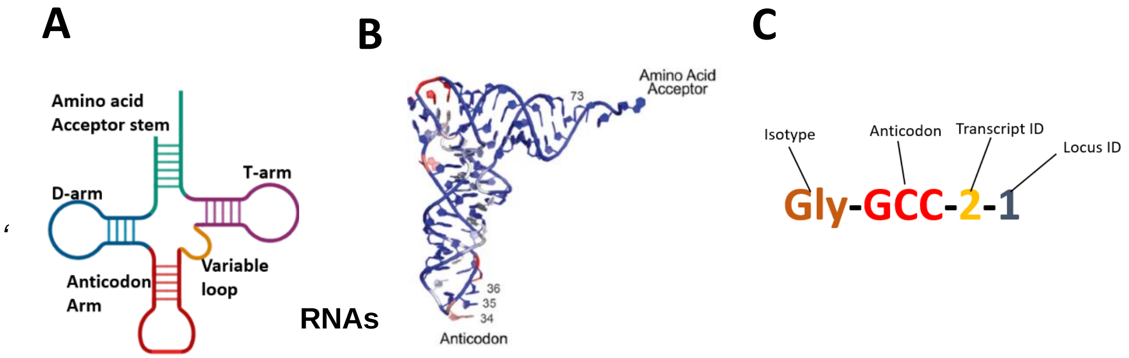
Dynamic mRNA decay plays an essential role in animal development. Biochemical studies agree that mRNA decay is largely regulated by sequence-encoded features, such as the 5'UTR, 3'UTR, and, most recently, the codon identity in the coding regions. Since the discovery of codon optimality-mediated mRNA (COMD) in eukaryotes less than a decade ago, subsequent works have established a mechanistic link between codon optimality and the canonical 5'-3' co-translational mRNA decay pathway. Traditionally, translation-coupled decay pathways were studied in the context of quality control of aberrant mRNAs. So, it remains incompletely understood the extent to which decay factors that sense slow decoding on non-optimal codons overlap with the canonical quality control pathways (such as no-go decay) for clearing poorly made mRNAs or if there exists an alternative mechanism for sensing the reduced elongation rates on normal mRNAs.

More relevant to my work is the outstanding question regarding the mechanism responsible for establishing the distinct and context-dependent effects of codons on mRNA stability in animals. The prevailing model for the origin of COMD is that the distinct stabilities of codons are correlated with their elongation rates which in turn are modulated by the abundance of cognate tRNAs. However, the cell-type matched and *in vivo* tRNA measurements are limited for animals. Although this model offers a temptingly straightforward explanation, still, it extrapolates from the tRNA-mediated codon optimality paradigm informed by studies in unicellular organisms [See section 1 of this chapter]. However, the evolution of multicellularity led to considerable divergence in the regulation of gene expression between unicellular organisms and metazoans [Britten and Davidson, 1969; Carroll, 2003]. So perhaps, the alternative explanation for codon-dependent elongation rates may be that ribosomes differentially decode codons by a mechanism that is independent of their tRNA concentrations. Hence, the proposed model of tRNA-modulated COMD, especially *in vivo*, necessitates further investigation

3. Primer on tRNA Biology:

3.1 General biophysical properties of tRNAs

tRNAs are one of the most abundant RNA molecules in the cell, accounting for about 8-10% of total cytosolic RNA. tRNAs are central to gene expression as they are the suppliers of amino acids in the protein synthesis process that decodes the genetic information in mRNA to proteins. tRNAs are short non-coding RNAs that have an average length of 75bp after maturation. tRNA genes are transcribed by RNA polymerase III (*POL3*), a highly evolutionarily conserved complex. All tRNA molecules fold into a conserved secondary structure that takes on a conserved cloverleaf conformation consisting of 3 stem-loops: D-loop, anticodon loop, and T-loop **(Figure 1A)**. These secondary structures further fold and stack onto themselves, mediated by van der Waal's forces, into the canonical L-shaped tertiary structure **(Figure 1B)**. This L-shaped tertiary structure is the substrate for aminoacyl synthetase (AARS). The short arm of the L-shape tertiary structure consists of the amino acid acceptor stem and the T-loop, and the long arm consists of the D-loop and anticodon stem-loop. The compact L-shape is also stabilized by various chemical moieties that are added during post-transcriptional tRNA maturation.



A: Secondary tRNA Structure **B)** tertiary tRNA structure **C)** Nomenclature of tRNAs reflects the fact that they are often found as multi-copy loci in the genome.

3.2 Canonical Regulation of tRNA biogenesis

In eukaryotes, *POL3* recruitment to tRNA genes is mediated by two general transcription factors, *TFIIIB* and *TFIIIC*. All tRNA genes contain two internal promoters, A- and B- boxes, which are specifically recognized and bound by a large multi-subunit protein complex, *TFIIIC*. *TFIIIC* recruits *TFIIIB,* which then recruits and physically binds *POL3*. Both in yeast and mammalian cells, *TFIIIB* consists of three subunits: *POL3*-specific subunit *BDP1*, *TFIIB*-related *Brf1*, and TATA-box binding protein (TBP), present also for the other two RNA polymerases. RNA *POL3* transcriptional activity, and thus, RNA biogenesis is positively regulated by pro-growth and proliferative stimuli (**Figure 1D**) TBP and Brf1 are post-translationally regulated by the three conserved mitogen-activated kinases of the *MAPK/ERK* pathway: *c-JNK, ERK, p38* kinase. Negative regulation of *POL3* by *MAF1* is the best characterized. mTORC1 kinase, the pioneering growth factor in nutrient sensing pathways, phosphorylates *MAF1* to its inactive form.; However, in nutrient deprivation/ stress response, *mTORC1* is inhibited, leading to dephosphorylation of MAF1. This dephosphorylated state of *MAF1* competitively binds to the Bfr1 subunit on TFIIIB, blocking RNA *POL3* recruitment and assembly (Wei et al., 2009; Vannini et al., 2010)

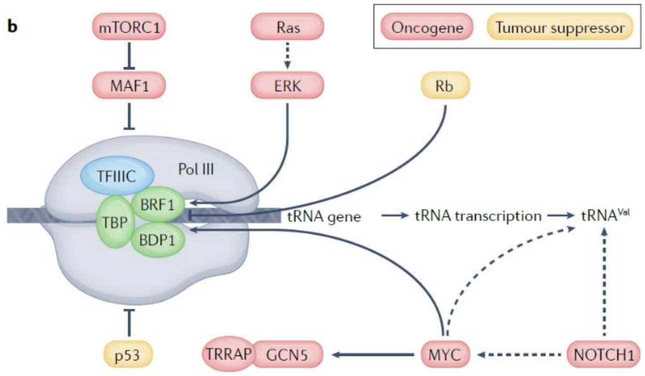


Figure 1D *POL3 Transcription of tRNA genes source: 'tRNA dysregulation and diseases,' Orellana et al., 2021*

Genome-wide chromatin profiling in animal cells agrees that tRNA transcription is developmentally regulated (Barski et al., 2010). For example, genome-wide RNA *POL3* occupancy is dynamically altered between the quiescent/inactive and proliferative states during human macrophage activation (Van Bortle et al., 2017). Thus, an outstanding question in tRNA biology is the mechanism of gene-specific regulation since all tRNA loci share the same internal promoters and general transcription factors. It is believed that tRNA genes are physically regulated by the chromatin context (Arimbasseri et al., 2016).

3.3 tRNA processing and maturation

After transcription, the pre-tRNA primary transcript undergoes further processing in the nucleus by enzymatic cleavage at the 5' and 3' termini and, where applicable, intron-removal in the nucleus. The 5' leader sequence (~10 bases) is cleaved by the ribozyme RNAase P (Carrara et al., 1989), while the 3' trailer sequence (a poly-U tract) is cleaved by the endonuclease RNase Z which exposes the crucial 3' AARS discriminator bases (Maraia et al., 2011). After 3' cleavage, the CCA-adding enzyme (ATP (CTP): tRNA nucleotidyltransferase) extends the 3' terminus with CCA, an essential recognition element for the AARS. The final major step of tRNA maturation involves various post-transcriptional editing, which occurs in both the nucleus and cytoplasm. Approximately 10-15% of tRNA nucleotides are modified, and post-transcriptional modifications regulate diverse aspects of tRNA metabolism, such as structural stability, decoding fidelity in translation, and recognition sites for enzymes (reviewed in Berg and Brandl, 2021).

3.4 Technical Innovations in cellular tRNA quantification

High-throughput sequencing of tRNAs remains an active area of innovation and has been pivotal in yielding fresh insights about tRNA regulation in multicellular development. Here, I will briefly summarize the history of tRNA quantification.

The traditional approach: Inferring tRNA levels variation.

The assumption that tRNA gene copy variation reasonably explains cytosolic tRNA levels seems to be valid for simple organisms (Ikemura 1981a; Dong 1996; Kanaya et al., 2001). For example, in *S.cerevisae,* genome-wide chromatin profiling of RNA *POL3* binding suggests all tRNA loci are actively transcribed (Harismendy et al., 2003), suggesting that tRNA gene copy alone may reasonably approximate cytosolic tRNA levels in these simple organisms. However, such linear relationships may not hold true in multicellular organisms characterized by dynamic chromatin states. Crosslinking and immunoprecipitation (CLIP) sequencing of *SSB*, a pre-tRNA binding protein, in HEK293 cells suggests that 40% of tRNA loci were transcriptionally silent (Gogakos et al., 2017).

Microarrays enabled genome-wide profiling of cytosolic tRNAs

In 1994, Affymetrix released the first commercial DNA microarray chip that, for the first time, enabled the profiling of thousands of mRNAs in one experiment (genome-wide) by using DNA probes that hybridize each cDNA in the library (Heather and Chain, 2016). A decade later, Tao Pan's lab developed the first microarray platform for genome-wide tRNA profiling that could distinguish between isoacceptors (Dittmar et al., 2005). Later on, the Pan lab uncovered differential tRNA expression between normal and breast cancer cell lines (Pavon-Eternod et al., 2009). This study propelled tRNAs into the spotlight as potential regulators of animal development (Goodarzi et al., 2016; Gingold et al., 2014). However, microarray-based methods impose certain limitations: 1) they require knowledge of a species' tRNA gene set beforehand due to the need for custom-made hybridization probes 2) cross-hybridization occludes the expression of tRNA genes that differ by less than 5bp to 8bp (Zaborske et al., 2009) and 2) they have a limited dynamic range of quantification and are less sensitive than RNAseq (Pang et al., 2014)

Next-Generation High-throughput tRNA sequencing

The current state-of-the-art for global tRNA profiling involves the cDNA library preparation designed for Illumina's second-generation next-generation sequencing (NGS) platform. In general, NGS is more sensitive, less labor-intensive (no need to design sequence-specific probes), and has higher discovery potential (no prior genomic data is needed) than microarrays. Nearly all of these tRNAseq methods share similar steps that were adapted from the protocol for barcoded small RNA cDNA library preparation ('small RNAseq') (Hafner et al., 2012). However, these second-generation tRNAseq protocols employ different biochemical approaches to address the biophysical properties of tRNAs that occlude adapter ligation and read-through by the reverse transcriptase, two essential steps for cDNA library preparation. For example, Hydro-tRNAseq applies a partial alkaline hydrolysis treatment to shear the tRNA fragments in order to overcome the stable secondary and tertiary structures, as well as to break the 3' aminoacyl-tRNA bond to enable ligation of the 3' adapter (Gokagos et al., 2017). To improve readthrough by the reverse transcriptase (RT), some methods apply a demethylase treatment using genetically modified *Alkb* demethylase (*E.coli)* to remove specific methylation modifications such as 1-methyladenosine, 6-methyladenosine, and 1-methylcytosine (Zheng et al., 2015; Pinkard et al., 2020). However, some modifications can be read through and result in non-random base misincorporations that are higher than expected by technical errors. Because post-transcriptional modifications are dynamic and essential to tRNA metabolism, these RT-misincorporation signatures can be identified at the post-read alignment step to make inferences about the tRNA epitranscriptome (Schwartz et al., 2018; Pinkard et al., 2020; Behrens et al., 2021).

Limitations of NGS tRNAseq

Studies that compare tRNA sequencing protocols on specific samples found a modest correlation between their measurements, likely due to protocol-specific biases (Pinkard, 2020; Bherens et al., 2020). A common technical bias among tRNAseq methods is the uneven read coverage wherein there is a higher read-depth at the 3' end than the 5' end, likely because the 5'end is generally more modified and/or structured. Another area for growth is the development of a standard 'off-the-shelf' bioinformatics workflow for analyzing tRNAseq libraries, which would improve reproducibility. Although the general steps for tRNAseq analysis parallel standard RNAseq - i.e., read trimming, read alignment, and post-alignment analysis - the biology of tRNA requires additional steps, but bioinformatic approaches vary. For example, one common strategy is to collapse all the identical tRNA genes in the reference set to keep only unique tRNAs that are no more than 95% to 100% similar (Hoffman et al., 2018) while other groups prepare a consensus reference set and performing SNP-aware alignment (Behrens et al., 2021).

4.1 Chapter Summary and Overview

In this review chapter, I emphasized the role of natural selection in shaping genomic and gene-specific codon usage patterns in service of optimizing gene expression, specifically post-transcriptional mRNA dynamics. As a result, the usage of a subset of codons improves mRNA translation and stability, and these codons are referred to as optimal codons. Presently, much of what we understand about the influence of codon optimality on gene expression regulation is largely informed by simpler organisms. Specific to eukaryotes, the interplay between codon optimality and gene expression is primarily derived from studies in the model fungus *S.cerevisiae*. In *S.cerevisiae*, codon bias in highly expressed correlates with tRNA abundance, which is indicative of selection for optimizing mRNA translation during protein synthesis. As an application, this genomic property positioned *S.cerevisiae* as an attractive 'bio-factory' for the heterologous production of proteins in the biotechnology industry [Kulagina et al., 2021]. *S.cerevisiae* is one of an estimated 1.5 million fungal species [Berbee and Taylor, 2017]. One of the gaps that my dissertation work addresses are the prevalence of natural selection on codon usage patterns in other fungal clades.

Until recently, it was assumed that the codon usage bias of metazoans is largely neutral, which led to the neglect of studying codon optimality as a potential regulatory grammar for mRNA dynamics in animal genetics. However, a handful of key studies that uncovered proliferation-specific and differentiation-specific codon usage in diverse human cell lines [Gingold et al., 2014] and codon optimality mediated mRNA decay (COMD), first in *S.cerevisiae* (Presynak et al., 2015), and later in model vertebrates and invertebrates (Bazzini et al., 2016; Mishma and Tomari, 2016; Burow et al., 2018) re-kindled an interest in codon optimality in animal development. Still, gaps remain. Among the handful of studies that characterized COMD in animal systems, the majority were based *in vitro.* Additionally, current models suggest that the distinct stabilities of codons are modulated by tRNA concentrations; however, genome-wide tRNA quantitation is limited in most species, more so in animals. To date, there are no published tRNA measurements for the model animal *Drosophila melanogaster* (fruitfly). As a neurobiology research group, we are generally interested in the post-transcription control of brain development. Collectively, these factors motivated us to investigate the regulation of tRNAs during neurogenesis in the *Drosophila* central nervous system using next-generation sequencing approaches and how changes in tRNA expression contribute to dynamic codon optimality and gene expression programs that support neural proliferation versus differentiation.

The study of the role of tRNAs and codon optimality in *normal* animal development is still in its early days. Hence, a major contribution of my dissertation work is establishing the *Drosophila* central nervous system as a model for studying differential tRNA regulation on cell fate determination in higher animals.

4.2 Dissertation research aims

My dissertation research studies the contribution of tRNAs and codon usage in shaping gene expression programs in animal developmental *Drosophila* neurogenesis) and in fungal evolution, thus covering the representatives from two of the four major branches of Eukarya.

In Chapter 2, I use experimental and bioinformatic approaches for investigating tRNA regulation in *Drosophila* neurogenesis. The objectives of this chapter were to:

> **Aim 1:** Measure the genome-wide tRNA repertoire in neuroblasts and post-mitotic neurons (addressed in Chapter 2)
>
> **Aim 2:** Determine if tRNA abundance contributes to tissue-specific codon optimality in neuroblasts and neurons (addressed in Chapter 2)

Chapter 3 is a methods-centric extension of Chapter 2 because tRNA sequencing is relatively recent, so there is no standardized workflow for analyzing tRNAseq data. So in Chapter 3, I compared tools for differential gene expression analysis for tRNAseq data. The objective of this chapter was to:

> **Aim 1:** Evaluate the performance of parametric and non-parametric methods for identifying differentially expressed tRNAs

Finally, Chapter 4 involves the large-scale analysis of genomic data from over 400 recently sequenced fungal species that are representatives of six out of the eight fungal phyla and eighteen taxonomic classes. The objectives of this chapter were to:

> **Aim 1:** Characterize kingdom-wide patterns of codon usage and genomic tRNA in Fungi
>
> **Aim 2**: Elucidate the evolutionary influences and functional implications of codon usage bias in Kingdom Fungi

References:

1. Akash, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics, 136:927-935

2. Al-Hawash, A. B., Zhang, X. & Ma, F. (2017). Strategies of codon optimization for high-level heterologous protein expression in microbial expression systems. Gene Rep. 9, 46–53

3. Allen SR, Stewart RK, Rogers M, Ruiz IJ, Cohen E, Laederach A, Counter CM, Sawyer JK, Fox DT. (2022). Distinct responses to rare codons in select Drosophila tissues. Elife., 6;11:e76893. doi: 10.7554/eLife.76893.

4. Andersson SG, Kurland CG. (1990). Codon preferences in free-living microorganisms. *Microbiol Rev* 54: 198–210.

5. Arimbasseri, A. G., & Maraia, R. J. (2016). RNA Polymerase III Advances: Structural and tRNA Functional Views. *Trends in biochemical sciences*, *41*(6), 546–559. https://doi.org/10.1016/j.tibs.2016.03.003

6. Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A. B., Birch, J., Cui, K., White, R. J., & Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nature structural & molecular biology*, *17*(5), 629–634. https://doi.org/10.1038/nsmb.1806

7. Bauer, K. E., Segura, I., Gaspar, I., Scheuss, V., Illig, C., Ammer, G., Hutten, S., Basyuk, E., Fernández-Moya, S. M., Ehses, J., Bertrand, E., & Kiebler, M. A. (2019). Live cell imaging reveals 3'-UTR dependent mRNA sorting to synapses. *Nature communications*, *10*(1), 3178. https://doi.org/10.1038/s41467-019-11123-x

8. Bazzini, A. A., Del Viso, F., Moreno-Mateos, M. A., Johnstone, T. G., Vejnar, C. E., Qin, Y., Yao, J., Khokha, M. K., & Giraldez, A. J. (2016). Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *The EMBO journal*, *35*(19), 2087–2103. https://doi.org/10.15252/embj.201694699

9. Behrens, A., Rodschinka, G., & Nedialkova, D. D. (2021). High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Molecular cell*, *81*(8), 1802–1815.e7. https://doi.org/10.1016/j.molcel.2021.01.028

10. Benisty, H., Weber, M., Hernandez-Alias, X., Schaefer, M. H., & Serrano, L. (2020). Mutation bias within oncogene families is related to proliferation-specific codon usage. *Proceedings of the National Academy of Sciences*, *117*(48), 30848-30856.

11. Berg, M. D., & Brandl, C. J. (2021). Transfer RNAs: diversity in form and function. *RNA biology*, *18*(3), 316–339. https://doi.org/10.1080/15476286.2020.1809197

12. Bolognani, F., & Perrone-Bizzozero, N. I. (2008). RNA-protein interactions and control of mRNA stability in neurons. *Journal of neuroscience research*, *86*(3), 481–489. https://doi.org/10.1002/jnr.21473

13. Britten, R. J., & Davidson, E. H. (1969). Gene Regulation for Higher Cells: A Theory: New facts regarding the organization of the genome provide clues to the nature of gene regulation. *Science*, *165*(3891), 349-357.

14. Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. *Nature* 325: 728–730.

15. Bulmer M. (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129: 897–907.

16. Burow, D. A., Martin, S., Quail, J. F., Alhusaini, N., Coller, J., & Cleary, M. D. (2018). Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila. *Cell reports*, *24*(7), 1704–1712. https://doi.org/10.1016/j.celrep.2018.07.039

17. Burow, D. A., Umeh-Garcia, M. C., True, M. B., Bakhaj, C. D., Ardell, D. H., & Cleary, M. D. (2015). Dynamic regulation of mRNA decay during neural development. *Neural development*, *10*, 11. https://doi.org/10.1186/s13064-015-0038-6

18. Carlini, D. B., & Stephan, W. (2003). In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. *Genetics*, *163*(1), 239-243.

19. Carrara, G., Calandra, P., Fruscoloni, P., Doria, M., & Tocchini-Valentini, G. P. (1989). Site selection by Xenopus laevis RNAase P. *Cell*, *58*(1), 37–45. https://doi.org/10.1016/0092-8674(89)90400-5

20. Carroll S. B. (2005). Evolution at two levels: on genes and form. *PLoS biology*, *3*(7), e245. https://doi.org/10.1371/journal.pbio.0030245

21. Chamary, J. V., & Hurst, L. D. (2005). Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else?. *Trends in Genetics*, *21*(5), 256-259.

22. Cheng, J., Maier, K. C., Avsec, Ž., Rus, P., & Gagneur, J. (2017). *Cis*-regulatory elements explain most of the mRNA stability variation across genes in yeast. *RNA (New York, N.Y.)*, *23*(11), 1648–1659. https://doi.org/10.1261/rna.062224.117

23. Collart, M. A., & Weiss, B. (2020). Ribosome pausing, a dangerous necessity for co-translational events. *Nucleic acids research*, *48*(3), 1043-1055.

24. Colliva, A., & Tongiorgi, E. (2021). Distinct role of 5'UTR sequences in dendritic trafficking of BDNF mRNA: additional mechanisms for the BDNF splice variants spatial code. *Molecular brain*, *14*(1), 10. https://doi.org/10.1186/s13041-020-00680-8

25. Crick F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563. https://doi.org/10.1038/227561a0

26. de Boer, E., Jasin, M., & Keeney, S. (2015). Local and sex-specific biases in crossover vs. noncrossover outcomes at meiotic recombination hot spots in mice. *Genes & development*, *29*(16), 1721-1733.

27. Dhindsa, R. S., Copeland, B. R., Mustoe, A. M., & Goldstein, D. B. (2020). Natural Selection Shapes Codon Usage in the Human Genome. *American journal of human genetics*, *107*(1), 83–95. https://doi.org/10.1016/j.ajhg.2020.05.011

28. Dittmar, K. A., Sørensen, M. A., Elf, J., Ehrenberg, M., & Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO reports*, *6*(2), 151–157. https://doi.org/10.1038/sj.embor.7400341

29. Dong, H., Nilsson, L., & Kurland, C. G. (1996). Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *Journal of molecular biology*, *260*(5), 649–663. https://doi.org/10.1006/jmbi.1996.0428

30. Duret L, Galtier N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annual Review of Genomics and Human Genetics 10:285–311. DOI: https://doi.org/10.1146/annurev-genom-082908-150001, PMID: 19630562

31. Duret L. (2002). **Current Opinion in Genetics & Development, 12:640–649**

32. Duret, L. (2000). tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends in Genetics, 16(7), 287-289.

33. Forrest, M. E., Pinkard, O., Martin, S., Sweet, T. J., Hanson, G., & Coller, J. (2020). Codon and amino acid content are associated with mRNA stability in mammalian cells. *PloS one*, *15*(2), e0228730. https://doi.org/10.1371/journal.pone.0228730

34. Friedel, C. C., & Dölken, L. (2009). Metabolic tagging and purification of nascent RNA: implications for transcriptomics. *Molecular bioSystems*, *5*(11), 1271–1278. https://doi.org/10.1039/b911233b

35. Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences*, *115*(21), E4940-E4949.

36. Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. (2018). Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Mol Biol Evol 35(5):1092–1103

37. Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., & Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Molecular biology and evolution*, *35*(5), 1092–1103. https://doi.org/10.1093/molbev/msy015.

38. Gingold, H., Tehler, D., Christoffersen, N. R., Nielsen, M. M., Asmar, F., Kooistra, S. M., ... & Pilpel, Y. (2014). A dual program for translation regulation in cellular proliferation and differentiation. *Cell*, *158*(6), 1281-1292.

39. Gogakos, T., Brown, M., Garzia, A., Meyer, C., Hafner, M., & Tuschl, T. (2017). Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. *Cell reports*, *20*(6), 1463–1475. https://doi.org/10.1016/j.celrep.2017.07.029

40. Goodarzi, H., Nguyen, H., Zhang, S., Dill, B. D., Molina, H., & Tavazoie, S. F. (2016). Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. *Cell*, *165*(6), 1416–1427. https://doi.org/10.1016/j.cell.2016.05.046

41. Goodenbour, J. M., & Pan, T. (2006). Diversity of tRNA genes in eukaryotes. *Nucleic acids research*, *34*(21), 6137-6146.

42. Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic acids research*, *10*(22), 7055–7074. https://doi.org/10.1093/nar/10.22.7055

43. Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic acids research*, *8*(1), r49–r62. https://doi.org/10.1093/nar/8.1.197-c

44. Hafner, M., Renwick, N., Farazi, T. A., Mihailović, A., Pena, J. T., & Tuschl, T. (2012). Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. *Methods (San Diego, Calif.)*, *58*(2), 164–170. https://doi.org/10.1016/j.ymeth.2012.07.030

45. Hanson G, Coller J. (2018). Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol 19(1):20–30

46. Harismendy O, Gendrel CG, Soularue P, Gidrol X, Sentenac A, Werner M, Lefebvre O. (2003). Genome-wide location of yeast RNA polymerase III transcription machinery. *EMBO J*. 22(18):4738-4747.

47. Heather JM, Chain B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1):1-8. doi:10.1016/j.ygeno.2015.11.003

48. Hershberg R, Petrov DA. (2008). Selection on codon bias. *Annu Rev Genet*. 42:287–299.

49. Higgs PG, Ran W. (2008). Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol. 2511:2279–2291

50. Hinnebusch, A. G., Ivanov, I. P., & Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science (New York, N.Y.)*, *352*(6292), 1413–1416. https://doi.org/10.1126/science.aad9868

51. Holland EC. (2004) Regulation of Translation and Cancer, Cell Cycle, 3:4, 450-453, DOI: 10.4161/cc.3.4.796

52. Hu, W., Sweet, T. J., Chamnongpol, S., Baker, K. E., & Coller, J. (2009). Co-translational mRNA decay in Saccharomyces cerevisiae. *Nature*, *461*(7261), 225–229. https://doi.org/10.1038/nature08265

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

54. Ikemura T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol*. 1584:573-597.

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

56. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. (2009). Science, 10;324(5924):218-23. doi: 10.1126/science.1168978.

57. Inouye, S., Sahara-Miura, Y., Sato, J. I. & Suzuki, T. Codon optimization of genes for efficient protein expression in mammalian cells by selection of only preferred human codons. (2015). Protein Expr. Purif. 109, 47–54

58. Johnson, E. L., Robinson, D. G., & Coller, H. A. (2017). Widespread changes in mRNA stability contribute to quiescence-specific gene expression patterns in a fibroblast model of quiescence. *BMC genomics*, *18*(1), 1-9.

59. Kames J, Alexaki A, Holcomb DD, Santana-Quintero LV, Athey JC, Hamasaki-Katagiri N, Katneni U, Golikov A, Ibla JC, Bar H, Kimchi-Sarfaty C .(2020). TissueCoCoPUTs: novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. J Mol Biol 432(11):3369–3378

60. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T: Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation

efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. (2001). J Mol Evol, 53:290-298.

61. Kimura, M. (1977). Nature 267, 275–276.

62. Kulagina, N., Besseau, S., Godon, C., Goldman, G. H., Papon, N., and Courdavault, V. (2021). Yeasts as biopharmaceutical production platforms. *Front. Fungal Biol.* 2, 733492. doi: 10.3389/ffunb.2021.733492

63. Liu Y, Beyer A, Aebersold R. (2016). On the dependency of cellular protein levels on mRNA abundance. Cell 165: 535 – 550

64. Lu P., Vogel C., Wang R., Yao X., MarcotteE.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol. 2007; 25: 117–24.

65. Lugowski, A., Nicholson, B., & Rissland, O. S. (2018). DRUID: a pipeline for transcriptome-wide measurements of mRNA stability. *RNA (New York, N.Y.)*, *24*(5), 623–632. https://doi.org/10.1261/rna.062877.117

66. Lynch, M., & Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences*, *112*(51), 15690-15695.

67. MA Sørensen, CG Kurland, S Pedersen, Codon usage determines translation rate in Escherichia coli. J Mol Biol 207, 365–377 (1989).

68. Machado HE, Lawrie DS, Petrov DA. Pervasive Strong Selection at the Level of Codon Usage Bias in *Drosophila melanogaster*.(2020). Genetics, 214(2):511-528. doi: 10.1534/genetics.119.302542.

69. Maraia, R. J., & Lamichhane, T. N. (2011). 3' processing of eukaryotic precursor tRNAs. *Wiley interdisciplinary reviews. RNA*, *2*(3), 362–375. https://doi.org/10.1002/wrna.64

70. Marck C, Grosjean H. 2002. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*. 810:1189–1232.

71. Mauger, D. M., Cabral, B. J., Presnyak, V., Su, S. V., Reid, D. W., Goodman, B., Link, K., Khatwani, N., Reynders, J., Moore, M. J., & McFadyen, I. J. (2019). mRNA structure regulates protein expression through changes in functional half-life. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(48), 24075–24083. https://doi.org/10.1073/pnas.1908052116

72. Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–564. https://doi.org/10.1073/pnas.74.2.560

73. Merianda T, Gomes C, Yoo S, Vuppalanchi D, Twiss JL. Axonal Localization of Neuritin/CPG15 mRNA in Neuronal Populations through Distinct 5′ and 3′ UTR

Elements. 2013. Journal of Neuroscience , 33 (34) 13735-13742; DOI: 10.1523/JNEUROSCI.0962-13.2013

74. Mishima, Y., & Tomari, Y. (2016). Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish. *Molecular cell*, *61*(6), 874–885. https://doi.org/10.1016/j.molcel.2016.02.027

75. Morton, B. R. (2003). The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. *Journal of molecular evolution*, *56*(5), 616-629.

76. Morton, B. R. (2022). Context-Dependent Mutation Dynamics, Not Selection, Explains the Codon Usage Bias of Most Angiosperm Chloroplast Genes. *Journal of molecular evolution*, *90*(1), 17-29.

77. Mouchiroud D, Gautier C, Bernardi G. 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. Journal of Molecular Evolution 27:311–320. DOI: https://doi.org/10.1007/ BF02101193, PMID: 3146641

78. Mugridge, J. S., Coller, J., & Gross, J. D. (2018). Structural and molecular mechanisms for the control of eukaryotic 5'-3' mRNA decay. *Nature structural & molecular biology*, *25*(12), 1077–1085. https://doi.org/10.1038/s41594-018-0164-z

79. Mugridge, J. S., Ziemniak, M., Jemielity, J., & Gross, J. D. (2016). Structural basis of mRNA-cap recognition by Dcp1-Dcp2. *Nature structural & molecular biology*, *23*(11), 987–994. https://doi.org/10.1038/nsmb.3301

80. Newton, D. C., Bevan, S. C., Choi, S., Robb, G. B., Millar, A., Wang, Y., & Marsden, P. A. (2003). Translational regulation of human neuronal nitric-oxide synthase by an alternatively spliced 5'-untranslated region leader exon. The Journal of biological chemistry, 278(1), 636–644. https://doi.org/10.1074/jbc.M209988200

81. Nomura, M., Gourse, R., & Baughman, G. (1984). Regulation of the synthesis of ribosomes and ribosomal components. *Annual review of biochemistry*, *53*(1), 75-117.

82. Novoa EM, Jungreis I, Jaillon O, Kellis M (2019) Elucidation of codon usage signatures across the domains of life. Mol Biol Evol 36(10):2328–2339

83. Orellana EA, Siegal E, Gregory RI. tRNA dysregulation and disease. Nat Rev Genet. 2022 Jun 9. doi: 10.1038/s41576-022-00501-9. Epub ahead of print. PMID: 35681060.

84. Pagani F, Raponi M, Baralle FE. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci USA 102, 6368–6372

85. Pang, Y. L., Abo, R., Levine, S. S., & Dedon, P. C. (2014). Diverse cell stresses induce unique patterns of tRNA up- and down-regulation: tRNA-seq for quantifying changes in tRNA copy number. *Nucleic acids research*, *42*(22), e170. https://doi.org/10.1093/nar/gku945

86. Passarelli, M. C., Pinzaru, A. M., Asgharian, H., Liberti, M. V., Heissel, S., Molina, H., Goodarzi, H., & Tavazoie, S. F. (2022). Leucyl-tRNA synthetase is a tumour suppressor in breast cancer and regulates codon-dependent translation dynamics. *Nature cell biology*, *24*(3), 307–315. https://doi.org/10.1038/s41556-022-00856-5

87. Pavon-Eternod, M., Gomes, S., Rosner, M. R., & Pan, T. (2013). Overexpression of initiator methionine tRNA leads to global reprogramming of tRNA expression and increased proliferation in human epithelial cells. *Rna*, *19*(4), 461-466.

88. Payne BL, Alvarez-Ponce D. Codon Usage Differences among Genes Expressed in Different Tissues of Drosophila melanogaster. Genome Biol Evol. 2019 Apr 1;11(4):1054-1065. doi: 10.1093/gbe/evz051. PMID: 30859203; PMCID: PMC6456009.

89. Pechmann, S., & Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature structural & molecular biology*, *20*(2), 237–243. https://doi.org/10.1038/nsmb.2466

90. Pereira LA, Munita R, González MP, Andrés ME (2017) Long 3'UTR of Nurr1 mRNAs is targeted by miRNAs in mesencephalic dopamine neurons. PLOS ONE 12(11): e0188177. https://doi.org/10.1371/journal.pone.0188177

91. Pinkard, O., McFarland, S., Sweet, T., & Coller, J. (2020). Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. *Nature communications*, *11*(1), 4104. https://doi.org/10.1038/s41467-020-17879-x

92. Ponger L, Duret L, Mouchiroud D: Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res 2001, 11:1854-1860.

93. Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H., & Dennis, P. P. (1979). Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(4), 1697–1701. https://doi.org/10.1073/pnas.76.4.1697

94. Pouyet, F., Mouchiroud, D., Duret, L., & Sémon, M. (2017). Recombination, meiotic expression and human codon usage. *Elife*, *6*, e27344.

95. Precup, J., Ulrich, A. K., Roopnarine, O., & Parker, J. (1989). Context specific misreading of phenylalanine codons. *Molecular and General Genetics MGG*, *218*(3), 397-401.

96. Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., & Coller, J. (2015). Codon

optimality is a major determinant of mRNA stability. *Cell*, *160*(6), 1111–1124. https://doi.org/10.1016/j.cell.2015.02.029

97. Radhakrishnan, A., & Green, R. (2016). Connections Underlying Translation and mRNA Stability. *Journal of molecular biology*, *428*(18), 3558–3564. https://doi.org/10.1016/j.jmb.2016.05.025

98. Raghavan, A., Ogilvie, R. L., Reilly, C., Abelson, M. L., Raghavan, S., Vasdewani, J., ... & Bohjanen, P. R. (2002). Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic acids research*, *30*(24), 5529-5538.

99. Randall, L. L., Josefsson, L. G., & Hardy, S. J. (1980). Novel intermediates in the synthesis of maltose-binding protein in Escherichia coli. *European journal of biochemistry*, *107*(2), 375–379. https://doi.org/10.1111/j.1432-1033.1980.tb06039.x

100. Richter JD, Zhao X. (2021). The molecular biology of FMRP: new insights into fragile X syndrome. Nat Rev Neurosci (4):209-222. doi: 10.1038/s41583-021-00432-0.

101. Rocha EP. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. (2004). Genome Res. 2004 Nov;14(11):2279-86. doi: 10.1101/gr.2896904

102. Romero, H., Zavala, A., & Musto, H. (2000). Codon usage in Chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic acids research*, *28*(10), 2084-2090.

103. Roy, B., & Jacobson, A. (2013). The intimate relationships of mRNA decay and translation. *Trends in genetics : TIG*, *29*(12), 691–699. https://doi.org/10.1016/j.tig.2013.09.002

104. Ruiz, L.M. , Armengol, G. , Habeych, E. , Orduz, S. A theoretical analysis of codon adaptation index of the Boophilus microplus bm86 gene directed to the optimization of a DNA vaccine. (2006). J Theor Biol, 239: 445–9.

105. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, *265*(5596), 687–695. https://doi.org/10.1038/265687a0

106. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. (2011). Global quantification of mammalian gene expression control. *Nature* 473: 337–342

107. Schwartz, M. H., Wang, H., Pan, J. N., Clark, W. C., Cui, S., Eckwahl, M. J., Pan, D. W., Parisien, M., Owens, S. M., Cheng, B. L., Martinez, K., Xu, J., Chang, E. B., Pan, T., & Eren, A. M. (2018). Microbiome characterization by

high-throughput transfer RNA sequencing and modification analysis. *Nature communications*, *9*(1), 5353. https://doi.org/10.1038/s41467-018-07675-z

108. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z., & Hwa, T. (2010). Interdependence of cell growth and gene expression: origins and consequences. *Science*, *330*(6007), 1099-1102.

109. Sha, Q. Q., Zhu, Y. Z., Li, S., Jiang, Y., Chen, L., Sun, X. H., Shen, L., Ou, X. H., & Fan, H. Y. (2020). Characterization of zygotic genome activation-dependent maternal mRNA clearance in mouse. *Nucleic acids research*, *48*(2), 879–894. https://doi.org/10.1093/nar/gkz1111

110. Shabalina, S. A., Ogurtsov, A. Y., & Spiridonov, N. A. (2006). A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic acids research*, *34*(8), 2428–2437. https://doi.org/10.1093/nar/gkl287

111. Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of molecular evolution*, *24*(1), 28-38.

112. Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, *15*(3), 1281–1295. https://doi.org/10.1093/nar/15.3.1281

113. Sharp, P. M., Stenico, M., Peden, J. F., & Lloyd, A. T. (1993). Codon usage: mutational bias, translational selection, or both?. *Biochemical Society Transactions*, *21*(4), 835-841.

114. Shu, H., Donnard, E., Liu, B., Jung, S., Wang, R., & Richter, J. D. (2020). FMRP links optimal codons to mRNA stability in neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30400–30411. https://doi.org/10.1073/pnas.2009161117

115. Smith NG, Eyre-Walker A. 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol Biol Evol. 18:982-986.

116. Spencer, P. S., Siller, E., Anderson, J. F., & Barral, J. M. (2012). Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *Journal of molecular biology*, *422*(3), 328–335. https://doi.org/10.1016/j.jmb.2012.06.010

117. Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, Wakamatsu A, et al. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res ;22:947–56.

118. Thomsen, S., Anders, S., Janga, S. C., Huber, W., & Alonso, C. R. (2010). Genome-wide analysis of mRNA decay patterns during early *Drosophila* development. *Genome biology*, *11*(9), 1-27.

119. Thoreen CC, Chantranupong L, Keys HR, Wang T, Gray NS, Sabatini DM. (2012). A unifying model for mTORC1-mediated regulation of mRNA translation. Nature 485:109–113 doi:10.1038/nature11083

120. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 1412:344–354

121. Tushev, G., Glock, C., Heumüller, M., Biever, A., Jovanovic, M., & Schuman, E. M. (2018). Alternative 3' UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. Neuron, 98(3), 495–511.e6. https://doi.org/10.1016/j.neuron.2018.03.030

122. Van Bortle, K., Phanstiel, D. H., & Snyder, M. P. (2017). Topological organization and dynamic regulation of human tRNA genes during macrophage differentiation. *Genome biology*, *18*(1), 180. https://doi.org/10.1186/s13059-017-1310-3

123. Vannini, A., Ringel, R., Kusser, A. G., Berninghausen, O., Kassavetis, G. A., & Cramer, P. (2010). Molecular basis of RNA polymerase III transcription repression by Maf1. *Cell*, *143*(1), 59–70. https://doi.org/10.1016/j.cell.2010.09.002

124. Varenne, S., & Lazdunski, C. (1986). Effect of distribution of unfavourable codons on the maximum rate of gene expression by an heterologous organism. *Journal of theoretical biology*, *120*(1), 99–110. https://doi.org/10.1016/s0022-5193(86)80020-0

125. Varenne, S., Buc, J., Lloubes, R., & Lazdunski, C. (1984). Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of molecular biology*, *180*(3), 549–576. https://doi.org/10.1016/0022-2836(84)90027-5

126. Vastenhouw, N. L., Cao, W. X., & Lipshitz, H. D. (2019). The maternal-to-zygotic transition revisited. *Development (Cambridge, England)*, *146*(11), dev161471. https://doi.org/10.1242/dev.161471

127. Veltri, A. J., D'Orazio, K. N., & Green, R. (2020). Make or break: the ribosome as a regulator of mRNA decay. *Cell research*, *30*(3), 195–196. https://doi.org/10.1038/s41422-019-0271-3

128. Vogel C, de Sousa Abreu R, Ko D, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* 6: 400

129. Wagner, A. (2005). Energy constraints on the evolution of gene expression. *Molecular biology and evolution*, *22*(6), 1365-1374.

130. Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., Zecha, J., Asplund, A., Li, L. H., Meng, C., Frejno, M., Schmidt, T., Schnatbaum, K., Wilhelm, M., Ponten, F., Uhlen, M., Gagneur, J., Hahne, H., & Kuster, B. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular systems biology*, *15*(2), e8503. https://doi.org/10.15252/msb.20188503

131. Warner JR. 1999. The economics of ribosome biosynthesis in yeast. Trends in Biochemical Sciences 24:437–440. DOI: https://doi.org/10.1016/S0968-0004(99)01460-7, PMID: 10542411

132. Wei, Y., Silke, J.R. & Xia, X. An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. Sci Rep 9, 3184 (2019). https://doi.org/10.1038/s41598-019-39369-x

133. Wei, Y., Tsang, C. K., & Zheng, X. F. (2009). Mechanisms of regulation of RNA polymerase III-dependent transcription by TORC1. *The EMBO journal*, *28*(15), 2220–2230. https://doi.org/10.1038/emboj.2009.179

134. Wint R, Salamov A, Grigoriev IV. (2022). Kingdom-wide analysis of fungal transcriptomes and tRNAs 1235 reveals conserved patterns of adaptive evolution. Mol Biol. Evol. 1236 https://doi.org/10.1093/molbev/msab372

135. Wu G, Culley DE, Zhang W. 2005. Predicted highly expressed genes in the genomes of Streptomyces coelicolor and Streptomyces avermitilis and the implications for their metabolism. Microbiology 151: 2175–2187. 10.1099/mic.0.27833-0

136. Yang, Z., and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol. Biol. Evol. 25, 568–579.

137. Yang, Z.H. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. Trends Ecol. Evol.15, 496–50

138. Zaborske J.M., Narasimhan J., Jiang L., Wek S.A., Dittmar K.A., Freimoser F., Pan T., Wek R.C.. Genome-wide analysis of tRNA charging and activation of the eIF2 kinase Gcn2p. J. Biol. Chem. 2009; 284:25254–25267

139. Zaidi SH, Denman R, Malter JS. 1994. Multiple proteins interact at a unique cis-element in the 39 UTR of amyloid precursor protein (APP) mRNA. J Biol Chem 269:24000–24007.

140. Zeng K, Charlesworth B. Studying patterns of recent evolution at synonymous sites and intronic sites in Drosophila melanogaster. J Mol Evol. 2010 Jan;70(1):116-28. doi: 10.1007/s00239-009-9314-6. Epub 2009 Dec 30. PMID: 20041239.

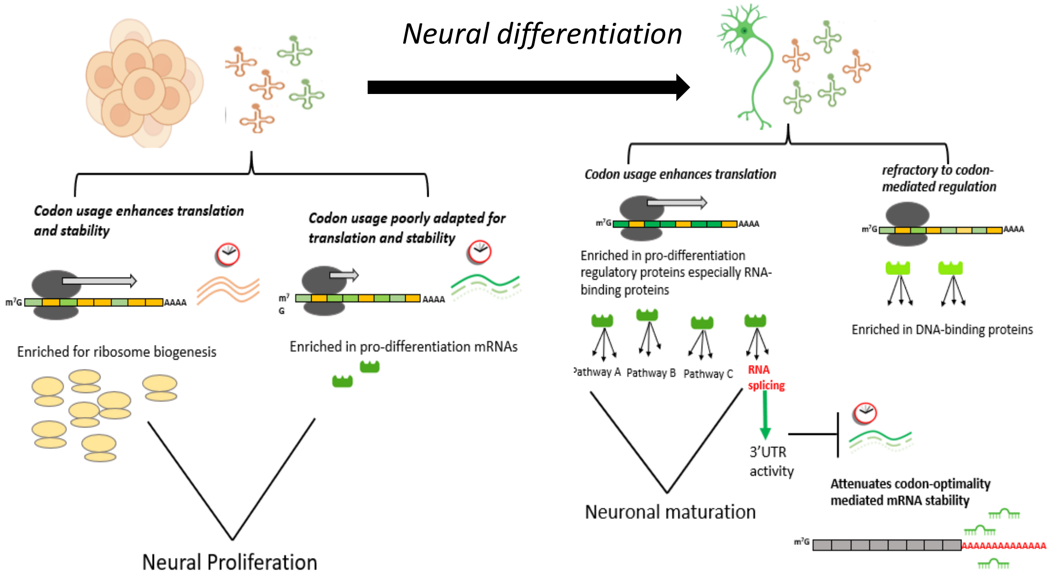
141. Zheng, G., Qin, Y., Clark, W. C., Dai, Q., Yi, C., He, C., Lambowitz, A. M., & Pan, T. (2015). Efficient and quantitative high-throughput tRNA

sequencing. *Nature methods*, *12*(9), 835–837. https://doi.org/10.1038/nmeth.3478

142. Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C. H., Fu, J., Chen, S., & Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(41), E6117–E6125. https://doi.org/10.1073/pnas.1606724113

143. Zhu Y, Neeman T, Yap VB, Huttley GA (2017) Statistical methods for identifying sequence motifs affecting point mutations. Genetics. https:// doi. org/ 10. 1534/ genetics. 116. 195677

Chapter 2: Dynamic changes in tRNA expression establish proliferation- and differentiation-specific codon optimality in neurogenesis


Neural differentiation
Codon usage enhances translation and stability
m^{7}G
AAAA
Enriched for ribosome biogenesis
Codon usage poorly adapted for translation and stability
m^{7}G
AAAA
Enriched in pro-differentiation mRNAs
Neural Proliferation
Codon usage enhances translation
m^{7}G
AAAA
Enriched in pro-differentiation regulatory proteins especially RNA-binding proteins
Pathway A
Pathway B
Pathway C
RNA splicing
3'UTR activity
Neuronal maturation
refractory to codon-mediated regulation
m^{7}G
AAAA
Enriched in DNA-binding proteins
Attenuates codon-optimality mediated mRNA stability
m^{7}G
AAAAAAAAAAAAAA

Background:

Research Gap: It is largely unexplored how the genome-wide regulation of transfer RNA (tRNAs) contributes to the genetic control of animal neurogenesis.

Key concepts: Cell differentiation, neurogenesis, post-transcription, tRNAs, codon optimality

Within higher animals, the central nervous system is the most complex and morphologically diverse organ. Healthy brain development requires precise spatiotemporal control of stem cell proliferation and differentiation. Neurogenesis is archetypal of cell lineage specification in that the dynamic remodeling of the genetic landscape orchestrates the differentiation of a small pool of neural progenitors into manifold subtypes of post-mitotic neurons and glia. Dysregulated neural stem cell proliferation is implicated in tumor initiation and growth, particularly in pediatric brain cancers (Maurange, 2020; Azzarelli et al.,2018). Hence, a major challenge in neurobiology is to unravel the genetic changes contributing to neural cell fate.

Regulation of mRNA is an essential aspect of gene expression as it is the mRNA repertoire that defines the repertoire of proteins that, in turn, establishes cellular identity and function. To sustain proliferation and progress through asymmetric cell division in a timely manner, neural progenitors must maintain high rates of global protein synthesis to meet their own metabolic needs in addition to regulating the production of differentiation-related proteins that are inherited by their progeny. On the other hand, post-mitotic neurons are characterized by low rates of global protein synthesis; instead, the polarized morphology imposes the need for local, activity-dependent protein synthesis at the synapses and dendrites (Buffington et al., 2014; Huber, 2000). Protein output is a function of mRNA steady-state levels. In turn, the mRNA steady-state level is dictated by the dynamic equilibrium between rates of synthesis (transcription) and degradation (Sun et al., 2012). Even so, much focus has been given to transcriptional control of mRNA dynamics; however, it is known that changes in transcription rate may be buffered against altering steady-state mRNA or protein levels (Trimmers and Tora, 2018; Swindell et al., 2015). Therefore, deciphering the dynamics of post-transcriptional processes– i.e., mRNA stability and translation efficiency – paints a complete picture of mRNA regulation across development.

Emerging studies support codon optimality as an influential *cis*-regulator of mRNA destruction in metazoans. Paradigm-shifting work in *S.cerevisiae* revealed that certain codons either stabilized (‘optimal codons’) or destabilized (‘non-optimal) their mRNAs. Furthermore, causality between codon identity and mRNA half-lives was established by monitoring changes in decay rates after re-coding with synonymous codons. To quantify this association the same study proposed the Codon Stabilization Coefficient (CSCs), which is the Pearson’s R correlation coefficient between codon frequency and mRNA half-lives (Presynyak et al., 2015). Since then, several animal studies have supported codon optimality as a major genetic determinant of mRNA degradation rates. Two groups independently elucidated the causal link between enrichment of non-optimal codons and destabilization of a subset of maternal mRNAs after zygote genome activation in *Xenopus laevis* (frog) and Danio *rerio* (zebrafish) (Bazzini et al., 2016; Mishima and Tomari, 2016). Reporter assays in transformed and normal human cell lines demonstrated that codon optimality strongly influences mRNA stability in a translation-dependent manner (Bazzini et al., 2019; Forrest et al., 2020). Using tissue-specific global profiling of mRNA decay, our lab discovered codon optimality as a major determinant of differential mRNA decay in the *Drosophila* embryo *in vivo*, a classical model for developmental biology (Burow et al., 2018). In *S.cerevisae*, the conserved mRNA decapping factor *Dhh1* was found to preferentially bind actively translated mRNAs enriched with slow decoding ribosomes (Radakrishnan et al., 2016). Ribosome profiling in yeast revealed a correlation between A-site occupancy – where aminoacylated tRNAs bind their cognate codons – and mRNA half-lives (Hanson and Coller, 2018). Collectively, these experiments lend evidence to how codon optimality underlies a general mechanism that couples mRNA stability to the translation elongation status. Still, evidence suggests that codon-dependent mRNA decay seems to be context-dependent and dynamically regulated during development. Although codon optimality was found to be a strong determinant of mRNA stability in *Drosophila* whole embryos, the stabilizing effect of these codons was attenuated – and, in some cases, reversed - in the embryonic nervous system. This led the authors to speculate if this developmental switch in codon-mediated mRNA decay is linked to altered tRNA expression upon embryonic neural differentiation (Burow 2018).

Traditionally regarded as housekeeping genes, several studies now support the dynamic regulation of tRNAs across development states (Ditmar, 2009; Gingold et al., 2014; Gogakos et al., 2017). Because the genomic dosage of tRNA genes varies widely, current models suggest that “optimal codons” have rapid elongation rates because they are decoded by abundant tRNAs. Combined mRNA sequencing and tRNA microarrays in hundreds of healthy and transformed human cell lines revealed proliferation and differentiation-specific signatures of codon-anticodon co-adaptation, suggesting a role for codon-mediated translation efficiency in cell-fate determination (Gingold et al., 2014). Because tRNAs are central to protein synthesis, distinct tRNA profiles should be informative about the proteomic landscape. In their systematic study of metastasis in breast cancer cell lines, Goodarzi et al. combined gain/loss-of-function of selected tRNAs and ribosome profiling to uncover concomitant changes between altered tRNA expression and protein elongation rates, as well as correlate changes in proteins enriched with cognate codons of the perturbed tRNAs (Goodarzi 2016). A caveat, however, is that inferences about gene expression programs in diseased states and cell lines may not recapitulate normal cellular development *in vivo*.

Recent bulk tRNA sequencing by Pinkard et al. characterized *in vivo* variation in tRNA levels between different regions of the adult mice brain (Pinkard et al.,2020). Ishimura et al. showed that double knock-down of neural-specific tRNA-Arg^{TCT} and ribosome recycling factor *Gtbp2* led to neurodegeneration in 3-day-old mice due to increasing ribosome pausing on mRNAs enriched with the cognate AGA codons. Indeed, aberrant tRNA processing and metabolism are implicated in over 50 neurodevelopmental and neurodegenerative diseases that are linked to aberrations in the translation machinery (Knight et al., 2020; Schaffer et al., 2019).   Still, changes in the tRNA landscape in neurogenesis and the regulatory consequences thereof remain largely uncharacterized. Moreover, most previous works profiled tRNAs under culture conditions, thus providing limited insights into the developmental role of tRNA variation *in vivo* and in non-diseased states.

To investigate how the interplay between tRNA expression and codon optimality contributes to proliferation and differentiation programs in neurogenesis, we leveraged high-throughput genomic assays and bioinformatics analysis to chart changes in mRNA stability and tRNA levels *in vivo* between populations of neural stem cells (neuroblasts) and post-mitotic neurons in the *Drosophila* larval central nervous system, a classical model of neurobiology. We demonstrate how cellular tRNA levels provide strong explanatory power for the codon stabilization coefficients (CSCs), thus providing evidence that the regulatory signal in codons is modulated by the variation in their cognate tRNA levels. We uncovered context-dependent codon optimality-mediated mRNA decay and translation efficiency, as well as functional group differences, that are oriented toward tissue-specific physiology. Specifically, the neuroblast tRNA pool establishes a codon optimality program that favors the stability and translation efficiency of mRNAs that are pro-proliferative (i.e., ribosome biogenesis and energy production); in contrast, the neuron tRNA pool supports the selective translation of a subset of RNA-binding proteins, many of which are known to regulate key neurogenic pathways, including alternative RNA splicing. Decades of work in the Drosophila central nervous system have yielded unmatched insights into the genetic control of neural proliferation and differentiation due to the suite of molecular tools that enable tissue-specific manipulation. Given that both translation elongation and the molecular aspect of neurogenesis (Bridi et al., 2020) are well-conserved in metazoans, our work has implications for how the regulation of tRNAs drives the translation control of neural development in other animals.

RESULTS

Figure 1: Distinct changes in the tRNA levels and post-transcriptional editing between neuroblasts and neurons

The cardinal importance of translation control of gene expression to neurological function is underscored by the prevalence of dysregulated translation in many neurodevelopmental diseases (Kapur et al., 2017). tRNA availability is a major determinant of translation efficiency, which directly modulates protein synthesis (Tuller et al., 2010). However, the regulation of tRNAs in neural proliferation and differentiation under physiologically normal conditions hitherto has not been explored. Here we set out to quantify the changes in tRNA landscape in *Drosophila* neurogenesis by sequencing mature tRNAs (Gokagos et al., 2014) from populations of neural progenitors (neuroblasts) and post-mitotic neurons. Neuron-enriched brains (90-95% neurons) (Homem and Knoblich, 2012) were dissected from late-stage larvae (120ALH)), and ectopic neuroblast brains were dissected from stage-matched *Insc-GAL4; UAS-aPKC*^{caax} mutants, whose neural progenitors can only undergo symmetric self-renewal(Lee et al., 2006). tRNAseq reads were aligned to a reference of high-confidence reference of mature tRNAs (addition of CCA added the 3'ends), in which identical sequences were collapsed (Methods). tRNA read counts were normalized using the trimmed mean of M-values (TMM), and differential gene expression analysis was performed with the non-parametric (empirical Bayes) tool, NOISeqBio, which was shown to better control the false-discovery rate between biological replicates(Tarazona et al., 2015)

Variation in tRNA composition at the Isodecoder level

First, we evaluated the potential regulation at the decoder level (**Figure 1A)**. i.e., tRNA genes that share the same anticodon but different sequence bodies. Here, we found significant differences in the levels of 11/84 nuclear tRNAs and 2/22 mitochondrial tRNAs between neuroblasts and neurons (FDR-adjusted q-value < 0.1; **Figure 1B**). Seven nuclear tRNA decoders are upregulated in neurons, with *Pro-CGG-1-x* having the greatest 2.5-fold enrichment in neurons compared to neuroblasts. Another neuron-upregulated *Gly-GCC-1-x* with a fold change of 2 and the most significant *q*-value was also found to be highly expressed in the adult murine CNS compared to non-CNS tissues (Pinkard et al., 2020). Conversely, three nuclear tRNA isodecoders were significantly upregulated in neuroblasts, in which *Arg-TCT-3-1*, with a 3.7-fold enrichment, is the most neuroblast-upregulated isodecoder.

Interestingly, 19 out of the 22 mitochondrial-tRNA genes have higher expression in the neuroblasts. This was rather surprising because mitochondrial respiration is believed to be suppressed in both *Drosophila* and mammalian stem cells, which rely mostly on glycolytic ATP production. However, emerging evidence agrees that other aspects of mitochondrial metabolism help regulate the proliferation and temporal patterning of neural stem cells (Iwata et 2021; den Amelie, 2019).

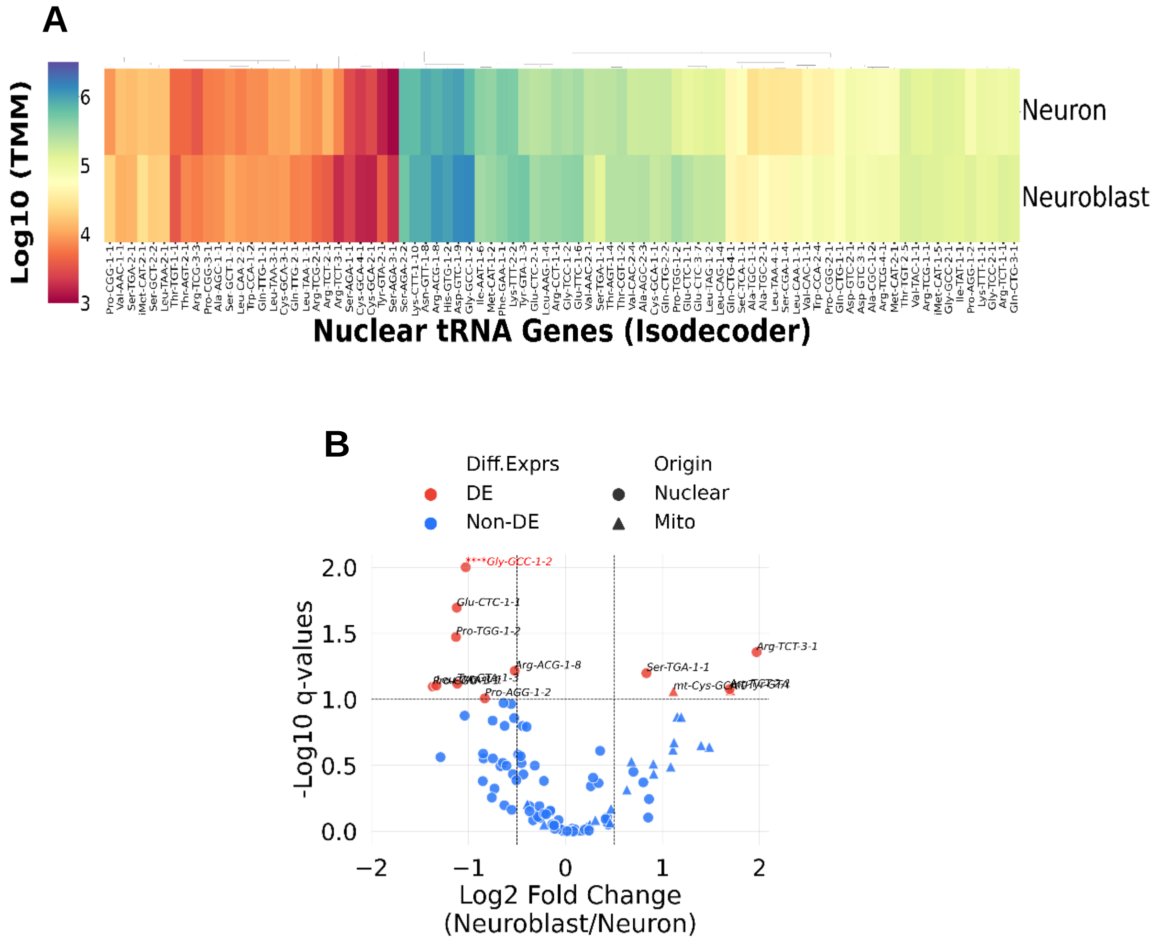


Figure 1 Genome-wide tRNA sequencing in *D.melanogaster* larval CNS of neuron-enriched (‘Neuron’) and neuroblast-enriched (‘Neuroblast’) brains (n=3 replicates/cell-type). **A)** Heatmap showing cellular tRNA composition at the isodecoder (gene) level. Normalized tRNA expression is based on Trimmed Mean of M-values (TMM).

B) Volcano plot shows 11 out of 84 nuclear tRNA isodecoder genes are significantly altered (FDR *q-value*<0.1) between neurons and neuroblast (red colored circles). 7 tRNA isodecoders were upregulated in neurons, and 3 tRNA isodecoders were significantly upregulated in neuroblasts. ****Gly-GCC-2-1* is highly significant (FDR *q-value*<0.001) but true q-value is not shown to avoid occluding other data points.

Mapping changes in the tRNA epitranscriptome between neuroblasts and neurons

At the isodecoder level, we further characterized changes in the levels of tRNA post-transcriptional modifications between neuroblasts and neurons. With an average of 15-25% of modified nucleosides, tRNA molecules bear the highest density of post-transcriptional modifications of all RNA types. Biochemical studies, primarily by mass spectrometry, have cataloged over 120 modified tRNA nucleosides thus far. Post-transcriptional modifications are essential to nearly all aspects of tRNA regulation, such as guiding proper folding of the tertiary structure, structural stability, fine-tuning decoding capacity and accuracy, aminoacylation efficiency, and acting as recognition elements for tRNA decay and trafficking. As a result, many modified sites are conserved across eukaryotic tRNAs **(Figure 1C)** (*reviewed in* Zhang et al., 2022). For example, in the adult mouse prefrontal cortex, the knockdown of *NSUN2*, which installs the 5-methylcytosine modification at the variable loop on serval tRNAs, led to a global decrease in *Gly-GCC* tRNAs and a concomitant decrease in the translation of Gly-rich proteins that resulted in defective neurotransmission and increased seizures (Blaze et al. 2021)

In library preparation, some tRNA modifications will arrest the reverse transcriptase (RT), whereas other types of modifications can be readthrough by the RT, either silently or resulting in non-random signatures of base misincorporation that is higher than expected by technical noise (Motorin and Marchand, 2021). Here, we take advantage of the latter behavior to profile the neural tRNA epitranscriptome and track changes in the levels of base editing across neurogenesis. Systematic analyses of next-generation sequence datasets (NGS) have shown that the average base substitution error rate ranges from 0.1% to 1% (Ma et al., 2019; Stoler et al., 2021). However, to achieve a more robust detection of variants, we required minimum coverage of 30 reads and set the background substitution error based on the spiked-in RNA oligonucleotide because we reasoned that since the spiked-in oligonucleotide is unmodified, any detected base mismatch will be technical in origin. Moreover, there are distinct steps in tRNA sequencing library preparation compared to standard mRNAseq on which the aforementioned analyses on sequence error rates are based. As such, we determined the variant frequency at the first 72 nucleotides for each isodecoder as the difference between the isodecoder base mismatch fraction and the highest mismatch fraction on the spiked-in oligonucleotide, which ranged from 1.0% to 3.3% across samples. The per base variant frequencies were then averaged across replicates to obtain the condition-specific variant frequencies. In the neuron condition, there are 1338 variant positions (>1%) out of a total of 6048 bases over the entire sequences **(Figure 1D),** and in the neuroblast, there are 1315 variant positions (>1%) out of a total 6048 bases (**Figure 1E).**

We next performed differential modification analysis. Modifications are also dynamically regulated in development, depending on intracellular factors such as the expression of the writer enzymes and the availability of certain precursor nutrients and metabolites (Asano et al., 2018; Schwarts et al., 2018; Frye et al., 2016).

We considered a tRNA site as differentially modified if it has an absolute change in variant frequency of at least 0.1 (10%) between neuron and neuroblast. In our dataset, 174 variant positions are differently modified, wherein 128 variant positions reported a higher level in neurons, whereas 46 variant positions were more edited in the neuroblast **(Figure 1F)**. The most upregulated variant in neurons compared to the neuroblast is at guanine at position 27 (G27) on *Ile-AAT-1-x*, having a different frequency of 0.48. G27 on tRNAs is known to carry the conserved modification *N^2N^2-methyl guanosine* (m^2_2G) or *N^2-methyl guanosine* (m^2G), deposited by eukaryotic tRNA methyltransferase 1 (Trm1) and is believed to stabilize the anticodon arm (Hori, 2014). Conversely, thymine at position 51 on *Gly-GCC-2-x* and thymine 58 on *Asp-GTC-2-x* tied for the most upregulated variants (0.3 difference) in neuroblast, but we cannot ascertain the identities of these modified nucleosides. We note, however, that these variant positions are adjacent to the variable loop on the T-arm, where uridine (thymine) bases are known to carry conserved modifications (Zhang et al., 2022).

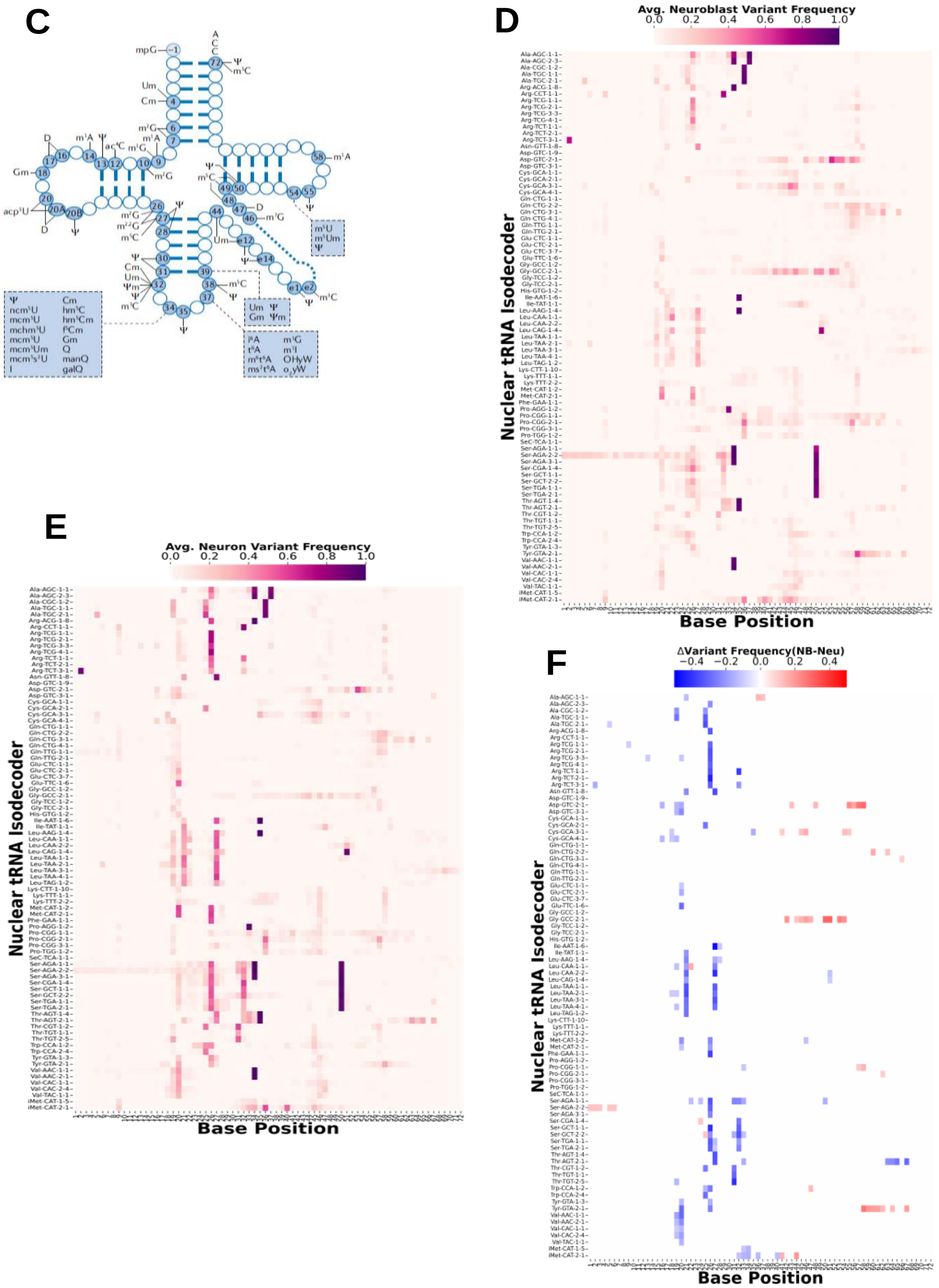


Figure 1: Estimating tRNA post-transcriptional levels based on modification-induced RT mismatch frequency

C) Schematic representation of the secondary structure of tRNA with post-transcriptionally modified residues (source: Suzuki, 2021) Heatmap variant fractions at the first 72 nucleotides of the mature tRNA transcripts in **D)** neuroblasts and **E)** neurons. Minimum coverage for variant is 30 reads and the background substitution error was to set to the highest base mismatch frequency (0.01 to 0.034) of the spike-in oligonucleotide that was added to each sample in the library preparation step. **F)** Heatmap shows sequence sites with 0.1 or more difference in variant fractions between neuroblasts and neurons.

tRNA Anticodon abundance changes across neurogenesis

Given that we are ultimately interested in the role of tRNAs in cell-type specific codon optimality, our analysis, hereafter, focuses on the nuclear anticodon pool. To quantify changes in the anticodon pool upon neural differentiation, we summed the reads of all tRNA isodecoders of the same anticodon and performed differential gene expression (Methods). 8 out of 45 tRNA anticodons differed significantly between the neuroblasts and neurons. 7 of the 8 differential expressed anticodons - *Pro-TGG, Pro-CGG, Gly-GCC, Glu-CTC, Tyr-GTA, Leu-CAA, Arg-ACG,* - were significantly upregulated in neurons; whereas only tRNA-Ser-TGA was significantly higher in neuroblasts **(Figure 1G).**

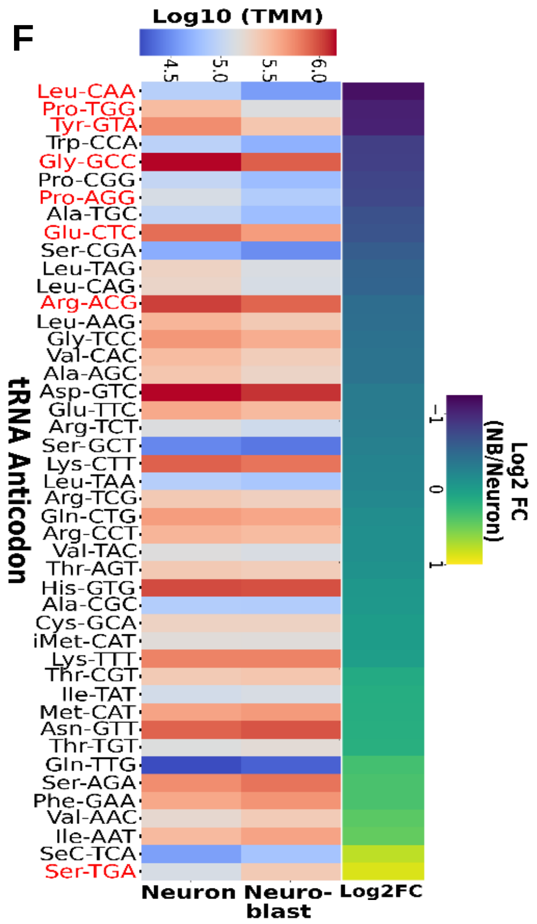


Figure 1G: Neural tRNA composition at the anticodon level Left: Heatmap showing the normalized levels of each nuclear tRNA anticodon group. Right: heatmap of the log base-2 fold changes anticodons levels between neuroblasts and neurons. levels of 7\8 anticodon were significantly higher in neurons, while 1/8 anticodon was significantly upregulated in neuroblasts. The differentially expressed anticodons (FDR *q-value*<0.1;) are highlighted in red on the y-axis.

Figure 2:Codon optimality differs in neuroblasts and neurons and affects cohorts of mRNAs that regulate developmental-specific functions

We next investigated the influence of codon-mediated mRNA decay on neural proliferation and differentiation. To this end, we obtained genome-wide mRNA decay rates from populations of neural progenitors ('neuroblast') and neurons by performing pulse-chase EC-tagging. Briefly, the larvae were fed 1 mM 5-ethynylcytosine (EC) for a 12-hour pulse and transferred to media containing 10 mM unmodified uridine for the following timepoints 3, 6, and 12-hour chase. After each chase timepoint, the total RNA was collected from the dissected brains, and the EC-tagged RNA was biotinylated and purified by a streptavidin pull-down and prepared for Illumina sequencing. mRNA decay rates were estimated by fitting an exponential curve through the chase timepoints (Methods).

To measure the influence of codon usage on mRNA stability, we calculated the codon stability coefficients (CSC) for each of the 61 sense codons as the Pearson's correlation coefficient between mRNA half-lives and codon frequency (Presynak et al., 2015). The neuron CSCs ranged from -0.17 to 0.24, and the neuroblast CSCs ranged from -0.19 to 0.19 (**Figure 2A)**. A positive CSC indicates that the codon is preferentially used in mRNAs with long half-lives (i.e., 'stabilizing' codon), whereas a negative CSC indicates that the codon is preferentially used in mRNAs with shorter half-lives (i.e. 'destabilizing' codon). In total, 27 codons exhibit altered stabilities (absolute difference in CSC>0.05) between neurons and neuroblasts. We also identified 5 "neuroblast-optimal" codons (CSC > 0.03 in neuroblasts and < 0.05 in neurons and a CSC difference > 0.05 between cell types) , and 2 "neuron optimal" codons, defined by the aforementioned rule **(highlighted on Figure 2A)**.

Next, we explored the biological influence of cell-type specific codon bias on mRNA stability. To this end, we performed gene ontology (GO) enrichment analysis on mRNAs from each cell type that are among the top 10% enriched with neuroblast-specific or neuron-specific optimal codons. 'Cytoplasmic translation' was the most significant GO category (FDR-adjusted p-value > 0.001) for the neuroblast-derived mRNAs enriched with neuroblast-optimal codons **(Figure 2A)**. On the other hand, neuron-derived mRNAs enriched with neuron-optimal codons reported top GO categories related to synaptic function **(Figure 2C)**. These trends show that codon optimality supports the stabilization of mRNAs that are relevant to the cell's physiology. Nonetheless, the influence of codon optimality is less impactful on neuron-specific mRNA decay, based on the fact that the average half-lives of these mRNAs only marginally changed compared to when they are expressed in the neuroblasts **(Figure 2C)**. This pattern of attenuated codon optimality in the larval nervous system aligns with Burow et al., who found that codon-mediated mRNA decay was attenuated in the embryonic nervous system (Burow et al., 2018).

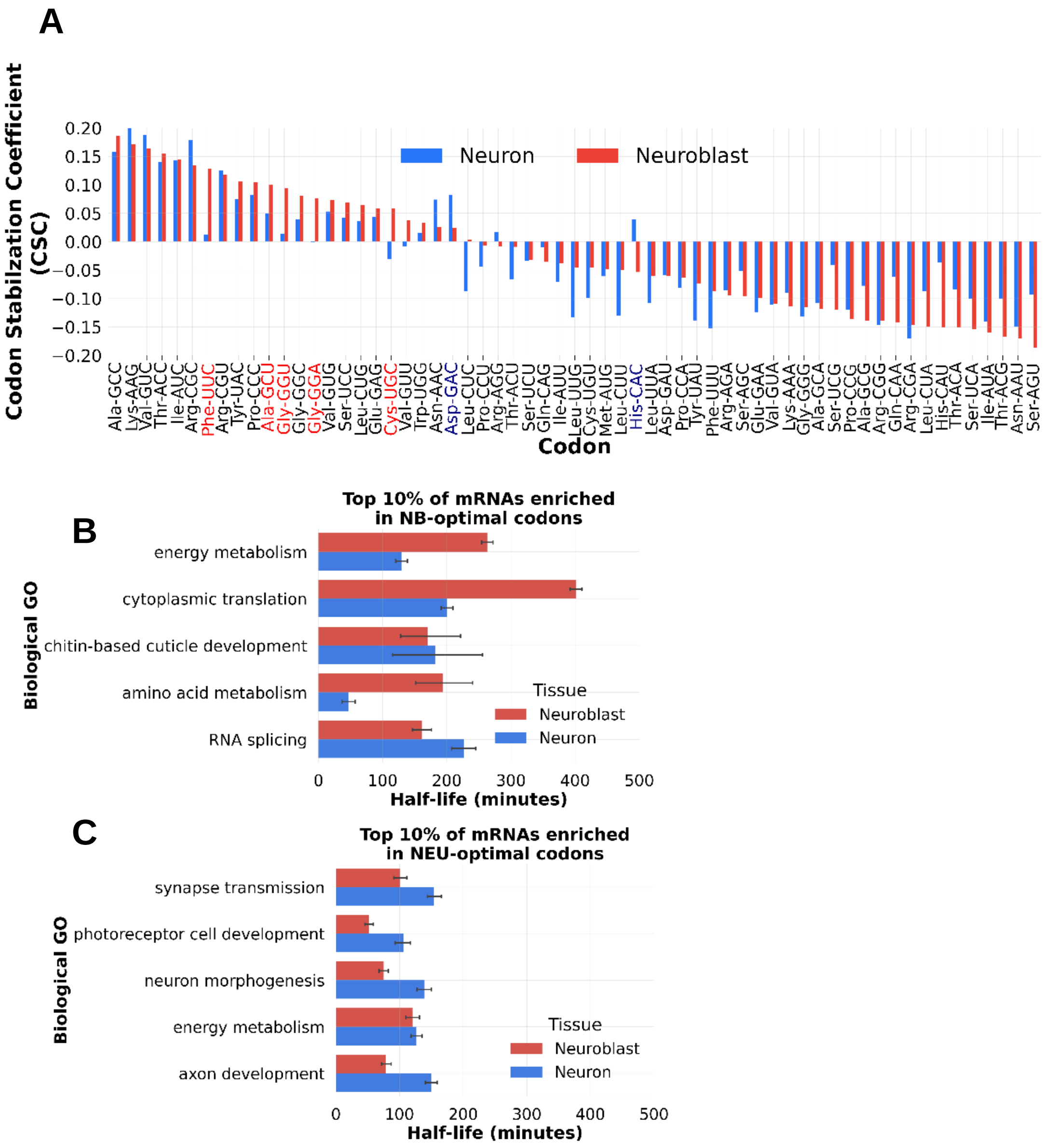


Figure 2: Distinct effects of codon-mediated mRNA decay between neuroblasts and neurons.

A) Barplot shows the paired values of the cell-type specific codon stabilization coefficient, CSC, which is based on the Pearson's correlation between codon enrichment and genome-wide half-lives of mRNAs measured in each cell-type. Codons with positive CSC are preferentially enriched in mRNAs with longer half-lives, in contrast to negative CSC codons which are used more often in mRNAs with shorter half-lives. On x-axis, codons highlighted in red= neuroblast optimal, and blue codons = neuron optimal. Cell-type specific optimal CSCs is defined by (CSC > 0.03 in condition cell-type A and < 0.05 in condition cell-type B and a CSC difference > 0.05 between cell types)

Cell-type specific codon optimality distinctly affect the stability of functionally related mRNAs. Barplots showing the cell-type specific half-lives (error bars= standard error)of the top 5 non-redundant gene ontology categories (all FDR adjusted p-value < 1e^{-5}), **B)** for neuroblast and **C)** neuron mRNAs that are among the top 10% enriched in neuroblast- or neuron-specific optimal codons.

Cell-type specific differences in the relationship between codon-mediated mRNA decay and cellular tRNA levels

Having observed that both tRNAs expression and the codon stabilization coefficients (CSCs) are altered between neuroblasts and neurons, we next investigated the role of tRNAs in codon-mediated mRNA stability. Experiments in *S.cerevisiae* supports the "stabilization-by-translation" model for how codon usage regulates mRNA degradation, wherein optimal codons enriched in stable mRNAs are expected to have faster elongation rates and, therefore, less likely to trigger quality control factors that sense stalled ribosomes (Hanson et al., 2018; Radakrishnan et al., 2016). Here, we observe a positive relationship between the CSCs and their cognate anticodon expression in both neuroblasts (Pearson's R= 0.43, p-value=1.8e^{-03}) and neurons (Pearson's R=0.53, p-value<1.08e^{-04}) **(Figure 2D; 2E).**

Since it is widely believed that proliferation is a major determinant of codon optimality (Gingold et al., 2014; Rochoa et al., 2003), we were not expecting a stronger linear correlation between anticodon levels and CSCs neurons than in neuroblast **(Figure 2D; 2E)**. However, a comparison to Watson-Crick anticodons fails to account for the reality that several codons are wobble translated by multiple tRNAs. Thus, to better reflect the total tRNA availability of a codon, we computed the tRNA adaptive index (tAI) that is widely used as a proxy for the translation efficiency of a codon, based on Watson-Crick and wobble base pairing (dos Reis et al., 2004). Because the original tAI relies on static tRNA copy number variation, we computed the cell-type specific tAI using the tRNA expression values, which offers a more dynamic representation of tRNA-mediated translation regulation across neural differentiation. We also normalized each codon's tAI within its amino acid family such that the most translationally optimal synonymous codons have a tAI equal to 1.0 (Methods). Again, we see that for both neuroblasts and neuron samples, there is a significant association, along with a large effect size, between a codon's translation efficiency (tAI) and its stabilizing influence on mRNA half-lives (Welch's T-test $P < 0.001$; Cohen's d>0.8) (**Figure 2F; 2G**). Broadly, this means that codons that are enriched in stable mRNAs (positive CSCs), on average, are better adapted to the cellular tRNA pool, based on higher median tAI, than those codons that are enriched in less stable mRNAs, i.e., negative CSCs. However, in neuroblasts, there is a stronger association between the tAI and CSCs (Cohen's d=1.44) compared to the neuron population (Cohen's d=1.00). Notably, in neuroblasts, there are twice as many codons with positive CSCs that are also the most translationally optimal (i.e., tAI=1.0) within their synonymous group compared to codons with the negative CSCs.

Since we noticed that neuroblasts exhibit a stronger association between tRNA availability and CSCs, we wondered if this trend is also context-dependent at the gene level. To monitor the relationship between mRNA stability and tRNA adaptiveness before and after neural differentiation, we selected mRNAs that are shared between neuroblasts and neurons based on having a minimum expression value of TPM>1 in both cell types. For each of the shared mRNAs, we computed the neural-specific tAI as the geometric mean of the neural-specific codon tAI represented in their CDS (Methods). There is a stronger linear correlation between mRNA half-lives and tAIs in neuroblasts (Pearson's R=0.38, *P*<2.16e^{-16}) (**Figure 2H**). But only a weak correlation in neurons (Pearson's R=0.10, *P*=2.72e^{-05}) (**Figure 2I**).

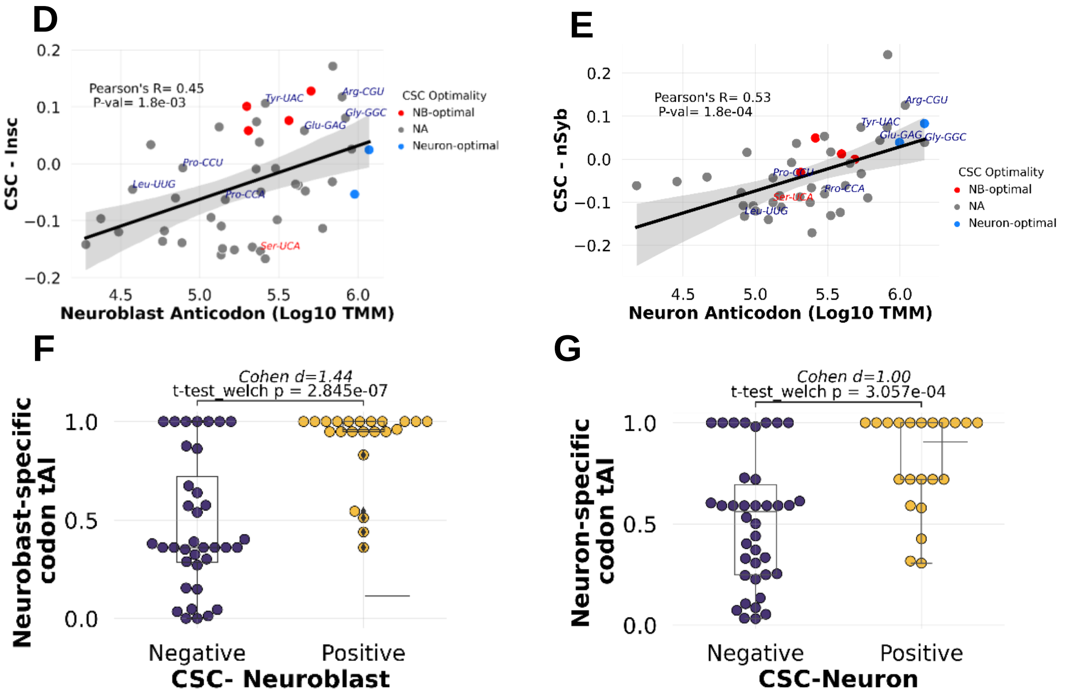


Figure 2: tRNA availability explains codon-mediated mRNA decay in the *Drosophila* CNS On

Scatterplots show a positive linear relationship between CSC and normalized anticodon levels in **D)** neuroblasts (Pearson' R=0.45,p-value<0.01) and **E)** neurons (Pearson' R=0.53,p-value<0.01).

F,G: Boxplots show large effect size (Cohen's d>1.0) between the stabilizing effect of codons and their total tRNA availability in each cell. On the y-axis is the tRNA adaptive index (tAI) that estimates the translation fitness or efficiency of each codon based on the availability of all of its cognate anticodon (both Watson-Crick and wobble pairings). For both neuroblasts and neurons, codons that are enriched in stable mRNAs are, on average, better translated than codons enriched in less stable mRNAs (Welch's T test, *P*<0.05).

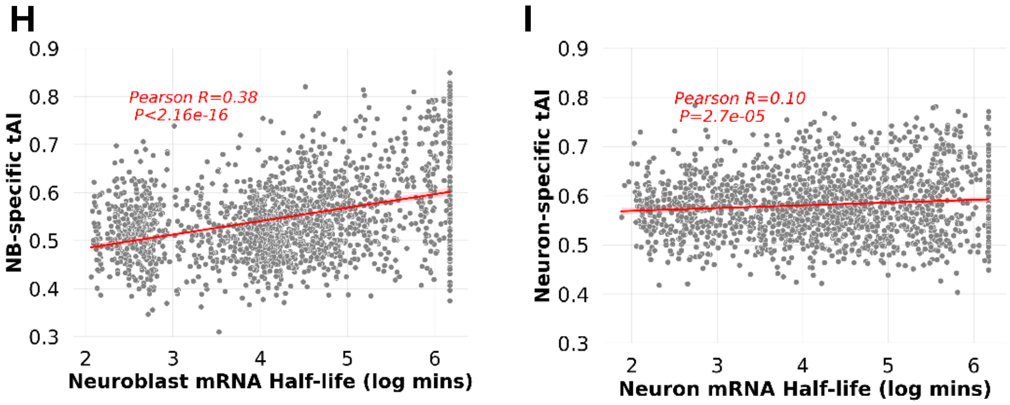


Figure 2: tRNA availability explains codon-mediated mRNA decay in the *Drosophila* CNS On

H,I: Scatterplot of gene-level tAI and mRNA half-lives. The gene-level tAI is computed as the geometric mean of the tAI of all codons within the coding sequence of the mRNA. tRNA adaptation better explains mRNA half-lives in (**H)** neuroblasts (Pearson's R=0.38, p-value<2.16e^{-16}) than in **I)** neurons (Pearson's R=0.10, p-value=2.7e^{-5})

tRNA epitranscriptome also shapes codon optimality: Inosine-34 modified tRNAs preferentially decode stable codons in neuroblasts

Modification in the anticodon stem profoundly impacts tRNA decoding fidelity and thus overall gene expression. The editing of adenosine to inosine by adenosine deaminases (hetADATs) at the wobble anticodon position of specific ANN tRNAs (A34-to-I34) is deeply conserved across eukaryotes. Inosine-34 modification expands the decoding capacity of ANN tRNAs from recognizing only U-ending codons to decoding C-ending and U-ending codons (Rafael-Yberns et al., 2019). Functional studies in fungal and bacterial systems showed that usage of Inosine-34 codons improved translation efficiency (Novoa et al., 2012; Lyu et al., 2020) and evolved as a major determinant of fungal gene expressivity (Wint et al., 2022). Since our previous analyses showed translation efficiency and mRNA stability are coupled, we wondered if inosine-34 tRNA decoding also contributed to codon-mediated mRNA decay in the *Drosophila* central nervous system. To this end, we compared the CSCs of the 16 codons that are decoded by inosine-34 edited tRNAs. hetADAT is constitutively expressed, and indeed we observed the characteristic high frequency of A-to-G mismatch (Peng et al., 2012) at the anticodon regions for all tRNAs that are known substrates of inosine-34 editing, namely: tRNA^{Thr}(AGU), tRNA^{Ile}(AAU), tRNA^{Pro}(AGG), tRNA^{Arg}(ACG), tRNA^{Leu}(AAG), tRNA^{Ala} (AGC), tRNA^{Val}(AAC), and tRNA^{Ser}(AGA) **(Figure 2J;2K)**. We found that Inosine-34 codons are overrepresented as positive CSCs (stable codons) in neuroblast (Fisher exact test P-value=6.7e^{-03}), but there is no significant difference in inosine-34 codon usage for neuron CSCs (Fisher *P* = 0.07) (**Figure 2L**).

This finding parallels recent ribosomal profiling experiments in human and mouse embryonic stem cells (ESCs) that elucidated how self-renewing hESCs preferentially used inosine-34 codons for enhancing translation efficiency in comparison to differentiating hESCs, as well as a downregulation of hetADATs upon hESC differentiation (Bornelov et al., 2019). These findings raise the possibility that other tRNA modifications in the anticodon stem may also be influencing codon stabilities.

In summary, our results support codon composition as a determinant of mRNA stability in neuroblasts but is weakly impacts mRNA decay in the neuron. The positive correlation between tRNA expression and CSCs agrees with the model that codon identity stabilizes mRNAs in a translation-dependent manner.

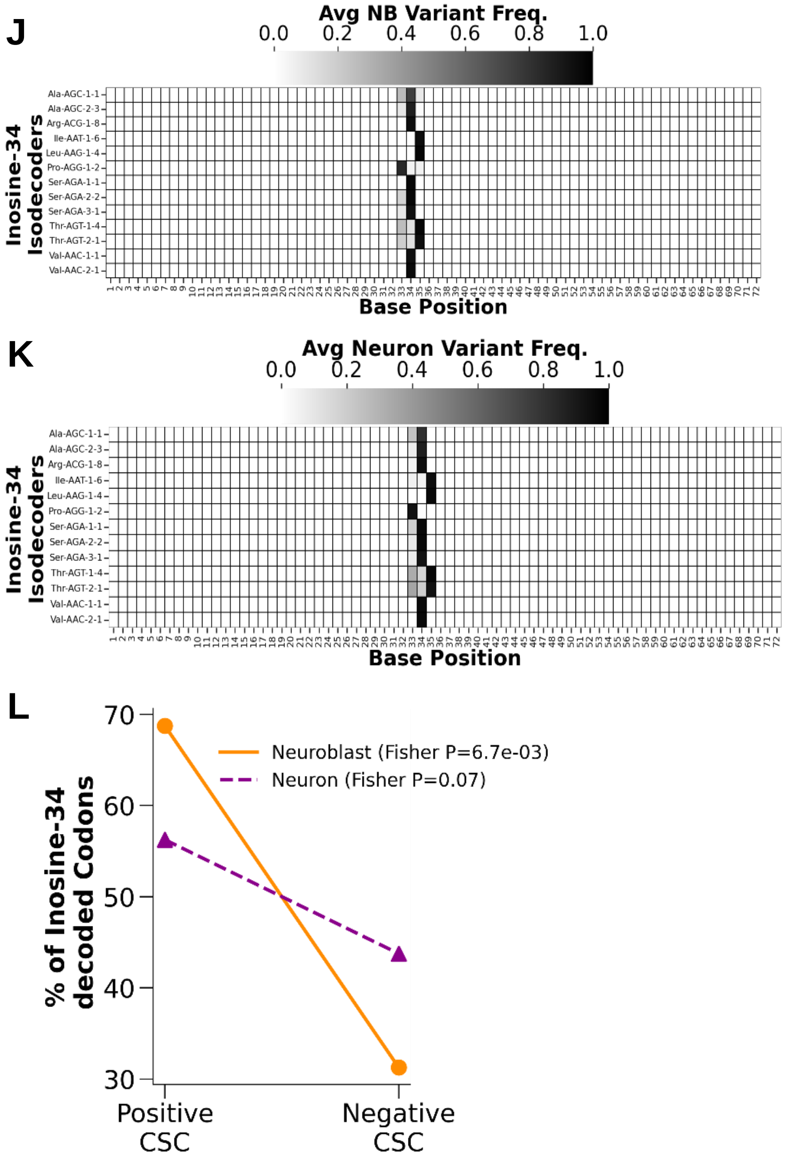


Figure 2: Preferential decoding of stable codons in neuroblasts by inosine-34 modified tRNAs

J,K: Modification heatmaps for the tRNAs that are known to be edited by hetADAT in the wobble anticodon position (all other variant sites are masked). Note that the Inosine-34 signal is not precisely at base 34 because the variable bases in the anticodon stem may lead to an offset by 1 base pair. However, the only other modified nucleoside in the anticodon region is found on the uridine of specific tRNAs [Zhang et al., 2022].

L: Comparison of the fraction of inosine-34 decoded codons (n=16) that match positive and negative CSCs in neuroblasts and neurons. Inosine-34 decoded codons are significantly enriched with positive CSCs in neuroblasts (Fisher's enrichment test *P*=6.7e^{-03}) but in neurons (Fisher's *P*=0.07).

Figure3: mRNAs expressed across neurogenesis exhibit proliferation-specific and differentiation-specific codon usage

An outstanding question in developmental biology is how stem cells are able to transcribe both pro-proliferative and pro-differentiation mRNAs whilst maintaining their own self-renewing phenotype. It has been shown that mammalian cells use distinct codon signatures in proliferation and differentiation genes, suggesting that codon bias may be exploited as a mechanism to distinctly co-regulate functionally related mRNAs (Gingold et al., 2014; Bornelov, 2019). Given the distinct tRNA expression profiles of neuroblasts and neurons, we sought to investigate if pro-differentiation and pro-proliferation transcripts employ distinct codon usage signatures, as this property would make them differently regulated by the cellular tRNA pool, depending on whether they are expressed in neural progenitors or the neurons.

We chose to focus on shared mRNAs because we wanted to monitor how the same set of mRNAs would be differently regulated due to the intrinsic properties of their coding regions. To directly elucidate the codon usage patterns of neural transcripts, we selected transcripts (n=5169 on) that have a non-zero expression (TPM>1) in both neuroblasts and neurons and are annotated on Flybase. We then calculated the relative codon frequencies such that each mRNA sequence is represented as a 59-dimensional vector. We normalized codon frequencies within the amino acid groups to mitigate confounding effects due to amino acid usage and gene length. To visualize how transcripts are distributed based on codon usage, we performed principal component analysis (PCA) followed by unsupervised clustering using Kmeans, where the optimal number of six clusters was determined using the elbow method (Methods). PCA on codon usage captured a total variance of 23%, with the first principal component, PC1, capturing 15% of the variation, and the second principal component, PC2, on codon usage captured 8% of the variation within the dataset (**Figure 3A**). Because PC1 captures most of the variance present in the data, we focused on mRNAs belonging to cluster 5 and cluster 4 that are projected at opposite poles of PC1, suggesting that these two mRNA sets are the most divergent in their codon usage. GO enrichment analysis revealed that cluster 4 mRNAs (n=373) are significantly enriched in growth-related terms related to ribosome biogenesis and energy metabolism (**Figure 3B**). It is noteworthy that ‘mitochondrial fusion’ is enriched in the ‘mitochondria’ term of Cluster 4 because the maintenance of fused mitochondria is essential to the proliferative capacity of stem cells, and disrupted mitochondrial fusion was shown to decrease self-renewal in mammalian and *Drosophila* neural stem cells (Dubal et al., 2022; Khacho et al., 2016). On the other hand, Cluster 5 mRNAs (n=219) at the positive pole of PC1 were significantly enriched with GO terms related to neural differentiation and neuronal function. Thus we assigned the negative pole of PC1 as capturing codon bias for ‘Proliferation’ and the positive pole as codon bias for ‘Differentiation.’

The enrichment of ribosomal and energy metabolism-related terms in Cluster 4 led us to calculate the popular codon bias metric, the codon adaptation index (CAI) (Sharp and Li, 1987). CAI quantifies the similarity of a gene's of synonymous codon usage to that of a reference set of highly expressed ribosomal protein genes because ribosomal proteins are usually constitutively expressed. Moreover, it is widely believed that the CDS of ribosomal proteins is under strong selection because it serves as an adaptation for maintaining high growth rates (Rochoa, 2003). We then re-colored the PCA plot according to gene CAI and observed that CAI decreases going from the negative pole ('Proliferation) of PC1 to the positive pole ('Differentiation') of PC1 (**Figure 3C**). The negative correlation between the 'Differentiation' pole and CAI suggests that the pro-differentiation genes avoid using codons that are favored by natural selection for adaptation to high-growth conditions. We then evaluated the contribution of each codon to the variation represented on PC1 by analyzing the PC1 loadings and identified GC-content as a major influence. G/C-ending codon usage grouped mRNAs along the negative pole of PC1, whereas mRNAs at the positive pole clustered based on A/U-ending codon usage (**Figure 3D**). Collectively, this PCA analysis clearly shows that, like mammalian cells, *Drosophila* proliferation mRNAs and differentiation mRNAs are differently codons biased. Still, biased codon usage patterns are shaped by neutral mutation bias and natural selection for translation efficiency (Sharp et al., 2010), the latter of which we will interrogate in the next section.

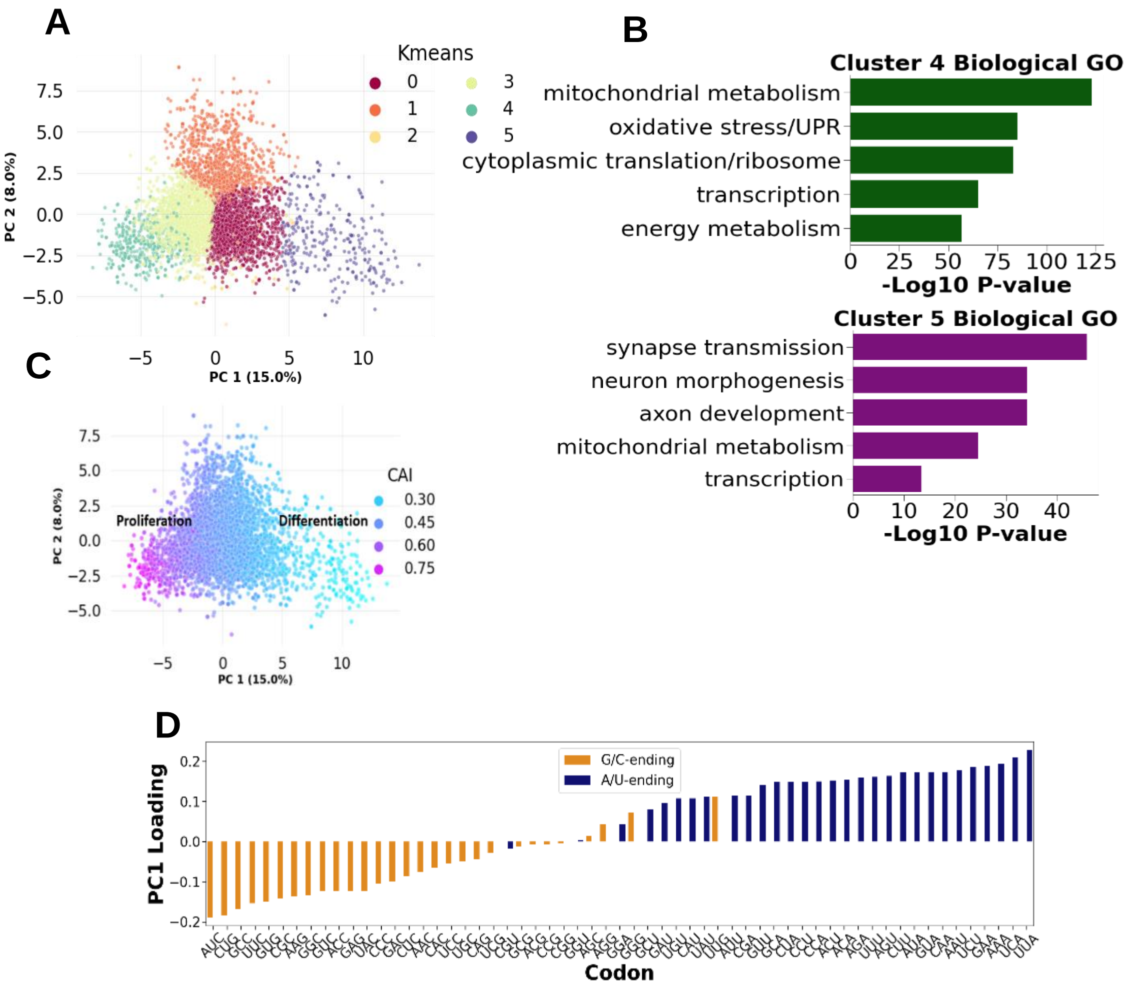


Figure 3: Divergent codon usage patterns between proliferation- and differentiation-related mRNAs expressed across neurogenesis. We identified shared mRNAs (n=5169 transcripts) based on their non-zero expression (TPM>1) in both neuroblasts and neurons.

A) PCA on their normalized codon frequencies (5169 x 59 matrix) captured a total variation of 23%, with the 15% variance captured by the first principal component, PC1. The unsupervised clustering by kmeans grouped the shared mRNAs into 6 clusters.

B) Shared mRNAs in clusters 4 and 5, both projected at opposite poles of PC1, were selected for gene ontology (GO) analysis. Bar plots show the top 20 GO terms for Cluster 4 mRNAs (n=373) and Cluster 5 mRNAs (n=219).

C) PCA plot from (A) recolored by each mRNA's Codon adaptation index (CAI), which measures codon bias of a gene based on its similarity to the synonymous codon usage of a reference set of genes that are constitutively and highly expressed in growth conditions.

D) Barplot showing the PC1 loading values which measures each codon's influence (direction and magnitude) on the variation captured by PC1. Bars for G/C-ending are colored in orange and bars for A/U-ending codons are colored in blue.

Figure 4: Differential tRNA expression establishes neural-specific translation dynamics

Having uncovered distinct codon usage patterns between proliferation and differentiation promoting mRNAs, we next monitored changes in codon-dependent translation efficiency, which we estimate based on the cell type-specific gene tRNA adaptive index (tAI). Experiments, from bacteria to yeast to mammals, have demonstrated how the direct perturbation of tRNA levels led to concomitant global changes in ribosome decoding rates and protein levels, thus providing concrete evidence that the tRNA landscape can be informative about the proteomic changes (Frumkin et al., 2018; Torrent et al., 2018; Goodarzi et al., 2016). Hierarchical clustering of shared neural mRNAs shows changes in tAI values based on the cell type. To gain insights into the physiological relevance of these changes, we highlighted the top 3 shared mRNAs with the greatest altered translation efficiency (tAI) in both directions (**Figure 4B)**. Here, the top 3 mRNAs with the greatest decline in tAI after neurogenesis all encode large ribosomal protein subunits. In contrast, the top 3 mRNAs with the greatest improvement in neuron tAI **—** namely, *caz, eIFH4, and Ars-2* **—** encoded highly conserved RNA-binding proteins. For example, the greatest increase in tAI encodes *caz,* which is essential for motor neuron development and is also the functional ortholog of the human RNA-binding protein fused-in-sarcoma (*FUS*), whose mutations are the primary cause of amyotrophic lateral sclerosis (Lou Gehrig’s disease).

Immediately from the heatmap, we noticed three distinct changes in translation adaptiveness (**Figure 4A**). Firstly, there is a conspicuous subset of mRNAs showing increased tAI by the neuron tRNA pool (blue in neuroblast and red in neuron). Thus we selected these ‘neuron-up’ mRNAs based on the top 10% difference in tAI (neuron-tAI - neuroblast-tAI). These neuron-upregulated mRNAs (n=517) are enriched with biological GO terms related to neural differentiation, such as ‘axonogenesis’ and ‘dendrite morphogenesis’ **(Figure 4C; bottom panel).** We also examined mRNAs that maintain similar high and low tAI across cell types. mRNAs that maintained a high tAI between neuroblasts and neurons (tAI>0.6 in both cell types) are more heterogenous top GO terms, having a combination of constitutive cellular processes (“housekeeping functions”) like ‘protein complex assembly’ and also nervous system related **(Figure 4C; top panel)**. Interestingly, we noticed that the mRNAs (n=411) that maintained a low tAI (neuron-tAI<0.5 and neuroblast-tAI<0.5 and non-overlapping with the previous mRNA sets) were as purely enriched for differentiation and neurogenesis, similar to the ‘neuron up’ group **(Figure 4C; middle panel).**

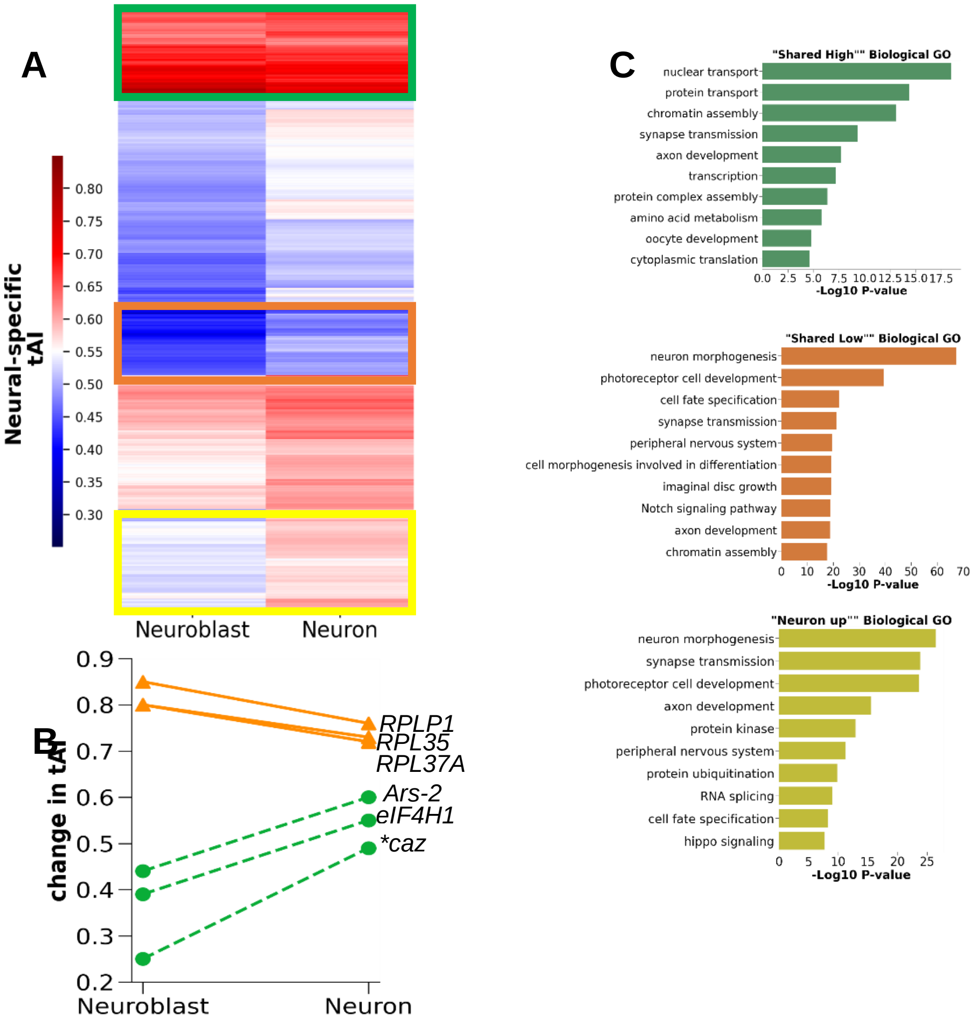


Figure 4: Altered tRNA pool changes the translation fitness of mRNAs across neurogenesis.

A: Hierarchical cluster heatmap comparing the cell-type specific tRNA adaptive index (tAI), a proxy for translation fitness for the cellular tRNA pool, for the mRNAs (n=5169) that are shared across neurons and neuroblasts (TPM>1). Highlighted are the 3 emergent patterns of tAI changes: "neuron-enhanced" mRNAs (top 10% of mRNAs with higher neuron-tAI than neuroblast-tAI; highlighted in yellow box), "shared-high" mRNAs (green box) that maintain a high tAI in both neurons and neuroblasts (neuron-tAI and neuroblast tAI >0.6) , and "shared-low" mRNAs (orange box) that maintained a low tAI in neuroblasts and neurons (neuron-tAI<0.5 and neuroblast-tAI <0.5).

B: The top 3 shared mRNAs with the largest change in translation efficiency (tAI) in both directions

C: Biological GO analysis showing the top 10 significantly enriched terms for each of the 3 clusters of mRNAs defined in (A). Summarily, both the "neuron-enhanced" (n=517 mRNAs; top panel) and the "shared-low" mRNAs (n=411 mRNAs; middle panel) encoded functions related to neural differentiation. In contrast, the GO for "shared-high" mRNAs (n=965 mRNAs; bottom panel) are heterogenous with the top 2 terms related to 'housekeeping functions'.

Distinct influence of natural selection between DNA-binding and RNA-binding pro-differentiation mRNAs by the neuron tRNA pool

In the previous section, the pro-differentiation mRNAs exhibited two distinct responses to changes in tRNA levels in neural proliferation and neural differentiation. For one group, their codon-dependent translation efficiency (tAI) was enhanced by the neuron tRNA, whereas the other group's tAI is refractory and instead maintained a constitutively low tAI between neuroblasts and neurons. This suggests that these mRNAs operate under distinct regulatory codes –i.e., codon dependent vs. codon independent.

We hypothesized that the subset of pro-differentiation transcripts that maintained a low tAI after neural differentiation ('constitutively low') is reflective of weaker translation selection compared to the pro-differentiation transcripts ('neuron-up') that exhibited enhanced tAI post-differentiation. Codon bias is a composite of neutral and adaptive evolutionary forces. A standard way of teasing a part of the influence of natural selection on coding sequences is to compare the tAI to the effective number of codons (ENC), a widely used metric that quantifies the deviation from equal synonymous codon usage (Wright 1990). The ENC values of genes range from 20 (extreme codon bias due to one codon type per amino acid) to 61 (no bias, equal usage of synonymous codons). Basically, the ENC was inspired by population genetics methods to measure the homozygosity of a class of synonymous codons. A significant anti-correlation is indicative that tRNA availability has evolutionarily shaped the variation in codon bias (dos Reis et al., 2002). The 'neuron-up' transcripts show strong and significant for translation selection (Pearson's R= -0.44, P=1.9e^{-26}), but the sign (Pearson's R= -0.02, P=0.6) of translation selection among the 'shared low' subset **(Figure 4D)**. We repeated this comparison using the neuroblast-specific tRNA availability, which attenuated the selection signal (Pearson's R= -0.39, P=1.5e^{-20}) for the 'neuron-up' mRNAs **(Figure 4C)**. This suggests that it is the neuron tRNA repertoire that is selectively driving the translation adaptation of this subset of pro-neurogenic mRNAs.

Would we have been able to detect these distinct evolutionary patterns in the nervous system had we relied on the traditional tRNA gene copy number to estimate tRNA abundance? For the longest time, it was believed that the influence of selection is very weak or even absent on the genome-wide codon usage patterns of animals, a viewpoint largely informed by the lack of correlation between frequently used codons (i.e., codon bias) and tRNA gene frequency in model animal species (Kanaya et al., 2001). When we repeated the same analysis instead using tRNA copy number frequency (static) to compute the gene tAI, the signal for natural selection was lost. Instead, the codon bias of the 'shared low' mRNAs was positively correlated with the gene copy number tAI **(Figure 4F)**. This specific analysis underscores the importance of making inferences based on direct tRNA measurements rather than gene copy numbers, which does not capture the reality of tissue-specific variation in tRNA composition in animals.

Next, we investigated if the difference in regulatory code – i.e. 'codon-dependent vs. no codon-dependent – is reflective of the distinct pathways they are potentially involved in. To this end, we performed Molecular GO Molecular enrichment. Both sets of mRNAs encoding nucleic acid and protein-binding domains indicate that these encode regulatory proteins. However, the 'Neuron up' mRNAs were more enriched from RNA-binding and protein-binding domains **(Figure 4G).** In contrast, the 'shared low' mRNAs were most significantly associated with DNA-binding **(Figure 4H).** These results raise further questions about the purpose of regulating RNA-binding proteins, as opposed to DNA-binding proteins, via tRNA-mediated codon optimality.

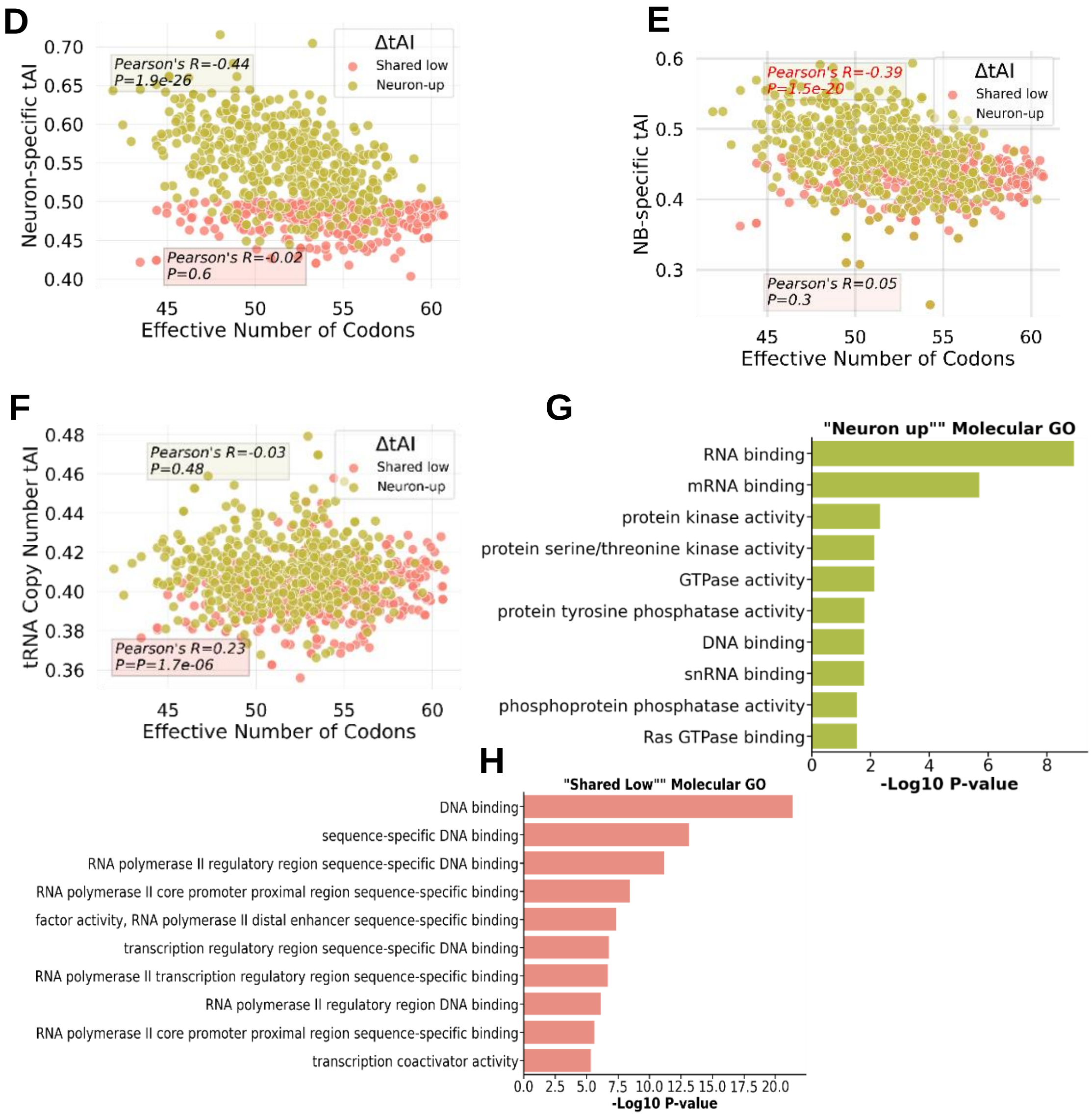


Figure 4: Pro-differentiation mRNAs that encode regulatory proteins evolved separate regulatory codes in the nervous system

D,E: pro-neural differentiation mRNAs exhibit distinct evolutionary signatures. On the x-axis is the effective number of codons, ENC that measures the deviation from equal synonymous usage of genes. Correlation between ENC and tAI indicates the influence of translation selection on codon usage bias. **D)** There is stronger signal for selection by the neuron tRNA pool than **E)** neuroblast tRNA pool. **F)** However, estimating tRNA availability using the traditional tRNA copy number does not produce a signal for selection on the ‘neuron up’ mRNAs

G,H: Molecular GO analyses shows that the mRNAs that exhibit increase their tAI by the neuron tRNA pool (‘neuron up’) tend to encode RNA-binding domains, in contrast to the enrichment of DNA-binding domains among the mRNAs whose tAI remained constitutively low throughout neurogenesis (‘shared low’) .

Interaction between altered tRNA expression and mRNA levels establishes tissue-specific translation dynamics

Lastly, we decided to take a systems-based approach to more realistically model how the interaction between tRNA and mRNA dynamics upon neurogenesis may lead to global changes in translation efficiency using a supply-demand framework. At the systems level, translation efficiency is regulated to ensure that the translation machinery (supply) can precisely fulfill the demands of protein synthesis to meet the cell's requirements. For a more systems-based representation of how anticodon-codon balance may control translation dynamics, we adopted the approach of (Pechmann and Frydman, 2013) to compute the tissue-specific translation efficiency of each shared transcript as the geometric mean of the ratio between codon-tAI (i.e., supply of cognate anticodons) and the codon frequencies weighted by mRNA expression, as this accounts for the fact that total codon demand is a function of mRNA abundance (Methods). Thus, the supply-demand ratio (SDR) integrates information from codon frequencies (static), tRNA abundance, and mRNA abundance (dynamic). A recent study of translation adaptation in normal and tumor human tissues showed that the SDR of a gene better correlated with the protein-to-mRNA ratios compared to tAI only (Hernandez-Alias et al., 2020).

We monitored changes in translation efficiency, as measured by the supply-demand ratio (SDR), after differentiation by computing the difference in SDR (ΔSDR) between neuroblast-specific SDR and neuron-specific SDR. As such, a positive ΔSDR signifies an improvement in translation efficiency post-differentiation. The ΔSDR values ranged from -0.23 (worst adapted to neuron translation) to 0.44 (best adapted to neuron translation) with and median of 0.033. Previously, we showed that proliferation-specific codon usage correlates with the CAI. We wondered how these sequence-intrinsic properties correlated with changes in translation efficiency in post-mitotic neurons. We found a significant negative correlation between post-differentiation ΔSDR and CAI (Pearson's R=-0.34, p-value<2.16 e^{-16}) (**Figure 4I)**. This relationship indicates that after neural differentiation, mRNAs that are enriched in the pro-proliferation codon usage are less optimized for translation by the neuron tRNA pool.

Next, to assess the biological relevance of the change in supply-demand balance upon neural differentiation, we performed GO analysis on the shared mRNAs among the top and top 10% of the ΔSDR values. mRNAs (n=436) with the most improved translation supply-demand post-differentiation yielded GO terms that purely relate to neurogenic development ((**Figure 4J; top panel)**. *cabeza (caz)* shows the greatest improvement in post-differentiation translation efficiency by a factor of 78% and 8.5 standard deviations from the mean. *Caz* is an RNA-binding protein that is essential for motor neuron development. Interestingly, *caper*, the conserved neuron-enriched alternative splicing factor, is among the top transcripts enhanced translation efficiency by the neuron tRNA pool. Conversely, the mRNAs that experience the greatest decline in translation supply-demand balance encoded both growth-related functions, such as ribosome' and 'energy,' and neurogenic functions, such 'axon development.'(**Figure 4J; bottom panel)**. The greatest decline (by a factor of 14%) in neuron translation efficiency is the encodes *lethal (3) 80Fj* (*l(3)80Fj*), the mammalian ortholog GCN1 (**Figure 4I)**. Notably, GCN1 is the activator of the well-characterized GCN2 (general control nonderepressible 2), the master regulator of protein homeostasis in the highly conserved eukaryotic integrated stress response. Although the precise role of GCN1 is less characterized, it was shown that knock-out GCN1 (but not knockout of GCN2) in mouse embryos led to growth retardation, increased mortality, and cell cycle arrest, suggesting that GCN1 is essential for normal cell growth independent of its role in the GCN2-mediated stress response (Yamazaki et al., 2020). While this set of analysis appears redundant with the former ΔtAI analysis (**Figure 4C)**, it is nonetheless noteworthy that ΔSDR values (which incorporates mRNA expression) recapitulated the GO results of the ΔtAI (only tRNA levels) values because it demonstrates that it is the regulation of tRNA levels that dominates the dynamic translation programs in development, without necessarily requiring major shifts in mRNA abundance. This may partially explain whys mRNA steady-state levels are only moderately correlated with protein abundance (Buccitelli and Selbach, 2020). Altogether, these results suggest that tRNA dynamics alter the translation program in a manner that supports the tissue type-specific phenotypes in neural proliferation and differentiation.

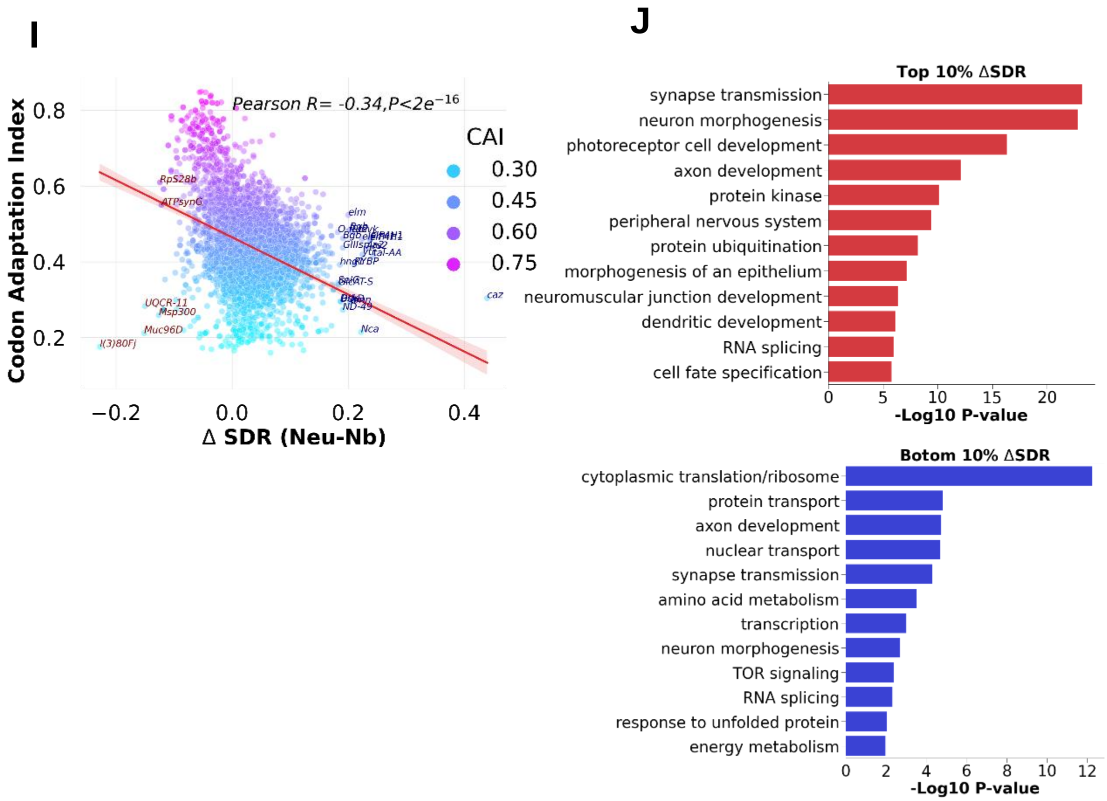


Figure 4: Altered tRNA pool changes the translation fitness of mRNAs across neurogenesis

I) Scatterplot highlights the negative relationship (Pearson's R=-0.34, *P<2.16e*$^{-16}$*)* between the post-differentiation change in translation supply-to-demand ratio (ΔSDR) of mRNAs, a proxy for translation efficiency described in the article, and the codon adaptation index (CAI), which was previously shown to correlate with proliferation-specific codon usage signatures

J) Biological GO enrichment of the mRNAs in top 10% most improved in translation supply-demand balance, ΔSDR, in the neurons **(top panel)** and mRNAs in the bottom 10% of post-differentiation ΔSDR **(bottom panel).**

Figure 5. Global codon usage patterns correlate with tRNA not just in time but also in space

As a final piece of analysis, we investigate positional-dependent effects on genome-wide spatial tRNA availability and CSCs to highlight more complex patterns of codon optimality. Across many species, including *Drosophila melanogaster,* computational analyses of genome-wide codon usage and tRNA gene copy number identified an evolutionarily conserved pattern, wherein the first 10-15 codons at the 5'terminus, on average, have lower tAI values (Tuller et al., 2010). Ribosomal occupancy patterns in *S.cerevisiae* later validated the accumulation of slow codons at the mRNA 5'terminus (Tuller et al., 2010a). Since codon ramps were initially inferred based on tRNA gene copy frequency, we wanted to explore if our experimentally determined tRNA abundance also generated this 'ramp' pattern. For the shared mRNAs in neuroblasts and neurons, we observed a ramp pattern due to low CSCs and tAI values in the first 10 codons **(Figure 5A;5B)**; however, the 3' end used more optimal codons **(Figure 5C; 5D)**. This spatial correlation further supports that tRNA availability constrains codon stabilities since our previous results show that codons with low tAI are also likely to be less stable. A similar positional bias was observed at first and last 50 codons of the open reading frames (ORFs) of maternal mRNAs in zebrafish embryogenesis, in which the 5' terminus codons were enriched with rare codons (low CAI) and the 3' codons were more optimal (higher CAI). Moreover, the synonymous substitution of 3' termini codons with non-optimal codons (lower CAI) of an EGFP reporter construct led to an increase in degradation (Mishma and Tomari, 2016).

Interaction between translation efficiency and local mRNA secondary structure at the 5' and 3' termini

The precise purpose of the selection on position-dependent codon bias remains elusive but appears to be context-dependent. Regarding selection for translation efficiency, two main reasons are proposed: 1) codon ramps promote efficient ribosome allocation (fewer ribosomal collisions) and fidelity of translation initiation, i.e., sufficient time for the pre-initiation translation complex (PIC) to find the start codon, thereby reducing the cost of mistranslation errors (Frumkin et al., 2017)   2) ramps, generally slow codons, influence co-translation protein folding (Tuller et al., 2010b; Pechmann and Frydman, 2013). In *S.cerevisiae*, the assay of over 30,000 variants of the first 10 codons (excluding the AUG start codon) of a green fluorescent protein (GFP) reporter showed that the codon ramps modulate protein yield(Osterman et al., 2020). Alternatively, the 5' and 3' codons may reflect selection against a stable local secondary structure (Goodman and Church, 2013). Local mRNA structure is predicted to be antagonistic to translation efficiency (Tuller et al., 2010b).

To further investigate the relationship between these 5' and 3' codon biases and local secondary structure in the neural mRNAs, we computed the minimum free energy, *ΔG,* of the first 93bps (i.e., 30 codons) of each of the shared mRNAs using the *RNAfold* package with default settings (Lorenz et al., 2011). A more negative *ΔG* represents a more stable predicted RNA structure. Then we examined the association between cell type-specific translation efficiency (SDR) according to the SDR deciles and the 5'/ 3' termini *ΔG*. As predicted, we found a significant negative association between the mRNA's translation efficiency and the stability of the secondary structure, with a stronger effect at the 3' terminus for both the neuron-specific and neuroblast-specific SDR **(Figure 5E-H)**. In other words, the best translationally adapted mRNAs tend to have a weaker 3' terminal secondary structure. We further found that, on average, the local structure a the 3' terminus is less stable than at the 5' terminus (Kolmogorov Smirnov test, $P=9.2^{-16}$) (**Figure 5I)**. Using reporter variants in *E.coli*, Frumkin et al. identified that the utilization of a strong secondary structure at the 5′ ends of the ORF could reduce fitness cost in protein synthesis (Frumkin et al., 2017). Taken together, this suggests that local mRNA secondary may also be influencing codon usage bias and, therefore, mRNA decay and translation dynamics in the *Drosophila* nervous system.

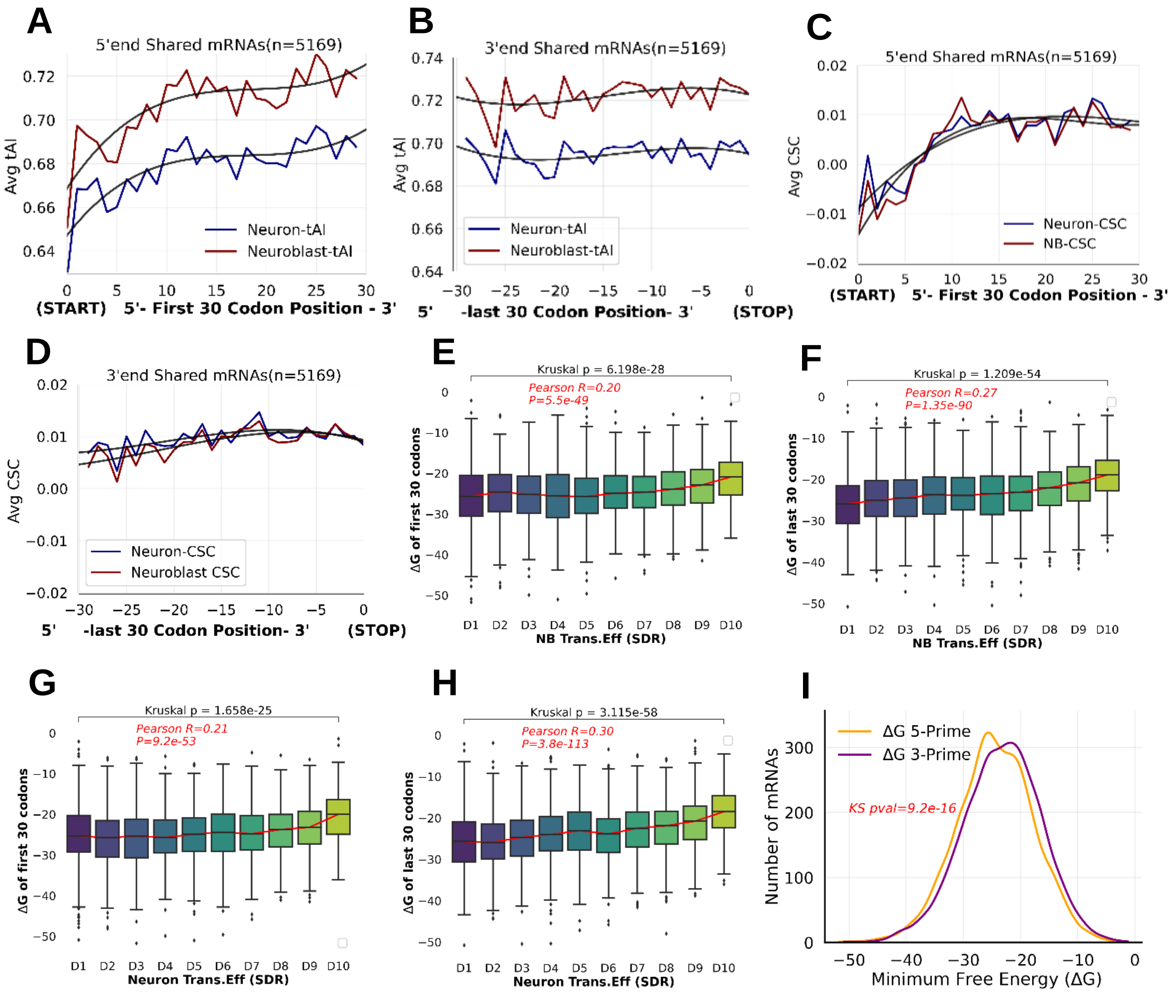


Figure 5: Position-dependent codon optimality and local mRNA secondary structure in neural shared mRNAs

A-D: For the mRNA shared between neuroblasts and neurons, the first 10 codons at the 5’terminus have, on average, lower tAI and CSC values compared to the codons at the 3’terminus.

E-H: Local mRNA secondary structural stability, *ΔG*, at the first and last 93nt in each mRNA was calculated using the RNAfold package with default settings. The more negative *ΔG*, the more stable the local structure. Overall, the mRNA’s translation efficiency negatively correlates with local structuredness, with stronger effect at the 3’ end of the mRNAs. **I)** On average, the 3’ local secondary structure of mRNAs is less stable than the 5’ secondary structure.

Discussion:

For the first time, we monitored the *in vivo* regulation of transfer RNAs (tRNA) and codon optimality dynamics in the most complex yet well-studied animal organ system, which is also well-conserved at the molecular and genetic levels. Here, we integrated data from genome-wide measurements of tRNA and mRNA pools to elucidate the regulatory role of codon identities on the post-transcriptional programs in neural progenitors (neuroblasts) and post-mitotic neuron representatives of proliferation and differentiated states. Codon optimality describes how codon identities constrain the fate of mRNAs in distinct ways. We identified distinct codon usage between stable and unstable transcripts in each cell type in support of codon optimality-mediated mRNA decay. We also observed several codons that changed their stabilizing properties between neuroblasts and neurons, which is indicative that codon optimality is dynamically regulated in development. In both neural cell types, stable mRNAs are enriched with cell-type specific optimal codons, and the mRNA function tends to align with the phenotype of the cell, consistent with observations in the *Drosophila* embryonic tissue (Burow et al., 2018). A major goal of this study was to address the basis of cell-type specific codon optimality. Although it is speculated that differential tRNA abundance establishes codon optimality, most animal studies lack cellular tRNA measurements. For the first time in an animal nervous system, we used high-throughput sequencing to map changes in the tRNA transcriptome and epitranscriptome *in vivo* between the proliferative and differentiated stages of neurogenesis. Our bioinformatics analyses reveal that the stabilizing effect of codons positively correlates with the cellular tRNA levels in neuroblasts and neurons, but the effect at the gene level attenuates in neurons. To investigate how codon optimality and tRNAs contribute to the dynamic translation programs in neurogenesis, we analyzed the codon usage patterns of mRNAs that are expressed in both neuroblasts. Like human cells (Gingold et al., 2014; Hernandez-Alias, 2020), *Drosophila* mRNAs involved in neural proliferation and differentiation exhibit distinct codon usage profiles, with strong GC-bias in the proliferation-oriented transcripts. Importantly, the codon usage of proliferation-oriented mRNAs makes them better adapted for translation by the neuroblast tRNA pool but less efficiently translated by the neuron tRNA pool upon differentiation. We show that although several pro-differentiation mRNAs are expressed in the neuroblasts, they have been distinctly shaped by natural selection for enhanced translation by the neuron tRNA pool. Altogether, our findings position codon optimality, as established by tRNA variation, as an influential genetic determinant of *Drosophila* neurogenesis as it establishes a mechanism that would enable coordinated regulation of functionally related mRNAs via the sharing of distinct codon profiles that, in turn, similarly constrains their dynamics upon cellular changes that are oriented toward the phenotype of the cell. Both early neurogenesis and translation elongation are well conserved between animals. Therefore our findings have implications for other metazoan systems.

Key Finding #1: *tRNA-mediated changes in codon optimality supports proliferation and differentiation in the nervous system in distinct ways*

A major goal of this study was to determine how context-dependent codon optimality is established during cell development. Codon-optimality mediated mRNA decay (COMD) which is a more recent mechanism of codon optimality. By superimposing data from independently obtained genome-wide measurements based on tRNA sequencing, mRNA sequencing, and mRNA decay in genetically identically neural populations, our findings demonstrate a global link between codon usage and tRNA availability, both on a frequential and spatial basis. Moreover, the variation in cellular tRNA availability (tAI) provided strong explanatory power for the differences in CSCs, as evidenced by the large effect size (Cohen's d>1.0).

The general mechanism that emerged from the study is that while the NB transcribes both pro-proliferation and differentiation mRNAs, the phenotypic impact of pro-differentiated mRNAs is likely suppressed/attenuated, in part because the codon usage of differentiation mRNAs makes them poorly adapted for translation by NB tRNA pool. However, after neurogenesis, the codon usage of a subset of pro-differentiation mRNAs, many of which are RNA binding proteins, are a better match to decoding by the neuron tRNA pool, thus experiencing a boost in their translation efficiency that may lead to proteostatic changes to establish developmental reprogramming. Nonetheless, codon optimality in these neural populations manifests in distinct and dynamic ways across development, in part due to the cell type-specific tRNA repertoires and physiology. Our results parallel a recent study in the mouse immune system that profiled changes in the tRNA and mRNA expression during antigenic activation of mouse T cells and demonstrated coordinated changes that promoted the up-regulation of T-cell proliferation mRNAs, after which the tRNA levels relaxed upon differentiation back to the basal level (Rak et al., 2021).

1.1 tRNA expression in neural progenitors supports proliferation by upregulating ribosome biogenesis and global translation

A small pool of nearly homogenous populations of neural progenitors must repeatedly divide to generate orders of magnitude more neuron and glial cells. The significant and strong association between tRNA levels and codon usage in explaining both codon-optimality mediated mRNA decay (COMD) and the variation in codon-dependent translation adaptiveness of mRNAs in *Drosophila* larval neuroblasts (NB) aligns with observations in proliferative systems, from bacteria (Rochoa et al.,), to fungi (Presynak et 2015; Tuller et al., 2010; Wint et al., 2022) to mammalian stem cells (Bornelov et 2019; Forrest et al., 2020) and the *Drosophila* embryos (Burow et al., 2018). Collectively, these data support the widely accepted 'stabilization-by-translation' model where variation in tRNA expression confers distinct elongation rates to their cognate codons such that optimal codons are more rapidly translated and thus less likely to recruit decay and recycling factors that are known to target stalled or paused mRNA-bound ribosomes (Radakrishnan et al., 2016; Buschauer et al., 2020).

The commonality between these studies and our present work, based on the GO functional analysis, is that tRNA-mediated codon optimality supports cell proliferation by upregulating ribosome biogenesis and cytoplasmic translation, which is indicative of the cellular need to sustain the accumulation of biomass and repeated rounds of cell division. It is well understood that the availability of free ribosomes, i.e., capacity for protein synthesis, is the primary limiting factor in actively dividing cells. On average, 20% of gene expression (up to 60% at the log phase in budding yeast) is invested in making ribosomal proteins and other core components of the translation machinery, thus enabling the synthesis of other proteins (Liebermeister et al., 2014; Warner et al., 2000). Thus, codon optimality represents a deeply conserved regulatory code for establishing gene expression programs under proliferative conditions.

1.2 tRNA-mediated codon optimality supports neuronal maturation by selectively upregulating the translation efficiency of regulatory proteins involved in neurogenic pathways such as RNA splicing

Here, we discuss the most interesting and novel set of results that emerged from the study in the neuron population. But first, the field has long appreciated the influence of codon optimality on post-transcriptional mRNA fate in proliferative conditions, especially in simpler organisms. Our contribution on this front was to present concrete evidence that it is the dynamic variation in cellular tRNA levels that both installs and transduces the distinct regulatory signals of the codon features in the coding region of mRNAs, especially in complex animals where codon optimality was long believed to be nearly absent. Thus, we do not find it surprising that a strong signal for tRNA-mediated codon optimality exists in the *Drosophila* larval neuroblasts because, broadly speaking, cell proliferation is oriented towards a rather general objective: produce lots of proteins to grow and exponentially increase the number of cells.

1.3 Codon-mediated mRNA decay is weaker in the post-mitotic neurons compared to neural progenitors.

If codon optimality is an adaptation of high-growth states (Gingold et al., 2014), then we expect codon optimality to minimally contribute to the genetic control of the development of differentiated tissues. Indeed, the previous work from our lab supported this view when it was observed that the effect of codon optimality on mRNA decay attenuates in the *Drosophila* embryonic nervous system, although their qPCR method – unlike this tRNA sequencing - failed to identify significant changes in the tRNAs between the whole embryos and the embryonic CNS. Still, we observed a similar but more complicated pattern in our present profiling of the larval CNS. Intriguingly, the variation in stabilizing effects of codons is nearly as strongly by neuron-specific tRNA availability (Cohen’s d=1.0) as observed in the neuroblasts (Cohen's d=1.44). Initially, this result looked encouraging. Yet, unlike in neuroblasts (Pearson's R=0.38****), the influence of codon optimality on mRNA stability unexpectedly did not persist to the gene level in neurons (Pearson's R=0.1**). So again, like *Burow et al., 2018*, we return to the question that motivated this research: why does the regulation of mRNA decay by codon optimality attenuate after neural differentiation? Our analysis of codon-dependent translation dynamics may have given us the answer.

1.4 Neuron tRNA pool promotes differentiation by enhancing the translation of RNA-binding proteins that promote miRNAs which potentially override codon-optimality-mediated mRNA stability

Although eukaryotic translation and mRNA decay are coupled since they share the same interaction on the mRNA, they are nonetheless still distinct pathways. We decided to explore not only the role of tRNA regulation in shaping specific mRNA decay programs but also the potential role of tRNA dynamic on differential mRNA translation. We chose to focus on shared mRNA transcripts, thereby holding constant the coding region (and thus codon usage) in order to observe how their translation dynamics would be altered by tRNA levels.

Differentiation is the inverse of proliferation, and the evolution of functional and morphological specialization necessitates distinct regulatory pathways. One-way animals achieve tissue-specific programs is via the regulation of RNA-binding proteins. Our data show that mRNAs encoding RNA-binding and protein-binding domains, i.e., post-transcriptional regulatory proteins – dominate the cohort of transcripts that likely experience enhanced translation by the neuron tRNA pool by virtue of their codon usage profiles. Regulatory networks mediated by RNA-binding proteins play a crucial role in nervous system development and maintenance, especially by generating functional diversity via alternative RNA splicing. Additionally, both the insect and mammalian nervous systems selectively express longer 3' UTR isoforms, compared to the rest of the body, via alternative polyadenylation (APA) (Blair et al., 2017; Oktaba et al., 2017). Many studies show that 3'UTRs contain *cis*-elements that are targets for RNA binding proteins (RBP) and microRNAs (miRNAs) that regulate mRNA degradation (Zaid, 1994; Dini Modigliani et al., 2014; Pereira et al., 2017) and neuronal subcellular mRNA trafficking and localization (Meer et al., 2012; Tushev et al., 2018; Bauer et al., 2019).

One mechanism of post-transcriptional regulation is 3'UTR-mediated mRNA destabilization via microRNAs (miRNA) that bind sites in the 3′ UTRs. So longer 3UTR isoforms are expected to have more miRNA binding sites (Bae et al., 2020). In our dataset, among the top 5 largest improvements in post-differentiation translation efficiency (tAI) is the *Ars-2* transcript (Arsenic resistance protein 2), a highly conserved gene that directs miRNA biosynthesis and maturation and directly binds to the CBP80 and Drosha complexes (Gruber et al., 2009; Gruber et al., 2011). Recently, in zebrafish embryogenesis, it was shown that miRNA antagonizes the stabilizing effect of optimal codons in quasidosage dependent manner. In this study, the researchers constructed reporter transcripts with varying degrees of codon optimality and found that miR-430/-427 destabilizing activity was strongest on transcripts with moderate codon optimality but less efficacious on transcripts with more extreme bias in codon optimality. Even more relevant to our findings, they also observed that, across a hundred maternal mRNAs, the enrichment of miRNA sites increases, and the stabilizing effect of codon optimality decreases (Medina-Muñoz et al., 2021). To put this all together, our results suggest that neuronal tRNA expression supports the enhanced translation of RBP, potentially increasing their protein output and thus their regulatory activities, one of which is the upregulation of 3'UTR-mediated mRNA stability that overrides the influence of codon optimality. Our findings raise the question of tRNA regulation in other differentiated tissues employs a similar mechanism of selective translation regulation of regulatory proteins, which in turn drives a developmental signal cascade leading to phenotypic changes. It would be interesting to see if this mode of tRNA-mediated regulation exists in the mammalian brain; however, to date, there are only two published *in vivo* tRNAseq from the mammalian CNS, both in mouse models (Blaze et al., 2021; Pinkard et al., 2020)

From the perspective of translation adaptation, we imagine that neuronal tRNA expression must balance the up-translation of regulatory genes with the down-translation of ribosomal genes, but not beyond a certain threshold since post-mitotic neurons still need to maintain constitutive protein synthesis and thus ribosomal biogenesis. This may explain why the post-differentiation increase in tAI of the top 3 mRNAs (all RBPs) is larger than the post-differentiation decrease in tAI of the top 3 mRNAs (all ribosomal proteins) **(Figure 4B)**.

What is the benefit of regulating cell development post-transcriptionally, particularly at the layer of mRNA translation? Because protein synthesis is the direct output of translation, the potential advantage of regulating mRNA fate at the stage of translation provides a faster way of modulating protein levels compared to more upstream control at the level of transcription. Borrowing from principles of computer information retrieval, translation regulation vs. transcriptional regulation is analogous to how access from RAM storage (translation) enables faster CPU computations (conversion of mRNA to protein) compared to loading instructions directly from disk storage (transcription). Nevertheless, the coding regions of mRNAs are under different functional constraints as well as different evolutionary pressures, and so only a subset of regulatory genes can benefit from improved translation efficiency due to changes in tRNA expression after neurogenesis. We showed that the mRNAs enriched in DNA-binding domains (DBPs) were refractory to codon-dependent translation efficiency by the neuron tRNA pool and also lacked a signal for translation selection by the cytosolic tRNAs. Thus, we wonder if there is any physiological purpose/advantage for the differential regulation of RBPs and DBPs in nervous system development.

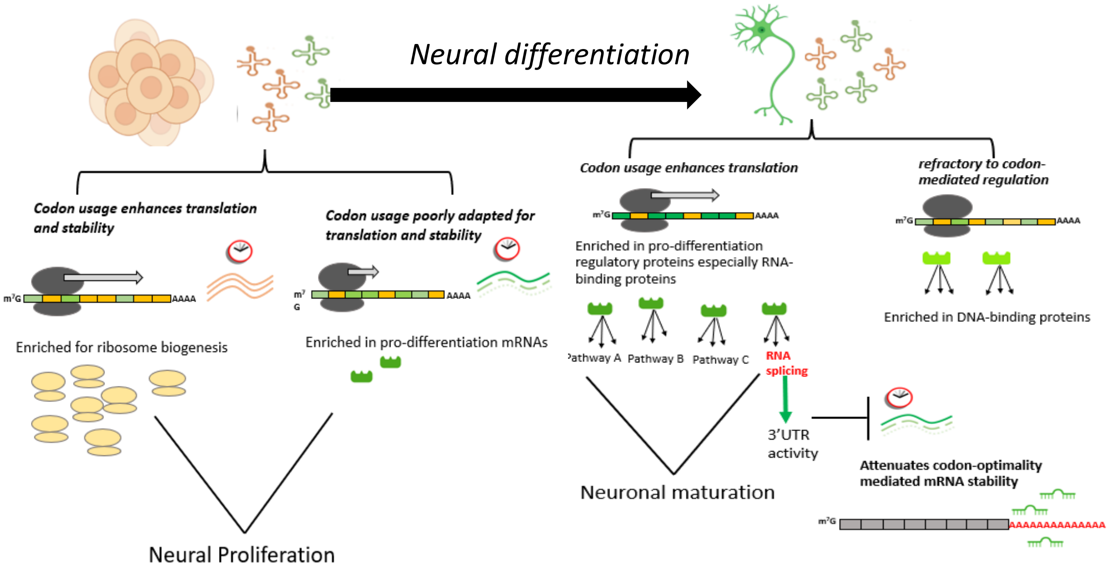

Neural differentiation
Codon usage enhances translation and stability
m7G
AAAA
Codon usage poorly adapted for translation and stability
m7G
AAAA
Enriched for ribosome biogenesis
Enriched in pro-differentiation mRNAs
Neural Proliferation
Codon usage enhances translation
m7G
AAAA
Enriched in pro-differentiation regulatory proteins especially RNA-binding proteins
Pathway A
Pathway B
Pathway C
RNA splicing
3'UTR activity
Neuronal maturation
refractory to codon-mediated regulation
m7G
AAAA
Enriched in DNA-binding proteins
Attenuates codon-optimality mediated mRNA stability
m7G
AAAAAAAAAAAAAA

Figure 6: Model of tRNA dynamics contributes to proliferation and differentiation programs in *Drosophila* larval neurogenesis:

General Model: tRNA regulation establishes cell-type specific codon optimality that facilitates coordinated regulation of functionally related mRNAs via the sharing of distinct codon profiles that similarly constrains their dynamics upon changes in tRNA levels in a way that is oriented toward the phenotype of the cell.

Neuroblast-specific tRNA repertoire establishes Neuroblast-specific codon optimality that regulates mRNA stability and translation efficiency of primarily ribosomal and energy metabolism mRNAs.

Neuron-specific tRN**A** repertoire supports the translation efficiency of pro-neurogenic mRNAs, including regulatory RNA-binding proteins that upregulate alternative RNA splicing as well as increased 3'UTR-mediated *cis-trans* interactions, which we speculate is one of the factors that partially explain the attenuation of codon optimality mediated decay in neurons.

Key Finding 2: Distinct variation in tRNA expression at different functional layers

2.1 Gene-specific regulation of tRNA isodecoders remains unsolved

Similar to tRNA sequencing in mammalian tissues (Pinkard et al., 2020), we observe that there is greater variation at the tRNA isodecoder level (greater range in fold change) compared to the anticodon level (smaller fold change range). This suggests that individual tRNA isodecoders are differently regulated. Still, the mechanisms by which identical tRNAs are distinctly regulated remain an open question because all tRNA loci share the same transcription machinery (RPOL3-TFIIIB-TFIIIC) and conserved internal promoters, known as A/B-box sequences. General tRNA biogenesis is majorly regulated via mTORC pathway, the conserved nutrient-sensing, and growth pathway in eukaryotic cells, by modulating the phosphorylation status of MAF1, the negative regulator of RPLO3 (Arimbasseri et 2016)

Interestingly, 45 % of *Drosophila melanogaster* tRNA loci are nested within the introns of longer protein-coding genes. A similar fraction is observed in mice and humans (Sagi et al., 2016). So, it is possible that transcriptional interference between RNA POL2 and RNA POL3 may also regulate the expression of specific ‘nested’ tRNA loci. Chromatin profiling experiments demonstrated how the presence of nested mammalian interspersed repeat (MIR) loci, also RPOL3 transcribed, led to the downregulation of its host protein-coding gene due to transcriptional interference (Yeganeh et al., 2017). So it is also possible that nested tRNAs are regulating their host protein-coding genes. One way of investigating this mechanism of tRNA loci regulation is via CRISPR/Cas9-mediated deletion of the nested tRNA loci.

2.2 Dynamic changes in the tRNA epitranscriptome

Variation in the tRNA epitranscriptome provides another layer of information about tRNA regulation. Dysregulation of tRNA modifying enzymes is increasingly identified as a driver of human diseases, most observed in neuro-pathologies and cancers, such as mutations in *PUS3, ADAT3, WDR4, NSUN2*, and *FTSJ1* enzymes (Blaze et al., 2021; Blanco et al., 2020; Delaunay et al., 2022; reviewed in Suzuki, 2021). In addition to discovering changes in tRNA levels, we inspected the modification profiles of the tRNA isodecoder and were able to identify modification signatures at conserved positions in the tRNAs. Many tRNA modifications are involved in housekeeping functions, such as stabilizing the secondary and tertiary tRNA structure. Hence, as expected, most modification frequencies did not change between the neural cell types. Still, we detected a total of 3% of sequence variants that exhibited at least a 10% difference between conditions.

2.3 Is the neuroblast tRNA epitranscriptome adapted for translation speed at the cost of translation fidelity?

What can our modification results reveal about cell-type specific tRNA adaptation? We speculate, based on two lines of evidence from our data, that the neuroblast (NB) tRNA epitranscriptome indicates a specific adaptation for translation speed. Under conditions of high growth, a proliferative cell may have to make a trade-off between translation accuracy and speed (Hausser et al., 2019; Wohlgemuth, 2011). Firstly, our data shows that Inosine-34 (I34) edited tRNAs significantly contribute to neuroblast-specific codon optimality but not in neurons. Especially, codons decoded by I34-tRNAs are enriched in the 'Proliferative' codon usage signature that we identified, and I34-tRNAs preferentially decode stable codons in the NB. The I: C wobble anticodon-codon pairing is weaker than the G: C bond due to one less hydrogen bond but still appears to be as robustly decoded (Hoernes et al., 2018). However, a weaker I: C bond would facilitate easier dissociation of the anticodon-codon pair at the ribosomal E-site, thus promoting faster translocation of the ribosome along the mRNA sequence. Secondly, we observed that the conserved G27 methylguanosine position on *Ile-AAT-1-6* is strongly hypomodified in NB compared to neurons. This is rather interesting because Ile-AAT also bears the inosine-34 modification, and its cognate codon, AUC, is shown to be the most influential feature on the 'Proliferation' axis of the PCA performed on the shared neural mRNAs (**Figure 3C).** Hypomodifications on the anticodon stem are predicted to destabilize the anticodon-codon stacking and lead to ribosome frameshifting errors which may result in premature translation termination (Mordret et al., 2019). However, a less rigid anticodon-codon pairing may potentially benefit faster ribosome translocation, albeit at the cost of more mistranslation errors. Interestingly, 'response to misfolded protein' and 'cellular response to oxidative stress' were among the top 20 Biological GO terms for the mRNAs selectively using the 'Proliferation' codon usage program, which exhibits better translation adaptation in NB. Taken together, perhaps the NB has mechanisms in place to ameliorate the proteotoxic stress arising from increased translation errors whilst benefiting from faster translation and, thus, increased availability of free ribosomes. In contrast, it is widely known that neurons are more sensitive to the accumulation of misfolded proteins because, unlike NB, neurons cannot dilute toxic metabolites by means of cell division.

3. Study Limitations

Several factors may explain why there is an incomplete relationship between our tRNA abundance and CSCs. The most straightforward reason is the lack of charged tRNA quantification since it is the concentration of aminoacylated tRNAs that ultimately dictates the codon decoding rates. We also lack information about the tRNA decay rates, which also contribute to tRNA steady-state levels. Secondly, from the standpoint of the genome-wide profiling of mRNA decay rates, it is likely that the chase time points (3hr, 6hr, 9hr, 12 hr) would not be able to detect a strong signal if any, of mRNAs with shorter half-lives. Forrest et al. also observed that amino acid usage also contributes to mRNA stability. However, in our analyses, we normalized away amino acid effects so as not to confound our results on codon optimality. Therefore, further exploration of the role of amino acid usage and free amino acid levels may provide a more complete picture. However, it must be noted that codon optimality is not predicated perfect correlation between tRNAs and codon usage because codon usage patterns are shaped by both natural selection and neutral forces, albeit at varying degrees of influence (dos Reis et al., 2004)

Inferring tRNA modification status based on RT-induced mismatch during sequence offers limited detection and is likely biased towards simple methylated nucleosides since they provide a less steric hindrance to readthrough by the RT. As such, this method is less likely to detect bulkier adducts as they would lead to increase RT-arrest and falls of. Moreover, the partial alkaline treatment specific to this method of tRNA library preparation may remove or alter the detection of some modifications. The better alternative for measuring tRNA modifications is mass spectrometry.

One major assumption we made is that the enhanced adaptation to the cellular tRNA pool maps proportionally to the change in protein abundance. While we do expect some agreement between changes in decoding rates by tRNAs and protein output (Goodarzi et al., 2016), there remains still additional steps involving the post-translation maturation of proteins. Genome-wide proteomic measurements would clarify the precise relationship between tRNA-mediated translation adaptation and upregulation of protein synthesis.

Concluding Remarks and Future Directions

My dissertation research contributes to the emerging field of how codon usage bias intersects with tRNA dynamics, i.e., codon optimality, to exert post-transcriptional control cell-fate determination. Specifically, this work positions tRNA regulation, and thus regulation of translation elongation, as a crucial checkpoint in cell development. Based on our data, we propose the model that neuronal tRNAs contribute to the regulation of cell differentiation by enhancing the translation, in a codon-dependent manner, of RNA-binding proteins that are crucial regulators of neurogenic development, such as alternative RNA splicing and axonogenesis. Moreover, several of the fly RNA-binding proteins are also functional orthologs in the human nervous system, including those implicated in neurological and neurodevelopmental disorders. This raises the possibility that the selective translation of regulatory RNA-binding proteins may explain the pervasiveness of defective tRNA metabolism in many neurodegenerative diseases.

The methods and analysis framework in service of this research question aptly represent the *zeitgeist* of "-omics" biological research. Advances in high-throughput molecular assays mean that for a given experiment, we are likely to measure many more data points/ observations than the research questions we have. Additionally, the exponential growth in computing power and storage enables the curation and access to petabytes of experimental data, thus creating opportunities to drive *in silico* research and yield novel insights beyond the purpose for which they were initially generated. In this study, we greatly benefit from the gene ontology information that has been curated from thousands and, in the case of *D.melanogaster,* decades of high-throughput functional studies. The richness and diversity in experimentally generated data also underscore the benefit of using model organisms, such as *Drosophila melanogaster*, to study basic biology.

Proposed future directions to the 'Neural tRNA Dynamics' story:

1) Elucidate the mechanism by which tissue-specific tRNA expression arises by assaying changes in tRNA transcription via RNA POL3 profiling and post-transcriptional dynamics, e.g., tRNA turnover. Given that tRNA post-transcriptional modifications are dynamically regulated, as my work alludes to, then a complete modification profiling by mass-spectrometry may yield insights into gene-specific or isodecoder-family-specific regulation of cytosolic tRNAs.
2) Experimentally validate causality between altered tRNA levels and mRNA decay through perturbation of specific tRNAs – such as the post-transcriptional knock-down of tRNAs via RNA-interference - and monitor concomitant changes in mRNA decay and translation dynamics. Also, examine changes in neural morphology via fluorescent imaging since my GO analyses of the codon-optimality responsive mRNAs showed strong enrichment for axon and synaptic growth.

3) Test the proposed model that tRNAs in neural differentiation upregulates 3'UTR and, by extension, miRNA activities that lead to overriding codon optimality mediated mRNA decay. One method is to manipulate the codon optimality of the predicted codon-dependent responsive RBPs. As an example, *Ars*-2, the crucial regulator of miRNA biogenesis, is predicted by our dataset to be one of the most responsive to codon-dependent translation regulation. To this end, one could leverage *Drosophila* genetics to 'knock in' (Bosch et al., 2020) synonymously recoded variants of *Ars-2*, with varying degrees of codon optimality, fused with a reporter tag, e.g., GFP. And then monitor the protein signal of the recoded *Ars-2* variants and assay genome-wide changes in miRNAs (Sabin et al., 2009; Hafner et al., 2011) and mRNA decay rates.

Materials and Methods:

1. Experimental preparation

Sample collection of Drosophila melanogaster larval brains

Flies were raised at 23-25^{o}C. Whole brains were dissected from late-stage larvae (120 hours after larval hatching) of wild-type flies (Oregon-R-P2), and transgenic ectopic-neuroblast mutants (UAS-aPKC^{caax}x Insc-Gal4) flies. A total of 3 biological replicates were obtained for each genotype.

tRNA sequencing (adapted from Gogakos et al., 2017 with minor changes)

All reagents should be RNAse-free. Isolate total RNA using standard Trizol extraction. Perform RNA quality control using 1% Agarose gel and nanodrop. Use the appropriate combination of synthetic RNA oligonucleotides, at lengths 19nt, 24nt, 35nt, 45nt, 61nt, 70nt, and 85nt, as size markers to guide size selection after each gel run. All adapters, primers, and size markers were purchased at Integrated DNA Technology (IDT).

Resolve 20ug of total RNA on a 15% denaturing Urea-SDS PAGE (Invitrogen) and excise gel that spans 60-100nt. Let gel rotate overnight in 350ul of 0.3M sodium acetate. Elute the RNA in 100% isopropanol and 1ul LPA on ice for 1 hour. Microcentrifuge at 20,000g to collect RNA. Perform 2 rounds of ethanol wash using 500ul 80% ethanol at 7500g for 5 minutes. Resuspend RNA in 12ul of RNAse-free water and subject it to a partial alkaline hydrolysis using 15uL-buffer of 100mM Na_2CO_3 and 100mM $NAHCO_3$ at 90^{o}C for 12 minutes (be as exact as possible). Resolve the hydrolyzed RNA on a 15% TBE Urea-denaturing SDS-PAGE gel (Invitrogen) and recover bands between 19-45nt. Dephosphorylate the hydrolyzed RNA using 10U calf intestinal phosphatase (New England Biolabs). Heat inactivate at 75^{o}C for 15 minutes. Perform 3' ligation by using 2ul of 10mM barcoded 3'adapter (26nt), 2ul of 10x truncated RNA Ligase2 (NEB), 6ul of 50% DMSO and 25% v/v PEG. Incubate the reaction at 16C for 6 hours and then overnight on the ice at 4^{o}C. Heat inactivate at 75^{o}C for 15 minutes. Rephosphorylate using 4ul of 10x T4 Polynucleotide Kinase (NEB), 6ul of 5mM ATP, and 2ul of 1mM DTT at 37^{o}C for an hour. Note the high concentration of ATP also inhibits the RnL2. Heat inactivate at 75^{o}C for 15 minutes. Perform ethanol-salt precipitation by adding 200ul of 100% ethanol and 6ul of 3M sodium acetate, and 1ul LPA. Store on the ice at -20^{o}C for at least 2 hours. Microcentrifuge at 20,000g to collect RNA. Perform 2 rounds of ethanol wash using 500ul 80% ethanol at 7500g for 5 minutes. Elute and Resuspend in 9ul of RNAse-free water. Perform 5'ligation by adding 1ul of 10x RNA ligase 1, 2ul of Ligase Buffer, 2ul of 10mM ATP, 6ul of 50% DMSO, and 2ul of 10mM 5'adapter (26nt) at 25^{0}C for 2 hours. On a 15% TBE Urea-denaturing SDS-PAGE gel (Invitrogen) and recover bands between 71nt and 120nt, corresponding to the ligated products. Prepare cDNA libraries as described in (Gogakos et al., 2017), with one exception of using 4ul of 10mM of DNTP instead of the 10ul and perform PCR amplification for 15 cycles. PCR amplified cDNA was sequenced on Illumina Novaseq6000.

Metabolic RNA labeling via pulse-chase EC-tagging and time course mRNAseq

5-EC was synthesized as previously described (Hida et al., 2017). For pulse-chase experiments, larvae were fed 1 mM EC for a 12-Hour pulse, and then transferred to media containing 10 mM unmodified uridine for the 3, 6, and 12-hour chase. A 6-hour collection of larvae was collected and left at 25°C until they were fed 1 mM 5EC at 72 hours after larval hatching (ALH) for 12 hours. After this pulse, larvae continued eating excess uridine until dissections were done at the final 12-hour chase time point at 96 hours ALH. Total RNA was extracted from crudely dissected central nervous system tissue using Trizol. A total of 25 µg of RNA was biotinylated using Click-iT Nascent RNA Capture reagents (ThermoFisher) and purified on Dynabeads MyOne Streptavidin T1 magnetic beads (ThermoFisher), as previously described ((Hida et al., 2017). After the final wash, beads containing captured RNA were used to make RNA Sequencing libraries using the Ovation® SoLo RNA-Seq System kit.

2. ***Statistical and Bioinformatics Analysis***

tRNA data sequencing analysis

Raw sequencing reads were processed using cutadapt (v3.7) to trim adaptor sequences and remove reads shorter than 12bp. Custom reference *D.melanogaster* tRNA transcriptome was generated by collapsing identical mature tRNA sequences from gtRNAdb (genome release 6.40) and adding a CCA at the 3' end of each sequence. tRNAseq reads were mapped to the reference tRNA transcriptome using subread-align (v.2.01; Liao et al.,2013) with the following parameters *'-t 1 -T 64 --multiMapping -B 5 -m 3 -I* ', allowing up to 3 mismatches and report the first 5 multi-mapped reads. Although bowtie2 is more popular in published tRNAseq studies, we found that subread is more consistent at assigning and handling multi-mapped reads (See *Chapter 3 Methods* for more details). For quality control, only aligned reads with MAPQ>=10 were retained. Read count summary was performed using subread's featureCounts, in which multi-mapped reads were fractionally split between their references. NOISeqBio in R(v.3.6) was used to perform read count normalization (TMM) and batch correction, and differential gene expression analysis. NOISeqBio uses non-parametric Bayesian methods to estimate differentially expressed genes in biological replicates and reports a q-value that is statistically equivalent to the Benjamini-Hochberg FDR. Although it recommends a threshold q-value of 0.2, we used a q-value of 0.1 instead. See Chapter 3 for more details on NOISeqBio performance.

Hypothesis testing and data visualization

All subsequent statistical tests and visualization were performed in python3 (v3.8) using the following packages: scipy, sklearn, pandas, matplotlib, seaborn, and statsmodel.

Gene Ontology analysis

Biological and molecular GO analysis was performed using *FlyEnrchr* (https://maayanlab.cloud/FlyEnrichr/)

3. ***Codon Optimality Metrics***

Codon Stabilization Coefficient, CSC

$$\mathrm{CSC}_c = \text{Pearson's R (codon frequency}_c\text{ , mRNA half-lives)}$$

tRNA adaptive Index (tAI)

The tAI of a codon was calculated according to dos Reis 2004, using the normalized tRNA expression levels (TMM) instead:

i. $tAI_c = \sum_{1:j} \left((1- s_{cj})^* \mathrm{tRNA\ TMM}_{cj} \right)$

Translation adaptive index of a codon c, tAI_c, is the weighted sum of all its j decoding tRNAs that read with an anticodon-codon binding affinity s_{cj}

The tAI of a gene is the geometric mean of the tAI values of the codons in its coding sequence.

$$tAI_g = (\Pi\ \mathrm{tAI}_c)^{1/L}$$

Normalized Codon Frequency (RSCU)

Relative synonymous codon usage (RSCU) is the ratio of observed usage to the expected uniform usage within its amino acid class. RSCU is invariant to sequence length or amino acid composition. The RSCU of the 59 degenerate codons was computed using custom python3 scripts (https://github.com/rhondene/Codon-Usage-in-Python) according to (Sharp and Li,1987). Six-fold amino acids (Leucine, Serine, Arginine) were split into 2-fold and 4-fold codon groups.

Translation Supply Demand Ratio (SDR)

i. Firstly, the codon-level translation supply-demand ratio (SDR_c) was calculated as
$$SDR_c = \mathrm{tAI}_c / \left(\sum(\text{mRNA\_exprs*normalized codfreq}_c)_{\text{normalized for amino acid}} \right)$$

ii. Then the mRNA level supply-demand, SDR, was calculated as the geometric mean of the codon SDR values:
$$SDR_g = (\Pi\ \mathrm{SDR}_c)^{1/L}$$

The effective number of codons and Codon Adaptation Index

The effective number of codons (N), which measures the degree of synonymous codon bias of gene or genome, and CAI were computed from coding sequences using DAMBE (Xia, 2017).

References

1. Aboukilila, M. Y., Sami, J. D., Wang, J., England, W., Spitale, R. C., & Cleary, M. D. (2020). Identification of novel regulators of dendrite arborization using cell type-specific RNA metabolic labeling. *PloS one*, *15*(12), e0240386. https://doi.org/10.1371/journal.pone.0240386
2. Arimbasseri, A. G., & Maraia, R. J. (2016). RNA Polymerase III Advances: Structural and tRNA Functional Views. *Trends in biochemical sciences*, *41*(6), 546–559. https://doi.org/10.1016/j.tibs.2016.03.003
3. Asano, K., Suzuki, T., Saito, A., Wei, F. Y., Ikeuchi, Y., Numata, T., Tanaka, R., Yamane, Y., Yamamoto, T., Goto, T., Kishita, Y., Murayama, K., Ohtake, A., Okazaki, Y., Tomizawa, K., Sakaguchi, Y., & Suzuki, T. (2018). Metabolic and chemical regulation of tRNA modification associated with taurine deficiency and human disease. *Nucleic acids research*, *46*(4), 1565–1583. https://doi.org/10.1093/nar/gky068
4. Azzarelli, R., Simons, B. D. and Philpott, A. (2018). The developmental origin of brain tumours: a cellular and molecular framework. Development 145, dev162693. doi:10.1242/dev.162693
5. Bae, B., & Miura, P. (2020). Emerging Roles for 3' UTRs in Neurons. *International journal of molecular sciences*, *21*(10), 3413. https://doi.org/10.3390/ijms21103413
6. Bauer, K. E., Segura, I., Gaspar, I., Scheuss, V., Illig, C., Ammer, G., Hutten, S., Basyuk, E., Fernández-Moya, S. M., Ehses, J., Bertrand, E., & Kiebler, M. A. (2019). Live cell imaging reveals 3'-UTR dependent mRNA sorting to synapses. *Nature communications*, *10*(1), 3178. https://doi.org/10.1038/s41467-019-11123-x
7. Bazzini AA, Del Viso F, Moreno-Mateos MA, Johnstone TG, Vejnar CE, Qin Y, Yao J, Khokha MK, and Giraldez AJ (2016). Codon identity regulates mRNA stability and translation efficiency during the maternal-tozygotic transition. *EMBO J*. 35, 2087–2103.
8. Blair, J. D., Hockemeyer, D., Doudna, J. A., Bateup, H. S., & Floor, S. N. (2017). Widespread Translational Remodeling during Human Neuronal Differentiation. *Cell reports*, *21*(7), 2005–2016. https://doi.org/10.1016/j.celrep.2017.10.095
9. Blanco, S., Dietmann, S., Flores, J. V., Hussain, S., Kutter, C., Humphreys, P., ... & Frye, M. (2014). Aberrant methylation of t RNA s links

cellular stress to neuro-developmental disorders. *The EMBO journal*, *33*(18), 2020-2039.

10. Blaze, J., Navickas, A., Phillips, H. L., Heissel, S., Plaza-Jennings, A., Miglani, S., Asgharian, H., Foo, M., Katanski, C. D., Watkins, C. P., Pennington, Z. T., Javidfar, B., Espeso-Gil, S., Rostandy, B., Alwaseem, H., Hahn, C. G., Molina, H., Cai, D. J., Pan, T., Yao, W. D., … Akbarian, S. (2021). Neuronal Nsun2 deficiency produces tRNA epitranscriptomic alterations and proteomic shifts impacting synaptic signaling and behavior. *Nature communications*, *12*(1), 4913. https://doi.org/10.1038/s41467-021-24969-x
11. Bornelöv S, Selmi T, Flad S, Dietmann S, Frye M (2019) Codon usage optimization in pluripotent embryonic stem cells. Genome biology 20: 119. pmid:31174582
12. Bosch, J. A., Colbeth, R., Zirin, J., & Perrimon, N. (2020). Gene Knock-Ins in *Drosophila* Using Homology-Independent Insertion of Universal Donor Plasmids. *Genetics*, *214*(1), 75–89. https://doi.org/10.1534/genetics.119.302819
13. Bridi, J. C., Ludlow, Z. N., Kottler, B., Hartmann, B., Vanden Broeck, L., Dearlove, J., Göker, M., Strausfeld, N. J., Callaerts, P., & Hirth, F. (2020). Ancestral regulatory mechanisms specify conserved midbrain circuitry in arthropods and vertebrates. Proceedings of the National Academy of Sciences of the United States of America, 117(32), 19544–19555. https://doi.org/10.1073/pnas.1918797117
14. Buccitelli, C., & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature reviews. Genetics*, *21*(10), 630–644. https://doi.org/10.1038/s41576-020-0258-4
15. Buffington, S. A., Huang, W., & Costa-Mattioli, M. (2014). Translational control in synaptic plasticity and cognitive dysfunction. *Annual review of neuroscience*, *37*, 17–38. https://doi.org/10.1146/annurev-neuro-071013-014100
16. Burow, D. A., Martin, S., Quail, J. F., Alhusaini, N., Coller, J., & Cleary, M. D. (2018). Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila. *Cell reports*, *24*(7), 1704–1712. https://doi.org/10.1016/j.celrep.2018.07.039
17. Burow DA, Umeh-Garcia MC, True MB, Bakhaj CD, Ardell DH, and Cleary MD (2015). Dynamic regulation of mRNA decay during neural development. *Neural Dev*. 10, 11
18. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. (2010). A role for codon order in translation dynamics. *Cell* 1412:355–367.
19. Colliva, A., Tongiorgi, E. Distinct role of 5′UTR sequences in dendritic trafficking of BDNF mRNA: additional mechanisms for the BDNF splice variants spatial code. Mol Brain 14, 10 (2021). https://doi.org/10.1186/s13041-020-00680-8

20. de Boer E, Jasin M, Keeney S. (2015). Local and sex-specific biases in crossover vs. noncrossover outcomes at meiotic recombination hot spots in mice. Genes & Development 29,1721–1733. https://doi.org/10.1101/ gad.265561.115
21. Delaunay, S., Pascual, G., Feng, B., Klann, K., Behm, M., Hotz-Wagenblatt, A., Richter, K., Zaoui, K., Herpel, E., Münch, C., Dietmann, S., Hess, J., Benitah, S. A., & Frye, M. (2022). Mitochondrial RNA modifications shape metabolic plasticity in metastasis. *Nature*, *607*(7919), 593–603. https://doi.org/10.1038/s41586-022-04898-5
22. Dini Modigliani, S.; Morlando, M.; Errichelli, L.; Sabatelli, M.; Bozzoni, I. (2014) An ALS-associated mutation in the FUS 302-UTR disrupts a microRNA-FUS regulatory circuitry. Nat. Commun. 5, 1–7.
23. Dittmar KA, Goodenbour JM, and Pan T (2006). Tissue-specific differences in human transfer RNA expression. *PLoS Genet*. 2, e221.
24. Hida N, Aboukilila MY, Burow DA, et al. (2017) EC-tagging allows cell type-specific RNA analysis. Nucleic Acids Res;45(15):e138
25. dos Reis, M., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research*, *32*(17), 5036–5044. https://doi.org/10.1093/nar/gkh834
26. Drummond, D. A., & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature reviews. Genetics*, *10*(10), 715–724. https://doi.org/10.1038/nrg2662
27. Dubal D, Moghe P, Verma RK, Uttekar B, Rikhy R.(2022).Mitochondrial fusion regulates proliferation and differentiation in the type II neuroblast lineage in Drosophila. *PLoS Genet* 18(2): e1010055. https://doi.org/10.1371/journal.pgen.1010055
28. Duret, L., & Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, *10*, 285–311. https://doi.org/10.1146/annurev-genom-082908-150001
29. Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G. D., Cox, J., Geiger, T., Lindner, A. B., & Pilpel, Y. (2019). Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. *Molecular cell*, *75*(3), 427–441.e5. https://doi.org/10.1016/j.molcel.2019.06.041
30. Forrest, M. E., Pinkard, O., Martin, S., Sweet, T. J., Hanson, G., & Coller, J. (2020). Codon and amino acid content are associated with mRNA stability in mammalian cells. *PloS one*, *15*(2), e0228730. https://doi.org/10.1371/journal.pone.0228730
31. Frumkin, I., Schirman, D., Rotman, A., Li, F., Zahavi, L., Mordret, E., Asraf, O., Wu, S., Levy, S. F., & Pilpel, Y. (2017). Gene Architectures that Minimize Cost of Gene Expression. *Molecular cell*, *65*(1), 142–153. https://doi.org/10.1016/j.molcel.2016.11.007
32. Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, Christophersen NS, Christensen LL, Borre M, Sørensen KD et al.,

(2014) A dual program for translation regulation in cellular proliferation and differentiation. *Cell*, 158: 1281 – 1292

33. Gogakos T, Brown M, Garzia A, Meyer C, Hafner M, Tuschl T. (2017). Characterizing expression and processing of precursor and mature human tRNAs by Hydro-tRNAseq and PAR-CLIP. *Cell Reports,* 20:1463 – 1475
34. Goodarzi H, Nguyen HCB, Zhang S, Dill BD, Molina H, Tavazoie SF. (2016). Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell,* 165:1416 – 1427
35. Goodenbour JM, Pan T .(2006), Diversity of tRNA genes in eukaryotes. *Nucleic Acids Res* 34:6137–6146
36. Goodman D.B., Church G.M., Kosuri S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. *Science*,342:475–479.
37. Gruber, J. J., Olejniczak, S. H., Yong, J., La Rocca, G., Dreyfuss, G., & Thompson, C. B. (2012). Ars2 promotes proper replication-dependent histone mRNA 3' end formation. *Molecular cell*, *45*(1), 87–98. https://doi.org/10.1016/j.molcel.2011.12.020
38. Gruber, J. J., Zatechka, D. S., Sabin, L. R., Yong, J., Lum, J. J., Kong, M., Zong, W. X., Zhang, Z., Lau, C. K., Rawlings, J., Cherry, S., Ihle, J. N., Dreyfuss, G., & Thompson, C. B. (2009). Ars2 links the nuclear cap-binding complex to RNA interference and cell proliferation. *Cell*, *138*(2), 328–339. https://doi.org/10.1016/j.cell.2009.04.046
39. Hafner M, Renwick N, Farazi TA, Mihailović A, Pena JT, Tuschl T. Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. Methods. 2012 Oct;58(2):164-70. doi: 10.1016/j.ymeth.2012.07.030. Epub 2012 Aug 7. PMID: 22885844; PMCID: PMC3508525.
40. Hausser, J., Mayo, A., Keren, L., & Alon, U. (2019). Central dogma rates and the trade-off between precision and economy in gene expression. *Nature communications*, 10(1), 1-15.
41. Hia, F., Yang, S. F., Shichino, Y., Yoshinaga, M., Murakawa, Y., Vandenbon, A., Fukao, A., Fujiwara, T., Landthaler, M., Natsume, T., Adachi, S., Iwasaki, S., & Takeuchi, O. (2019). Codon bias confers stability to human mRNAs. *EMBO reports*, *20*(11), e48220. https://doi.org/10.15252/embr.201948220
42. Hilgers, V., Lemke, S. B., & Levine, M. (2012). ELAV mediates 3' UTR extension in the Drosophila nervous system. Genes & development, 26(20), 2259–2264. https://doi.org/10.1101/gad.199653.112.
43. Hoernes, T. P., Faserl, K., Juen, M. A., Kremser, J., Gasser, C., Fuchs, E., Shi, X., Siewert, A., Lindner, H., Kreutz, C., Micura, R., Joseph, S., Höbartner, C., Westhof, E., Hüttenhofer, A., & Erlacher, M. D. (2018). Translation of non-standard codon nucleotides reveals minimal requirements for codon-anticodon interactions. *Nature communications*, *9*(1), 4865. https://doi.org/10.1038/s41467-018-07321-8

44. Iwata, R., & Vanderhaeghen, P. (2021). Regulatory roles of mitochondria and metabolism in neurogenesis. Current opinion in neurobiology, 69, 231-240.
45. Huber, K. M., Kayser, M. S., & Bear, M. F. (2000). Role for rapid dendritic protein synthesis in hippocampal mGluR-dependent long-term depression. *Science (New York, N.Y.)*, *288*(5469), 1254–1257. https://doi.org/10.1126/science.288.5469.1254
46. Kapur, M., Ganguly, A., Nagy, G., Adamson, S. I., Chuang, J. H., Frankel, W. N., & Ackerman, S. L. (2020). Expression of the Neuronal tRNA n-Tr20 Regulates Synaptic Transmission and Seizure Susceptibility. *Neuron*, *108*(1), 193–208.e9. https://doi.org/10.1016/j.neuron.2020.07.023
47. Kapur, M., Monaghan, C. E., & Ackerman, S. L. (2017). Regulation of mRNA Translation in Neurons-A Matter of Life and Death. *Neuron*, *96*(3), 616–637. https://doi.org/10.1016/j.neuron.2017.09.057
48. Knight, J., Garland, G., Pöyry, T., Mead, E., Vlahov, N., Sfakianos, A., Grosso, S., De-Lima-Hedayioglu, F., Mallucci, G. R., von der Haar, T., Smales, C. M., Sansom, O. J., & Willis, A. E. (2020). Control of translation elongation in health and disease. *Disease models & mechanisms*, *13*(3), dmm043208. https://doi.org/10.1242/dmm.043208
49. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, *44*(W1), W90–W97. https://doi.org/10.1093/nar/gkw377
50. Lee, C.Y., Robinson, K.J., and Doe, C.Q. (2006). Lgl, Pins and aPKC regulate neuroblast self-renewal versus differentiation. *Nature* 439: 594–598.
51. Lee, S., Sato, Y., & Nixon, R. A. (2011). Lysosomal proteolysis inhibition selectively disrupts axonal transport of degradative organelles and causes an Alzheimer's-like axonal dystrophy. Journal of Neuroscience, 31(21), 7817-7830
52. Leśniewska E, Boguta M. (2017). Novel layers of RNA polymerase III control affecting tRNA gene transcription in eukaryotes, Open Biol.7170001170001http://doi.org/10.1098/rsob.170001
53. Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, *41*(10), e108. https://doi.org/10.1093/nar/gkt214
54. Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C. *et al.,.* (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6:**26 https://doi.org/10.1186/1748-7188-6-26
55. Lyu X, Yang Q, Li L, Dang Y, Zhou Z, Chen S, et al.,. (2020) Adaptation of codon usage to tRNA I34 modification controls translation kinetics and proteome landscape. PLoS Genet 16(6): e1008836. https://doi.org/10.1371/journal.pgen.1008836

56. Frye, M., Jaffrey, S. R., Pan, T., Rechavi, G., & Suzuki, T. (2016). RNA modifications: what have we learned and where are we headed?. *Nature reviews. Genetics*, *17*(6), 365–372. https://doi.org/10.1038/nrg.2016.47
57. Khacho, M., Clark, A., Svoboda, D. S., Azzi, J., MacLaurin, J. G., Meghaizel, C., Sesaki, H., Lagace, D. C., Germain, M., Harper, M. E., Park, D. S., & Slack, R. S. (2016). Mitochondrial Dynamics Impacts Stem Cell Identity and Fate Decisions by Regulating a Nuclear Transcriptional Program. *Cell stem cell*, *19*(2), 232–247. https://doi.org/10.1016/j.stem.2016.04.015
58. Medina-Muñoz, S. G., Kushawah, G., Castellano, L. A., Diez, M., DeVore, M. L., Salazar, M. J. B., & Bazzini, A. A. (2021). Crosstalk between codon optimality and cis-regulatory elements dictates mRNA stability. Genome biology, 22(1), 1-23.
59. Meer, E. J., Wang, D. O., Kim, S., Barr, I., Guo, F., & Martin, K. C. (2012). Identification of a cis-acting element that localizes mRNA to synapses. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(12), 4639–4644. https://doi.org/10.1073/pnas.1116269109
60. Mishima Y, and Tomari Y (2016). Codon usage and 3^0 UTR length determine maternal mRNA stability in zebrafish. *Mol. Cell* 61, 874–885
61. Motorin, Y., & Marchand, V. (2021). Analysis of RNA Modifications by Second- and Third-Generation Deep Sequencing: 2020 Update. *Genes*, *12*(2), 278. https://doi.org/10.3390/genes12020278
62. Nadkarni P. M. (2002). An introduction to information retrieval: applications in genomics. *The pharmacogenomics journal*, *2*(2), 96–102. https://doi.org/10.1038/sj.tpj.6500084
63. Narula A, Ellis J, Taliaferro JM, Rissland OS. (2019). Coding regions affect mRNA stability in human cells. *RNA* 25:1751–1764
64. Niimi, T. et al. (1994). Recognition of the anticodon loop of $tRNA^{Ile1}$ by isoleucyl-transfer RNA synthetase from *Escherichia coli*. *Nucleosides Nucleotides Nucleic Acids* 13, 1231–1237
65. Novoa, E. M., Pavon-Eternod, M., Pan, T., & Ribas de Pouplana, L. (2012). A role for tRNA modifications in genome structure and codon usage. *Cell*, *149*(1), 202–213. https://doi.org/10.1016/j.cell.2012.01.050
66. Oktaba, K., Zhang, W., Lotz, T. S., Jun, D. J., Lemke, S. B., Ng, S. P., Esposito, E., Levine, M., & Hilgers, V. (2015). ELAV links paused Pol II to alternative polyadenylation in the Drosophila nervous system. *Molecular cell*, *57*(2), 341–348. https://doi.org/10.1016/j.molcel.2014.11.024
67. Osterman, I. A., Chervontseva, Z. S., Evfratov, S. A., Sorokina, A. V., Rodin, V. A., Rubtsova, M. P., Komarova, E. S., Zatsepin, T. S., Kabilov, M. R., Bogdanov, A. A., Gelfand, M. S., Dontsova, O. A., & Sergiev, P. V. (2020). Translation at first sight: the influence of leading codons. *Nucleic acids research*, *48*(12), 6931–6942. https://doi.org/10.1093/nar/gkaa430
68. Pan T. (2018). Modifications and functional genomics of human transfer RNA.Cell Res 28: 395

69. Pechmann S, Frydman J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 20: 237 – 243
70. Peng, Z., Cheng, Y., Tan, B. C., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., Guo, J., Dong, Z., Liang, Y., Bao, L., & Wang, J. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology*, *30*(3), 253–260. https://doi.org/10.1038/nbt.2122
71. Pereira LA, Munita R, González MP, Andrés ME.(2017). Long 3'UTR of Nurr1 mRNAs is targeted by miRNAs in mesencephalic dopamine neurons. PLOS ONE 12(11): e0188177. https://doi.org/10.1371/journal.pone.0188177
72. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, and Coller J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell* 160, 1111–1124.
73. Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, and Coller J (2016). The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. *Cell* 167, 122–132.
74. Rafels-Ybern, À., Torres, A. G., Camacho, N., Herencia-Ropero, A., Roura Frigolé, H., Wulff, T. F., Raboteg, M., Bordons, A., Grau-Bove, X., Ruiz-Trillo, I., & Ribas de Pouplana, L. (2019). The Expansion of Inosine at the Wobble Position of tRNAs, and Its Role in the Evolution of Proteomes. *Molecular biology and evolution*, *36*(4), 650–662. https://doi.org/10.1093/molbev/msy245
75. Rak, R., Polonsky, M., Eizenberg-Magar, I., Mo, Y., Sakaguchi, Y., Mizrahi, O., Nachshon, A., Reich-Zeliger, S., Stern-Ginossar, N., Dahan, O., Suzuki, T., Friedman, N., & Pilpel, Y. (2021). Dynamic changes in tRNA modifications and abundance during T cell activation. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(42), e2106556118. https://doi.org/10.1073/pnas.2106556118
76. Rudinger-Thirion J, Lescure A, Paulus C, Frugier M.(2011) Misfolded human tRNA isodecoder binds and neutralizes a 3′UTR-embedded Alu element. *Proceedings of the National Academy of Sciences of the United States of America*, 108:E794–E802
77. Sabin, L. R., Zhou, R., Gruber, J. J., Lukinova, N., Bambina, S., Berman, A., Lau, C. K., Thompson, C. B., & Cherry, S. (2009). Ars2 regulates both miRNA- and siRNA- dependent silencing and suppresses RNA virus infection in Drosophila. *Cell*, *138*(2), 340–351. https://doi.org/10.1016/j.cell.2009.04.045
78. Sagi, D., Rak, R., Gingold, H., Adir, I., Maayan, G., Dahan, O., Broday, L., Pilpel, Y., & Rechavi, O. (2016). Tissue- and Time-Specific Expression of Otherwise Identical tRNA Genes. *PLoS genetics*, *12*(8), e1006264. https://doi.org/10.1371/journal.pgen.1006264

79. Kostinski, S., & Reuveni, S. (2020). Ribosome Composition Maximizes Cellular Growth Rates in E. coli. *Physical review letters*, *125*(2), 028103. https://doi.org/10.1103/PhysRevLett.125.028103
80. Schaffer, A. E., Pinkard, O., & Coller, J. M. (2019). tRNA Metabolism and Neurodevelopmental Disorders. *Annual review of genomics and human genetics*, *20*, 359–387. https://doi.org/10.1146/annurev-genom-083118-015334
81. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342. https://doi.org/10.1038/nature10098Stoler
82. Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR genomics and bioinformatics*, *3*(1), lqab019. https://doi.org/10.1093/nargab/lqab019
83. Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivière, L., Maier, K. C., Seizl, M., Tresch, A., & Cramer, P. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome research*, *22*(7), 1350–1359. https://doi.org/10.1101/gr.130161.111
84. Suzuki T. (2021). The expanding world of tRNA modifications and their disease relevance. *Nature reviews. Molecular cell biology*, *22*(6), 375–392. https://doi.org/10.1038/s41580-021-00342-0
85. Suzuki, T., Yashiro, Y., Kikuchi, I., Ishigami, Y., Saito, H., Matsuzawa, I., Okada, S., Mito, M., Iwasaki, S., Ma, D., Zhao, X., Asano, K., Lin, H., Kirino, Y., Sakaguchi, Y., & Suzuki, T. (2020). Complete chemical structures of human mitochondrial tRNAs. *Nature communications*, *11*(1), 4269. https://doi.org/10.1038/s41467-020-18068-6
86. Swartling, F. J., Čančer, M., Frantz, A., Weishaupt, H., & Persson, A. I. (2015). Deregulated proliferation and differentiation in brain tumors. *Cell and tissue research*, *359*(1), 225–254. https://doi.org/10.1007/s00441-014-2046-y
87. Swindell, W. R., Remmer, H. A., Sarkar, M. K., Xing, X., Barnes, D. H., Wolterink, L., Voorhees, J. J., Nair, R. P., Johnston, A., Elder, J. T., & Gudjonsson, J. E. (2015). Proteogenomic analysis of psoriasis reveals discordant and concordant changes in mRNA and protein abundance. *Genome medicine*, *7*(1), 86. https://doi.org/10.1186/s13073-015-0208-5
88. Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic acids research*, *43*(21), e140. https://doi.org/10.1093/nar/gkv711
89. Timmers, H., & Tora, L. (2018). Transcript Buffering: A Balancing Act between mRNA Synthesis and mRNA Degradation. *Molecular cell*, *72*(1), 10–17. https://doi.org/10.1016/j.molcel.2018.08.023Titus MB, Wright EG, Bono JM, Poliakon AK, Goldstein BR, Super MK, Young LA, Manaj M,

90. Titus, M. B., Wright, E. G., Bono, J. M., Poliakon, A. K., Goldstein, B. R., Super, M. K., Young, L. A., Manaj, M., Litchford, M., Reist, N. E., Killian, D. J., & Olesnicky, E. C. (2021). The conserved alternative splicing factor caper regulates neuromuscular phenotypes during development and aging. *Developmental biology*, *473*, 15–32. https://doi.org/10.1016/j.ydbio.2021.01.011
91. Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A., & Madan Babu, M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science signaling*, *11*(546), eaat6409. https://doi.org/10.1126/scisignal.aat6409
92. Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3645–3650. https://doi.org/10.1073/pnas.0909910107
93. Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., & Ziv-Ukelson, M. (2011). Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome biology*, *12*(11), R110. https://doi.org/10.1186/gb-2011-12-11-r110
94. Tushev, G., Glock, C., Heumüller, M., Biever, A., Jovanovic, M., & Schuman, E. M. (2018). Alternative 3' UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. Neuron, 98(3), 495–511.e6. https://doi.org/10.1016/j.neuron.2018.03.030
95. Van Bortle, K., Phanstiel, D. H., & Snyder, M. P. (2017). Topological organization and dynamic regulation of human tRNA genes during macrophage differentiation. *Genome biology*, *18*(1), 180. https://doi.org/10.1186/s13059-017-1310-3
96. Liebermeister, W., Noor, E., Flamholz, A., Davidi, D., Bernhardt, J., & Milo, R. (2014). Visual account of protein investment in cellular functions. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8488–8493. https://doi.org/10.1073/pnas.1314810111
97. Warner J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences*, *24*(11), 437–440. https://doi.org/10.1016/s0968-0004(99)01460-7
98. Wohlgemuth, I., Pohl, C., Mittelstaet, J., Konevega, A. L., & Rodnina, M. V. (2011). Evolutionary optimization of speed and accuracy of decoding on the ribosome. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1580), 2979-2986.
99. Wright F. 1990 The ‘effective number of codons’ used in a gene. *Gene.* 871: 23–29.
100. Xia X. (2017). DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. *The Journal of heredity*, *108*(4), 431–437. https://doi.org/10.1093/jhered/esx033

101. Yamazaki, H., Kasai, S., Mimura, J., Ye, P., Inose-Maruyama, A., Tanji, K., Wakabayashi, K., Mizuno, S., Sugiyama, F., Takahashi, S., Sato, T., Ozaki, T., Cavener, D. R., Yamamoto, M., & Itoh, K. (2020). Ribosome binding protein GCN1 regulates the cell cycle and cell proliferation and is essential for the embryonic development of mice. *PLoS genetics*, *16*(4), e1008693. https://doi.org/10.1371/journal.pgen.1008693

102. Yeganeh, M., Praz, V., Cousin, P., & Hernandez, N. (2017). Transcriptional interference by RNA polymerase III affects expression of the *Polr3e* gene. *Genes & development*, *31*(4), 413–421. https://doi.org/10.1101/gad.293324.116

103. Zaidi SH, Denman R, Malter JS. (1994). Multiple proteins interact at a unique cis-element in the 39 UTR of amyloid precursor protein (APP) mRNA. J Biol Chem 269:24000–24007.

104. Zhang W, Foo M, Eren AM, Pan T. (2022). tRNA modification dynamics from individual organisms to metaepitranscriptomics of microbiomes. *Molecular Cell* , 82(5):891-906. DOI: 10.1016/j.molcel.2021.12.007

Chapter 3: Parametric and Non-Parametric Estimation of Differential tRNA expression analysis

Background

High-throughput RNA sequencing remains the technique of choice for estimating genome-wide levels of gene expression. In a typical bulk RNAseq experiment, RNA molecules are extracted from a biological sample (usual tissues), sheared into smaller fragments, reverse transcribed into cDNA, and amplified. These cDNA fragments are sequenced, often generating millions of sequenced reads per sample. The expression of a genomic feature (gene, transcript, etc.) is estimated based on the number of reads mapped to that region. Thus, RNAseq characterizes the biological status of the profiled tissue by providing a snapshot of the transcriptome. In comparative RNAseq experiments, the primary motivation is to associate changes in gene expression to the distinct biological outcomes between two or more conditions, e.g., healthy vs. diseased tissue A vs. tissue B. Because the distribution of mapped reads is due to both biological and technical variability, a central goal of RNAseq analysis is to determine those genomic features that are *biologically* differentially expressed between conditions.

General formulation of modeling RNAseq count data

Methods for gene expression estimation usually represent RNAseq input data as a count matrix. Let there be $\boldsymbol{N}$ samples (or replicates) per biological condition and a total of $\boldsymbol{G}$ genes (features/variables) in the reference genome. Thus, the entire RNAseq count dataset is an $\boldsymbol{N}$x$\boldsymbol{G}$ matrix, where $\boldsymbol{R_{gi}}$ denotes the number of reads that maps to a gene $\boldsymbol{g}$ from sample $\boldsymbol{i}$. To allow for direct comparison between samples, the mapped read counts are normalized prior to DGE estimation [Dillies et al.,2013; Evans and Stoebel, 2018].

Parametric modeling of RNAseq count data using the negative binomial distribution

Parametric-based methods infer differential gene expression by assuming that the input read counts $\boldsymbol{R_{gi}}$ is drawn from a specific parametric distribution D. Because RNAseq experiments generate count data, continuous distributions such as the normal distribution are not suitable; instead, parametric methods model count data using discrete distributions, such as the Poisson and Negative Binomial. However, because the mean read count from biological replicates is smaller than the variance, the negative binomial is widely adopted for regression of count responses. As such, the negative binomial (NB) model is used by the two most common parametric tools for differential gene expression analysis, i.e., DEseq2 [Love et al., 2014] and edgeR [Robinson and Oshlack, 2010]. Both methods perform univariate negative binomial regression on each gene to estimate the gene count R_{gi} (response variable) [Equation 4.1].

$$\mathbf{R_{gi} \sim NB(\mu_{gi}, \boldsymbol{\phi}_{g})} \qquad \text{[Eq. 4.1]}$$

$\boldsymbol{\mu_g}$ is the normalized mean gene counts for a sample ***i***. Importantly, $\boldsymbol{\phi_g}$ is the NB dispersion parameter that describes the biological variability of the gene, and this parameter is estimated during model fitting. After model fitting, parametric tests such as the Wald's or Fisher's exact test is employed to test for significantly differentially expressed genes.

Non-parametric methods for estimating RNAseq gene expression

Non-parametric approaches to DGE analysis do not assume the underlying distribution to model the count data but build an empirical distribution from the count data [Li and Tibshirani, 2013; Tarazona et al., 2015]. Here, I summarize the non-parametric tool, NOISeqBio, which is optimized specifically for biological replicates [Taranoza et al., 2015].

Let R^{A}_{g} and R^{B}_{g} represent the counts of a gene in two distinct biological conditions, *A* and *B*. NOISeqBio quantifies differential expression based on the two statistics: the log fold change, L_g, and the absolute difference, D_g, of the normalized mean gene counts [Equation 4.2].

$$L_g = \log_2 (R^{A}_{gi} / R^{B}_{gi}), \text{ log ratio} \qquad \text{Eq. 4.2}$$
$$D_g = |R^{A}_{gi} - R^{B}_{gi}|, \text{ absolute difference}$$

The L_g and D_g values are corrected for biological variability (L^*, D^*) and pooled together to compute the differential expression parameter, Θ. They consider the probability distribution of Θ as a mixture distribution of genes with an altered expression between conditions, f_a, and genes with an invariant expression between conditions, f_o (null). [Equation 4.3]

$$\Theta = (L^* + D^*)/2$$
$$f(\Theta) = p_o f_o(\Theta) + p_a f_a(\Theta), \qquad \text{Eq 4.3}$$

Where p_o is the probability of non-differential expression between conditions and p_a is the probability that a gene is differentially expressed. To build a null noise distribution, the null scores, Θ_o, are estimated using the assumption that there is no change in expression between conditions. The probability density functions of $f_o((\Theta)$ and f_a (Θ) are estimated using Gaussian kernel density estimator [Taranoza et al., 2016].

Finally, given a gene with Θ, the posterior probability of differential expression is computed using Bayes' rule [Equation 4.4]. According to a proof formulated by

Efron et al. (2001), **q=**1- $p_a(\Theta_g)$ mathematically approximates the Benjamini-Hochberg false-discovery rate (FDR).

$$p_a(\Theta_g) = 1- [\ p_o\ (f_o(\Theta_g)/\ f_a(\Theta_g)\ )\ ] \quad \text{Eq. 4.4}$$

Performance of parametric and non-parametric methods for DGE analysis

All RNAseq methods were developed with messenger RNAseq data in mind but are readily adopted for small RNAseq and tRNAseq gene expression analysis. Most published tRNAseq analyses utilize DEseq2, although it has not been tested if tRNAseq datasets are a good fit for the negative binomial model. Moreover, there are thousands of mRNA types (features) compared to a couple of hundred tRNA genes in the animal genomes. That is, a tRNAseq dataset has much fewer features than a mRNAseq dataset. Altogether, these factors motivated me to compare the performance of the parametric negative binomial model (DEseq2) and non-parametric data-adaptive method NOISeqBio on my tRNAseq count data derived in Chapter 2 of this thesis. There is no ground truth to determine which approach is more accurate in reality; however, I perform goodness-of-fit tests to the negative binomial distribution to gain insights into the extent to which my tRNAseq count data deviates from expectation.

Materials and Methods:

Neural tRNAseq samples

3 biological replicates of neuron-derived tRNAs and neuroblast-derived tRNAs were prepared from *D.melanogaster* larval brains using the HydrotRNAseq protocol as outlined in the Methods section of Chapter 2 of this dissertation.

Read alignment with subread versus bowtie2

Libraries were aligned to mature tRNA transcriptome using subread aligner (Liao et al., 2013). On a technical note, I compared the performance of bowtie2 (Langmead and Salzberg, 2012), the most widely used aligner in tRNAseq studies, and subread. Specifically, I became interested in how both aligners handled multimapping reads because of the nature of tRNA gene families, wherein non-identical isodecoders (same anticodon but different sequence body) and isoacceptors (same amino acid but different anticodon) may still share identical sub=sequences.

Both aligners are comparable in the total reads mapped; however, I find that subread is more consistent in assigning multimapped reads as well in handling multimapped reads from the SAM file. Regarding the multimapping mode using the -k parameter for reporting the k best alignments, the bowtie2 documentation states, "*Bowtie 2 does not "find" alignments in any specific order, so for reads that have more than N distinct, valid alignments, Bowtie 2 does not guarantee that the N alignments reported are the best possible in terms of alignment score*." (source: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#k-mode-search-for-one-or-more-alignments-report-each). Moreover, bowtie2 provides minimal information about multimapped reads as it only assigns reads a MAPQ score of 0 or 255. In contrast, subread's alignment strategy is explicitly designed to handle multimapping based on its 'seed and vote' strategy; wherein, if there are *n* best alignments of a read, then the read is assigned to those locations, and you can use the -B to control the final number of alignments to report in the SAM file. In comparison, subread provides a straightforward way to identify and retrieved multimapped or uniquely mapped reads using the 'NH:i:n" and "HI:i:n" SAM tags. Finally, subread has a dedicated 'small RNAseq' alignment mode which was designed with microRNAs in mind, and the average read lengths of micro-RNAseq overlap with the read lengths in this tRNAseq protocol since HydrotRNAseq was adapted based on the small-RNAseq protocol for microRNAs (Hafner et al., 2011).

Read Count Normalization

For quality control, alignments with MAPQ>=10 are retained in the input for read count summarization using featureCounts(Liao et al., 2013). Multimapped reads were proportionally assigned to their references. DESeq2 uses performs count normalization using its median-of-ratios in which counts are divided by sample-specific size factors that are determined by the median ratio of gene counts relative to the geometric mean per gene.

NOISeqBio includes three commonly used normalization methods: RPKM, TMM, and Upper Quartile. For this tRNAseq dataset, the type of normalization did not affect the results, so I chose the TMM (Trimmed Mean of M-values; Robinson and Oshlack, 2010).

Differential gene expression analysis

Differential gene expression analysis by DESeq2 was performed using the default settings, which include *fitType =‘parametric’* to estimate the negative binomial dispersion parameters. To control for batch effects, the date of preparation for each library was included as a factor in the design matrix:

DESeqDataSetFromMatrix(countData=countData,colData=metaData, design=~Exp_date+Tissue)

In NOISeqBio, batch correction was performed using the ARSyn method by supplying the date of library preparation, similar to DESeq2. Then differential gene expression analysis by was performed with the default settings. *q=0.9* was chosen as the cut-off probability for differential expression since the FDR adjusted p-value of DESeq2 equals 0.1, and for NOISeqBio, 1-q is equivalent to the Benjamini-Hochberg FDR value.

Goodness-of-fit test for the negative binomial distribution.

The Pearson chi-squared goodness-of-fit test is often employed for discrete distributions such as Poisson and binomial, where the variance is a function of the mean. However, the assumptions of Pearson chi-squared goodness of fit do not extend to negative binomial regression because of the additional dispersion parameter (Pierce and Schafer, 1986; Mi and Schafer, 2015). To assess how well the tRNAseq fits the negative binomial model, I used the tweeDEseq package in R (Esnaola et al., 2013) to perform goodness-of-fit tests of each tRNA gene to the Poisson-Tweedie (PT) family of distributions. The PT distributions include Poisson, negative binomial, and Polya-Aeppli (geometric Poisson). tweeDEseq estimates the parameter of each PT distribution using iterative Newtonian maximum likelihood estimation. Each tRNA gene is a 1x6 dimensional vector of the DESeq2 normalized counts.

Quantile-quantile plots using the DESeq2 estimated dispersion parameter.

As a complementary approach, I used the probplot() function from the scipy.stats package (v1.90) in python3 (v3.8) to generate quantile-quantile plots for each tRNA gene-based the negative binomial dispersion parameter estimated by Deseq2 fitting. The *n* and *p* parameters required for the negative binomial option in scipy.stats.probplot were computed as follows:

$$\sigma^2 = \mu + \phi \mu^2$$

$$p = \mu / \sigma^2$$

$$n = \mu^2 / (\sigma^2 – \mu)$$

where ϕ is the DESeq2 estimated dispersion parameter, σ^2 is the variance, μ is the DESeq2 mean normalized counts (per condition), n is the number of successes, and p is the probability of success.

Results:

Here, I sought to compare the number of differentially expressed genes (DEGs) from my neural hydrotRNAseq dataset (n=106 genes) that are determined to be differentially expressed by DESeq2 (parametric negative binomial) and NOISeqBio (non-parametric) (**Figure 1A, 1B)**. Using an FDR-adjusted p-value threshold of 0.1, DESeq2 reported a total of 5 genes as differentially expressed, whereas NOISeqBio reported 13 genes are differentially expressed (threshold probability q=0.9). *Arg-TCT-3-1* is the only DEG detected by DESeq that overlaps with NOISeqBio DEGs (**Figure 1C)**. Interestingly, all the tRNAs identified by DESeq2 as DEG are single-locus genes. Recall that many tRNA genes bear multiple identical loci throughout the genome. Thus the reference transcriptome for this study is built by collapsing identical tRNA sequences.

Most tRNA gene counts do not fit the negative binomial distribution

Most tRNAseq studies are interested in the abundance of nuclear tRNA genes for downstream analyses. DESeq2 reported only 1 nuclear tRNA as DEG in comparison to NOISeqBio, which yielded 11 nuclear tRNAs as DEG. So, I wondered about the extent to which tRNA gene counts fit the negative binomial distribution, and its poor fit may explain why DESeq2 failed to identify more nuclear tRNAs as DEG. To this end, I performed goodness-of-fit tests for each tRNA gene, based on the DESeq2 normalized counts, to the Poisson-Tweedie (PT) family of distributions for modeling count data [Esnaola et al.,2013]. Briefly, this method applies maximum likelihood estimation of the shape parameter for the negative binomial, Poisson, and Poyla-Aeppli (geometric Poisson). 10 out of the 106 tRNA genes fit the negative binomial distribution (p-value>0.2). The Poisson distribution was the best fit (i.e., highest p-value) for *Arg-TCT-3-1*, the only nuclear tRNA that is DEG by DESeq2. In total, 90 genes either could not converge to a shape parameter, or they converged but did not fit either of the PT distributions **(Figure 1D)**. Notably, 10 out of 11 nuclear tRNAs that are DEG by NOISeqBio did not fit either of the PT distributions.

Finally, as a qualitative evaluation of goodness-of-fit to the negative binomial, I used the dispersion parameters estimated by DESeq2 to construct a quantile-quantile plot for each of the 11 nuclear tRNAs that are DEG by NOISeqBio, which also includes *Arg-TCT-3-1* that is reported as DEG by DESeq2. Quantile-quantile plot is a graphical method for evaluating if data conforms to a reference distribution. The closer the data points are to the 45^{o} line, the more similar the dataset is to the reference distribution. Between conditions, *Arg-TCT-3-1* and *Ser-TGA-1-1* are the most consistent in their closeness to the reference line (**Figure 1E, 1F)**

Discussion:

Here, I compared parametric and non-parametric methods for determining differential gene expression from tRNAseq data comprising biological replicates and tested the assumption that tRNAseq count data conforms to the negative binomial distribution that is widely used by popular differential gene expression tools such as DESeq2. Maximum likelihood estimation could only fit 10/106 tRNA genes to the negative binomial model. Only 1 nuclear tRNA gene was differentially expressed by DESeq2, compared to 11 nuclear tRNA genes by NOISeqBio, a non-parametric method that does not make underlying assumptions about the data. I believe these results may be informative for the implementation of tRNAseq bioinformatics pipelines.

In general, a large number of hypothesis tests (one per gene/transcript) is performed in a single RNA-Seq study. Thus, a longstanding statistical challenge of modeling RNAseq data is the *curse of dimensionality,* i.e., limited statistical power to detect truly differentially expressed genes because of the simultaneous regression fitting on hundreds or thousands of genes (features) from a small number of sample replicates. Increasing the sample size increases statistical power; however, due to the cost and labor of RNAseq library preparation, many studies are limited to 3-5 replicates per condition. NOISeqBio reported more nuclear tRNAs as differentially expressed (n=11) compared to DESeq2 (n=1). The advantage of parametric-based methods is that even with a small sample size, they demonstrate robust statistical power for estimating differentially expressed genes [Yu et al., 2020]. However, previous studies have shown that the violation of the distributional assumptions leads to poor performance and reduced sensitivity, especially if there are outliers [Li and Tibshirani, 2013] and genes with low count [Bullard et al., 2010]. Therefore, the reliability of parametric-based methods for DGE is heavily influenced by the kind of parametric statistical model. So perhaps, the low number of differentially expressed tRNAs by DESeq2 is due to the generally poor fit of the count data to the negative binomial model. On the other hand, successful convergence of the iterative maximum likelihood estimation is influenced by the type of optimization algorithm and thus may affect the results of the goodness-of-fit tests.

Still, no ground-truth tRNAseq dataset is available, so it remains uncertain which differential gene expression tool is more accurate. Nevertheless, I draw on two key bioinformatics studies that compared the performance of popular parametric and non-parametric tools, inclusive of DESeq2 and NOISeqBio, on mRNAseq datasets comprising biological replicates. Importantly, both studies had access to ground-truth datasets for evaluating the sensitivity (fraction of true DEG that is identified as DEG) and precision (fraction of DEG that is truly DEGs). In both works, the authors highlight the problem of previous RNAseq benchmark studies relying on simulated data and/or technical replicates for evaluating DEG workflows (Bacarella et al., 2018; Stupnikov et al., 2021). Stupnikov and colleagues reported that NOISeqBio was the most robust at controlling the FDR across different sample sizes, sequencing depth, and expression levels, in comparison to Deseq2 on biological replicates [Stupnikov et al., 2021]. A similar result was observed in the comparison across different read depths and sample sizes (Bacarella et al., 2018); however, they note that for sample numbers below five, NOISeqBio exhibits high sensitivity (recall more true DEG) but at the cost of low precision (more false positives). Thus, using a low sample size in NOISeqBio may increase the type I error rate (false positives). Regarding my tRNAseq, I surmise that the NOISeqBio false positive DEGs are likely those whose theta values,Θ_a, overlap with the noise distribution f_o (**Figure 1C)**.

In conclusion, the hydro-tRNAseq dataset in this study is a poor fit to the negative binomial model and related Poisson-Tweedie count distributions, which would partially explain why non-parametric NOISeqBio yields more statistically differentially expressed tRNAs than DESeq2. As more studies profiling tRNA gene expression is published, it would be worthwhile to extend this kind of analysis to multiple tRNAseq datasets from different species as well as different tRNAseq library preparation protocols.

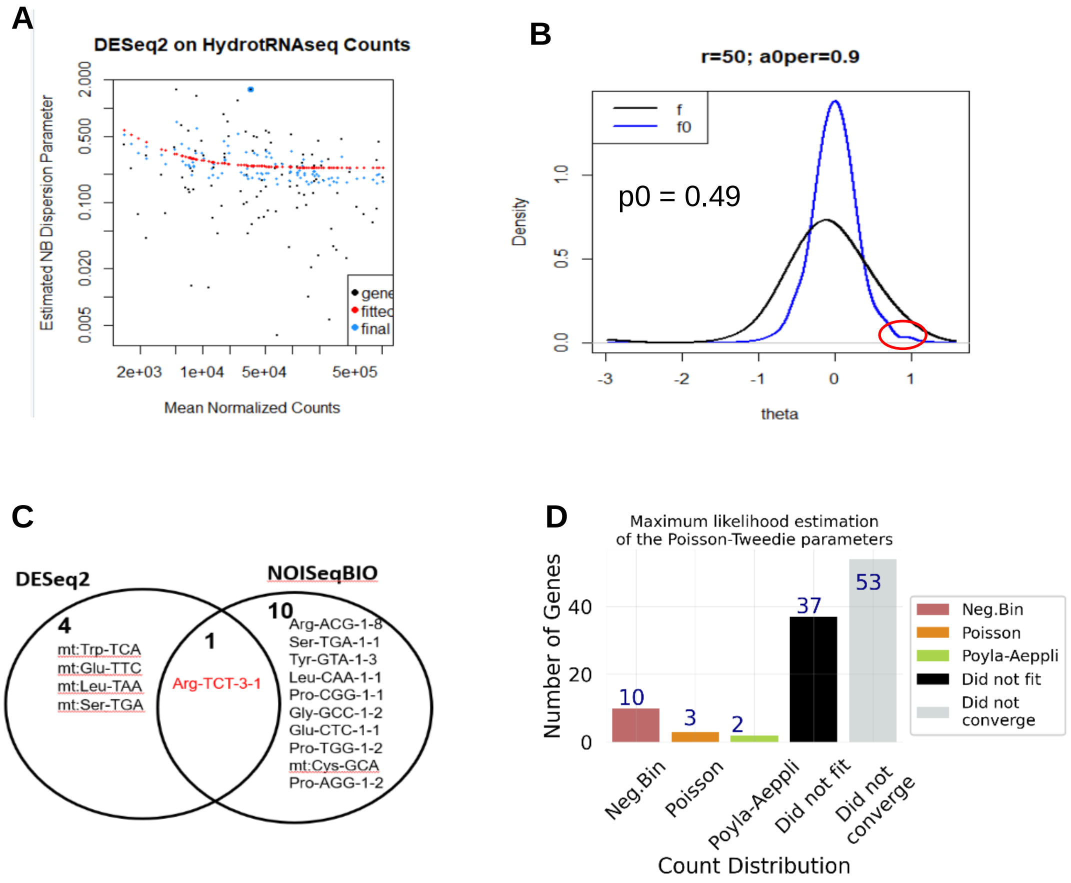


Figure 1: Differential Gene Expression using DESeq2 (parametric) and NOISeqBio (non-parametric)

A: DESeq2 gene-wise mean-dispersion plot of the fitted negative binomial model for the *Drosophila* larval CNS hydro-tRNAseq dataset (n=106 genes) comprising 3 biological replicates from neuron and neuroblast tissues.

B: NOISeqBio's empirical distributions modelling data noise, $f_{o,}$ and variation in gene expression between conditions, f_{a}, p0 is the estimated probability of non-differentially expressed. Red circle highlights the differential expression parameter that overlaps with the null noise distribution.

C: Venn diagram showing overlap of differentially expressed tRNA genes reported by DESeq2 (FDR-adjusted p-value = 0.1) and NOISeqBio (posterior probability threshold =0.9; equivalent to FDR=0.1).

D: Barplot showing the number of tRNA genes (based on DESeq2 normalized counts) that fit the count distributions from the Poisson-Tweedie family. Distribution parameters were estimated via Newton-maximum likelihood estimation.

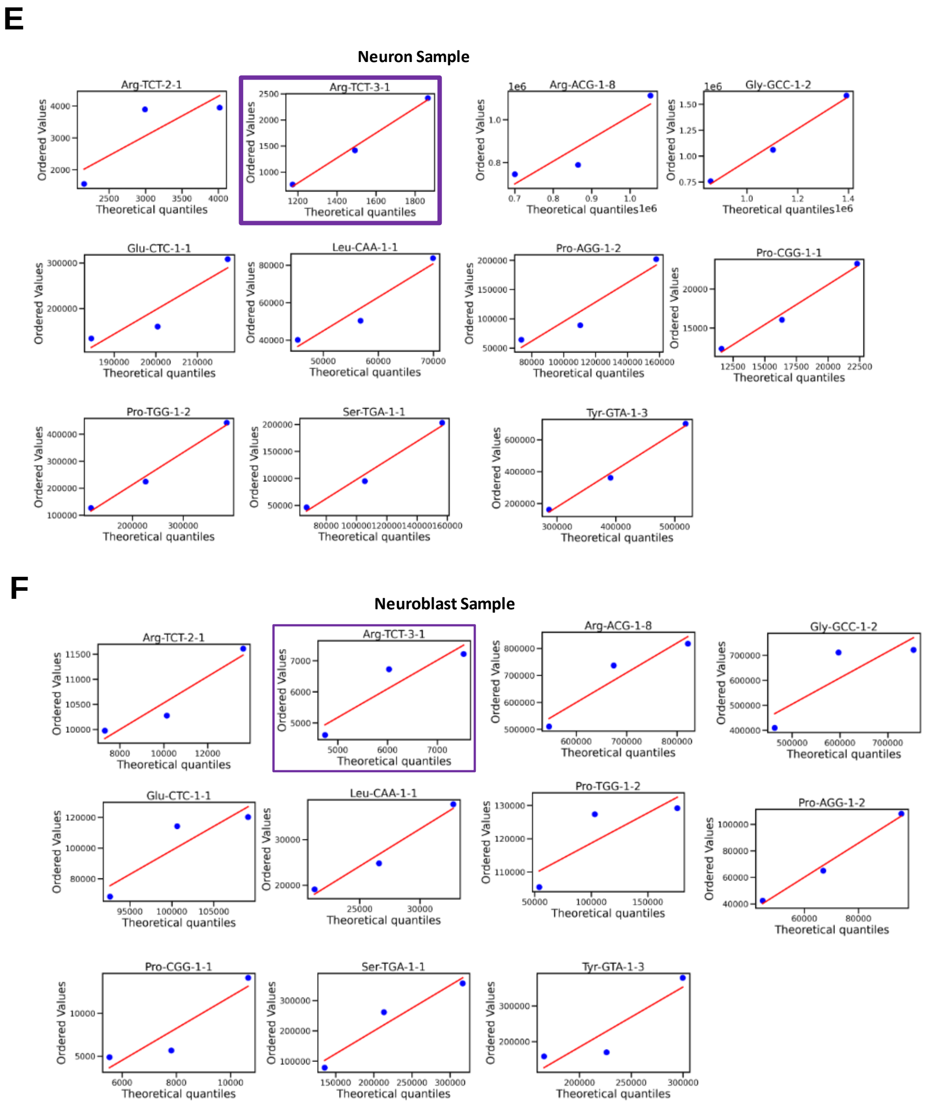


Figure 1: Differential Gene Expression using DESeq2 (parametric) and NOISeqBio (non-parametric)

E,F: Quantile-quantile plots visualizing the fit between the normalized counts of nuclear tRNA genes that are differentially expressed by DESeq2 and NOISeqBio, and the theoretical negative binomial model that is constructed using the dispersion parameter estimated by DESeq2 estimated .

References

1. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1-21.
2. Baccarella, A., Williams, C. R., Parrish, J. Z., & Kim, C. C. (2018). Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. *BMC bioinformatics*, *19*(1), 1-12.
3. Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11: 94. doi: 10.1186/1471-2164-10-221.
4. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics. 14(6):671-83.
5. Efron,B., Tibshirani,R., Storey,J.D. and Tusher,V. (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.
6. Esnaola M, Puig P, Gonzalez D, Castelo R, Gonzalez JR (2013). “A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments.” *BMC Bioinformatics*, **14**, 254.
7. Evans, C., Hardin, J., & Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in bioinformatics*, *19*(5), 776–792. https://doi.org/10.1093/bib/bbx008
8. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923
9. Li J. Tibshirani R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat. Methods Med. Res., 22, 519–536.
10. Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, *41*(10), e108. https://doi.org/10.1093/nar/gkt214
11. Mi, G., Di, Y., & Schafer, D. W. (2015). Goodness-of-fit tests and model diagnostics for negative binomial regression of RNA sequencing

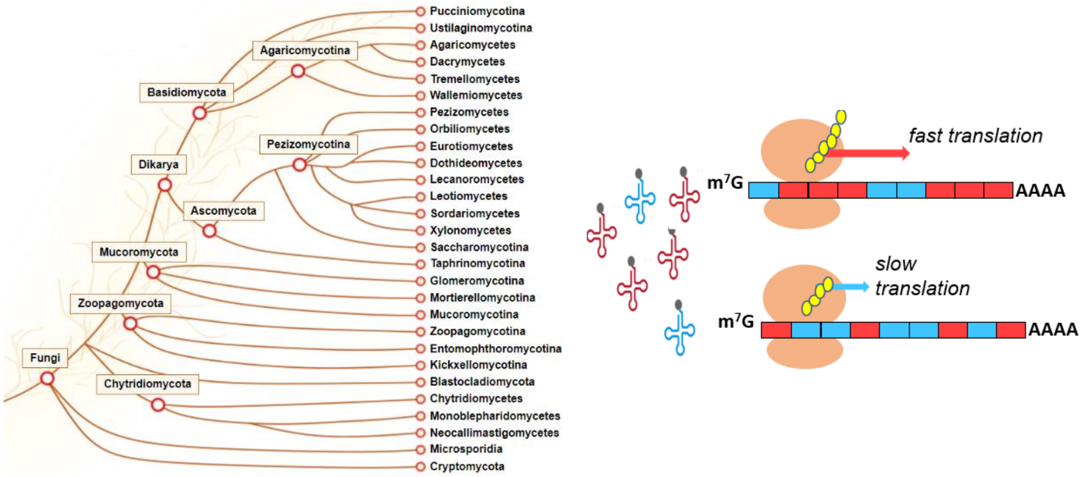
data. *PloS one*, *10*(3), e0119254. https://doi.org/10.1371/journal.pone.0119254

12. Pierce DA, Schafer DW (1986) Residuals in Generalized Linear Models. *Journal of the American Statistical Association* 81: 977–986. 10.1080/01621459.1986.10478361
13. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, *26*(1), 139-140
14. Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F, Kennedy R, Salto-Tellez M, Prise KM, McArt DG. Robustness of differential gene expression analysis of RNA-seq. Computational and structural biotechnology journal. 2021 Jan 1;19:3470-81.
15. Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic acids research*, *43*(21), e140. https://doi.org/10.1093/nar/gkv711
16. Yu, L., Fernandez, S., & Brock, G. (2020). Power analysis for RNA-Seq differential expression studies using generalized linear mixed effects models. *BMC bioinformatics*, *21*(1), 198. https://doi.org/10.1186/s12859-020-3541-7

Chapter 4: Kingdom-wide Analysis of Fungal Transcriptomes and tRNAs Reveals Conserved Patterns of Adaptive Evolution[*]

**Original publication*: Wint R, Salamov A, Grigoriev IV. 2022. Kingdom-wide analysis of fungal transcriptomes and tRNAs 1235 reveals conserved patterns of adaptive evolution. Mol Biol. Evol. 1236 https://doi.org/10.1093/molbev/msab372

Key words: codon usage, tRNA, translation, selection, fungi, gene expression, macroevolution, machine learning



Background

The billion-year-old kingdom Fungi, comprising at least 1.5 million species, is deeply intertwined with the diversification and maintenance of terrestrial ecosystems (Berbee and Taylor, 2017). Paleo-botanical studies credit mutualistic symbiosis for the successful colonization of land by primitive plants – resulting in the greening of the Earth that facilitated the evolution of more complex animal forms (Field et al., 2015). Indeed, 90% of extant plant species still rely on mycorrhizal fungi for nutrient uptake and resistance to pathogens and abiotic stressors (Eva et al., 2018). Many fungi are also pathogens of plants, fungi, and animals and pose an emerging medical threat to humans (Janbon et al., 2019). The diversity of fungal bioproducts is leveraged in biotechnology to manufacture commercial enzymes, medicines, and even biofuel (Sepala et al., 2017). Therefore, a comprehensive understanding of the evolution of fungal genomes and traits is valuable for several applications.

Protein-coding genes evolved codon usage bias due to the combined but uneven effects of adaptive and non-adaptive influences. Thus, codon usage analysis is an established framework for studying the evolution of protein-coding genes. Studies in model fungi agree on codon usage bias as an adaptation for fine-tuning gene expression levels; however, such knowledge is lacking for most other fungi. Comparative studies aim to disentangle the trait variation due to shared ancestry versus adaptation. Because of common descent, phenotypic traits from closely related species are likely to violate the identically and independently distributed requirement of standard regression tests, which risks an increase in type I errors. Phylogenetic comparative methods (PCMs) are regression algorithms that account for phylogenetic signals in comparative trait data (Felenstein,1985). The phylogenetic signal is the tendency of closely related species to exhibit greater similarities in traits than other species when sampled randomly from the same phylogenetic tree. The strength and direction of the phylogenetic signal are used to infer whether trait variation exhibits signs of evolution due to genetic drift, stabilizing selection, or divergent or convergent evolution (Blomberg et al., 2003; Pagel et al.,1999). PCMs have been applied to interrogate macroevolutionary questions such as the evolution of fungal modes of nutrition (James, 2006), the evolution of physiological and behavioral traits in primates (Kamilar and Cooper, 2010), plant-pollinator co-evolution (Smith et al.,2010), trait evolution by adaptive radiation in reptiles and avians (Pichereira-Donoso et al., 2015; McEntee et al., 2018). However, the application of PCMs is rather limited in cross-species codon usage studies (Sharp, 2010; LaBella et al., 2019). Recent large-scale sequencing projects have advanced our understanding of fungal phylogeny (Grigoriev et al., 2014; Ahrendt et al., 2018), thereby broadening the scope for comparative studies.

Here, we aimed to detail the evolutionary and functional underpinnings of codon usage variation in Kingdom Fungi by analyzing transcriptomic and tRNA data from over 400 representative species that are distributed across 18 taxonomic classes and 6 major phyla (Spatafora et al., 2017). Principal component analysis of codon usage frequencies effectively separated the species into respective sub-kingdoms, with the rare codons AUA^{Ile} and GGG^{Gly} driving the codon-specific variation. Using phylogenetic reconstruction methods, we inferred the macroevolutionary processes, including adaptive mechanisms, that explain the change in codon usage and tRNA patterns over time. We also performed genome-level analyses to examine the relationship between codon usage, tRNA supply, and gene expression levels. Phylogenetic signals of codon frequencies and genomic tRNA abundance were weaker than expected by genetic drift and phylogenetic relatedness. Yet, most genomes converged toward translation bias, wherein the most abundant mRNAs are enriched with codons for major tRNAs, in contrast to the low abundant mRNAs having greater codon bias for minor tRNAs. Finally, given the prevalence of adaptive codon usage, we present a neural network, *Codon2Vec*, that directly takes the coding sequences as input to reliably predict expression (median accuracy of 83.8% ±0.05). Altogether, our results support that natural selection for the efficiency of mRNA translation is a conserved influence among fungi.

RESULTS

1. ***Codon usage bias is evolutionarily correlated with the usage of GC-ending codons***

We obtained whole transcriptomes and predicted tRNA genes from 459 species sampled from six out of the eight recognized fungal phyla (Methods). Namely, 52 species belonging to the four early-diverging phyla of *Chytridiomycota, Blastidiomycota, Zoopagomycota, Mucoromycota,* and 408 species from the two dikarya phyla *Basidiomycota and Ascomycota*. Dikarya is the more species-rich sub-kingdom comprising 98% of all fungi - but 90% of our dataset - and is characterized by a more complex sexual lifecycle (Stajich et al., 2009).

We measured the degree of codon usage bias by computing the effective number of codons, ENC, for each species (Wright, 1990). ENC ranges from 20 to 61, where 20 represents an extreme bias of using only one codon per amino acid, while 61 represents uniform synonymous codon usage, that is, no bias. The mean ENC values ranged from 32.8 (high bias) to 56.9 (weak bias). To visualize the macroevolutionary pattern of codon usage bias, we applied continuous maximum likelihood ancestral state reconstruction (Revell, 2013) that projected the species' ENC values onto a pruned phylogenetic tree. The ancestral reconstruction shows that the more biased genomes accumulate in the early-diverging lineages (Figure 1A), with the most codon-biased genomes occurring in *Neocallimastigomycota,* the earliest diverging class of free-living fungi (Berbee and Taylor, 2017). Also, there is more fluctuation in codon bias along the upper branches that slows down upon the divergence of *Agaricomycotina*, the largest class (~70%) in *Basidiomycota*. Similarly, species in *Ascomycota* exhibit less variation in their codon bias. Variation in the GC-content at the third codon position (GC3%) is closely linked to codon usage bias since all degenerate amino acids allow for silent G or C substitutions. The mean GC3% ranges from 10.6% to 85.1%, with a median of 57%. Overall, early-diverging fungi exhibit, on average, lower GC3% but more variability among individual values **(Figure 1A).**

Next, we assessed the evolutionary relationship between codon usage bias and GC3% using phylogenetic independent contrast (PIC). PIC regression corrects for phylogenetic non-independence by using the contrasts between nodes instead of the trait values directly (Garland et al., 1992). For the entire tree, the PIC model *ENC~GC3* yielded a negative coefficient of -13.51 (adjusted R^2 =10.9%, p-value=2.69e^{-12}). Because PIC is calculated without an intercept term, the R^2 coefficient is the square of Pearson's R correlation coefficient. Therefore, codon usage bias and GC3% are moderately correlated (Pearson's R=32.6%). Although it may be reasonable to assume the evolutionary GC3-bias is driven by the usage of G/C-ending codons, it was found that the usage of certain G/C-ending codons was negatively correlated with GC3-bias in some plants and prokaryotes (Palidwor et al., 2010). To evaluate the relationship between codon usage and GC3-bias, we computed the phylogenetic-corrected Pearson's correlation between individual codon frequencies (normalized for amino acid usage), GC3%, and ENC separately (Figure 1B). The usage of all G/C-ending codons is positively correlated with GC3%, whereas all A/U-ending codons are anticorrelated with GC3%. All G/C-ending codons negatively correlated with ENC, which means that the increase in usage of G/C-ending codons correlates with an increase in codon bias. Conversely, the usage of all A/U-ending negatively correlates with codon bias. Interestingly, we obtained different correlation values between codon bias and the normalized codon frequencies with and without phylogenetic correction. Without the correction, some A/U-codons positively correlate with codon bias, and some G/C-ending codons negatively correlate with codon bias (Figure 1C). This discrepancy suggests that codon frequencies also have a phylogenetic signal.

Since the relationship between codon bias and GC3% seems inverted for the early-diverging and dikaryic lineages (Figure 1), we split the tree into the separate sub-kingdoms and re-evaluated the phylogenetic correlation between codon usage bias and GC3%. Codon bias and GC3% are evolutionarily anticorrelated in the early-diverging subtree (coefficient=21.1, R^2 =32.6%, p-value=7.16e^{-6}, number of tips=51 species). In contrast, dikarya species are positively GC3%-biased (coefficient= -29.8, R^2 =49.3%, p-value<2.16e^{-16}; number of tips=367).

Fitting macroevolutionary models to codon usage bias

Phenotypic variation among extant species is a confluence of shared ancestry and responses to neutral and adaptive processes. Interspecies codon usage bias is widely held to explain by neutral drift (Grantham et al., 1980). To determine the pattern of evolution that best explains codon usage bias, we fitted 4 different likelihood models of macroevolution to the ENC and GC3 values: 1) Brownian motion (drift/random walk), 2) Ornstein-Uhlenbeck (fluctuating directional selection), 3) early-burst (exponential decrease in trait variation over time), and 4) delta (rate shifted Brownian motion) (Pennell et al., 2014). Notably, Brownian motion is the null hypothesis of genetic drift that models interspecies trait data as a random walk (Felestein, 1985). Based on the goodness-of-fit Akaike information criterion (AIC) scores, the early-burst model, which simulates adaptive radiation, best explained the phylogenetic variation of both codon bias and GC3% (Supp. Table 2). Macroevolution by adaptive radiation is characterized by higher rates of trait evolution early in a clade's history, followed by an exponential decline through time (Simpson, 1953).

2. Codon-level macroevolutionary analysis reveals codon frequencies as mostly deviate from genetic drift.

Since variation in the frequencies of synonymous codons underlies codon usage bias, we examined the macroevolutionary trends of the individual codons. First, we quantified transcriptome-wide relative synonymous codon usage (RSCU) of the 59 degenerate codons (Sharp et al., 1986). RSCU=1 means codons are used according to neutral or uniform expectation. Importantly, RSCU normalizes codon frequencies within their amino acid class which minimizes amino acid composition effects. To characterize which codons fungi generally prefer for making proteins, we quantified the most (highest RSCU) and least (lowest RSCU) preferred codons. Overall, C-ending codons consistently had the highest transcriptomic representation across the amino acid types (Figures S1A, S1C).

To summarize the variation of interspecies codon usage, we performed multivariate analysis using principal component analysis (PCA) on the 59 RSCU x 459 species matrix. The first two principal components explained 82% of the interspecies variation (Figure 2A). PC1 (78% explained variance) separated species according to differences in GC-content at the third codon position (GC3%), wherein loadings of G/C- and A/U- ending codons are equally but inversely correlated to PC1 (Figure 2B). This finding aligns with previous work that establishes variation in G+C content as the major determinant of interspecies differences in CUB (Chen et al., 2004; Novoa et al., 2019). The second principal component, PC2 (4.0% explained variance is driven by differences in individual codon frequencies, with the strongest signal due to the rare codons GGG^{Gly} and AUA^{Ile} (Figure 2B; S1B). Notably, PCA separated the species into their sub-kingdoms (Figure 2A).

The sub-kingdom clustering by the PCA led us to measure the extent to which phylogenetic effect (i.e., phylogenetic relatedness) underlies the choice of codon representation in the transcriptome. To this end, we computed the Blomberg's *K* statistic (Methods) of the normalized codon frequencies. Blomberg's *K* measures the strength and direction of trait evolution relative to that expected under the Brownian motion model that considers the phylogenetic distance as the only predictor of trait similarity among species (Blomberg et al., 2003). All codons reported statistically significant phylogenetic signals (p-values<0.05). However, the strength and direction of evolution, even among synonymous codons, varied (Figure 2C). 34/59 codons exhibited a low phylogenetic signal (K<1), suggesting variation due to convergent evolution (Revell et al., 2008; Kamilar and Cooper, 2013). 10 out of the 59 codons followed the expected Brownian process (K=1) of genetic drift. 15 out of the 59 codons exhibited high phylogenetic signal (K>1) indicative of either stabilizing selection or low rates of evolution (Blomberg et al., 2003). We also fitted different models of macroevolution to the individual codon frequencies. Like genomic CUB, adaptive radiation was the best fitting model for all the 59 degenerate codons (Supp. Data). Taken together, these findings highlight that individual codons follow different modes of evolution. Importantly, the frequencies of 49 out of 59 codons are not fully explained by phylogenetic relatedness that is expected under genetic drift.

3. Identification of phylogenetically rare tRNAs and strong evolutionary preference for Inosine34-modified tRNAs

Considering that the frequencies of most codons deviated from genetic drift, the next logical step was to analyze the tRNA gene sets since codon usage is widely believed to co-evolve with tRNA supply in several species (Sharp et al., 2010). Both the number of distinct tRNA anticodon types and total tRNA genes (tRNAome) vary widely across the 459 genomes under study. The median number of distinct anticodon types is 44, ranging from a maximum of 58 in *Ascobolus immerses* and a minimum of 18 in *Sporobolomyces linderae* (Figure S2A). Like all previously studied genomes, no species in our dataset possessed the full theoretical complement of 61 tRNA anticodon families (Marck and Grosjean, 2002). Interestingly, we identified 11 species possessing less than 30 anticodons, which is the theoretical minimum for decoding the standard genetic code (Marck and Grosjean, 2002). The median tRNAome is 144 genes, with a minimum of 24 and a maximum of 3481. *A.immersus* and *Melampsora allii-populina* both possess extreme tRNAomes of 3481 and 2216 genes, respectively.

Next, we measured the phylogenetic signal of the copy number for tRNAs that are cognate to the 59 degenerate sense codons. Like most codons, tRNAs also exhibited a phylogenetic signal that is lower than expected by drift, with K ranging from 0.02 to 0.70. These weak phylogenetic signals are consistent with tRNA gene dosage as evolutionarily labile (Heurto et al., 2010). 44 out of 59 sense tRNAs yielded a statistically significant phylogenetic signal (p<0.05). The lack of phylogenetic signal (p>0.05) in the remaining 15 tRNAs implies that they either evolved completely independent of phylogeny or are mostly absent in the fungal genomes since entire tRNA families are known to be extinct in certain clades (Rak et al., 2018). To this end, we identified 19 tRNA anticodon types that rarely occur among the fungal genomes, three of which are nonsense suppressors (tRNA^{Sup} (UUA), tRNA^{Sup} (CUA), tRNA^{SeC} (UCA)) (Figure 3A). 14 out of the 16 rare sense tRNAs overlapped with the 15 tRNAs that lacked phylogenetic signal (Figure 3B). Only tRNA^{Ile} (UAU) is prevalent yet lacks a phylogenetic signal. In other words, close relatives are no similar in their genomic copies of tRNA^{Ile} (UAU) than if they were randomly placed on the tree. This finding may be explained by highly accelerated birth-death evolution or anticodon shifts of tRNA^{Ile} (UAU) along the phylogeny (Velandia-Huerto et al., 2016).

Detection of selenocysteine-tRNAs in dikarya genomes

Here, we would like to report the detection of selenocysteine tRNA (SeC-tRNA). At the time of this finding, tRNAs corresponding to the 21st amino-acid selenocysteine were considered absent in all fungi (Lobanov et al., 2007) until Mariotti et al. uncovered the presence of tRNA^{SeC} (UCA) in nine early-diverging fungi (Mariotti et al., 2019). However,, all three of our Sec-tRNA positive fungi – *Rhodocollybia butyracea, Sugiyamaella americana,* and *Lollipopaia minuta* – are dikarya from *Basidiomycota* and *Ascomycota* phyla (Supplemental Table 3). We identified the presence of tRNA^{SeC} (UCA) in these three genomes based on overlapping results from at least one of the general-purpose tRNA gene finders, tRNAscanSE2.0 (Chan and Lowe, 2106) or aragorn1.2.38 (Laslett and Canback, 2004), and the specialized tRNA^{SeC} gene finder Secmarker (Santesmasses et al., 2017). As a negative control, we repeated the analysis on the well-studied fungal genomes of *S.cerevisiae* and *N.crassa*. Even with the unrealistically relaxed parameters, tRNA^{SeC} (UCA) was not detected in either genome.

Next, we examined the prevalence of inosine-modified tRNAs. Adenosine-to-inosine (6-deaminated adenosine) conversion is the most common post-transcriptional editing in eukaryotic RNAs (Nishikura, 2016). In eukaryotes, A34-to-I34 conversion is restricted among the eight tRNA types: tRNA^{Thr}(AGU), tRNA^{Ile}(AAU), tRNA^{Pro}(AGG), tRNA^{Arg}(ACG), tRNA^{Leu}(AAG), tRNA^{Ala} (AGC), tRNA^{Val}(AAC), and tRNA^{Ser}(AGA). Inosine-34 tRNAs (INN) decode both NNC and NNU codons in eukaryotes (Rafels-Ybern et al., 2017). Although both INN and GNN tRNAs decode C-ending codons, the I: C anticodon-codon bond is known to be less stable than the G: C bond (Hoernes et al.; 2018). Yet, we found that for the amino acids that are recognized by isoacceptor pairs of GNN and a putatively inosinylated ANN, the GNN iso-acceptor is mostly absent within the genomes (Figure 3C). For example, tRNA^{Leu} (GAG) is the rarest tRNA, being predicted in only 1/459 species (Figure 3A), yet its Watson-Crick cognate codon CUC is the commonly most preferred for encoding leucine (347/459 species; Supp. Figure 2A). Likewise, the usage of the AUC codon is frequently the most preferred for isoleucine, yet its cognate tRNA^{Ile} (GAU) is rarely present (16/459 species). According to wobble rules, both Leu-CUC and Ile-AUC codons are decoded by inosine-modified tRNA^{Leu} (AAG) and tRNA^{Ile} (AAU), respectively. In contrast, when the ANN tRNA is not a target of inosinylation, the GNN iso-acceptor is far more prevalent (Figure 3D). As previously mentioned, genomes in our dataset are mostly biased for NNC codons (S2A), so this finding suggests that inosine-modified tRNAs are positively selected for in fungi. To summarize, phylogenetic comparative analyses revealed that the interspecies variation of codon usage bias and individual codon frequencies do not support genetic drift as the dominant mode of evolution.

4. Signatures of neutral and adaptive evolution on intra-genomic codon usage bias

Having analyzed codon usage patterns at the macroevolutionary scale, we next sought to disentangle the signatures of adaptive and neutral evolution on within-genome codon usage bias. At the organismal level, codon usage bias is a composite of drift, neutral and adaptive mutational bias (dos Reis, 2009). To determine whether codon usage bias is driven solely by GC-compositional mutational bias in each species, we compared the empirical effective number of codons (ENC) of all coding sequences to their theoretical ENC that is expected under the neutral-mutational model. The neutral-mutational model is the null hypothesis that selection pressure does not act on the synonymous third codon position sites; rather, codon bias is only a function of GC3% (Wright 1990). The codon usage bias of 458 out of 459 species deviated significantly from neutral expectation (paired Wilcoxon signed-rank test p-values<0.05; Figure 4A). Next, we assessed each species' fit to neutral expectation by computing the R^2 between empirical and theoretical ENC of all coding sequences within each genome (Figure 4B). The R^2 values ranged from 0.0001 to 0.88. 70 genomes, mostly dikarya, reported an R^2 value of at least 0.75 ('Very Strong'), which indicates that their codon usage bias is largely influenced by neutral mutational bias. Notably, early-diverging species make up 12% of the dataset, but 28% of the genomes with 'Weak' neutral mutational bias.

Fungal transcriptome-wide codon usage and tRNA copy number are positively correlated

Another signature of natural selection is the correlation between codon frequencies and the supply of cognate tRNAs (Sharp et al., 2010). To explore this, we computed Spearman's rank coefficient between transcriptome-wide relative codon frequencies (RSCU) and tRNA gene copy number. 313/459 species yielded significant and positive correlations (p-values<0.05; Fig 4C). Those species with non-significant correlation tend to possess single copy tRNA gene sets. Across the species, the most overrepresented codons – highest transcriptome-wide RSCU - generally match tRNAs with higher copy number (mean tRNA gene copy number of 5.3) compared to most underrepresented codons (mean of 1.3 tRNA gene copy number, one-sided paired Wilcoxon signed ranked test p-value = $1.05e^{-54}$; Figure 4D).

In summary, we showed that variation in the genome-level codon usage bias is influenced by both neutral (GC3-composition) and adaptive mutational bias (cognate tRNAs).

5. Differential adaptation to the genomic tRNA abundance underlies expression-linked codon usage bias

Here, we explored the functional implications of adaptive codon usage bias by analyzing the contribution of gene expression to codon preferences. Because we have RNAseq data for most species in our dataset (432/459), we could empirically investigate expression-linked codon usage bias. To this end, we filtered the top 10% ('high') and bottom 10% ('low') of expressed coding sequences as the working dataset for each species since directional selection acts at the extremes. We observed that for most species, PCA on the 59-dimensional matrix of codon frequencies (RSCU) separated the genes according to expression level (Figure 5A; Supp Data). The trend suggests that gene expression level is a driver of codon usage patterns.

Not an artifact: PCA arch of high expressed genes caused by strong deviation from neutral compositional bias

Intriguingly, the pattern of only the high expressed genes clustering as an arch in *Z. heterogamous* also appeared in 27 other species (left panel Figure 5A; Supp Data). Guttman or arch effect in dimensionality reduction techniques, such as PCA or correspondence analysis, is observed when the first two transformed axes are curvilinear because the structure of the data is dominated by a single latent variable that gradually shifts from one extreme to another, i.e., the data points lie on a gradient (Diaconis et al., 2008; Ngyuen et al., 2019). Given that codon usage is influenced by both directional neutral and selection pressures, we hypothesized that the latent gradient underlying the high expressed genes represents the shift in influence of neutral to selection pressures. To explore this in *Z. heterogamous*, we generated an ENC-GC3 neutrality plot in which the standard curve is the expected relationship between ENC and GC3% when codon usage is solely explained by neutral compositional bias (Wright 1990). The neutrality plot confirmed our hypothesis as the genes become more C3-biased with increasing distance from the neutral curve (Figures 5B). However, the low expressed gene set did not exhibit such marked deviation from neutral codon usage bias (Figures 5C). Indeed, the latent gradient responsible for the arch is increased C3 usage and decreased A3 usage (Figure 5D; S3A-C). We identified a similar pattern for 19 of the remaining 27 'arch' species in which their high expressed genes also lie on a C3% gradient when projected onto the first two principal components (Figure S3G) and deviating from the expected neutral curve (Figure S3H). This led us to revisit the compositional bias analysis performed in the previous section, where we found that 24 out of these 28 species, including *Z.heterogamus,* fell in the 'Weak' category (Figure 4B, S3D). Additionally, 24/28 of them are AU3-biased (mean GC3 <50%), and 22 of them are early-diverging fungi (Figure S3D). Together, these results suggest that selection is particularly stronger in the highly expressed genes of these species.

High expressed genes preferentially use codons decoded by major tRNAs but avoid codons decoded by minor tRNAs

We then asked if the divergent codon preference between the high and low expressed genes is related to adaptation to translation efficiency (Duret, 2000). To this end, we measured the fraction of preferred and non-preferred codons that are decoded by major and minor tRNAs per species. Preferred codons are significantly enriched (higher RSCU) in the high expression gene set, whereas non-preferred codons are enriched (higher RSCU) in the low expression set (Benjamini-Hochberg adjusted p-values <0.05) (Yannai et al., 2018). We observed that C-ending codons are mostly preferred by high expressed genes compared to A-ending codons in low expressed genes (Figure S3E). Major and minor tRNAs have the highest and lowest copy numbers, respectively, within an amino acid class. Overall, highly expressed genes preferentially use codons that are decoded by major tRNAs, which is indicative of selection for rapidly translated codons (Figure 5E). On average, 43% of preferred codons are recognized by major tRNAs compared to 24% by minor tRNAs ((Figure 5E; paired Wilcoxon signed-rank test p-value = 4.02e^{-66}). Conversely, non-preferred codons better matched minor tRNAs (mean fraction = 34%) than major tRNAs (Figure 5F; mean fraction = 18%; Wilcoxon p-value=1.43 e^{-64};).

Because we identified the widespread preference for inosine-modified tRNAs, we extended our analysis to account for inosine-34 wobble decoding. This resulted in a marked increase in the mean fraction of preferred decoded by major tRNAs from 43% to 66% (Wilcoxon p-value = 7.96e^{-70}), but the mean fraction of preferred codons matching minor tRNAs remained the same. For example, the match rate in *Z.heterogamus* rose from 56% to 81%. In 84% of the species, at least 50% of their preferred codons are cognate to their major tRNAs when inosine34 decoding is considered (Figure 5E). However, the inclusion of I34 modification did not substantially alter the fraction of non-preferred codons decoded by minor tRNAs or major tRNAs (18%; Wilcoxon p-value = 3.86e^{-62}) (Figure 5F). Therefore, the codon bias in high expressed genes can be partially explained by selective usage of inosine34 decoded codons. These results align with experiments in mammalian and bacterial systems that demonstrated improved agreement between codon usage and tRNA abundance when I34 modification is accounted for and that transcripts with codon compositions that matched I34 tRNAs were more efficiently translated (Novoa et al., 2012).

Prevalence of codon optimization for translation in high expressed genes shows signs of convergent evolution

To quantify the association between expression-linked codon bias and adaptation to the tRNA supply in a genome, we derived the translation bias score (TBS) (Methods). Formally, TBS is the difference between the fractions of preferred codons and non-preferred codons for major tRNAs normalized by their sum. A TBS of +1 indicates that the codon bias of highly expressed genes confers them exclusive access to the most abundant tRNAs in the cellular pool compared to the lowest expressed genes; whereas a TBS of 0 means that there is no competition for major tRNAs between the high and lowest expressed gene sets. Among the 432 species analyzed, the mean TBS is +0.40 when restricted to Watson-Crick pairing, and a mean of +0.54 when I34 wobble is considered (Figure 5G). That is, on average, 54% more of the major tRNAs are decoding high expressed genes than the low expressed genes, which we interpret as selection for translation speed. However, there are a few species that possess negative TBS meaning their low expressed genes are more codon biased for major tRNAs.

We wondered if the skewness of translation bias scores (TBS) toward positive values is a consequence of phylogenetic relatedness since species richness is unevenly distributed along our fungal tree. To measure the strength of the phylogenetic effect of the TBS values, we computed Blomberg's K statistic, which yielded K=0.12 for Watson-Crick pairing and K=0.18 for inosine-34 wobble decoding (both p-values = 0.01). These low *K* values indicate that distantly related species have more similar translation bias scores than expected by phylogenetic distance, a pattern often attributed to convergent evolution (Revell et al., 2008). As a complementary approach, testing different macroevolutionary models supports the Ornstein-Uhlenbeck process of fluctuating directional selection as the best fit for translation bias (Figure 5H). Additionally, the ancestral state reconstruction shows that similarly high translation bias is distributed across multiple and distant lineages (Figure S3F). Therefore, both phylogenetic methods agree on adaptive codon usage bias - at least in the context of gene expression - as a realization of convergent evolution. Altogether, the concordance between tRNA supply and expression-linked codon preferences supports selection on codon usage for translation accuracy and speed in fungi.

6: *Codon2Vec* neural network for predicting the expression of coding sequences

Building predictive models for gene expression remains a pertinent challenge in genomics. Inspired by natural language processing models (Mikolov et al., 2013), we implemented a 3-layer neural network -Codon2Vec - that predicts expression class ('high' or 'low') directly from input coding sequences (Figure 6A; Methods). A neural network is a supervised algorithm that can model complex non-linear patterns that underlie the data. The first layer of Codon2Vec performs featurization of input sequences by mapping each codon type to a real-valued vector or 'embeddings' in Euclidean space. The codon embeddings are adjusted during model training to minimize the error between the predicted and ground truth labels.

To achieve a balanced dataset, we trained Codon2vec on the coding sequences from the top and bottom 10% expression. The training data was split into 70:20:10 for training: validation: test sets. Model selection was determined based on the training and validation sets. Final predictive performance was reported on the hold-out test set using the metrics: misclassification error, area-under-the-receiver-operator-characteristic curve (AUC-ROC), sensitivity, and specificity (Methods). An AUC-ROC of 0.5 indicates that a model failed at learning and instead makes random predictions. When applied separately to 300 different species, Codon2Vec achieved a high median AUC-ROC score of 83.8% (Figure 6C). Randomizing the association between the input and class labels ablated Codon2Vec's discriminative power and drove the AUC-ROC to 0.5 or random predictions (Figures 6C; S4D).

We hypothesized that the model's decision boundary is the differential codon bias that exists between the sequences in the high expression and low expression classes. To this end, we computed the difference between the mean ENC of the expression classes (DOM-ENC) and measured the Spearman's rank correlation between the species-specific AUC-ROC scores and the DOM-ENCs, but there was no significant correlation (R=0.1, p-value = 0.074). Since codons are the features, we repeated the procedure using the frequency of optimal codon (Ikemura 1981) (DOM-FOP). This resulted in a significant and positive correlation (Spearman's rank coefficient R=0.45, p-value= 1.05e^{-16}) (Figure 6D). We interpret this as the model performing better on genomes that have a wider margin of optimal codon content between the high and low expressed genes. As a sanity check to see if the length of coding sequences was a confounding variable, we found no significant correlation (R=0.1, p-value=0.0755). Remarkably, Codon2Vec learned the intrinsic differences in optimal codon content between high and low expressed genes even though we did not explicitly provide this property.

Discussion

Much of our understanding of codon usage is inferred from collating findings across single-species analyses. To better detail the evolutionary mechanisms that have shaped codon usage patterns through time, we employed a phylogenetic comparative approach to analyze hundreds of representative species that span the major phyla of Kingdom Fungi. We showed neglecting the phylogenetic effect can lead to different conclusions about the influence of individual codons on the degree of codon bias (Figure 1C). Our macroevolutionary analyses support, contrary to the widely held neutral-drift hypothesis, adaptive mechanisms as the driver of interspecies codon usage patterns in fungi. Fitting of various likelihood models of trait evolution to our 452-taxa phylogenetic tree showed that variation in codon usage bias and GC3% best fit the pattern generated by adaptive radiation. Additionally, the phylogenetic effect on most codon frequencies was found to be stronger or weaker than expected by random drift, a sign that is usually interpreted as stabilizing selection or convergent evolution respectively (Losos 2011, Revel et al., 2008). Adaptive codon usage was also evident in the genome-level analyses. Gene expression level and codon usage were broadly correlated as principal component analysis separated the highest and lowest expressed genes based on their codon usage patterns. In some species, primarily early-diverging, the deviation of high expressed genes from compositional bias was strong enough to dominate the signal captured by both principal components resulting in a Guttman effect. Since differential codon bias could arise by a neutral mechanism such as GC-biased gene conversion (Marais, 2003), we demonstrated how this prevalent pattern of expression-linked codon bias reflected differences in translation efficiency. Broadly, the high expressed genes preferentially used codons matching the most abundant tRNAs, whereas the low expressed genes were more biased for codons read by the least abundant tRNAs. Moreover, the pervasive trend of codon optimization of high expressed genes for translation efficiency, which we quantified using our translation bias scores, infers convergent evolution as the phylogenetic effect was significantly weaker than expected by Brownian motion trait evolution. Altogether, these findings are consistent with the influence of natural selection on codon usage to promote translation efficiency. Our results on the prevalence of adaptive codon usage bias in fungi are consistent with the recent sub-phylum-wide codon usage analysis of *Sacchoromycotina* budding yeasts (LaBella et al.,2019).

Macroevolutionary analyses of codon usage reveal the influence of adaptive mechanisms

Although claims about codon usage are usually based on single-species analyses, we believed that inferences about the mode and tempo of codon usage macroevolution would further elucidate the adaptive significance of codon usage patterns. Principally, we found that the tempo of codon usage bias (CUB), and the evolutionarily correlated GC3%, in our 452-taxa tree best follow the pattern of adaptive radiation. Other fungal phylogenomic studies, primarily in mushroom-forming (*Agaricomycetes*) lineages, have also reported evidence of adaptative radiation for certain morphological traits (Varga et al., 2018; Gaya et al., 2015; Nagy et al., 2012). Various hypotheses exist for the intrinsic and ecological drivers of fungal radiations, including the evolution of complex fruiting bodies (Varga et al., 2018), transition to mutualism (Sanchez-Garcia and Matheny, 2017), and defense mechanisms (Gaya et al., 2015; Nagy et al., 2012). Previous studies have linked CUB to ecological specialization (Roller et al., 2013; Botzman et al., 2011). Badet et al. uncovered that generalist parasitic fungi are more codon biased than non-parasitic fungi (Badet et al., 2017). Our finding raises the question of what ecological opportunities underlie the macroevolution of codon usage bias (CUB). Visual inspection of our ancestral reconstruction shows that the decrease in the variability of CUB coincides with the divergence within *Basidiomycota*. *Basidiomycota* (club fungi) comprises about one-third of all fungi (Stajich et al., 2009). Saprophytic *Agaricomycotina* accounts for two-thirds of basidiomycetic fungi, whereas the two minor but earlier diverged classes - *Puccinomycotina* and *Ustilaginomycotina* – are mostly plant parasites (Mao et al., 2019). Relatedly, ancestral reconstruction of fungal nutritional modes showed that parasitism is non-randomly distributed along the tree and more prevalent in earlier-diverged lineages (James et al., 2006). Taken together, the evolution of

CUB in fungi may be connected to lifestyle adaptation. We believe that a deeper study of the macroevolutionary relationship between CUB and the various ecological specialization in fungi is needed.

Selection for translation efficiency may explain convergent codon usage in fungi

Macroevolutionary analyses revealed that variation in synonymous codons is mostly convergent. The normalized frequencies of 34/59 codons yielded significantly low Blomberg's K values, indicating distantly related lineages are more similar in their codon choices than expected phylogenetic relatedness, i.e., convergence (Kamilar and Cooper, 2013). Causes of convergent evolution are generally attributed to shared constraints (molecular/ genetic/ physiological/ ecological, etc.) that limit or bias the production of phenotypic variation or, to a lesser extent, random chance (Losos, 2011a). Here, we reason that the macroevolutionary convergence of codon usage frequencies reflect the shared constraints imposed by neutral - for example, GC-compositional bias - and adaptive pressures, for example, selection for balancing codon representation with tRNA supply (Figures 4B, 4C, 5E). Moreover, the maximum likelihood best fits the translation bias scores to the Ornstein-Uhlenbeck process of optima-directed trait evolution (Butler et al., 2004), lends further evidence that adaptative mechanisms have influenced codon usage patterns over time. That is, the convergence of highly expressed genes being codon-biased for the most abundant tRNAs compared to low expressed genes suggests that efficient protein synthesis is one of the selective optima that has constrained fungal codon usage. This assertion is consistent with genome-engineering experiments that first demonstrated how codon optimization in highly expressed genes exerts global effects on cellular fitness by promoting rapid turnover of free ribosomes enabling translation initiation on other transcripts (Frumkin et al., 2018).

Selection for translation efficiency is expected to favor those codons with better anticodon-codon pairing kinetics (Higgs and Petrov, 2008). Possibly, the weaker I: C anticodon-codon bond (Hoernes et al., 2018) promotes faster dissociation of the discharged tRNA from the ribosomal E-site, leading to less ribosome pausing and more available free ribosomes. This model may explain the conserved preference for inosine-modified tRNAs (INN) over GNN isoacceptors (Figure 3C), especially the general bias for tRNA^{INN} decoded codons – primarily C-ending – observed in high expressed genes (Figure 5E). In light of this, a component of the general GC3-bias among fungi may actually be a C3-bias due to selection for INN tRNAs.

Here, we highlight the limitations of our study and potential areas for improvement. We assumed that the tRNA concentration scale with tRNA copy number, which is the general case in unicellular organisms, for example, chromatin profiling in *S.cerevisiae* revealed all tRNA genes as transcriptionally active (Harismendy et al., 2003). Like all statistical models, inference from phylogenetic comparative methods (PCM) is constrained by assumptions and uncertainty. The main assumptions of PCMs are: 1) the phylogenetic tree is accurate, 2) all the extant taxa are represented 3) there is measurement error in the trait data. Our 452-taxa tree does not preclude biased inference due to uneven taxon sampling since we used a limited number of representative species per clade. While we are using the best available molecular tree, given the vastness of the kingdom Fungi and ongoing sequencing campaigns, we foresee updates in the fungal phylogeny (Ahrendt et al., 2018). Lastly, various evolutionary processes may give rise to the same phylogenetic pattern, and current macroevolutionary models may be limited in their capability to capture more complex patterns of trait evolution (Losos et al., 2011b).

In a minority of species, the low expressed genes were more codon biased for the major tRNAs (Figure 5G). This rather counterintuitive finding joins two previous works that challenge the default view that selection is reserved for codon usage of highly expressed genes (Zhou et al., 2009; Yannai et al., 2018). Codon usage in low expressed CDS may be influenced by selection for mRNA structure, mRNA stability to support sufficient protein production, or co-translational protein folding of structural sites that are sensitive to translation speed or accuracy (Zouridis et al., 2008; Zhou et al., 2009). Other than translation efficiency, the co-variation between codon usage bias and gene expression levels may reflect selection for mRNA stability as certain codons mitigate ribosomal stalling, as observed in *S.cerevisiae* (Presynak et al., 2015), or linked transcriptional efficiency as seen in *N. crassa (*Zhou et al., 2016; Zhao et al., 2021).

We also identified fungi possessing less than the theoretical minimum of 30 tRNA anticodons required for standard mRNA translation (Marck and Grosjean et al., 2002). Interestingly, 7 out of these 11 species are mutualistic symbionts - *Sporobolomyces linderae, Cenococcum geophilum, Meliniomyces bicolor, Neocallimastix californiae* – or pathogens/parasites - *Teratosphaeria nubilosa, Mixia osmundae, Elsinoe ampelina*. Perhaps their reduced tRNA gene set reflect lifestyle adaptations such as selection for rapid DNA replication, or importing necessary tRNA molecules from the host - a rare mechanism for eukaryotes that was first observed in plasmodium parasites (Bour et al., 2016). This mechanism may explain how *Mixia osmundae* maintains survival as a biotrophic intracellular parasite in plants (Toome et al., 2014)*.* Also, these minimalist fungi may also employ promiscuous super-wobbling decoding, as observed in plastomes of vascular plants (Rogalski et al., 2008). Therefore, these minimalist symbionts would make ideal candidates for studying non-standard translation of the genetic code and the co-evolution of decoding strategies within a eukaryotic host-symbiont pair.

Codon2Vec: Addition of a sequence-to-expression model to the functional codon usage toolkit

We believe the value of Codon2Vec is twofold. Because the model is trained on whole coding sequences, it learns a more biologically meaningful representation of the codon usage patterns. Codon order, such as codon-pair bias, has been shown to also contribute to the protein yield of highly expressed genes (Cannarozzi et al., 2010; Gamble et al., 2016). But standard codon-based approaches are limited in capturing effects due to codon frequency. Because the model algorithm represents codons as vectors ('embeddings') in Euclidean space, in principle, contextually related codons are projected close together in embedding space (Mikolov et al., 2013). Secondly, unlike standard approaches, Codon2Vec is not restricted to a pre-defined reference set of genes. Moreover, Codon2Vec bypasses the need for artisanal feature selection since it extracts information directly from sequences and expression data, and the function that maps codons to real-valued vectors is also learned during training. Although neural networks are regarded as decision ' black-boxes,' we showed that the model is at least learning to classify coding sequences based on differences in codon optimality. However, neural networks can learn complex functions, so there may be other sequence/codon usage properties that it may have learned. Embedding neural networks have also been used for other biological applications, like predicting chemical and physicochemical properties from protein sequences (Yang et al., 2018) and gene annotation (Du et al., 2018). Once trained on the host's gene expression data, Codon2Vec can then serve as an oracle to guide the codon optimization of exogenous genes. A nice follow-up would be to experimentally validate Codon2Vec's predictions in optimizing heterologous gene expression systems.

In conclusion, by combining genomics and macroevolutionary analyses, we characterized the significance of and prevalence of adaptive processes in shaping fungal protein-coding genes. In the age of 'big genomics data,' it would be interesting to see if similar macroevolutionary modes and mechanisms explain interspecific codon usage variation in other clades.

FIGURE LEGEND

Figure 1: Inferring the tempo and mode of the evolution of codon usage bias and GC3% in fungi

Ancestral reconstruction of the codon usage bias, measured by the mean effective number of codons (ENC) and GC3-content projected onto a pruned fungal phylogenetic tree (number of tips =417 species). Color gradient represents the trait values for species at the tips and estimated trait values for internal nodes. Species with higher codon usage bias, i.e., lower ENC and low GC3%, primarily accumulate in the early-diverging lineages. The size of the tree obscures tip labels, but greater details are available in supplementary data. Photo credits: https://mycocosm.jgi.doe.gov/mycocosm/

B: Cluster heatmap showing phylogenetic corrected Pearson’s R correlation between normalized codon usage, GC3-content (GC3%), and ENC, which is the inverse of codon bias. G/C-ending codons are all positively correlated with GC3% and codon bias, whereas A/U-ending codons are all negatively correlated with GC3% and codon bias.

C: Scatterplot showing Pearson’s R correlation coefficients between individual codon frequencies and codon bias (ENC) with and without correcting for phylogenetic signal.

Figure 2: Phylogenetic analysis of transcriptome-wide codon usage frequencies

A: Principal component analysis on RSCU matrix clusters species into the dikarya and early-diverging sub-kingdoms, primarily along the axis of PC2. Each dot is a species whose color and shape represent its sub-kingdom.

B: PCA biplot of the loadings showing the contribution of each codon to PC1 and PC2. Codons are colored based on the G/C or A/U composition of the third base. A/U and G/U-ending codons equally but inversely contribute to the PC1 score of a species. On the other hand, differences in PC2 species scores primarily correlate with the usage of the rare codons, GGG^{Gly} and AUA^{Ile} (see also S1D).

C: Stripplot shows the variation in the phylogenetic signal of the frequencies of degenerate codons. 49/59 codons reported Blomberg’s K not equal to 1, suggesting they evolved at a rate less than or greater than the rate expected by genetic drift modeled by Brownian motion (BM). All p-values are statistically significant (<0.05).

Figure 3: Analysis of tRNA gene composition across the phylogeny

A: Prevalence of each tRNA anticodon type across the 459 fungal genomes based on tRNAscanSE2.0 predictions. Low-quality tRNA and pseudogenes with a covariance score below 50.0 are not included. Rare tRNAs are highlighted in red.

B: Overlap between tRNAs rarely present in fungal genomes and tRNAs with non-significant phylogenetic signal (p-value >0.05 for Blomberg's K). $tRNA^{Ile}$(UAU) is the only phylogenetically non-significant tRNA that is not rare. This suggests that this tRNA gene's evolution is de-coupled from phylogeny.

C, D: Evolutionary bias for the inosine-34 modification**.** For amino acids decoded by both ANN and GNN tRNAs, when the first anticodon position is a target of A-to-I editing, the INN tRNAs are more prevalent while the GNN isoacceptor is rare. **D**: However, if the ANN tRNA is not a target of A-to-I editing, then the GNN isoacceptor is more prevalent, and the ANN is rare.

Figure 4: Signatures of mutational bias and natural selection on within-genome codon usage bias

A: Deviation of genomic ENC values from Wright's neutral mutational model**.** The outer histogram shows the distribution of the mean difference between the empirical ENC and theoretical ENC that is expected by GC3-compositional bias measured in each of the 459 species. Inset displays the Wilcoxon signed-rank p-values (log base 10) that measure the significance of deviation. 458/459 species reported significant p-values (right of red dashed line). Both y-axes represent the number of species.

B. Variation in the influence of neutral pressures on species' codon usage bias. Swarmplot of species' R^2 values (n points =459 species) that measures the fit between empirical codon usage bias ('ENC_obs') and theoretical codon usage bias ('ENC_theo') expected solely due to GC3-compositional bias, grouped by 'Very Strong' (R^2 >=0.75), 'Strong' (0.75> R^2 >=0.5), 'Moderate' (0.5 > R^2 >=0.25) and 'Weak' (R^2<0.25).

C: Codon frequency correlates with tRNA copy number. The histogram shows the distribution of Spearman *ρ* correlation coefficient between RSCU and cognate tRNA gene copy number. 313/459 species reported statistically significant p-values (p<0.05). The black line is the mean correlation coefficient.

D: Histogram showing differences in the tRNA copy numbers for the codons with highest and lowest representation (RSCU) in the transcriptome over all the 459 species **Inset**: Distribution of the copy numbers of tRNAs (y-axis) for the codons with highest (red) and least (blue) transcriptome-wide RSCU. The most frequently used codons are decoded by tRNAs with higher copy numbers, whereas the least frequent codons are decoded low copy-number tRNAs.

Figure 5: Expression-linked codon usage bias correlates with tRNA supply

A: High and low expressed genes exhibit different codon usage patterns. Example of PCA applied to the codon usage 59-dimensional RSCU matrix of the top (‘high’) and bottom (‘low’) 10% expressed genes. Each dot is a gene. The right panel is the more common cluster pattern. The left panel depicts Guttman (‘arch’) effect in the high expressed genes of *Z.heterogamus*.

B-D: ENC-GC3 plot of *Z.heterogamus* genes elucidates source of PCA arch effect. The solid red line in scatterplots B and C represents the expected curve when codon usage bias is only affected by neutral mutation pressure. **B)** Codon usage bias of high expressed genes deviate strongly from neutral mutation pressure while becoming more C3-biased but **C)** low expressed genes better fit expected neutrality. **D)** Arch effect captured the variation in C3% due to neutral and selection pressures.

E and F: Highly expressed genes are biased for translationally optimal codons. E) Fraction of preferred codons (significantly enriched in the top 10% expressed genes) that are decoded by major (most abundant) and minor (least abundant) tRNA isoacceptor per genome (n=432 species) with and without the inclusion of inosine-34 modification. F) Paired comparison of the fraction of non-preferred codons (significantly enriched in the bottom 10% expressed genes) recognized by major and minor tRNAs.

G: Distribution of the translation bias scores (TBS) across all 432 species. TBS is the difference between the fraction codons enriched in the 10% highest versus lowest expressed transcripts normalized by their sum. The positively skewed distribution indicates that the highest expressed transcripts are generally codon-biased for rapid translation.

H: Akaike information criterion (AIC) goodness-of-fit evaluation of the various maximum likelihood macroevolutionary models fitted to translation bias scores (n species tips=396). ‘BM’=Brownian motion, ‘OU’= Orhnstein-Uhlenbeck. The OU model of fluctuation directional selection yielded the lowest AIC, and therefore the best fit model.

Figures 6: Neural network uses codons as features to predict gene expression.

A: Schema of Codon2Vec. Codon2vec is a fully connected multi-layer neural network that uses an embedding layer to transform codons in the input coding sequences to a real-valued vector. The final output of the model is a vector of probabilities for each gene expression class (i.e., “high” or “low”). Detailed description in Methods.

B: Codon2Vec’s performance on a single species. Model performance is evaluated on the hold-out test sets based on the area under the curve of the receiver-operating-characteristics curve (AUC-ROC). An AUC score of 0.5 (dashed line) represents random predictions.

C: Model generalizability: Codon2Vec achieves high predictive performance on 300 different species. However, shuffling of ground truth labels independent of input sequences ablated Codon2Vec’s ability to learn meaningful associations.

D: Codon2Vec‘s prediction accuracy (AUC-ROC) positively and significantly correlates with differential usage of optimal codons between high and low expressed genes. The frequency of optimal codons (FOP) is another standard CUB metric (Ikemura 1981; Methods).

MATERIAL AND METHODS

Genomic Data Acquisition

All 459 whole fungal transcriptomes, as well as the corresponding RNA-sequencing expression data and eukaryotic cluster of orthologous (KOGs) annotations, were downloaded from the Joint Genome Institute's Mycocosm database (https://mycocosm.jgi.doe.gov). Only coding sequences (CDS) longer than 150bp with annotated start and stop codons were retained for downstream analysis.

tRNA gene prediction: tRNA genes were predicted with tRNA-scanSE2.0 (Lowe and Chan 2016) with eukaryotic-specific parameters. For quality control, only high-confidence tRNA genes with a covariance score of at least 50 were retained for analyses. tRNA gene copy number was used as the proxy for tRNA abundance.

Seleno-cysteine tRNA identification

High-confidence tRNA genes are assigned a tRNA-scanSE score of 50 and over. After applying the cut-off covariance score of 50, 6 genomes still retained high-scoring Sec-tRNA genes. To independently validate these tRNAscanSE predictions, these genomes were re-analyzed with another highly accurate but more conservative general-purpose tRNA gene finder aragorn1.2.38 (Laslett & Canback, 2004) using eukaryotic-specific parameters and a Sec-tRNA specific gene finder Secmarker (2015 Guigo). The final SeC positive species were taken as an overlap of any of these general gene-finders with the specialized Secmarker program.

Codon Usage Metrics

The effective number of codons (ENC), which measures the degree of synonymous codon bias of gene or genome, was computed from coding sequences using CodonW 1.4.4 (Peden, 1995; Wright 1990).

The theoretical ENC is the expected value estimated solely based on GC3% due to neutral mutational bias. The theoretical ENC of a gene *g* that is only influenced by GC3-compositional bias was computed according to (Wright 1990) using custom python3 code.

$ENC_{theo} = 2 + GC3_g + (29 / GC3_g^2 + (1-GC3_g)^2)$

G+C composition at 3rd codon position (GC3-content): GC3%was computed using CodonW 1.4.4 (Peden 1995)

Relative synonymous codon usage (RSCU) is the ratio of observed usage to the expected uniform usage within its amino acid class. RSCU is invariant to sequence length or amino acid composition. The RSCU of the 59 degenerate codons was computed using custom python3 scripts according to (Sharp and

Li,1987). Six-fold amino acids (Leucine, Serine, Arginine) were split into 2-fold and 4-fold codon groups.

Preferred and Non-preferred codons were selected in each species based on the top and bottom 10% of expressed coding sequences. A codon is considered preferred if its RSCU value was significantly higher in the highly expressed CDS set (Mann-Whitney *U* test, Benjamini-Hochberg adjusted p-value <0.05). Conversely, a non-preferred codon reported a significantly higher RSCU in the low expressed CDS set. With this definition, more than one synonymous codon of an amino acid may be preferred or non-preferred.

Frequency of optimal codons (Fop): Fop was computed according to (Ikemura 1981) using custom python3 scripts based on optimal codons derived from a reference set of top 30 highly expressed ribosomal genes.

Translation Bias Score, TBS: We introduce our TBS to measure the extent to which the codon usage of an organism’s gene expression reflects adaptation for the cellular tRNA supply based on the equation:

((fraction of preferred codons decoded by major tRNAs) – (fraction of non-preferred codons decoded by major tRNAs)) / (sum of the fractions)

Major tRNAs are the most abundant tRNA isoacceptor within an amino acid class, either based on gene copy number or tRNA concentration. In this paper, we used gene copy number as a proxy for cellular tRNA concentrations.

Comparative Phylogenetic Calculations

We downloaded from the fungal phylogenetic tree from Joint Genome Institute Mycocosm (https://mycocosm.jgi.doe.gov). The phylogenetic tree was then pruned using Dendropy package (v4.40) in python3.7. Taxonomic ranks were obtained from National Center for Biotechnology Information (NCBI).

Phylogenetic Independent Contrasts (PIC) models the trait covariation according to the formula Y = β X +ε, where Y and X are traits and β is the evolutionary correlation coefficient that quantifies the degree of coevolution between traits X and Y. Phylogenetic independent contrast was computed with the picante package in R (Kembel et al., 2010).

Both Maximum likelihood continuous ancestral trait reconstruction was performed using contMap(model=Brownian motion) function from the package phytools (Revell 2011). Blomberg’s K statistic were computed using the phylosig() function in phytools library (Revell 2012) implementation in R which p-values are calculated based on 100 permutations.

Fitting and evaluation of maximum likelihood evolutionary models for continuous character evolution were performed using the geiger library (Pennell et al., 2014) in R.

Correlation, Hypothesis and Multivariate analyses

All correlation and significance tests were done using the scipy (v1.3.1) and statsmodels(0.10.1) libraries in python3.7. Principal component analysis (PCA) was performed using the scikit-learn (v0.21.3; Pedregrosa et al., 2011) in python3.7.

Supervised Neural Network Codon2Vec

Codon2Vec is an artificial neural network (ANN) that learns the species-specific dependency between the codon composition (features) of a coding sequence (CDS) and expression level. A neural network is a class of machine learning algorithms that uses layers of interconnected computation nodes to learn complex patterns that underlie the data. We implemented and trained Codon2Vec using the keras (v2.2.4) (Chollet 2015) with tensorflow v1.8 backend, and scikit-learn libraries in Python3.7.

Available as a command-line tool for download at this github repository.

Dataset Collection and Preprocessing

We selected CDS from the top and bottom 10% of expression distribution relative to the mean expression. Each CDS was represented as a vector of codons in sequence. Then each unique codon is assigned a unique integer (*tokenization)* such that each CDS becomes recoded as a vector of integers. Finally, the lengths of the CDS were set to a fixed size of 2000, either by trimming longer sequences or padding with zeroes. The input and output data were shuffled and partitioned into 70% training set, 20% validation set and 10% test set. The training set is used to learn the model weights, whereas the validation set is used to fine-tune the model's generalizability by evaluating whether the model is over- or under-fitting on data it was not trained on. The final evaluation is performed on the test set.

Model training

Codon2Vec is a feedforward ANN with 3 fully connected computation layers - embedding layer, rectified linear unit (ReLu) activation layer and sigmoid output layer. We also incorporated "drop-out" regularization to reduce overfitting. Each layer is described in more detail in the subsequent paragraphs.

Learning optimized model weights

A node is the fundamental computation unit of an ANN. In a fully connected ANN, all nodes of a layer receive the *weighted* output of each node from the previous layer. The weights (**W**) represent the relative importance of a node to the model performance. During the forward pass of training, the input (**X**) undergoes a series of matrix multiplications and non-linear transformations (ϕ) as it flows sequentially between nodes in each layer until the predicted output is generated in the final layer.

Generating prediction: $\hat{\boldsymbol{Y}} = \phi\,(W^T X)$, where X and W are matrices

Model weights (**W)** were randomly initialized based on the Glorot uniform distribution. We chose the binary cross-entropy loss as the optimization objective that computes the error between the predicted output ($\widehat{Y}$) and ground truth (**Y**).

Loss(Y, $\widehat{Y}$) = - (Y * log($\widehat{Y}$) + (1 - Y) * log(1 - $\widehat{Y}$))

where ($\widehat{Y}$) ∈ {0,1} and Y is binary encoded as 0 or 1

In the backward pass of training, the contribution of the current set of weights (W) to the model error is computed by taking the partial derivative of the loss function with respect to each layer's weights (W). The weights are then updated by their gradients in the direction that minimizes the loss function. Weights were tuned via backpropagation using the Adam optimization, a variant of stochastic gradient descent based on adaptive learning (Kingma et al., 2014).

Model architecture

The first layer serves as feature extraction by learning the weights that map each of the 64 unique codon features to a meaningful dense real-valued 4-dimensional vector (embeddings) in Euclidean space. The number 4 is a hyperparameter. The advantages of embedding representation are: 1) the features that maximize model performance are learned directly from the CDS 2) the numerical transformation of codons makes modeling amenable to neural networks, and 2) it is more computationally efficient than the alternative one-hot representation as each codon would have been assigned a 1x64 dimensional sparse vector of mostly zeroes compared to Codon2Vec's 1x4 dimensional dense vector.

The second layer applies the ReLu activation function to the weighted sum of outputs from the embedding layer. The ReLu function is a widely preferred non-linear transformation for the inner (hidden) layers of neural networks because it speeds up convergence, and is robust to vanishing gradients.

ReLu (x) = max(0, x)

Finally, the output layer applies the sigmoid function that maps the continuous values to a real value between 0 and 1, such that the final output is a 2-dimensional vector of the prediction probabilities for each expression class.

Model Evaluation

We evaluated Codon2Vec's predictive performance using misclassification error, sensitivity, specificity and precision on the test set. Let TP, TN, FP, FN denote true positives, true negatives, false positives and false negatives, respectively:

Misclassification error = 1 - (TP+TN) / (TP+TN+FP+FN)

Sensitivity = TP/(TP+FN)

Specificity = TN / (TN+FP)

Precision = TP/ (TP+FP)

Furthermore, we plotted the receiver-operating characteristic curves and calculated the area under the ROC curve (AUC-ROC).

Data availability

All custom python3 and R scripts will be available https://github.com/rhondene/Adaptive_Codon_Usage_Kingdom_Fungi

Acknowledgments

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Wint R, Salamov A, Grigoriev IV. 2022. Kingdom-wide analysis of fungal transcriptomes and tRNAs 1235 reveals conserved patterns of adaptive evolution. Mol Biol. Evol. 1236 https://doi.org/10.1093/molbev/msab372
2. Ahrendt SR, Quandt CA, Ciobanu D, Clum A, Salamo, A, Andreopoulos B, Cheng JF, Woyke T, Pelin A, Henrissat B, Reynolds, NK, Benny GL, Smith ME, James TY, Grigoriev IV, et al.,. 2018. Leveraging single-cell genomics to expand the fungal tree of life. *Nat Microbiol.* 3:1417–1428.
3. Badet T, Peyraud R, Mbengue M, Navaud O, Derbyshire M, Oliver RP, Barbacci A, Raffaele, S. 2017. Codon optimization underpins generalist parasitism in fungi. *eLife*, 6:e22472. 2
4. Beaulieu JM, Jhwueng D-C, Boettiger C, O'Meara BC. 2012. Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*. 66: 2369–2383.
5. Berbee ML, James T Y, Strullu-Derrien C. 2017. Early diverging fungi: diversity and impact at the dawn of terrestrial life*. Ann. Rev. Microbiol*. 71:41–60.
6. Blomberg SP, Garland T, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution.* 57:717–745.
7. Botzman M, Margalit H. 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles*. Genome Biol*. 1210: R109.
8. Bour T, Mahmoudi N, Kapps D, Thiberge S, Bargieri D, Ménard R, Frugier M. Apicomplexa-specific tRip facilitates import of exogenous tRNAs into malaria parasites. 2016. *Proc Natl Acad Sci USA*. 11317:4717-4722.
9. Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*. 1293: 897–907.
10. Butler MA, King AA. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *Am Nat*. 2004;164: 683–695.
11. Butler MA and King AA. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am. Nat.* 164:683–695.
12. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. *Cell* 1412:355–367.

13. Chan PP, Lowe TM. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* 44D1: D184–D189.

14. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. USA.* 101:3480–3485.

15. Chollet F. 2015. keras. *GitHub*. https://github.com/fchollet/keras

16. Disconis P, Goel S, Holmes S. 2008. Horseshoes in multidimensional scaling and local kernel methods. *Ann Appl Stat*. 2:777– 807.

17. dos Reis M, Wernisch L. 2004. Estimating translational selection in eukaryotic genomes. *Mol Biol Evol*. 262: 451–461.

18. Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet*. 167:287–289.

19. Felsenstein, J. 1985 Phylogenies and the comparative method. *Am Nat.* 125, 1–15

20. Field KJ, Rimington WR, Bidartondo MI, Allinson KE, Beerling DJ, Cameron DD, Duckett JG, Leake JR, Pressel S. 2015. First evidence of mutualism between ancient plant lineages Haplomitriopsida liverworts and Mucoromycotina fungi and its response to simulated Palaeozoic changes in atmospheric CO2. *New Phytol*. 205: 743-756.

21. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. 2018.Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci USA.* 1521: E4940-E4949.

22. Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. 2016. Adjacent codons act in concert to modulate translation efficiency in yeast. *Cell*. 166(3):679-690.

23. Garland Jr T, Harvey PH, Ives AR. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* 41:18-32.

24. Gaya E, Fernández-Brime S, Vargas R, Lachlan RF, Gueidan C, Ramírez-Mejía M, Lutzoni F. 2015. The adaptive radiation of lichen-forming Teloschistaceae is associated with sunscreening pigments and a bark-to-rock substrate shift. *Proc Natl Acad Sci U S A.* 112(37):11600-11605.

25. Gingold H, Pilpel Y. 2011. Determinants of translation efficiency and accuracy. *Mol Syst Biol*. 7:481.

26. Grantham R, Gautier C, Gouy M, Mercier R, Pav A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acid Res*. 81: r49–r62.

27. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I. 2014. MycoCosm portal: Gearing up for 1000 Fungal Genomes. *Nuc Acids Res*. 421: D699-D704.

28. Harismendy O, Gendrel CG, Soularue P, Gidrol X, Sentenac A, Werner M, Lefebvre O. 2003. Genome-wide location of yeast RNA polymerase III transcription machinery. *EMBO J*. 22(18):4738-4747.

29. Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet*. 42:287–299.

30. Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol. 2511:2279–2291.

31. Hoernes TP, Faserl K, Juen MA, Kremser J, Gasser C, Fuchs E, Shi X, Siewert A, Lindner H, Kreutz C, Micura R, Joseph S, Höbartner C, Westhof E, Hüttenhofer A, Erlacher MD*.* 2018. Translation of non-standard codon nucleotides reveals minimal requirements for codon-anticodon interactions. *Nat Comm.* 9:4865.

32. Ikemura T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol*. 1584:573-597.

33. James T, Kauff F, Schoch C., *et al.,.* 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature.* 443, 818–822.

34. Janbon G, Quintin J, Lanternier F, d'Enfert C*.* 2019. Studying fungal pathogens of humans and fungal infections: fungal diversity and diversity of approaches. *Genes Immun.* 20: 403–414.

35. Kamilar JM, Cooper N. 2013. Phylogenetic signal in primate behaviour, ecology and life history. *Phil Trans R Soc B*. 368: 20120341.

36. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*. 26:1463–1464.

37. LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas. A 2019. Variation and selection on codon usage bias across an entire subphylum. *PLoS Genet*.157: e1008304.

38. Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*. 321:111-116.

39. Lobanov AV, Fomenko DE, Zhang Y, Sengupta A, Hatfield DL, Gladyshev VN. 2007. Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol*. 8:R198

40. Losos JB. 2011. Convergence, adaptation, and constraint. *Evolution*. 65, 1827-1840.

41. Losos, J.B. 2011. Seeing the forest for the trees: the limitations of phylogenies in comparative biology: American Society of Naturalists Address. *Am. Nat*. 177:709–727.

42. Mao H and Wang, H. 2019. Resolution of deep divergence of club fungi phylum Basidiomycota*. Synth. Syst. Biotechnol*. 4:225–231.

43. Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 196:330–338.

44. Marck C, Grosjean H. 2002. tRNAomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*. 810:1189–1232.

45. Mariotti M, Salinas G, Gabaldón T, Gladyshev VN. Utilization of selenocysteine in early-branching fungal phyla*.* 2019. Utilization of selenocysteine in early-branching fungal phyla. *Nat Microbiol.* 4:759–765.

46. McEntee JP, Tobias JA, Sheard C, Burleigh JG. 2018. Tempo and timing of ecological trait divergence in bird speciation*. Nat Ecol Evol*. 2:1120–1127.

47. Nagy LG, Házi J, Szappanos B, Kocsubé S, Bálint B, Rákhely G, Vágvölgyi C, Papp T. 2012. The evolution of defense mechanisms correlate with the explosive diversification of autodigesting Coprinellus mushrooms Agaricales Fungi. *Syst Biol* .61:595–607

48. Naranjo-Ortiz MA, Gabaldon T. 2019. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biol Rev Camb Philos Soc.* 94: 2101-2137.

49. Nguyen LH and Holmes, S. 2019. Ten quick tips for effective dimensionality reduction. *PLoS computational biology*. 15(6): e1006907.

50. Nishikura, K. 2016. A-to-I editing of coding and non-coding RNAs by ADARs. Nat. Rev. *Mol. Cell Biol.* 17, 83–96.

51. Novoa EM, Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. 2012. A role for tRNA modifications in genome structure and codon usage. *Cell.* 1491:202-213.

52. Novoa EM, Jungreis I, Jaillon O, Kellis M. 2019. Elucidation of Codon Usage Signatures across the Domains of Life. *Mol Biol Evol.* 3610:2328-2339.

53. Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature*. 401:877–884.

54. Palidwor GA, Perkins TJ, Xia X. 2010. A general model of codon bias due to GC mutational bias. *PLoS One*. 5:13431.

55. Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol*. 20(2):237-243.

56. Peden JF. 1999. Analysis of codon usage. PhD Thesis, University of Nottingham.

57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: Machine Learning in Python. *JMLR*; 12: 2825-2830.

58. Pennell M and Harmon LJ. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann. N. Y. Acad. Sci*. 1289:90–105.

59. Pennell MW, Eastman JM Slater GJ, Brown JW, Uyeda JC, Fitzjohn RG, Harmon LJ. 2014. Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*15, 2216–2218.

60. Pichereira-Donoso D, Harvey LP, Ruta M. 2015. What defines an adaptive radiation? Macroevolutionary diversification dynamics of an exceptionally species-rich continental lizard radiation. *BMC Evol Biol*. 15:153.

61. Powell JR, Moriyama E. 1997. Evolution of codon usage in *Drosophila*. *Proc Natl Acad Sci USA.* 94:7784–7790.

62. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J. 2015. Codon optimality is a major determinant of mRNA stability. *Cell*. 160:1111–1124.

63. R Development Core Team. 2011. R: A Language and Environment for Statistical Computing.

64. Rafels-Ybern A, Torres AG, Grau-Bove X, Ruiz-Trillo I, Ribas de Pouplana L. 2018. Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla. *RNA Biology*. 15(4-5):500-507.

65. Rak R, Dahan O, Pilpel Y. 2018. The Couplers of Genomics and Proteomics. *Annu Rev Cell Dev Biol*. 34:239-264.

66. Revell LJ, Harmon LJ, Collar DC. 2008. Phylogenetic signal, evolutionary process, and rate. *Syst. Biol*. 57, 591–601.

67. Revell LJ. Two new graphical methods for mapping trait evolution on phylogenies. 2013. *Methods Ecol Evol*. 4:754–9.

68. Revell, L.J. 2012 Phytools: phylogenetic tools for comparative biology and other things. *Methods in Ecology and Evolution*. 3: 217-223.

69. Rogalski M, Karcher D, Bock R. 2008. Superwobbling facilitates translation with reduced tRNA sets*. Nat. Struct. Mol. Biol*.; 15:192-198

70. Roller M, Lucić V, Nagy I, Perica T, Vlahovicek K. 2013. Environmental shaping of codon usage and functional adaptation across microbial communities. *Nucleic Acids Res*. 4119:8842–8852.

71. Sánchez-García M and Matheny PB. 2017. Is the switch to an ectomycorrhizal state an evolutionary key innovation in mushroom-forming fungi? A case study in the Tricholomatineae Agaricales. *Evolution*. 711:51-65.

72. Santesmasses D, Mariotti M, Guigó R. 2017. Computational identification of the selenocysteine tRNA tRNASec in genomes. *PLoS Comput Biol*. 132:e1005383.

73. Sepala S, Wilken E, Knop D, Solomon K, O'Malley M. 2017. The importance of sourcing enzymes from non-conventional fungi for metabolic engineering and biomass breakdown. *Metabolic Engineering*. 44:45-59.

74. Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci*. 3651544:1203–1212.

75. Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*. 1413:5125–5143.

76. Simpson GG. 1953. The major features of evolution. New York: Columbia University Press.

77. Smith SD. 2010. Using phylogenetics to detect pollinator-mediated floral evolution. *New Phytologist*. 188: 354–363.

78. Spatafora JW, Aime MC, Grigoriev IV, Martin F, Stajich JE, Blackwell M. 2017. The fungal tree of life: from molecular systematics to genome-scale phylogenies. *Microbiol Spectr* 5: FUNK-0053-2016

79. Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW. 2009. The fungi. *Curr Biol*. 1918: R840-R845.

80. Supek F and Vlahoviček, K. 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*. 11:463.

81. Toome M, Ohm RA, Riley RW, James TY, Lazarus KL, Henrissat B, Albu S, Boyd A, Chow J, Clum A, Heller G, Lipzen A, Nolan M, Sandor L,

Zvenigorodsky N, Grigoriev IV, Spatafora JW, Aime MC. 2014. Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen Mixia osmundae. *New Phyt.* 202(2):554-564.

82. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*. 1412:344–354.

83. Varga T, Krizsán K, Földi C, Dima B, Sánchez-García M, Sánchez-Ramírez S, Szöllősi GJ, Szarkándi JG, Papp V, Albert L, Andreopoulos W, Angelini C, Antonín V, Barry KW, *et al.,*. 2019. Megaphylogeny resolves global patterns of mushroom evolution. *Nat Ecol Evol.* 3:668–678.

84. Velandia-Huerto CA, Berkemer SJ, Hoffmann A, Retzlaff N, Romero Marroquín LC, Hernández-Rosales M, Stadler PF, Bermúdez-Santana CI. 2016. Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies. *BMC Genomics*. 17(1):617.

85. Wright F. 1990 The ‘effective number of codons’ used in a gene. *Gene.* 871: 23–29.

86. Yang KK, Wu Z, Bedbrook CN, Arnold FH. 2018. Learned protein embeddings for machine learning. *Bioinformatics*. 1:7.

87. Yannai A, Katz S, Hershberg R. 2018. The codon usage of lowly expressed genes is subject to natural selection. *Genome Biol Evol*. 105: 1237–1246.

88. Yona AH, Bloom-Ackermann Z, Frumkin I, Hanson-Smith V, Charpak-Amikam Y, Feng Q, Boeke JD, Dahan O, Pilpel Y. 2013. tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* 2: e01339.

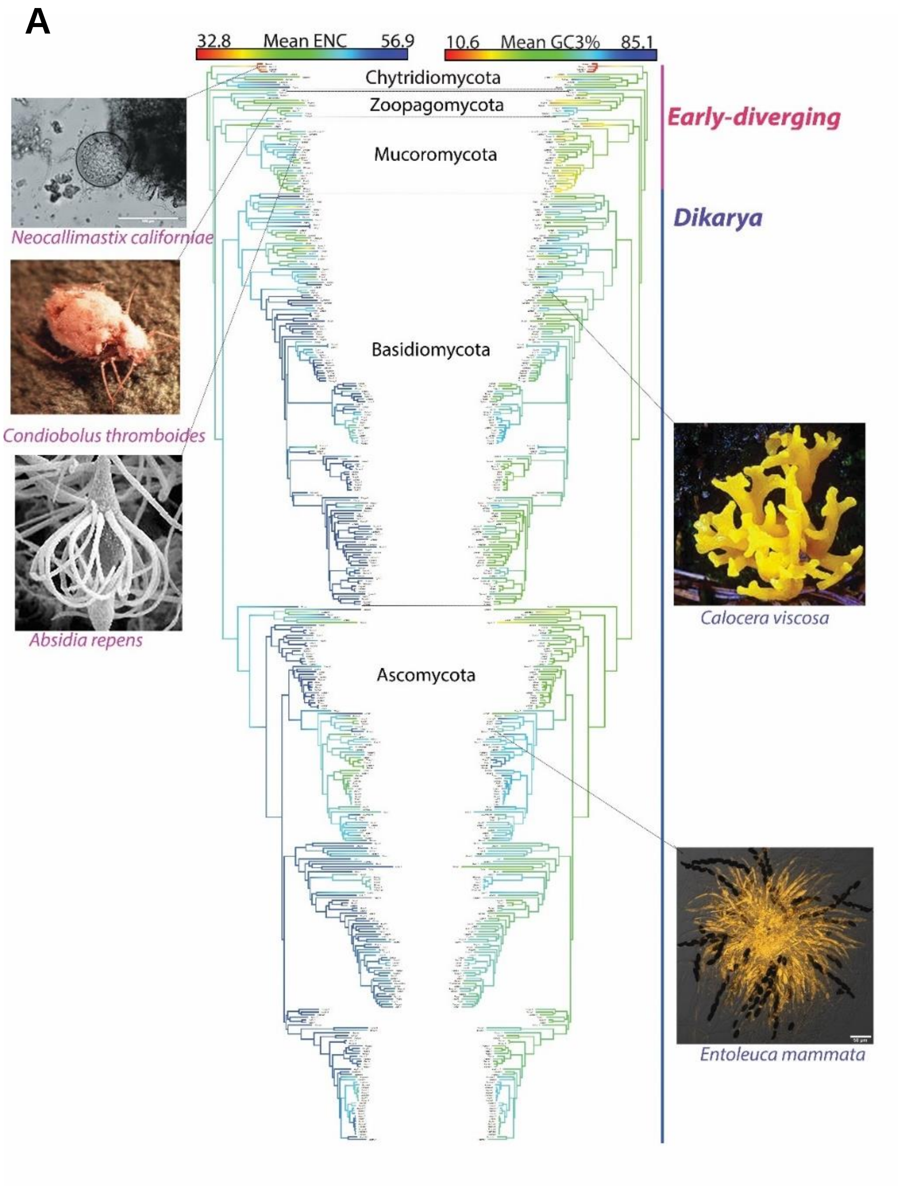
89. Zhao F, Zhou Z, Dang Y, Na H, Adam C, Lipzen A, Ng V, Grigoriev IV, Liu Y. 2021. Genome-wide role of codon usage on transcription and identification of potential regulators. *Proc Natl Acad Sci U S A.* 1186: e2022590118.

90. Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol.* 267:1571–1580.

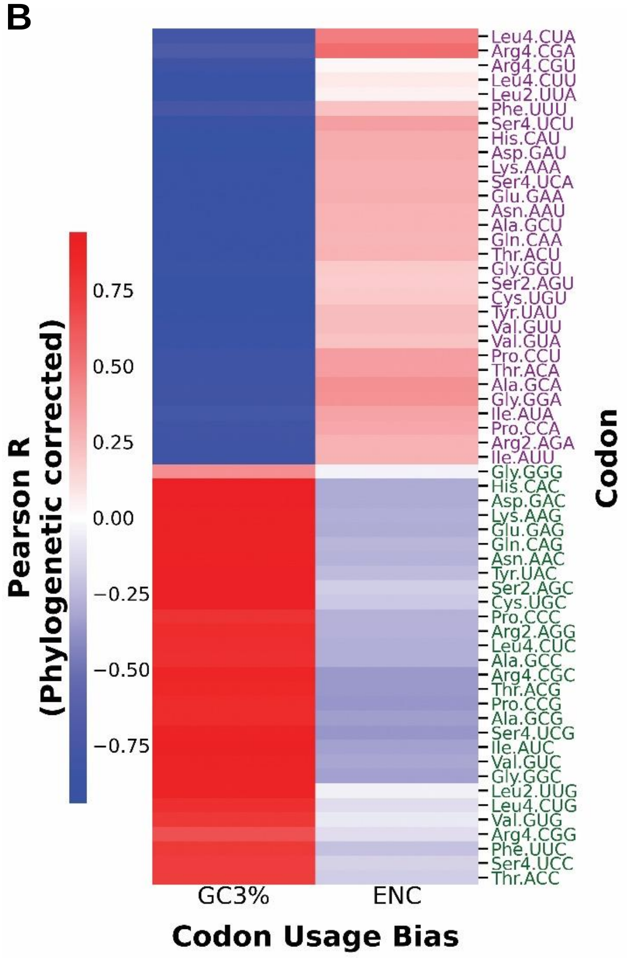
91. Zhou Z, Dang Y, Zhou M, Li L, Yu C, Fu J, Chen S, Liu Y. 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl Acad. Sci. USA.* 113: E6117-E6125.

92. Zouridis H. and Hatzimanikatis V. 2008. Effects of codon distributions and tRNA competition on protein translation*. Biophys. J.* 95:1018–1033.

Figure 1


A
32.8
Mean ENC
56.9
10.6
Mean GC3%
85.1
Chytridiomycota
Zoopagomycota
Mucoromycota
Basidiomycota
Ascomycota
Early-diverging
Dikarya
Neocallimastix californiae
Condiobolus thromboides
Absidia repens
Calocera viscosa
Entoleuca mammata

B


Pearson R
(Phylogenetic corrected)
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Leu4.CUA
Arg4.CGA
Arg4.CGU
Leu4.CUU
Leu2.UUA
Phe.UUU
Ser4.UCU
His.CAU
Asp.GAU
Lys.AAA
Ser4.UCA
Glu.GAA
Asn.AAU
Ala.GCU
Gln.CAA
Thr.ACU
Gly.GGU
Ser2.AGU
Cys.UGU
Tyr.UAU
Val.GUU
Val.GUA
Pro.CCU
Thr.ACA
Ala.GCA
Gly.GGA
Ile.AUA
Pro.CCA
Arg2.AGA
Ile.AUU
Gly.GGG
His.CAC
Asp.GAC
Lys.AAG
Glu.GAG
Gln.CAG
Asn.AAC
Tyr.UAC
Ser2.AGC
Cys.UGC
Pro.CCC
Arg2.AGG
Leu4.CUC
Ala.GCC
Arg4.CGC
Thr.ACG
Pro.CCG
Ala.GCG
Ser4.UCG
Ile.AUC
Val.GUC
Gly.GGC
Leu2.UUG
Leu4.CUG
Val.GUG
Arg4.CGG
Phe.UUC
Ser4.UCC
Thr.ACC
Codon
GC3%
ENC
Codon Usage Bias

C

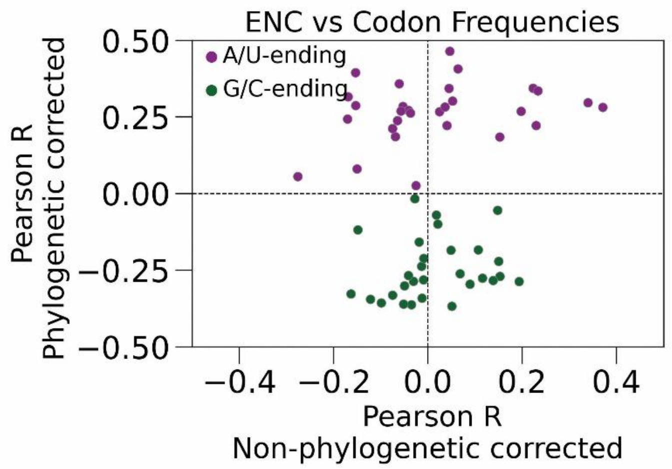

ENC vs Codon Frequencies
● A/U-ending
● G/C-ending
0.50
0.25
0.00
−0.25
−0.50
Pearson R
Phylogenetic corrected
−0.4
−0.2
0.0
0.2
0.4
Pearson R
Non-phylogenetic corrected

Figure 2

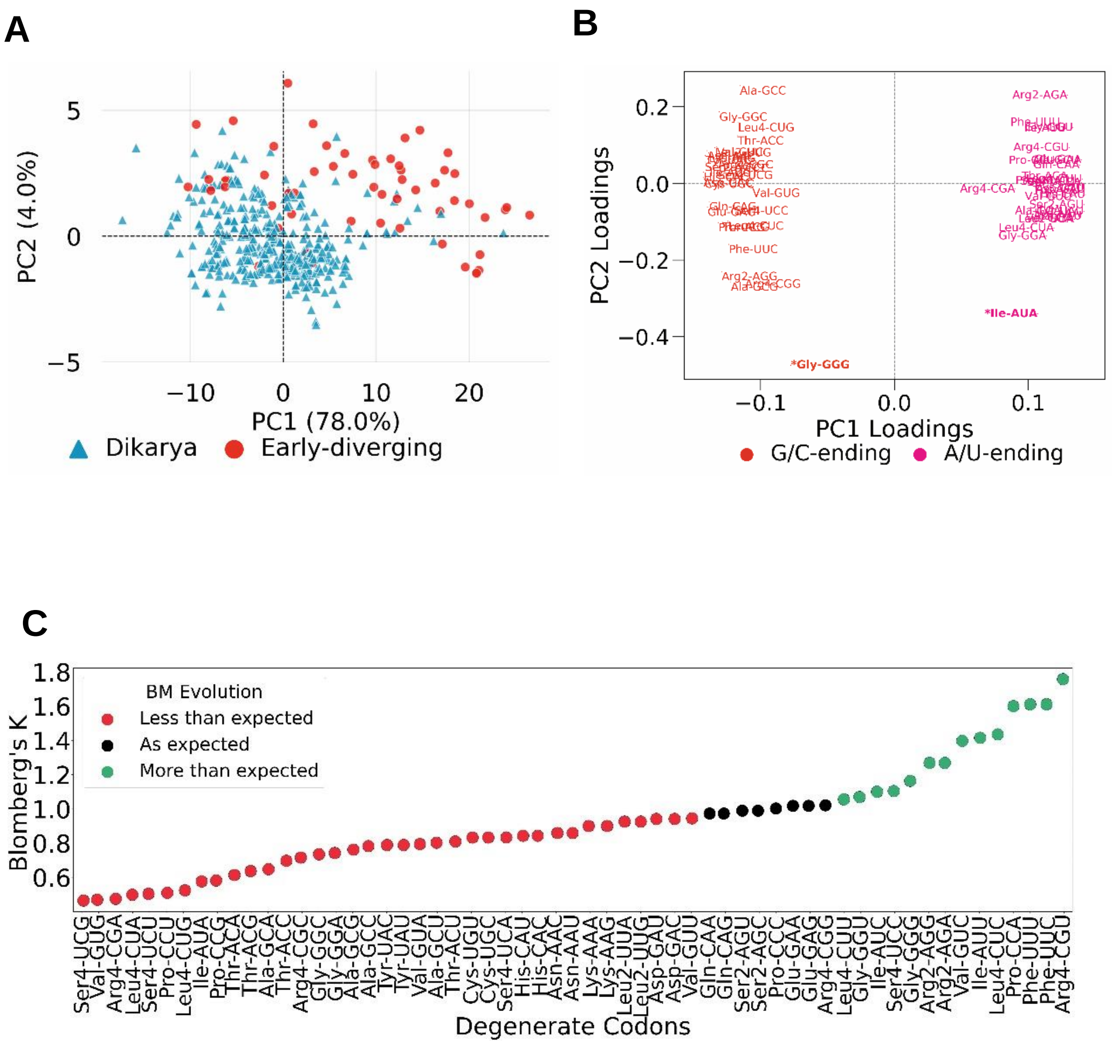


Figure 3

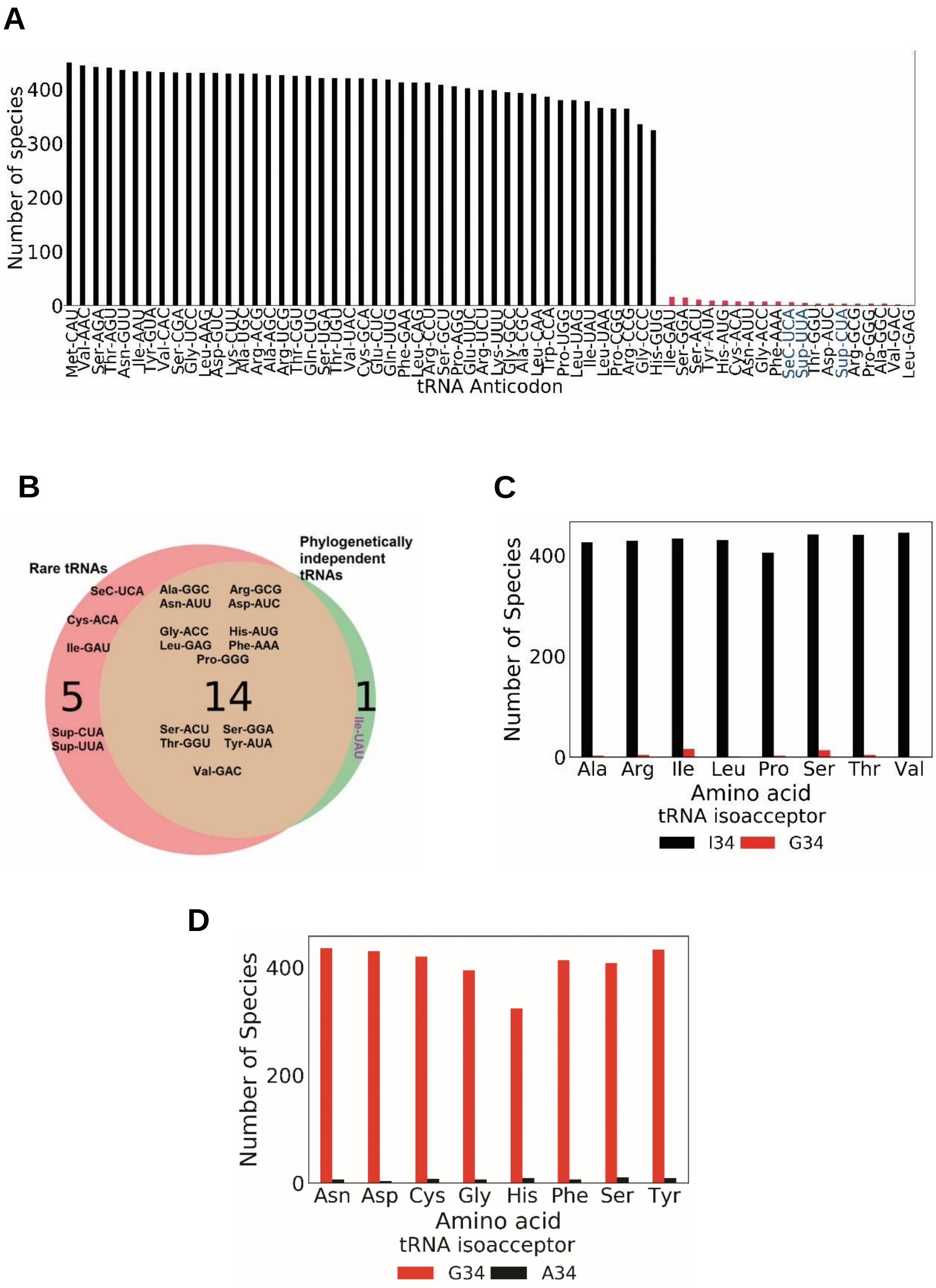

A
Number of species
0
100
200
300
400
Met-CAU
Val-AAC
Ser-AGA
Thr-AGU
Asn-GUU
Ile-AAU
Tyr-GUA
Val-CAC
Ser-CGA
Gly-UCC
Leu-AAG
Asp-GUC
Lys-CUU
Ala-UGC
Arg-ACG
Ala-AGC
Arg-UCG
Thr-CGU
Gln-CUG
Ser-UGA
Thr-UGU
Val-UAC
Cys-GCA
Glu-CUC
Gln-UUG
Phe-GAA
Leu-CAG
Arg-CCU
Ser-GCU
Pro-AGG
Glu-UUC
Arg-UCU
Lys-UUU
Gly-GCC
Ala-CGC
Leu-CAA
Trp-CCA
Pro-UGG
Leu-UAG
Ile-UAU
Leu-UAA
Pro-CGG
Arg-CCG
Gly-CCC
His-GUG
Ile-GAU
Ser-GGA
Ser-ACU
Tyr-AUA
His-AUG
Cys-ACA
Asn-AUU
Gly-ACC
Phe-AAA
SeC-UCA
Sup-UUA
Thr-GGU
Asp-AUC
Sup-CUA
Arg-GCG
Pro-GGG
Ala-GGC
Val-GAC
Leu-GAG
tRNA Anticodon
B
Rare tRNAs
Phylogenetically independent tRNAs
SeC-UCA
Cys-ACA
Ile-GAU
5
Sup-CUA
Sup-UUA
Ala-GGC
Asn-AUU
Arg-GCG
Asp-AUC
Gly-ACC
Leu-GAG
His-AUG
Phe-AAA
Pro-GGG
14
Ser-ACU
Thr-GGU
Ser-GGA
Tyr-AUA
Val-GAC
1
Ile-UAU
C
Number of Species
0
200
400
Ala
Arg
Ile
Leu
Pro
Ser
Thr
Val
Amino acid
tRNA isoacceptor
I34
G34
D
Number of Species
0
200
400
Asn
Asp
Cys
Gly
His
Phe
Ser
Tyr
Amino acid
tRNA isoacceptor
G34
A34

Figure 4

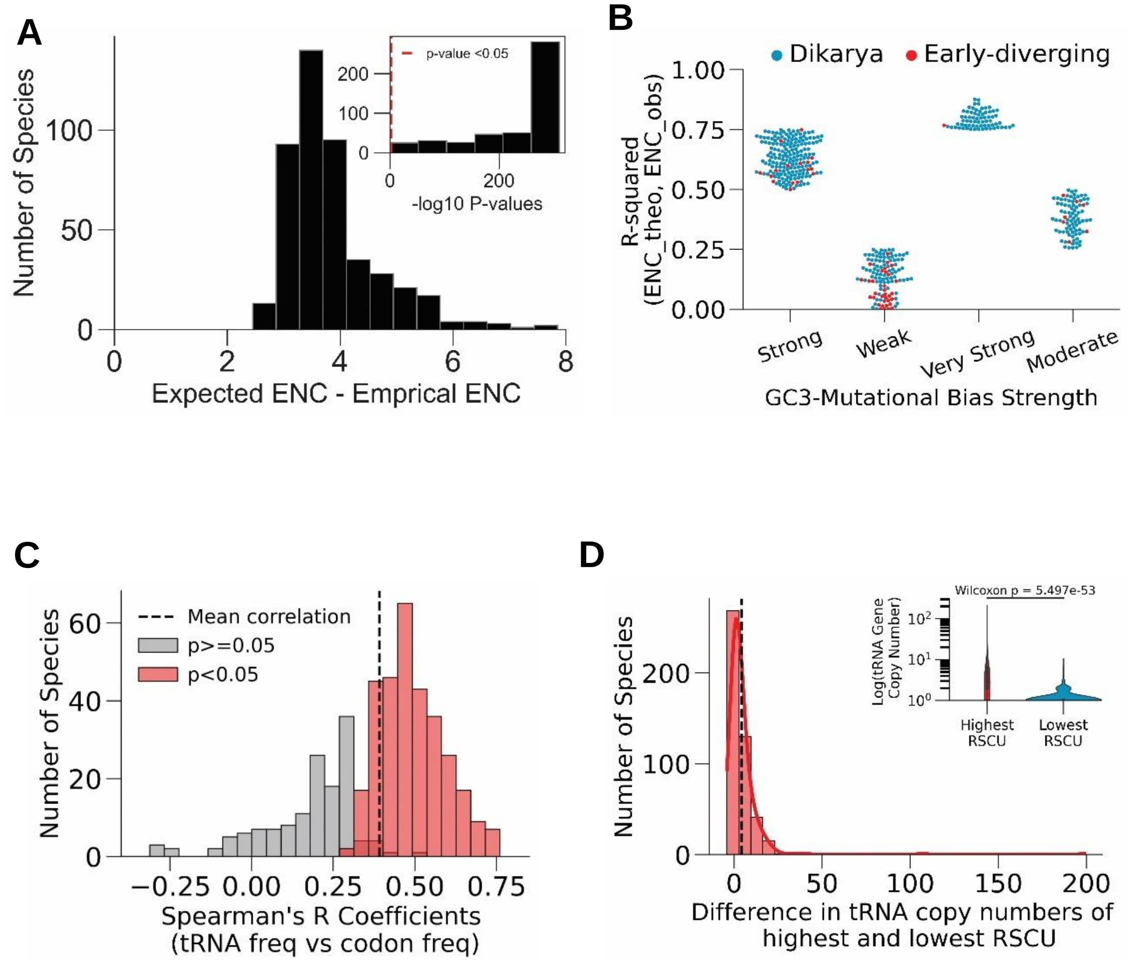

A
Number of Species
0
50
100
0
2
4
6
8
Expected ENC - Emprical ENC
p-value <0.05
200
100
0
0
200
-log10 P-values
B
Dikarya
Early-diverging
R-squared
(ENC_theo, ENC_obs)
1.00
0.75
0.50
0.25
0.00
Strong
Weak
Very Strong
Moderate
GC3-Mutational Bias Strength
C
Mean correlation
p>=0.05
p<0.05
Number of Species
60
40
20
0
−0.25
0.00
0.25
0.50
0.75
Spearman's R Coefficients
(tRNA freq vs codon freq)
D
Number of Species
200
100
0
0
50
100
150
200
Difference in tRNA copy numbers of
highest and lowest RSCU
Wilcoxon p = 5.497e-53
Log(tRNA Gene
Copy Number)
10^2
10^1
10^0
Highest
RSCU
Lowest
RSCU

Figure 5

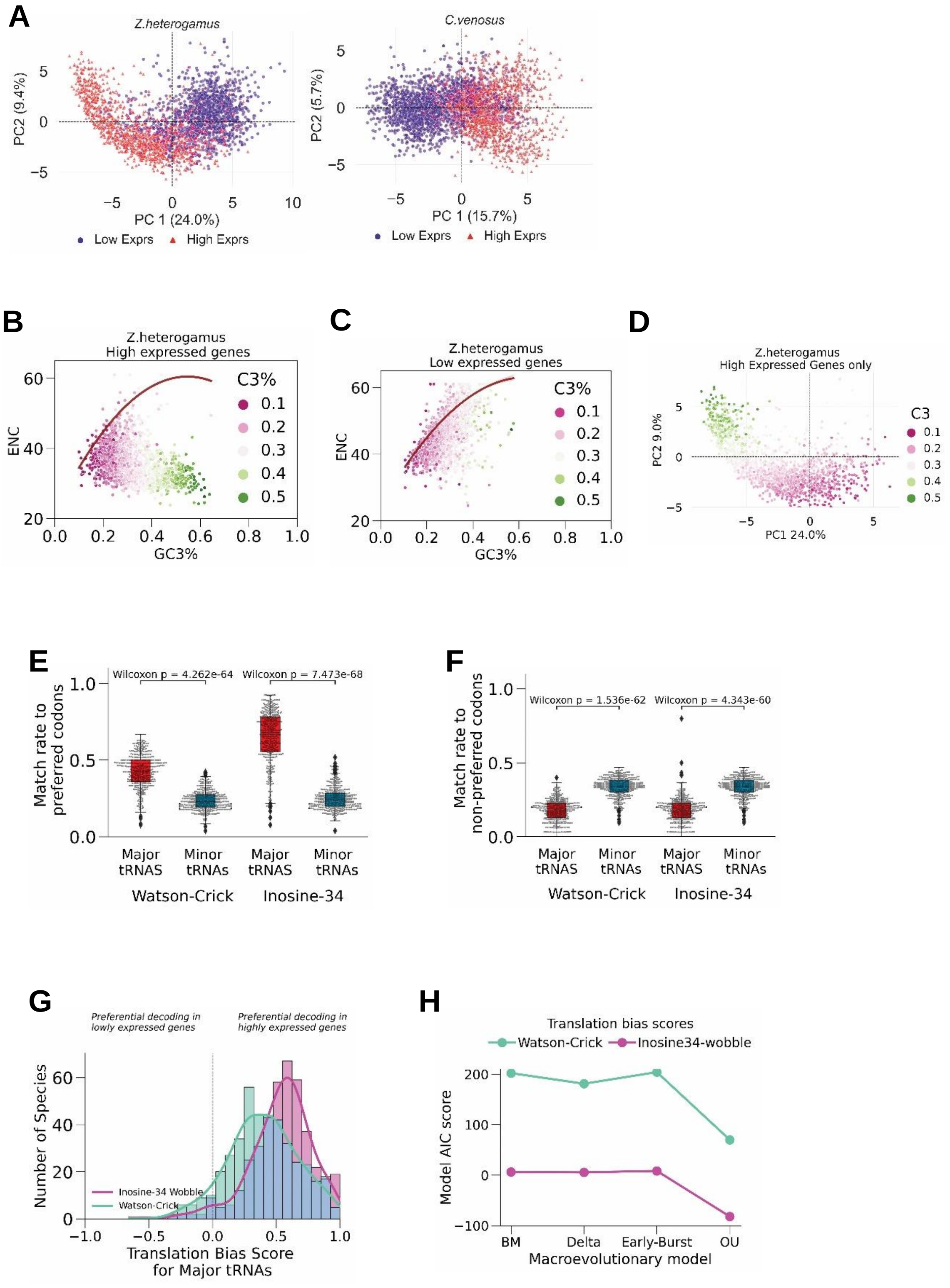
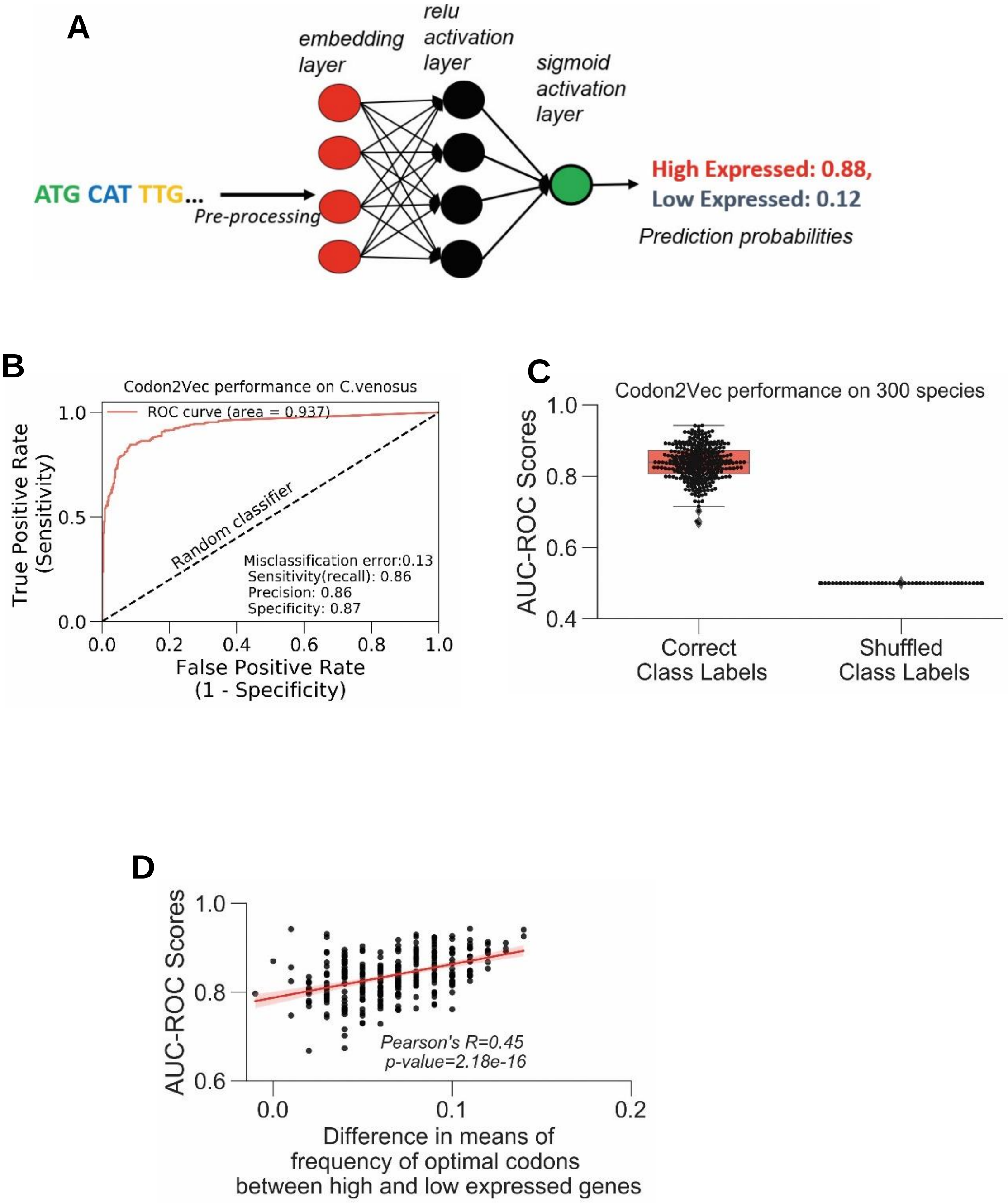

A
Z.heterogamus
PC2 (9.4%)
PC 1 (24.0%)
Low Exprs
High Exprs
C.venosus
PC2 (5.7%)
PC 1 (15.7%)
Low Exprs
High Exprs
B
Z.heterogamus
High expressed genes
ENC
GC3%
C3%
0.1
0.2
0.3
0.4
0.5
C
Z.heterogamus
Low expressed genes
ENC
GC3%
C3%
D
Z.heterogamus
High Expressed Genes only
PC2 9.0%
PC1 24.0%
C3
E
Wilcoxon p = 4.262e-64
Wilcoxon p = 7.473e-68
Match rate to preferred codons
Major tRNAS
Minor tRNAs
Major tRNAS
Minor tRNAs
Watson-Crick
Inosine-34
F
Wilcoxon p = 1.536e-62
Wilcoxon p = 4.343e-60
Match rate to non-preferred codons
Major tRNAS
Minor tRNAs
Major tRNAS
Minor tRNAs
Watson-Crick
Inosine-34
G
Preferential decoding in lowly expressed genes
Preferential decoding in highly expressed genes
Number of Species
Inosine-34 Wobble
Watson-Crick
Translation Bias Score for Major tRNAs
H
Translation bias scores
Watson-Crick
Inosine34-wobble
Model AIC score
BM
Delta
Early-Burst
OU
Macroevolutionary model

Figure 6


A
embedding layer
relu activation layer
sigmoid activation layer
ATG CAT TTG...
Pre-processing
High Expressed: 0.88,
Low Expressed: 0.12
Prediction probabilities
B
Codon2Vec performance on C.venosus
ROC curve (area = 0.937)
Random classifier
Misclassification error:0.13
Sensitivity(recall): 0.86
Precision: 0.86
Specificity: 0.87
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
C
Codon2Vec performance on 300 species
AUC-ROC Scores
1.0
0.8
0.6
0.4
Correct Class Labels
Shuffled Class Labels
D
AUC-ROC Scores
1.0
0.8
0.6
Pearson's R=0.45
p-value=2.18e-16
0.0
0.1
0.2
Difference in means of frequency of optimal codons between high and low expressed genes

Supplemental Figures and Legend

Figure S1: Summary of the most and least preferred codons in fungal genomes

A-B: Most and least preferred codons in the 21 degenerate amino acid group (6-box amino acids are disaggregated). **A) "**Most preferred" codons have the highest transcriptome-wide relative synonymous codon usage (RSCU) in their degenerate amino acid family. **B) "**Least preferred" codons have the lowest transcriptome-wide RSCU in their degenerate amino acid family. With these definitions, multiple codons may be 'most' or 'least' preferred in the transcriptomes.

C: Summary of the preference of NNX codons is preferred (highest RSCU) and no-preferred (lowest RSCU) in the species' transcriptomes.

D: Heat map showing the distribution of the relative loadings for each codon on the first three principal components.

Figure S2: Variation in tRNA composition

A: Distribution of the number of distinct tRNA anticodon types across the 459 species

B: Distribution of the log base 10 of the total number of tRNA genes (tRNAome)

Figure S3: Expression-linked codon usage bias and adaptation for translation efficiency

A-C: Scatterplots showing only the high expressed genes after principal component analysis was performed on 59-RSCU matrix of high and low expressed genes of *Z.heterogamus*. **A)** High expressed genes colored by fraction of A-ending codons (A3%), **B)** G-ending codons and **C)** U-ending codons. The latent gradient underlying the arch effect is change in A3% and C3% (Figure 5D).

D: The GC3-content and goodness-of-fit to GC3-compositional bias of the 27 species whose high expressed genes formed an arch pattern (Guttman effect) when PCA was applied to RSCU of their low and high expressed genes. On the x-axis, early-diverging species are colored in red, and dikarya in blue.

E: Identity and distribution of codons that are preferred according to transcript abundance. Stripplot showing codons are preferred (significantly higher RSCU in high expressed genes than low expressed genes) and non-preferred (significantly higher RSCU in low expressed genes) across the 420 species with available RNAseq data.

F: Projection of translation bias scores onto the fungal tree shows that codon bias in high expressed genes for major tRNAs emerges in multiple and across distantly related lineages.

G: Scatterplots of the 27 species whose high expressed genes form an arched cluster when projected on the first two principal component axes after PCA was performed on the 59-RSCU matrix of their high and low expressed genes.

H: To assess the strength of composition bias on the codon usage of the high expressed genes in **Figure S3G**, we made an ENC-GC3 plot where the standard curve is the expected relationship between codon bias (ENC) and GC-content at the 3rd codon position (GC3%) in the absence of selection [Wright 1990].

Figure S4: Architecture and performance of Codon2Vec neural network

A: Illustration of how Codon2vec's preprocesses coding sequences in order to extract codons are features.

B: Distribution of the mean CDS lengths across the 459 genomes.

C-E: CodonVec's performance on conditionally and constitutively expressed genes at different growth-stages: To test how the model responds to expression dynamics, we used growth-stage specific (blastospore and hyphal) RNAseq data of *Metarhizium anisopliae* (Iwanicki et al, 2020) to separately train growth-stage specific models .Conditionally (genes expressed as high in one condition but low in the other) and constitutively expressed genes (genes with fold change= between conditions) were held-out test sequences used for model prediction **C)** The blastospore-specific model, 'Blastospore-C2V', predicted the blastospore high expressed test sequences with higher probabilities that the hyphal-specific model (Hyphae-C2V). **D)** Conversely, the hyphal-specific model predicted high expressed hyphal sequences with higher probabilities. **E)** . Both stage-specific models performed similarly on genes that did change expression between the blastospore and hyphal stage. Prediction probabilities are mean of 30 iterations.

Table S1: Phyla level analysis of the evolutionary relationship between codon bias and GC3-content

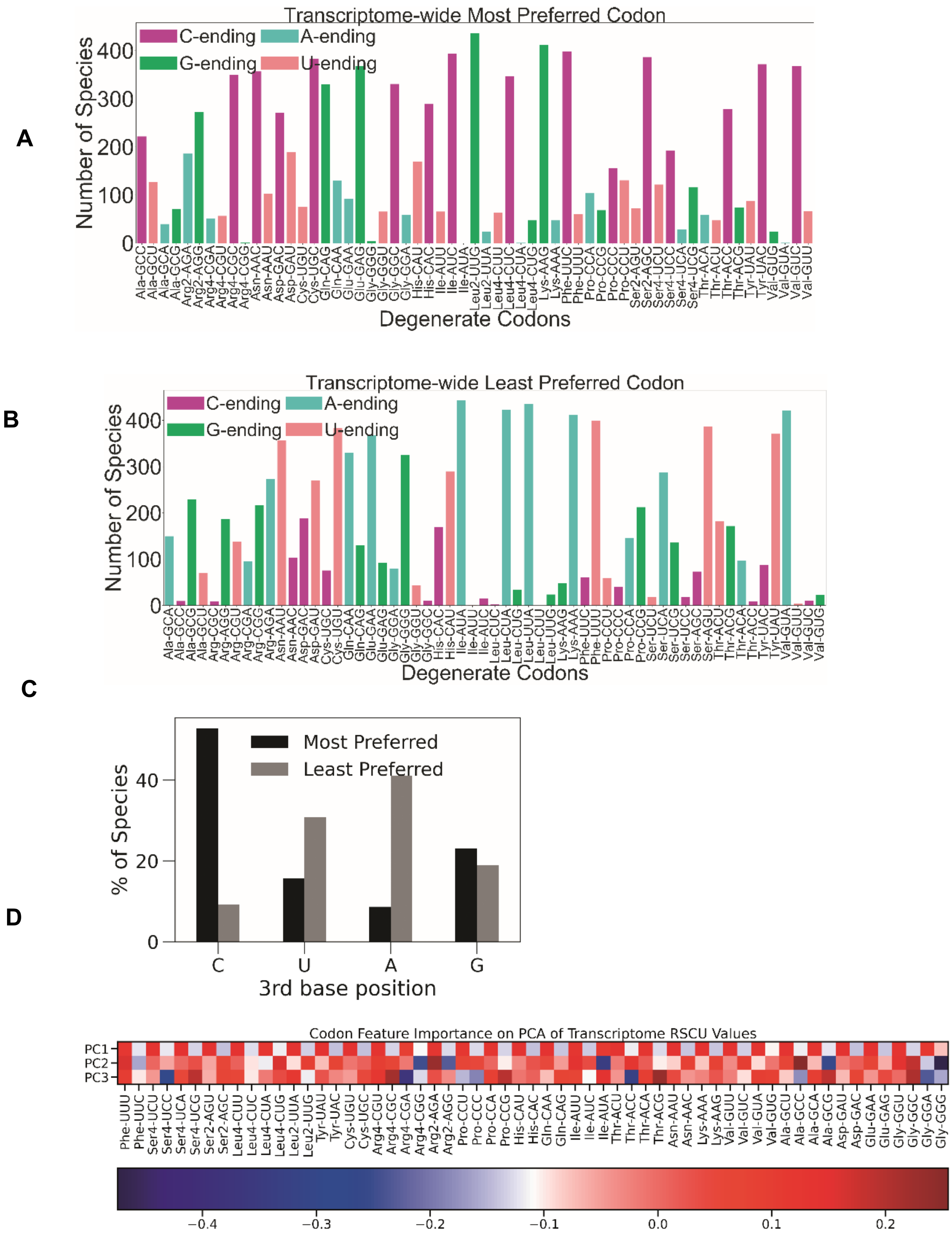
Summary of regression results of phylogenetic independent contrasts (PIC) between effective number of codons (ENC) and GC3-content (ENC~GC3) to individual superphyla and sub-phyla that have at least 10 species to ensure sufficient statistical power to detect phylogenetic signal [Blomberg et al 2003]. Phyla level results are consistent with sub-kingdom patterns in that all dikaryic phyla show positive correlation between ENC and GC3%, whereas all the early-diverging phyla are negatively GC3-biased

Table S2. Summary of the relative fit of various macroevolutionary models for codon usage bias and GC3-content. The early-burst model reports the lowest Akaike information criterion (AIC) score making it the best-fitting model for both genomic codon usage bias, as measured by ENC, and GC3-content. The early-burst model detects trait evolution by adaptive radiation.

Table S3:

Table showing the species with selenocysteine tRNA (SeC-tRNA) predicted by tRNAscanSE2.0 after low quality covariance scores (<50) were removed. The SeC-tRNA predictions are compare to another general purpose tRNA gene finder (Aragorn) and selenocysteine tRNA specific tool (SeCmaker)

Figure S1


A
Transcriptome-wide Most Preferred Codon
C-ending
A-ending
G-ending
U-ending
Number of Species
0
100
200
300
400
Ala-GCC Ala-GCU Ala-GCA Ala-GCG Arg2-AGA Arg2-AGG Arg4-CGA Arg4-CGU Arg4-CGC Arg4-CGG Asn-AAC Asn-AAU Asp-GAC Asp-GAU Cys-UGU Cys-UGC Gln-CAG Gln-CAA Glu-GAA Glu-GAG Gly-GGG Gly-GGU Gly-GGC Gly-GGA His-CAU His-CAC Ile-AUU Ile-AUC Ile-AUA Leu2-UUG Leu2-UUA Leu4-CUU Leu4-CUC Leu4-CUA Leu4-CUG Lys-AAG Lys-AAA Phe-UUC Phe-UUU Pro-CCA Pro-CCG Pro-CCC Pro-CCU Ser2-AGU Ser2-AGC Ser4-UCU Ser4-UCC Ser4-UCA Ser4-UCG Thr-ACA Thr-ACU Thr-ACC Thr-ACG Tyr-UAU Tyr-UAC Val-GUG Val-GUA Val-GUC Val-GUU
Degenerate Codons
B
Transcriptome-wide Least Preferred Codon
C-ending
A-ending
G-ending
U-ending
Number of Species
0
100
200
300
400
Ala-GCA Ala-GCC Ala-GCG Ala-GCU Arg-CGC Arg-AGG Arg-CGU Arg-CGA Arg-CGG Arg-AGA Asn-AAU Asn-AAC Asp-GAC Asp-GAU Cys-UGC Cys-UGU Gln-CAA Gln-CAG Glu-GAA Glu-GAG Gly-GGA Gly-GGG Gly-GGU Gly-GGC His-CAC His-CAU Ile-AUA Ile-AUU Ile-AUC Leu-CUC Leu-CUA Leu-CUG Leu-UUA Leu-CUU Leu-UUG Lys-AAG Lys-AAA Phe-UUC Phe-UUU Pro-CCU Pro-CCC Pro-CCA Pro-CCG Ser-UCU Ser-UCA Ser-UCG Ser-UCC Ser-AGC Ser-AGU Thr-ACU Thr-ACG Thr-ACA Thr-ACC Tyr-UAC Tyr-UAU Val-GUA Val-GUU Val-GUC Val-GUG
Degenerate Codons
C
Most Preferred
Least Preferred
% of Species
0
20
40
C
U
A
G
3rd base position
D
Codon Feature Importance on PCA of Transcriptome RSCU Values
PC1
PC2
PC3
Phe-UUU Phe-UUC Ser4-UCU Ser4-UCC Ser4-UCA Ser4-UCG Ser2-AGU Ser2-AGC Leu4-CUU Leu4-CUC Leu4-CUA Leu4-CUG Leu2-UUA Leu2-UUG Tyr-UAU Tyr-UAC Cys-UGU Cys-UGC Arg4-CGU Arg4-CGC Arg4-CGA Arg4-CGG Arg2-AGA Arg2-AGG Pro-CCU Pro-CCC Pro-CCA Pro-CCG His-CAU His-CAC Gln-CAA Gln-CAG Ile-AUU Ile-AUC Ile-AUA Thr-ACU Thr-ACC Thr-ACA Thr-ACG Asn-AAU Asn-AAC Lys-AAA Lys-AAG Val-GUU Val-GUC Val-GUA Val-GUG Ala-GCU Ala-GCC Ala-GCA Ala-GCG Asp-GAU Asp-GAC Glu-GAA Glu-GAG Gly-GGU Gly-GGC Gly-GGA Gly-GGG
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2

Figure

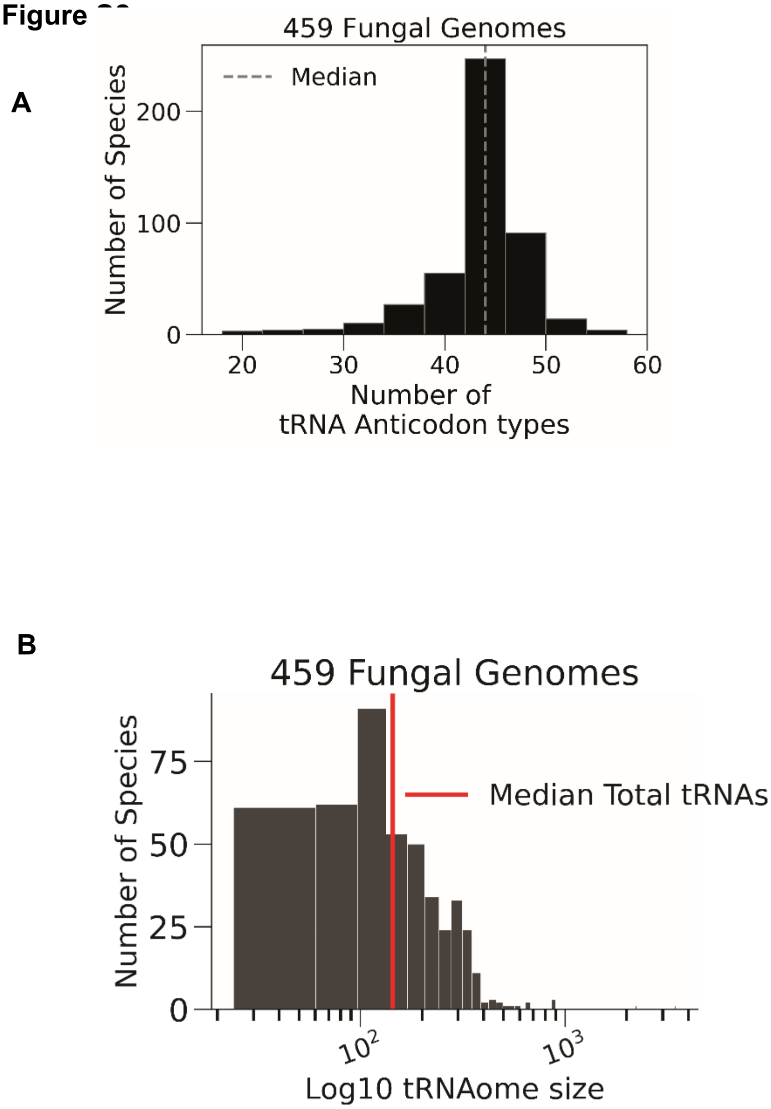
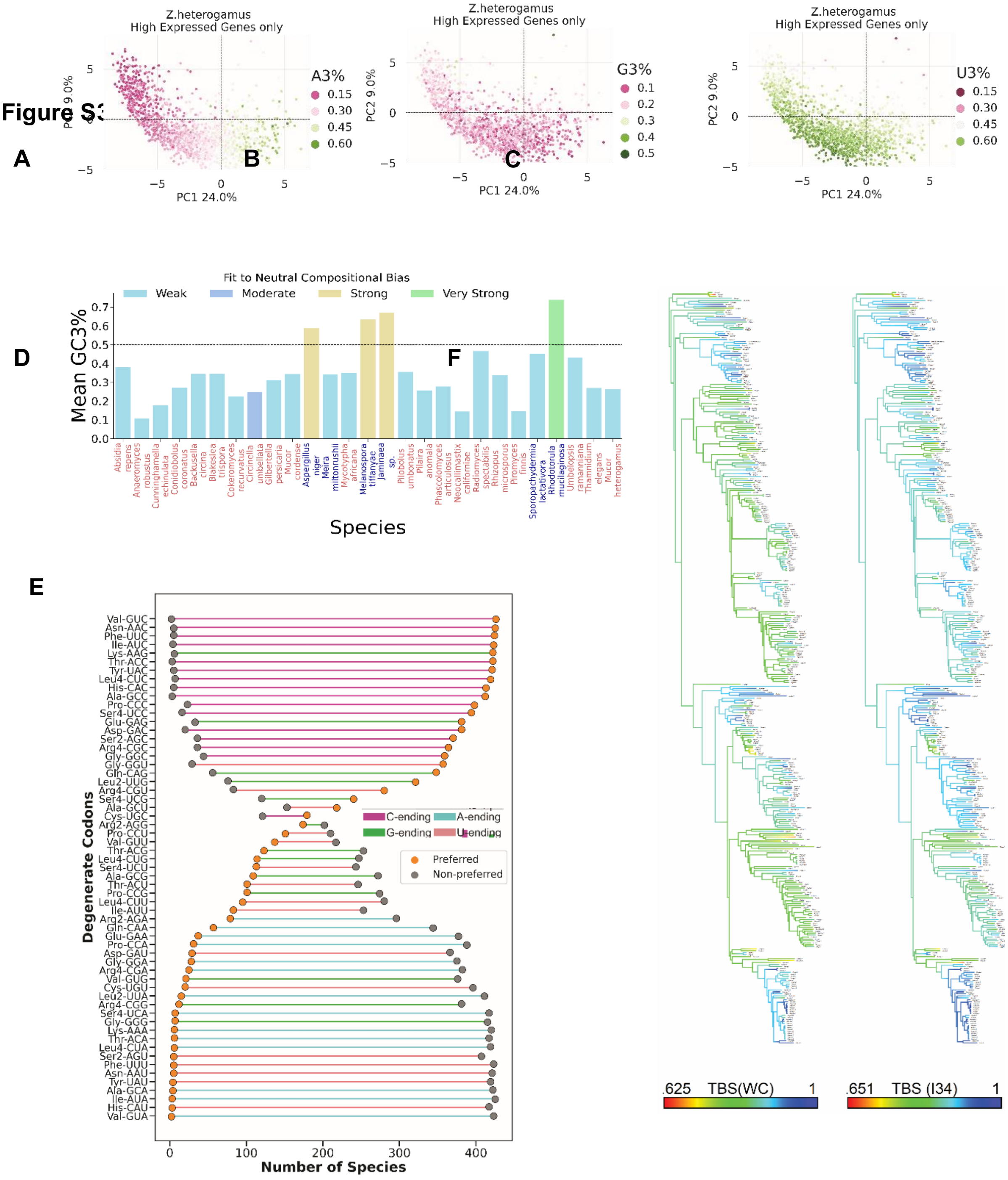
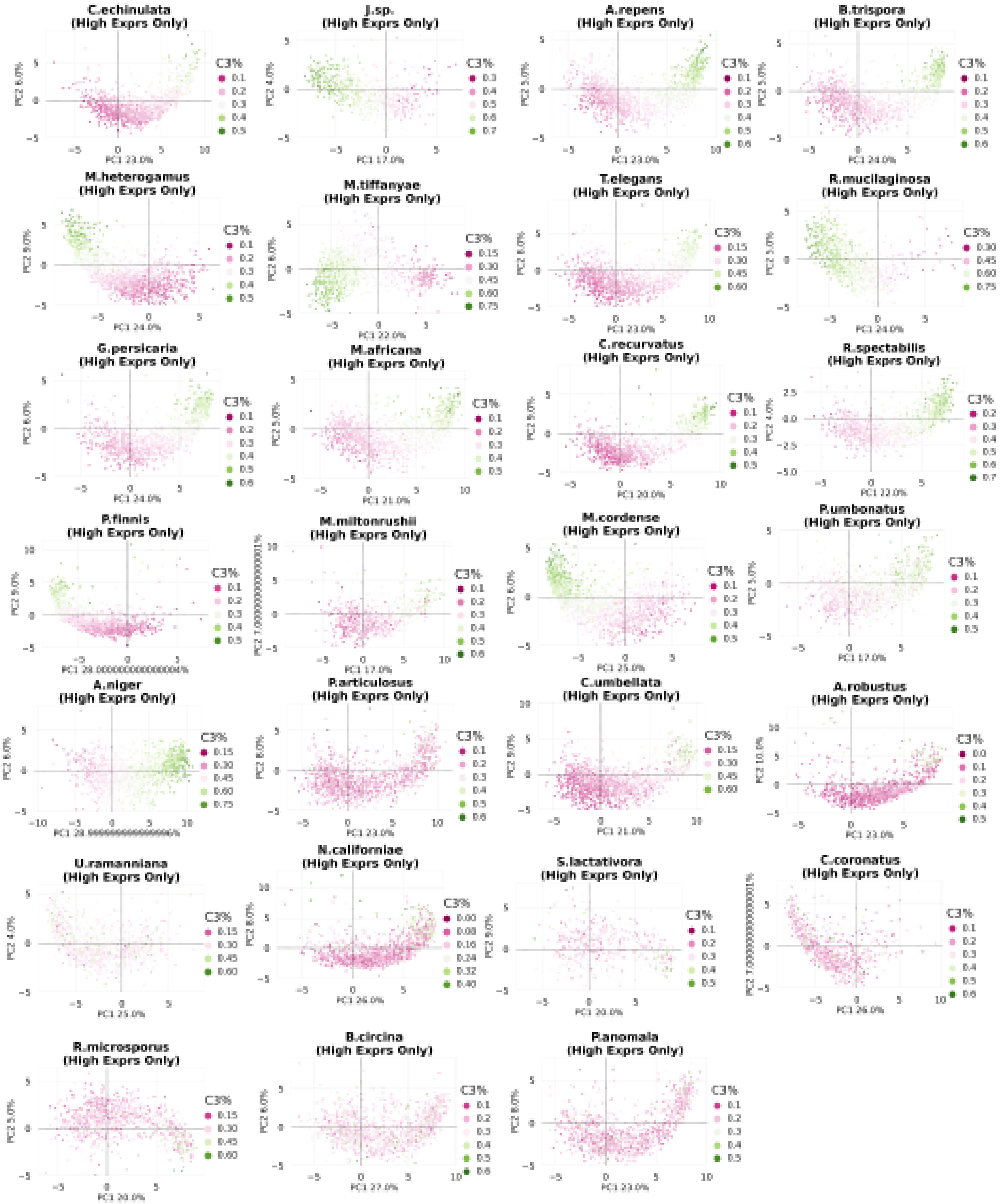

A
459 Fungal Genomes
Median
Number of Species
0
100
200
20
30
40
50
60
Number of
tRNA Anticodon types
B
459 Fungal Genomes
Median Total tRNAs
Number of Species
0
25
50
75
10^2
10^3
Log10 tRNAome size

Figure S3


A
Z.heterogamus
High Expressed Genes only
PC2 9.0%
PC1 24.0%
A3%
0.15
0.30
0.45
0.60
B
Z.heterogamus
High Expressed Genes only
G3%
0.1
0.2
0.3
0.4
0.5
C
Z.heterogamus
High Expressed Genes only
U3%
0.15
0.30
0.45
0.60
5
0
−5
D
Fit to Neutral Compositional Bias
Weak
Moderate
Strong
Very Strong
Mean GC3%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Absidia repens
Anaeromyces robustus
Cunninghamella echinulata
Conidiobolus coronatus
Backusella circina
Blakeslea trispora
Cokeromyces recurvatus
Circinella umbellata
Gilbertella persicaria
Mucor cordense
Aspergillus niger
Meira miltonrushii
Mycotypha africana
Melanospora tiffanyae
Jaminaea sp.
Pilobolus umbonatus
Pilaira anomala
Phascolomyces articulosus
Neocallimastix californiae
Radiomyces spectabilis
Rhizopus microsporus
Piromyces finnis
Sporopachydermia lactativora
Rhodotorula mucilaginosa
Umbelopsis ramanniana
Thamnidium elegans
Mucor heterogamus
Species
E
Degenerate Codons
Val-GUC
Asn-AAC
Phe-UUC
Ile-AUC
Lys-AAG
Thr-ACC
Tyr-UAC
Leu4-CUC
His-CAC
Ala-GCC
Pro-CCC
Ser4-UCC
Glu-GAG
Asp-GAC
Ser2-AGC
Arg4-CGC
Gly-GGC
Gly-GGU
Gln-CAG
Leu2-UUG
Arg4-CGU
Ser4-UCG
Ala-GCU
Cys-UGC
Arg2-AGG
Pro-CCU
Val-GUU
Thr-ACG
Leu4-CUG
Ser4-UCU
Ala-GCG
Thr-ACU
Pro-CCG
Leu4-CUU
Ile-AUU
Arg2-AGA
Gln-CAA
Glu-GAA
Pro-CCA
Asp-GAU
Gly-GGA
Arg4-CGA
Val-GUG
Cys-UGU
Leu2-UUA
Arg4-CGG
Ser4-UCA
Gly-GGG
Lys-AAA
Thr-ACA
Leu4-CUA
Ser2-AGU
Phe-UUU
Asn-AAU
Tyr-UAU
Ala-GCA
Ile-AUA
His-CAU
Val-GUA
C-ending
A-ending
G-ending
U-ending
Preferred
Non-preferred
0
100
200
300
400
Number of Species
F
.625 TBS(WC) 1
.651 TBS (I34) 1

G


C.echinulata (High Exprs Only)
PC2 6.0%
PC1 23.0%
C3% 0.1 0.2 0.3 0.4 0.5
J.sp. (High Exprs Only)
PC2 4.0%
PC1 17.0%
C3% 0.3 0.4 0.5 0.6 0.7
A.repens (High Exprs Only)
PC2 5.0%
PC1 23.0%
C3% 0.1 0.2 0.3 0.4 0.5 0.6
B.trispora (High Exprs Only)
PC2 5.0%
PC1 24.0%
C3% 0.1 0.2 0.3 0.4 0.5 0.6
M.heterogamus (High Exprs Only)
PC2 9.0%
PC1 24.0%
C3% 0.1 0.2 0.3 0.4 0.5
M.tiffanyae (High Exprs Only)
PC2 6.0%
PC1 22.0%
C3% 0.15 0.30 0.45 0.60 0.75
T.elegans (High Exprs Only)
PC2 8.0%
PC1 23.0%
C3% 0.15 0.30 0.45 0.60
R.mucilaginosa (High Exprs Only)
PC2 5.0%
PC1 24.0%
C3% 0.30 0.45 0.60 0.75
G.persicaria (High Exprs Only)
PC2 6.0%
PC1 24.0%
C3% 0.1 0.2 0.3 0.4 0.5 0.6
M.africana (High Exprs Only)
PC2 5.0%
PC1 21.0%
C3% 0.1 0.2 0.3 0.4 0.5
C.recurvatus (High Exprs Only)
PC2 9.0%
PC1 20.0%
C3% 0.1 0.2 0.3 0.4 0.5
R.spectabilis (High Exprs Only)
PC2 4.0%
PC1 22.0%
C3% 0.2 0.3 0.4 0.5 0.6 0.7
P.finnis (High Exprs Only)
PC2 9.0%
PC1 28.000000000000004%
C3% 0.1 0.2 0.3 0.4 0.5
M.miltonrushii (High Exprs Only)
PC2 7.000000000000001%
PC1 17.0%
C3% 0.1 0.2 0.3 0.4 0.5 0.6
M.cordense (High Exprs Only)
PC2 6.0%
PC1 25.0%
C3% 0.1 0.2 0.3 0.4 0.5
P.umbonatus (High Exprs Only)
PC2 5.0%
PC1 17.0%
C3% 0.1 0.2 0.3 0.4 0.5
A.niger (High Exprs Only)
PC2 6.0%
PC1 28.999999999999996%
C3% 0.15 0.30 0.45 0.60 0.75
P.articulosus (High Exprs Only)
PC2 8.0%
PC1 23.0%
C3% 0.1 0.2 0.3 0.4 0.5 0.6
C.umbellata (High Exprs Only)
PC2 9.0%
PC1 21.0%
C3% 0.15 0.30 0.45 0.60
A.robustus (High Exprs Only)
PC2 10.0%
PC1 23.0%
C3% 0.0 0.1 0.2 0.3 0.4 0.5
U.ramanniana (High Exprs Only)
PC2 4.0%
PC1 25.0%
C3% 0.15 0.30 0.45 0.60
N.californiae (High Exprs Only)
PC2 8.0%
PC1 26.0%
C3% 0.00 0.08 0.16 0.24 0.32 0.40
S.lactativora (High Exprs Only)
PC2 8.0%
PC1 20.0%
C3% 0.1 0.2 0.3 0.4 0.5
C.coronatus (High Exprs Only)
PC2 7.000000000000001%
PC1 26.0%
C3% 0.1 0.2 0.3 0.4 0.5 0.6
R.microsporus (High Exprs Only)
PC2 5.0%
PC1 20.0%
C3% 0.15 0.30 0.45 0.60
B.circina (High Exprs Only)
PC2 6.0%
PC1 27.0%
C3% 0.1 0.2 0.3 0.4 0.5 0.6
P.anomala (High Exprs Only)
PC2 8.0%
PC1 23.0%
C3% 0.1 0.2 0.3 0.4 0.5

H

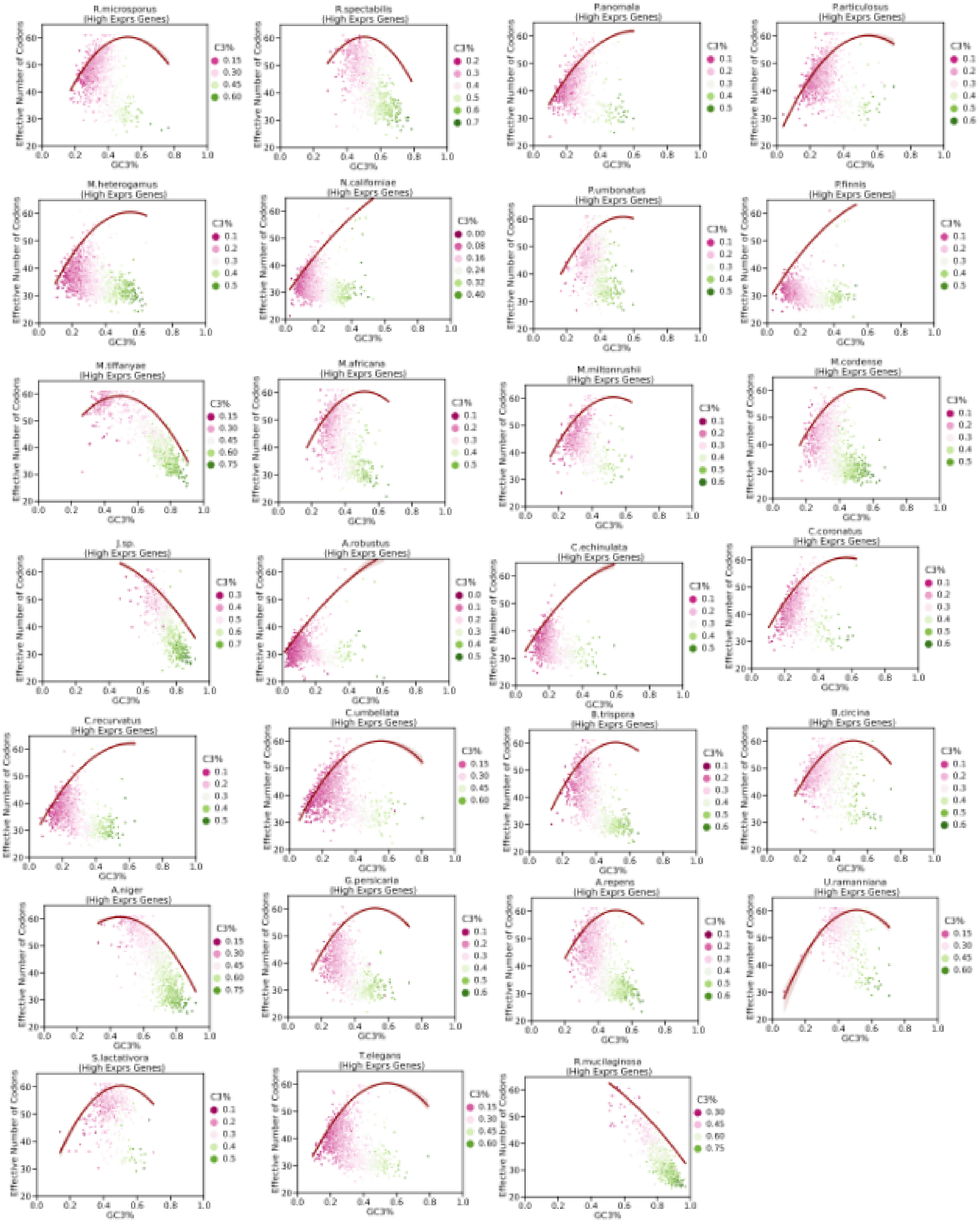
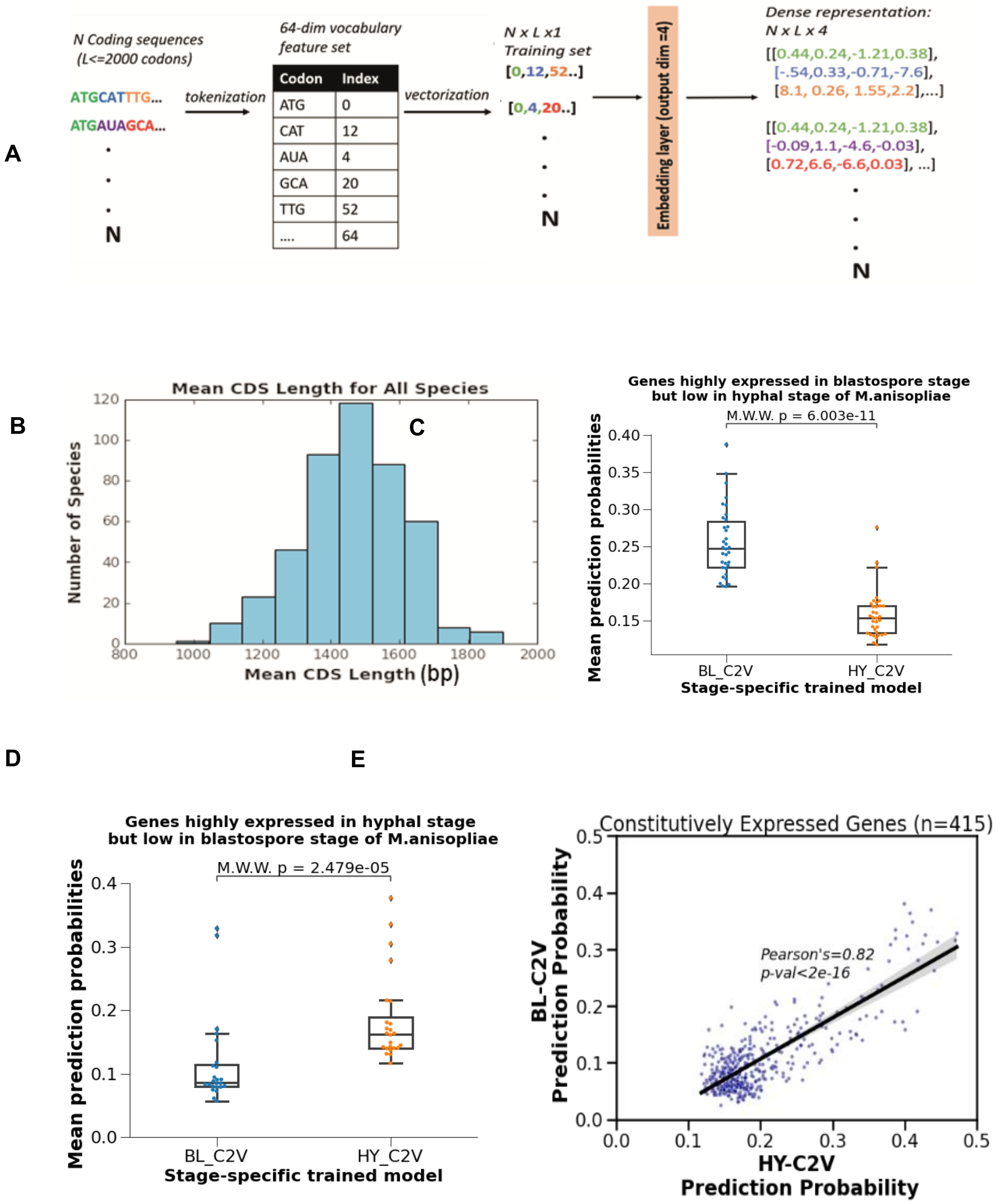

R.microsporus (High Exprs Genes)
C3% 0.15 0.30 0.45 0.60
R.spectabilis (High Exprs Genes)
C3% 0.2 0.3 0.4 0.5 0.6 0.7
P.anomala (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
P.articulosus (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5 0.6
M.heterogamus (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
N.californiae (High Exprs Genes)
C3% 0.00 0.08 0.16 0.24 0.32 0.40
P.umbonatus (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
P.finnis (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
M.tiffanyae (High Exprs Genes)
C3% 0.15 0.30 0.45 0.60 0.75
M.africana (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
M.miltonrushii (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5 0.6
M.cordense (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
J.sp. (High Exprs Genes)
C3% 0.3 0.4 0.5 0.6 0.7
A.robustus (High Exprs Genes)
C3% 0.0 0.1 0.2 0.3 0.4 0.5
C.echinulata (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
C.coronatus (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5 0.6
C.recurvatus (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
C.umbellata (High Exprs Genes)
C3% 0.15 0.30 0.45 0.60
B.trispora (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5 0.6
B.circina (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5 0.6
A.niger (High Exprs Genes)
C3% 0.15 0.30 0.45 0.60 0.75
G.persicaria (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5 0.6
A.repens (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5 0.6
U.ramanniana (High Exprs Genes)
C3% 0.15 0.30 0.45 0.60
S.lactativora (High Exprs Genes)
C3% 0.1 0.2 0.3 0.4 0.5
T.elegans (High Exprs Genes)
C3% 0.15 0.30 0.45 0.60
R.mucilaginosa (High Exprs Genes)
C3% 0.30 0.45 0.60 0.75
Effective Number of Codons
20 30 40 50 60
GC3%
0.0 0.2 0.4 0.6 0.8 1.0

Figure S4


A
N Coding sequences
(L<=2000 codons)
ATGCATTTG...
ATGAUAGCA...
N
tokenization
64-dim vocabulary
feature set
Codon
Index
ATG
0
CAT
12
AUA
4
GCA
20
TTG
52
....
64
vectorization
N x L x1
Training set
[0,12,52..]
[0,4,20..]
N
Embedding layer (output dim =4)
Dense representation:
N x L x 4
[[0.44,0.24,-1.21,0.38],
[-.54,0.33,-0.71,-7.6],
[8.1, 0.26, 1.55,2.2],...]
[[0.44,0.24,-1.21,0.38],
[-0.09,1.1,-4.6,-0.03],
[0.72,6.6,-6.6,0.03], ...]
N
B
Mean CDS Length for All Species
Number of Species
0
20
40
60
80
100
120
800
1000
1200
1400
1600
1800
2000
Mean CDS Length (bp)
C
Genes highly expressed in blastospore stage
but low in hyphal stage of M.anisopliae
M.W.W. p = 6.003e-11
Mean prediction probabilities
0.40
0.35
0.30
0.25
0.20
0.15
BL_C2V
HY_C2V
Stage-specific trained model
D
E
Genes highly expressed in hyphal stage
but low in blastospore stage of M.anisopliae
M.W.W. p = 2.479e-05
Mean prediction probabilities
0.4
0.3
0.2
0.1
0.0
BL_C2V
HY_C2V
Stage-specific trained model
Constitutively Expressed Genes (n=415)
BL-C2V
Prediction Probability
0.5
0.4
0.3
0.2
0.1
0.0
Pearson's=0.82
p-val<2e-16
0.0
0.1
0.2
0.3
0.4
0.5
HY-C2V
Prediction Probability

Supplemental Table 1

Clade (n leaves)	PIC (ENC~GC3) results: Coeff, R^2, p-value	Evolutionary relationship between codon bias and GC3%
Ascomycota (n=206)	-32.2, R^2 = 43.0%, p-val< 2e^{-16}	Positively correlated
Pezzizomycotina (n=202)	-39.8, R^2=68.0%, pval < 2e^{-16}	Positively correlated
Basidiomycota (n=160)	-28.6, R^2= 59.5%, pval<2e^{-16}	Positively correlated
Agaricomycotina (n=131)	-30.9, R^2 = 64.9%, pval< 2e-16	Positively correlated
Ustilaginomycotina(n=10)	-23.6, R^2 = 42.2%, pval = 0.025	Positively correlated
Pucciniomycontina (n=19)	-28.4, R^2 = 58.7%, pval = 7.9e^{-5}	Positively correlated
Mucoromycota(n=30)	27.9, R^2 = 51.3%, pval = 1.07e-^{06}	Negatively correlated
Zoopagomycota (n=12)	8.7, R^2 = 3.2e-3 %, pval = 0.33	Negatively correlated, but not significant.

Supplemental Table 2

	Akaike Information Criteria Scores			
Codon Usage Trait	Brownian Motion	Ornstein-Uhlenbeck	Early-burst	Delta
ENC	1896.7	1896.7	1878.1	1882.1
GC3	-1177.8	-1177.8	-1237.8	-1220.1

Supplemental Table 3

	Number of Selenocysteine tRNAs predicted		
Genomes	tRNAscanSE2.0	Aragorn1.2.38	SecMarker
Rhodocollybia butyracea	1	2	2
Sugiyamaella americana	1	1	1
Lollipopaia minuta	1	0	1
Aspergillus heteromorphus	1	1	0
Rhizophagus cerebriforme	2	1	0
Tothia fuscella	3	3	0