Translational control by differential expression of tRNAs

Xavier Hernandez Alias

TESI DOCTORAL UPF / 2022

DIRECTORS DE LA TESI

Dr. Luis Serrano Pubul
Centre de Regulació Genòmica, Programa de Biologia Sintètica i de Sistemes

Dr. Martin H. Schaefer
IEO European Institute of Oncology IRCCS, Department of Experimental Oncology

PROGRAMA DE BIOLOGIA SINTÈTICA I DE SISTEMES
CENTRE DE REGULACIÓ GENÒMICA

DEPARTAMENT DE MEDICINA I CIÈNCIES DE LA VIDA
UNIVERSITAT POMPEU FABRA

Quan escoltem les preguntes i n'acollim l'embat,
i en fem un món, una passió, un combat,
i les fem fructificar i repartim la riquesa
del seu do de claror i del seu fruit de benestar,
¡com ens omple les hores pensar, investigar!,
¡que intensa la vida feta afany de recerca!

Dr. David Jou Mirabent
L'Eco de Sitges, Núm. 6636

Acknowledgements

I would like to start by thanking all the amazing people that have accompanied me throughout this adventure, with their discussions, their contributions, their support; this thesis is also yours. First of all, a very special thank you to my supervisors, Luis and Martin, for taking me as a PhD student. The door to your office was always open for guidance and advice during these years, both scientific and personal. You transmitted your passion for science, steered me through the challenges, comforted me at the troubles, and supported me in all moments. Luis, thank you for your full availability in your packed agenda, your outside-the-box thinking, your to-the-point feedback, and, especially, for your trust. Martin, thank you for always being so close and approachable despite the distance, for taking me as part of your own group, for your critical and down-to-earth thinking, for your energy and positiveness. You have nurtured the person and scientist I am today.

Furthermore, I would like to extend this acknowledgement to the support and valuable feedback from the members of my thesis advisory committee: Manu Irimia, Eva Novoa, and Fran Supek.

A very special thank you also goes to all present and former members of Serrano's lab, for the fascinating discussions, the afterwork "stadiums", the beach volleyball matches, the retreats, the celebrations, and, most importantly, the amazingly pleasant atmosphere we shared. A la Hannah, amb qui he tingut l'honor i el plaer de compartir projectes i formar un "codon usage" tàndem immillorable; he estat molt afortunat d'heretar la teva línia de recerca i d'aprendre al costat d'una científica exemplar de cap a peus. A Ludo, por hacer mi doctorado mucho más feliz; no podría haber tenido una compañera de viaje mejor, en los buenos y los malos momentos. Al Damiano, amb qui una pandèmia ens va clausurar en un pis de 60 m^2 durant mesos; gràcies per les converses, els sopars, les videotrucades, les compres compartits. Al Miquel, per

unir-te al grup de les òmiques i contagiar la teva passió i motivació per la ciència, fins i tot després d'haver-me suportat ja a Zürich. Al Samuel i el Marc, per ser els meus referents en modelatge i "machine learning". To Sarah (and Nathalie), for setting the highest standards of biology; I am so grateful for the science we shared. And to so many great labmates with whom I have been so lucky to share this thesis: Irene, Javi G, Yamile, Caro, Raül R, Raúl B, Cristina, Javi D, Ari, Quim, Atilla, Ania, Eli, Laura, Leandro, Claire, Carlos, Rocco, Daniel, Maria, Dan, Alícia, Rahma, Òscar, Eva, Sira, Anas, Violeta.

I would also like to thank the amazing CRG PhD community, especially the 2018 promotion, you have made this shared journey so much relieving and fun. Thank you Álvaro, Fede, Morghan, Ivan, Sonia, Queralt, Laasya, Leila, Andrea, Julia, and so many others. A very sincere thank you to the training and academic office, Imma, Anna and Damjana, for your support during the whole PhD, and for the training and teaching opportunities. A la Reyes i la Magalí, pel vostre suport logístic i estar sempre disposades a ajudar. Al departament de comunicació, per comptar sempre amb mi per a divulgar la ciència.

Moreover, I would like to acknowledge the amazing months I spent at the Tao Pan's group in Chicago. Thank you Tao, Chris K, Chris W, Noah, Adam, and the whole team. Albeit short, you received me with open arms and I enjoyed every scientific and social time with you.

By finishing this PhD, I would also like to thank all the people that have mentored, educated and guided me. Al Josep Corominas, la Carme Taulés i la Sílvia Pasqual, i a tots els mestres de l'Escola Pia Sitges, per haver-me encès la flama de la passió per la ciència. A la Marta Cascante, el Josep Maria Fernández, el Joan Carles Ferrer, per haver-me obert les portes a la Bioquímica des de la Universitat de Barcelona. To Uwe Sauer, Michiel Karrenbelt, and Dimitris Christodoulou, for their enlightening critical thinking and for introducing me into the world of research at the ETH Zürich.

No puc deixar d'agrair a totes les persones que m'han acompanyat i donat suport directament o indirecta. Als "crujitas", per tots els sopars, escapades, festes; per ser-hi sempre des de petits. Als "enciams", pels retrobaments i les batalletes post-Bioquímiques. Als "ex-CJT", per ser la millor vàlvula d'escapament i sempre un suport indefallent; gràcies a CJT per haver-nos unit i educat. A la Moixiganga de Sitges, per ajudar-me a desconnectar participant de la cultura sitgetana i acollir-me en una colla vibrant i enèrgica. Als ex-companys de Zürich, per seguir compartint experiències plegats malgrat la vida ens ha desperdigat pel món.

Per acabar, vull fer un agraïment infinit a la meva família, per ser el meu pal de paller, sempre i en qualsevol moment. Als meus pares i la meva germana, per acompanyar-me, donar-me suport, esperonar-me, cuidar-me i animar-me a cada pas de la meva carrera. A l'Ava i la Iaia, per ser les RRPP més incondicionals de la meva recerca. A la "secta", per ser-hi sempre, tenint-nos cura els uns dels altres. Al Guim, que amb el seu naixement i alegria m'ha fet el padrí més cofoi del món. I a tota la família al complet. Us estimo.

Abstract

Although different tissues showcase differences in codon usage and anticodon tRNA repertoires, the codon-anticodon co-adaptation of multicellular eukaryotes is not completely understood. On the one hand, coding sequences are determined by manifold overlapping factors (codons, mRNA stability, splicing, etc.) and, on the other hand, tRNAs are intricately regulated at multiple levels (expression, modification, aminoacylation, fragmentation). In this thesis, we uncover the importance of tRNA and codon usage on mRNA translation and its tissue-specificity applying a systems biology approach to human high-throughput datasets. First, analyzing the tRNA abundance in over 8,000 tumor and healthy samples unveil that the variability of the tRNA pool is largely related to the proliferative state across tissues. We investigate the correspondence between tRNAs and human diseases, including cancer and viral infections. By leveraging proteomics and transcriptomics datasets, we next identify how transcripts in different tissues have distinct codon preferences. Finally, we discover a regulatory mechanism of tissue-specific translation through the coordination of tRNA modification patterns and tRNA aminoacylation. Altogether, this work not only provides evidence of tissue-specific tRNA expression and protein synthesis, but also makes this knowledge applicable to the development of tissue-targeted therapeutics.

Keywords: functional genomics, translation, codon usage, tRNA, machine learning.

Resum

Malgrat teixits diferents presenten diferències en l'ús de codons i en els repertoris d'anticodons dels ARNt, el coneixement sobre la co-adaptació entre codó-anticodó en eucariotes multicel·lulars és incomplet. D'una banda, les seqüències codificants depenen de diversos factors superposats (codons, estabilitat dels ARNm, empalmament, etc.) i, de l'altra, els ARNt són regulats de forma complexa a múltiples nivells (expressió, modificació, aminoacilació, fragmentació). En aquesta tesi, descobrim la importància dels ARNt i l'ús de codons en la traducció d'ARNm i la seva especificitat de teixit emprant mètodes de biologia de sistemes en conjunts massius de dades. Primer, l'anàlisi de l'abundància dels ARNt en més de 8.000 mostres sanes i tumorals revela que la variabilitat del conjunt d'ARNt està lligada a l'estat proliferatiu dels teixits. Investiguem la correspondència entre els ARNt i malalties humanes, com ara el càncer i les infeccions víriques. Utilitzant dades de proteòmica i transcriptòmica, a continuació identifiquem com els transcrits de teixits diferents tenen preferències de codons diferents. Finalment, descobrim un mecanisme de regulació de la traducció específica de teixit a través de la coordinació dels patrons de modificació i d'aminoacilació dels ARNt. En conclusió, aquest treball no només aporta evidència sobre l'especificitat de teixit en l'expressió d'ARNt i en la síntesi de proteïnes, sinó que també contribueix al desenvolupament de teràpies dirigides a teixits.

Conceptes clau: genòmica funcional, traducció, ús de codó, ARNt, aprenentatge automàtic.

List of publications

Hernandez-Alias X, Benisty H, Schaefer MH, Serrano L. Translational efficiency across healthy and tumor tissues is proliferation-related. Molecular Systems Biology. 2020;16(3):e9275.
Related to Chapter 3

Ghose R, Aranguren-Ibáñez Á, Arecco N, Balboa D, Bataller M, Beltran S, et al. From research to rapid response: mass COVID-19 testing by volunteers at the Centre for Genomic Regulation. F1000Research. 2020;9:1336.

Benisty H, Weber M, **Hernandez-Alias X**, Schaefer MH, Serrano L. Mutation bias within oncogene families is related to proliferation-specific codon usage. Proc. Natl. Acad. Sci. U.S.A. 2020;117(48):30848–56.

Delgado Blanco J, **Hernandez-Alias X**, Cianferoni D, Serrano L. *In silico* mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. PLOS Computational Biology. 2020;16(12):e1008450.

Head SA, **Hernandez-Alias X**, Yang J-S, Ciampi L, Beltran-Sastre V, Torres-Méndez A, et al. Silencing of SRRM4 suppresses microexon inclusion and promotes tumor growth across cancers. PLOS Biology. 2021;19(2):e3001138.

Hernandez-Alias X, Benisty H, Schaefer MH, Serrano L. Translational adaptation of human viruses to the tissues they infect. Cell Reports. 2021;34(11):108872.
Related to Chapter 4

Hernandez-Alias X, Katanski CD, Zhang W, Assari M, Watkins CP, Schaefer MH, Serrano L, Pan T. Single-molecule tRNA-seq analysis reveals coordination of tRNA modification and charging and fragmentation. Nucleic Acids Research. 2022; in press.
Related to Chapter 6

Benisty H, **Hernandez-Alias X**, Weber M, Anglada-Girotto M, Mantica F, Radusky L, et al. Evolutionary conservation of A/T-ending codons reflects co-regulation of expression and complex formation. bioRxiv. 2022;2022.01.17.475622.

Hernandez-Alias X, Benisty H, Serrano L, Schaefer MH. Using protein-per-mRNA differences among human tissues in codon optimization. bioRxiv. 2022;2022.03.22.485268.
Related to Chapter 5

Preface

In February 2001, Nature and Science published the first draft of the human genome. It constituted the beginning of a new era where sequencing would become routine in research and clinical practice. However, how far are we from the initial prospects of the Human Genome Project? In the announcement press release, the White House advocated:

> *Scientists will be able to use the working draft of the human genome to: (1) Alert patients that they are at risk for certain diseases; (2) Reliably predict the course of disease; (3) Precisely diagnose disease and ensure that the most effective treatment is used; and (4) Develop new, more effective treatments at the molecular level.*
>
> President Bill Clinton (June 26th, 2000)

More than twenty years later, we are still far from achieving 100% of these expectations, but why? Stemming from the classic experiments by Beadle and Tatum in *Neurospora crassa* (1941), biological research has followed for decades the "one gene - one protein - one function" paradigm (i.e. molecular biology paradigm), which is based on two assumptions: (1) genotype and phenotype are directly linked, and (2) proteins are organized in linear pathways, in which downstream functions are determined by upstream elements. From this perspective, one would expect that knowing the entire human genome with over 20,000 genes, we should now be able to know all the subsequent proteins and their function. In fact, technological advances allow us to identify, quantify and perturb genomes, transcriptomes, proteomes, metabolomes; but the links between genotype and phenotype have scarcely been established. At the start of this thesis, it was thus apparent that life was more than the sum of its parts, and therefore we needed a

holistic approach to the genotype-phenotype challenge: the systems biology paradigm.

In systems biology, we take systems as a network of nodes and edges, and thus place importance not only to the parts but also to their connections. With this in mind, we can then think of phenotypes as the outcome of a certain network state, and predict its response to specific internal (e.g. a gene knock-out) or external (e.g. a nutrient deprivation) perturbations. This has been the framework that has guided my thesis.

In this coming of age of the classic central dogma of biology, the flow of information along DNA-RNA-protein is no longer linear and unidirectional. DNA is accurately structured and epigenetically modified, then RNAs produced in transcription are also finely processed, spliced, structured, modified and degraded and have manifold upstream and downstream roles both as coding and non-coding sequences, and finally proteins are synthesized during translation, modified, translocated, folded individually or forming complexes, among many other interconnected processes. Within this broad picture, my thesis focuses on the regulation of translational elongation during protein synthesis, and more specifically on the role of tRNAs in decoding different three-letter combinations of nucleotides along coding sequences.

In the first chapter of the thesis, I introduce the necessary concepts and background of this work, as well as describe the state of the art. In the second chapter, I define the main objectives of the thesis. In chapters 3-6, all the performed studies with their corresponding methods and results are exposed and discussed. Finally, in the last chapter, I provide a general overview and discussion of the research in the light of the starting objectives.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1. The Central Dogma of molecular biology

The typical human body is composed of dozens of different tissues and hundreds of cell types. While they all share a common genotype, their gene expression needs to be finely regulated at many levels to showcase distinct phenotypes and functions. As first postulated by Crick in 1958, the *central dogma* of molecular biology defines that genetic information at the DNA is first copied into a messenger RNA (mRNA) molecule during *transcription*, which is finally used as a template for protein synthesis during *translation* (1). While many variations of the central dogma exist (mRNA splicing, non-coding RNAs, retrotranscription in RNA viruses, etc.), the overall flow of genetic information from DNA to proteins is universal among eukaryotes, archaea and prokaryotes (2).

Among coding sequences, the term *gene expression* encompasses the full process from reading the genetic instructions of DNA until producing a functional protein (3). It constitutes a highly complex process with many intricate regulatory mechanisms throughout multiple information layers: chromatin architecture, mRNA processing and export, RNA and protein interactions, metabolic allosterism, etc. In this

section, I will summarize the general interplay between DNA, RNA and proteins (**Fig. 1.1**).

1.1.1. From DNA to mRNA: Transcription

The first step of gene expression is the copy of a DNA gene into a single-stranded RNA nucleotide sequence—the *transcript*—during the process of transcription (**Fig. 1.1**). Among all transcripts of a typical human cell, the most abundant species are noncoding RNAs (ncRNA), whose functions are the RNAs themselves and hence are not translated into proteins (2). These include ribosomal RNAs (rRNA), which form the structure of ribosomes to synthesize proteins; transfer RNAs (tRNA), which are couplers between nucleotides and amino acids (see section 1.3); small nuclear RNAs (snRNA), which participate in RNA splicing and other nuclear processes; small nucleolar RNAs (snoRNA), which process and modify rRNAs; or microRNAs (miRNA), which inhibit expression of specific mRNAs and induce their degradation. Globally, over 90% of the transcriptome is composed of rRNAs and tRNAs, while mRNAs comprise only about 3-5% of total RNAs (2).

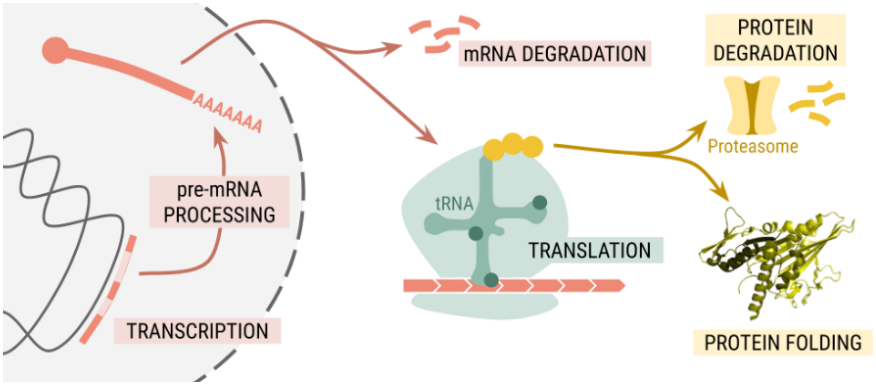


Figure 1.1. A schematic version of the gene expression pathway.
The expression of protein-coding genes starts with the transcription by RNA Pol II in the nucleus. The resulting pre-mRNA is then spliced, 5' capped, polyadenylated, and exported to the cytosol, where ribosomes catalyze the mRNA translation. The synthesized protein can finally adopt their functional three-dimensional structure. During all this process, quality control and turnover mechanisms exist which determine mRNA and protein half-lives.

The enzymes that catalyze the phosphodiester bond between ribonucleotides during transcription are called *RNA polymerases*. Eukaryotes have three RNA polymerases: RNA polymerase I is in charge of transcribing 5.8S, 18S and 28S rRNA genes; RNA polymerase II transcribes most human genes, including all protein-coding ones; and RNA polymerase III is responsible for transcribing tRNAs and some other ncRNAs (2). To initiate the synthesis of mRNAs, RNA polymerase II requires the binding of *transcription factors* to correctly position at the promoter, unwind the double-stranded DNA, and start the transcript elongation. Furthermore, this process occurs in the context of histone-associated DNA in the form of chromatin, which is regulated by chromatin remodeling complexes and epigenetic marks.

After transcription, eukaryotic mRNAs—so-called precursor mRNAs or pre-mRNAs—undergo several processing steps: *capping* of the 5′ ends with 7-methylguanosine, removal of introns and junction of exon sequences by *RNA splicing*, and *polyadenylation* of 3′ ends (**Fig. 1.1**) (2). Fully processed transcripts can then be exported from the nucleus to the cytosol and translated into protein.

1.1.1.1. mRNA structure and degradation

The structure of single-stranded mRNA molecules is dependent on the interactions between nucleotide base pairs, which can form extremely complex and functional secondary and tertiary structures. Highly structured transcripts are generally associated with longer half-lives (4), with the exception of strong structures near the start codon, since ribosomes require more energy to initiate translation and thus leave the molecule unprotected from degradation (5,6).

The half-life of transcripts can be extremely variable, ranging between the many hours of highly abundant genes, and the few minutes of short-lived genes that need to be expressed in bursts (3,7). Mechanisms responsible for degradation can be classified into two categories: mRNA surveillance mechanisms that degrade improperly processed transcripts, and general mRNA turnover mechanisms. The first group includes the

nonsense-mediated decay (NMD)—degrades mRNA containing premature stop codons, generally because of incorrect splicing (8)—, *no-go decay* (NGD)—degrades transcripts containing stalled ribosomes due to damaged bases or blocking secondary structures (9)—, and the *non-stop decay* (NSD)—degrades mRNAs not containing a stop codon (10).

mRNA turnover is mainly determined by the gradual poly-A tail shortening of transcripts (2). When the poly-A length is reduced below ~25 bases, the transcript is decapped and degraded by 5'-exonucleases. In some cases, the decapping and degradation of mRNAs can also happen through deadenylation-independent pathways (7). Some transcripts can be degraded by an endonucleolytic cleavage, which is the case of RNA interference pathways (11).

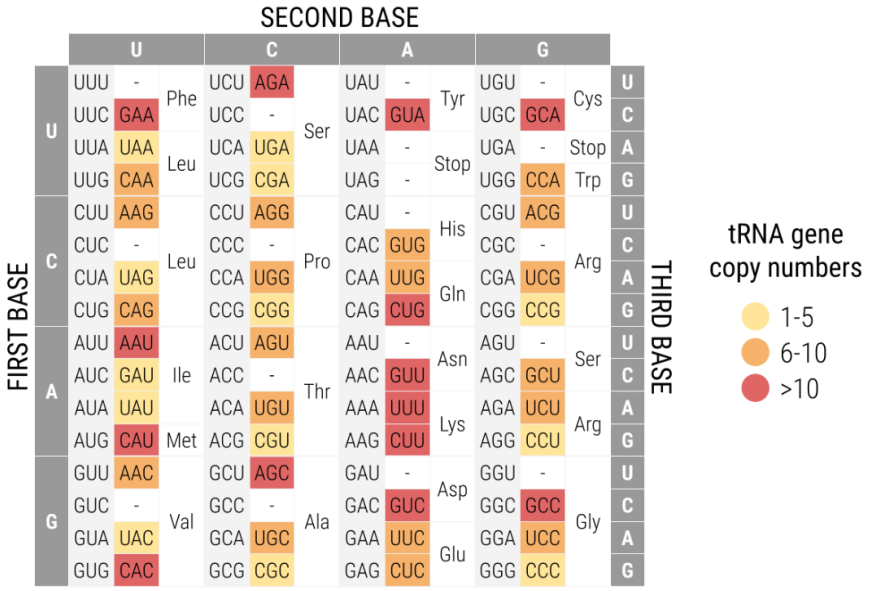


Figure 1.2. The universal genetic code and their cognate tRNA gene copy numbers in the human genome.
The genetic code establishes the correspondence between the 20 amino acids (in white) and the 64 codons (in light gray), which include three stop and 61 amino-acid encoding codons. Amino-acid encoding codons are recognized during translation by the anticodon of cognate tRNAs, which are variably abundant along the human genome (12) (see colors). The base pairing between the third base of codons and the first base of anticodons often follows non-Watson-Crick rules (see section 1.3.6).

1.1.2. From mRNA to protein: Translation

Once the mRNA is exported from the nucleus to the cytosol, the ribosomes are responsible for translating the nucleotide sequence into another sequence of chemically distinct residues: amino acids (**Fig. 1.1**) (2). In consequence, the code composed of 4 nucleotides (A, C, G, T/U) needs to be translated into 20 amino acids, which precludes a 1-to-1 correspondence. In this combinatorial problem, nucleotides are grouped in 4^3=64 possible triplets—so-called *codons*—that are unambiguously assigned to the 20 amino acids. These rules constitute the *genetic code* (**Fig. 1.2**), which is universal in all life forms with very few exceptions for some particular codons, such as mitochondrial genes (13).

Among all 64 codons, 61 of them encode for amino acids and the other three are stop codons (a.k.a. non-sense or termination codons), which signal the termination of translation. Out of 20 amino acids, 18 of them can therefore be encoded by two or more codons, which are called *synonymous*. During the process of translation, the adaptor molecules responsible for recognizing the mRNA codons and binding the corresponding amino acids are *transfer RNAs* (tRNA, see section 1.3) (2). Briefly, among the full tRNA structure, three bases in one of their loops constitute the anticodon, which is in charge of pairing with the complementary codon. However, in the human genome, there are only 46 different anticodons that need to recognize 61 distinct codons (**Fig. 1.2**) (12). In consequence, some codons require non-Watson-Crick base-pairing rules—so-called *wobble* base-pairing—which tolerate a mismatch at the third codon position (14). Nucleotide modifications of tRNAs play an important role in determining these base-pairing rules.

Eukaryotic ribosomes consist of a 60S large subunit—formed by the 5S, 28S and 5.8S rRNAs and ~49 proteins—and a 40S small subunit—formed by the 18S rRNA and ~33 proteins (**Fig. 1.3**) (2). The two subunits are separately assembled in the nucleolus and then exported to the cytoplasm. Ribosomes contain four RNA binding sites; one for the mRNA and three for tRNAs (aminoacyl or *A site*, peptidyl

or *P site*, and exit or *E site*). During mRNA translation, the small subunit contains the recognition interface between mRNA codons and tRNA anticodons, while the large subunit catalyzes the formation of peptide bonds, at a rate of ~5 amino acids per second (15,16). As a result, the *translation efficiency* is defined as the rate of protein production per mRNA transcript.

1.1.2.1. Translation initiation

The AUG codon for methionine constitutes the start codon of coding sequences. Therefore, mRNA translation starts with the assembly on the start codon of the *80S initiation complex* containing the two ribosomal subunits, the $tRNA_i^{Met}$, and the mRNA (**Fig. 1.3**) (7). Specifically, the small ribosomal subunit first binds to the $tRNA_i^{Met}$ at the P site—$tRNA_i^{Met}$ differs from elongating $tRNA^{Met}$, which can only bind to the A site. Next, this complex recognizes the 5' end of mRNAs and starts scanning the sequence until finding the first downstream AUG, generally within the first 100 nucleotides. At this point, the large ribosomal subunit joins the complex. The formation of the initiation complex is coordinated by the eukaryotic translation *initiation factors* (eIFs). While translation initiation is generally the rate-limiting step in protein synthesis, it can be affected by ribosome stalling and collisions during elongation (17,18).

1.1.2.2. Translation elongation

Once the initiation complex is formed, the peptide chain starts to elongate, which is mediated by specialized proteins called *elongation factors* (EFs) (7). In particular, the initiation complex contains an occupied P site with the $tRNA_i^{Met}$, while the adjacent A site is empty and available for binding to another tRNA. Therefore, aminoacylated tRNAs enter the A site in association with EF1α·GTP as a *ternary complex* (**Fig. 1.3**). When the incoming tRNA correctly base-pairs with the corresponding mRNA codon, GTP is hydrolyzed allowing a tight codon-anticodon binding. The peptidyl-transferase reaction then takes place between the P-site peptide and the A-site amino acid. Upon hydrolysis of EF2·GTP, the ribosome translocates a distance of one

codon along the transcript; the E site contains the non-aminoacylated $tRNA_i^{Met}$, the P site contains the peptidyl-tRNA, and the A site is available again to repeat the process. In the following round, the E-site tRNA will be ejected upon the EF1α·GTP hydrolysis. During the elongation process, the nascent polypeptide slides through a channel in the large ribosomal subunit (7).

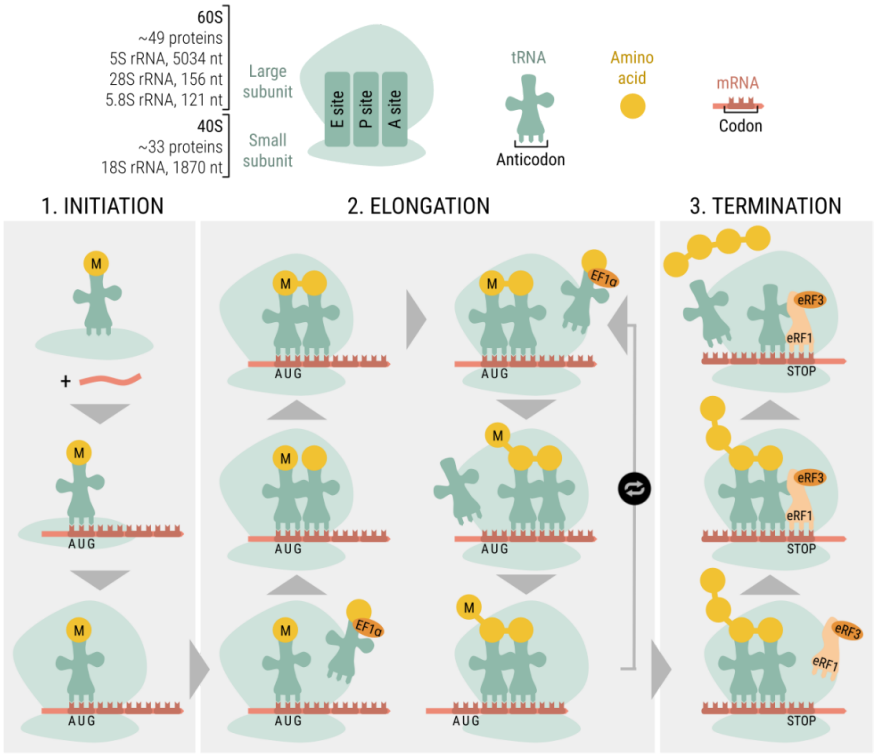


Figure 1.3. mRNA translation in eukaryotes.
Elements required for translation include ribosomes, aminoacylated tRNAs and the mRNA. In initiation, by orchestration of several eIFs, the complex formed by the small subunit and $tRNA_i^{Met}$ starts scanning the mRNA until the initiator AUG codon is recognized, and then the large unit is incorporated. In elongation, three sequential steps are successively repeated: the ternary complex with the cognate tRNA base pairs with the A-site codon, the peptide bond is formed, and the ribosome translocates to the next codon. When the stop codon is reached, the tRNA-like eRF1 binds at the A site and the eRF3 cleaves the peptidyl-tRNA bond of the last incorporated tRNA.

1.1.2.3. Translation termination

There stop codons (UAG, UAA, UGA) exist in the genetic code, which need to be recognized by the elongating ribosome to terminate translation. Two protein *eukaryotic release factors* (eRFs) have been identified: eRF1 resembles the structure of a tRNA and recognizes the stop codons, eRF3 then promotes cleavage of the peptidyl-tRNA bond (**Fig. 1.3**) (7).

To increase the rate of mRNA translation, multiple ribosomes are actively translating one single mRNA molecule simultaneously, generating polyribosomes or *polysomes* (7). Moreover, the mRNA folds in translation forming a circular 3D structure that positions 5' and 3' ends closeby, and hence the ribosomal subunits and translation factors can be efficiently recycled.

1.1.2.4. Protein degradation

The lifespan of proteins can be extremely variable, since some proteins are required in short bursts while others remain for the entire human life, such as the crystallin proteins of the eye (7). There exist two main degradation mechanisms in the cell: *lysosomes* and *proteasomes* (19,20). The first consist of vesicles with an acidic lumen containing hydrolytic enzymes, which can degrade damaged organelles and extracellular proteins. On the other hand, most damaged, misfolded or regulated intracellular proteins are enzymatically polyubiquitinated, which targets them to the proteasome, an abundant ATP-dependent protease complex.

1.1.3. The regulatory landscape of gene expression

As described in this first section of the introduction, the gene expression pathway involves an intricate interplay of mechanisms that expand from the gene to the functional protein. Therefore, each step in this process can be regulated to determine the cell phenotype. With the advent of simultaneous and system-wide measurements of transcriptomes and proteomes, modeling efforts indicate that ~50% of protein variability

across human tissues is determined at the post-transcriptional level, and hence cannot be explained by mRNAs alone (21).

The gene expression pathway is hierarchical, and the expressed products are successively amplified from DNA-mRNA-protein. Therefore, while transcripts do not generally exceed the few thousands per cell, the dynamic range of proteins can reach ~10^8 molecules/cell, which can hence be more widely regulated (3). In fact, recent studies suggest that post-transcriptional regulation has an important buffering role, denoising the intrinsic variability of transcriptional bursts and thus allowing a more robust control of gene expression (3,22). Furthermore, given that translation is at the last steps of gene expression, its regulation can generate much faster cellular responses than transcriptional control (23).

1.2. The determinants of coding sequences

While synonymous codons lead to identical amino acid sequences, they are not uniformly distributed in the genome or between different organisms—i.e. *codon usage bias* (CUB) (17,24,25). The CUB can have a huge impact on gene expression at multiple levels, as observed both in endogenous and heterologous proteins (26–28). In particular, coding sequences are determined by multiple overlapping codes, including determinants of transcription, epigenetics, splicing, folding, interactome or translation (29). In this section, I will focus on the translation-related effects of CUB in the human genome.

1.2.1. Codon composition along the human genome

The human genome has an average GC content of 40.5% (genome assembly hg38), but this percentage increases to 51.1% for protein-coding sequences and 55.3% considering only the GC content at the third or wobble position of codons (GC3), which is the variable position between synonymous codons (**Fig. 1.2**) (30). Humans, as most mammal species, are therefore biased towards G/C-ending codons (**Fig. 1.4A**). Commonly used codons—mostly G/C-ending—are called

preferred or optimal codons; the others are referred to as rare or nonoptimal codons.

The GC content varies widely between chromosomal regions or isochores, which is mostly caused by *GC-biased gene conversion* (31). During meiotic recombination, biases in double strand break sites and mismatch repair mechanisms favor the transmission of G/C over A/T alleles. In consequence, isochores with higher recombination rates, which correspond with higher expression levels in meiosis (32), showcase higher G/C-ending CUB due to GC-biased gene conversion.

On the other hand, variation of CUB across genes is closely related to mRNA translation at multiple levels (25). Optimal codons are associated with higher translation elongation and initiation rates (33,34), higher translation efficiency (28), and lower premature termination (35). In fact, changes in CUB of reporter genes in HeLa cells show up to 46-fold differences in protein levels (27), which has been widely exploited in biotechnology. For instance, in heterologous gene expression, codon optimization is commonly applied to design sequences that resemble the codon usage of highly expressed genes of the host (36). Mispairing between tRNAs and codons can also hamper *translation fidelity*, and a trade-off exists between translation speed and error rates (37). How these effects of CUB in translation are mediated by human tRNAomes will be further discussed in section 1.3.6.

Recent evidence shows that CUB not only has an effect on protein synthesis, but also on *mRNA stability* (38–40). Stable mRNA transcripts are enriched in optimal codons, decoded by highly abundant and aminoacylated tRNAs, while rare codons lead to mRNA decay. In yeast, the DEAD-box helicase Dhh1 was identified to bind to slow-moving ribosomes and trigger the mRNA degradation (41).

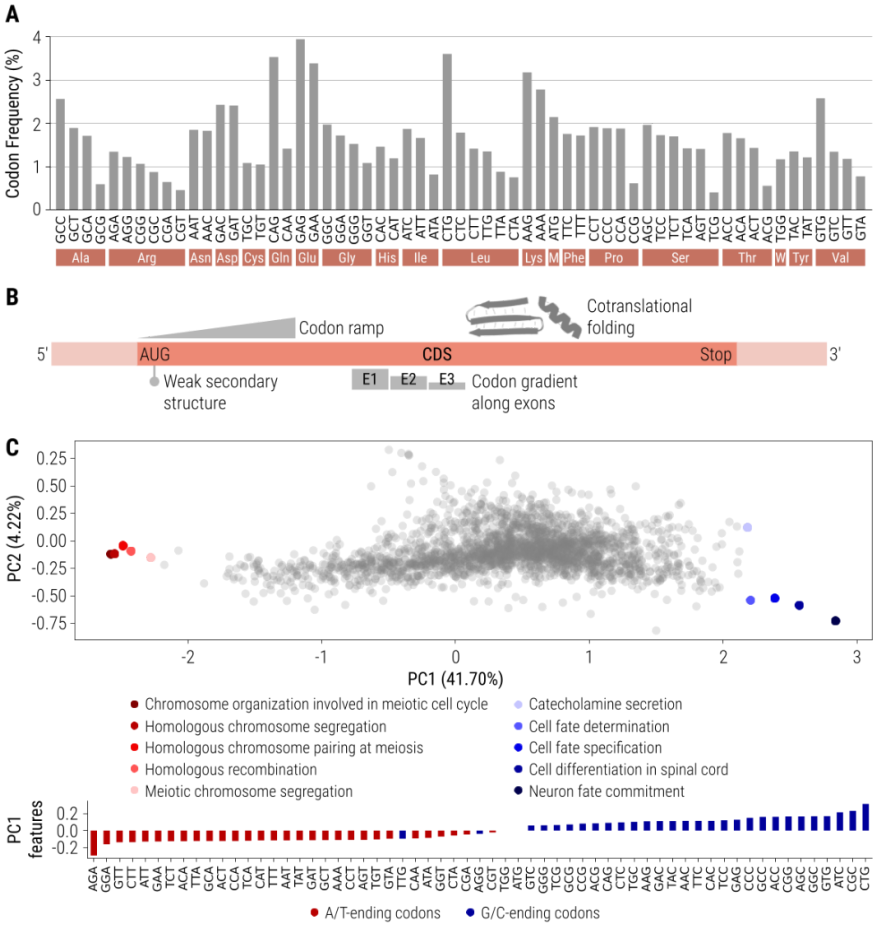


Figure 1.4. Codon composition of the human genome.
(A) Frequencies of all 61 amino-acid encoding codons in the human genome from the CoCoPUTs database (30), ordered by the amino acid they encode. (B) Overlapping factors that shape the codon usage along genes. (C) Principal Component Analysis of the Relative Synonymous Codon Usage (RSCU) of human genes. On this projection, the average RSCU of all GO Biological Process classes with >40 genes are depicted, similar to Gingold et al. (42). Top and bottom classes are colored and labeled. The contribution of each codon to the PC1 component is shown at the bottom, where bars are colored based on the A/T vs G/C wobble base of codons.

1.2.2. Codon composition within genes

In addition to CUB differences across genes, there are several factors that determine the choice of synonymous codons within genes (**Fig. 1.4B**) (24,43). First, coding sequences select for or against certain *regulatory motifs*, such as microRNA binding sites, splicing control elements, or nucleosome positioning (44). Interestingly, single-exon genes and 5′-exons of multi-exon genes are generally more G/C-rich (27).

On the other hand, a gradient from nonoptimal to optimal codons exist at 5′-proximal ends of coding sequences by two main mechanisms: (a) strong *mRNA secondary structures* can hamper translation initiation (5), and (b) a *codon ramp* of rare codons slows translation elongation at the beginning of genes to avoid ribosomal traffic jams (45).

Folding of most proteins occurs cotranslationally *in vivo*, and hence mRNA translation kinetics can alter *protein folding* and structure (25). In general, weakly structured or disordered domains require longer time to fold and are hence more sensitive to translation kinetics, while well-structured domains fold more rapidly and robustly. For instance, low-complexity proteins in humans, mostly related to cell adhesion functions, are sensitive to altered tRNA modifications such as inosine (46). Moreover, the conservation of clusters of optimal and nonoptimal codons is associated with the protein secondary structure (47,48). In consequence, the codon from which an amino acid is translated can determine its dihedral angle within the protein backbone (49).

1.2.3. Codon composition across conditions

Given that CUB affects mRNA translation and there exists CUB variability across genes, multiple studies have proposed that CUB can functionally regulate translation of specific genes in certain conditions (14,17). For instance, in yeast, the codon usage of the transcriptome changes under different environmental stresses (50), optimal codons are particularly enriched in important pathways such as glycolysis (38), and

this codon adaptation across metabolic pathways differs between aerobic versus anaerobic species (51).

In human cells, subsets of genes sharing a common GO function showcase differential CUB, with the two extremes corresponding to genes involved in cell proliferation and differentiation (**Fig. 1.4C**) (42). In particular, proliferation genes are A/T-rich, while G/C-ending codons are abundant in differentiation counterparts. Differential ribosomal pausing on specific codons is actually detected along the mitotic cell cycle in human cells (52). Several studies have proposed the dynamic regulation of tRNAomes to coordinate these changes in proliferation codon usage (42,53).

Similarly, differences in CUB across human tissues have been reported (54,55). Genes with similar synonymous codon usages tend to share common functions, similar expression patterns across tissues, and coordinated interacting partners (56,57).

1.2.4. CUB: neutral or selected?

The existence of CUB across genes and organisms has long been recognized and studied from an evolutionary perspective (58). While there is no doubt that differences in synonymous codon usage have an impact on gene expression, the evolutionary origin of CUB is debated between two hypotheses (24). On the one hand, the *mutational pressure* shapes genomes as a result of biases in nucleotide mutation rates or repair mechanisms, which cause neutral synonymous changes without any impact on fitness. For instance, in mammals, the GC-biased gene conversion is believed to explain much of the existing variability in GC content along the genome (32). On the other hand, *natural selection* states that synonymous changes can be advantageous or detrimental to organisms and are therefore positively or negatively selected in evolution. This is a major contributor in short-lived species with large effective population sizes (59). In fact, the two hypotheses are not mutually exclusive, and hence the actual human CUB is likely a combination of both mutational and selective forces (58).

1.3. Transfer RNA

Transfer RNAs constitute the second most abundant RNA species of the human cell after rRNA, accounting for around 15% of total RNA (7). Despite their transcriptomic abundance, tRNA genes occupy <0.002% of the human genome, with a total of 429 loci (12). These genes encode for 46 different *isoacceptors*, i.e. tRNA species with a distinct anticodon (**Fig. 1.2**). Therefore, for each tRNA isoacceptor family, there is an average of nine gene copies or *isodecoders*, which can differ slightly in their sequence. Furthermore, tRNA genes are not uniformly distributed along the human genome, but they are clustered both linearly and three-dimensionally in the nucleolus (2).

In evolutionary terms, huge differences in the tRNA gene copy number exist among organisms, with higher numbers generally related to larger genomes (14). Moreover, the variability in tRNA composition between kingdoms is associated with their tRNA modification pattern and their genomic codon usage (60,61).

The mitochondrial genome also encodes for their own set of tRNAs, which consist of 22 single-copy distinct isoacceptors (14). Compared to the cytosolic counterparts, this reduced set of tRNAs showcase differences in structure, modification patterns, and codon-anticodon base pairing rules; their aminoacylation is also catalyzed by nuclearly-encoded mitochondria-specific aminoacyl-tRNA synthetases. Given the low redundancy of the mitochondrial genome, alterations in their tRNA sequences and machinery have been more frequently related to mistranslation and human diseases (62).

1.3.1. tRNA transcription

RNA Pol III is the enzyme responsible for transcribing tRNA genes in eukaryotes, which is formed of 17 subunits (see section 1.1.1). The promoter of tRNA genes is internally located within their sequence and consists of two motifs known as the *A box* and the *B box* (**Fig. 1.5**) (63). In fact, tRNA isodecoders whose B box deviates from the consensus

sequence showcase lower expression levels (64). These sequences are recognized by TFIIIC, which then recruits TFIIIB and RNA Pol III and initiates transcription. After termination, which requires a stretch of T residues, RNA Pol III can be recycled and transcription reinitiated, hence allowing enhanced transcription. The elevated transcription rate of tRNAs leads to high transcription-associated mutagenesis, while at the same time experiencing a strong purifying selection (65).

In basal growth conditions, HEK293T cells express only around half of all their tRNA genes, while the rest remain silent; they are suggested to be tRNA pseudogenes or play other extra-transcriptional roles (66). On the other hand, more than half of the detectable tRNAs are differentially expressed at the isodecoder level among human cell lines (67). However, at the isoacceptor level, tRNA pools appear more stable (67,68).

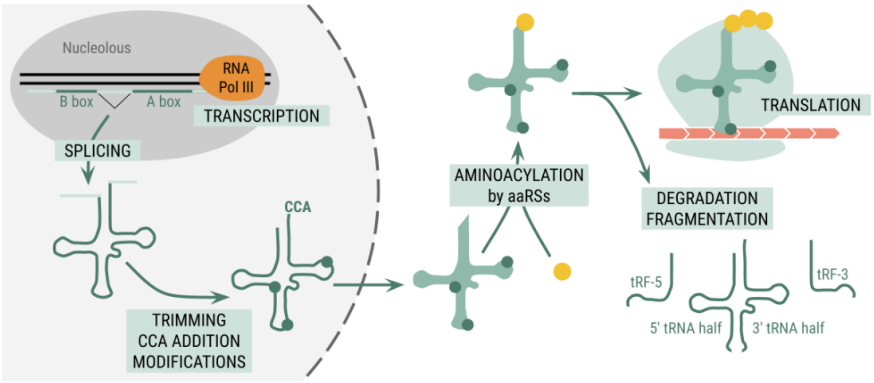


Figure 1.5. tRNA biosynthesis and turnover.
Transcription of tRNA genes is mediated by RNA Pol III in the nucleolus, which is regulated by transcription factors, epigenetic marks and chromatin organization. After transcription, pre-tRNAs require trimming of 5' leader and 3' trailer sequences, splicing of introns (if present), addition of a CCA tail, and export to the cytosol. tRNAs are post-transcriptionally modified during the biosynthesis process, either at the nucleus or the cytosol. Once tRNA species have been correctly aminoacylated by their cognate amino acids, they can be used in mRNA translation. The regulated fragmentation of tRNAs can generate specific tRFs with diverse functions beyond protein synthesis.

Transcription by RNA Pol III can be regulated by factors such as MAF1, which is regulated by mTORC1 phosphorylation and represses tRNA transcription under nutrient stress (69,70). Other transcription factors that regulate RNA Pol III include SOX4, Myc, p53, or Rb (71–74). RNA Pol II has also recently been reported to regulate tRNA transcription by Pol III (75). Similar to protein-coding genes, the chromatin status of tRNA genes can also regulate their transcription. For instance, activatory and inhibitory histone modifications and DNA methylation are associated with changes in tRNA expression and can change in development, aging, and cancer (42,76–78). Furthemore, the topological organization of tRNA genes within linear clusters and 3D domains coordinates transcription during cellular differentiation (78).

After transcription, precursor tRNAs (*pre-tRNA*) contain a 5' leader sequence, a 3' trailer sequence and, in 5% of human tRNA genes, an intron at residue 37 (14). Therefore, before leaving the nucleus, pre-tRNA undergo cleavage of the leader and trailer sequences by RNase P and RNase Z endoribonucleases, respectively; splicing of introns; and addition of a *CCA tail* by the template-independent CCA-adding enzyme (**Fig. 1.5**) (63). Interestingly, the 3' trailer sequence is the binding site of the La protein, which facilitates the proper folding and prevents exonuclease digestion. While not much is known on the dynamic regulation of tRNA processing mechanisms, mutations in this machinery can cause neurodevelopmental diseases (79).

1.3.2. tRNA structure

In 1965, Holley et al. determined the first structure of a tRNA, a yeast $tRNA^{Ala}$ (80); the secondary and 3D structure of tRNAs is generally conserved across the three kingdoms of life. In humans, the length of cytoplasmic tRNAs range between 73 and 97 nucleotides (12), which fold forming a cloverleaf-like secondary structure (**Fig. 1.6A**). Based on this canonical structure, tRNA sequences are numbered starting from base 1, which is the 5′ base that pairs with base 72, until the CCA tail, which corresponds to positions 74, 75, 76 (63). The tRNA secondary

structure consists of four arms: the *acceptor stem* (a.k.a. A stem), the *D stem-loop* (a.k.a. dihydrouridine stem-loop), the *anticodon stem-loop* (a.k.a. C stem-loop, which contains the anticodon at positions 34, 35, 36), and the *T stem-loop* (a.k.a. TΨC stem-loop, where Ψ refers to pseudouridine). In some tRNAs, mainly $tRNA^{Ser}$ and $tRNA^{Leu}$, an additional arm (*variable stem-loop*) extends between base 45 and 46, whose nucleotides are numbered with an "e".

In three dimensions, tRNAs fold into an L-shaped tertiary structure composed of two domains joined at right angles, which is maintained through non-canonical pairing between T and D arms (**Fig. 1.6B**) (81). The acceptor domain is formed by the A and T arms folding into a minihelix, while the C and D arms form the anticodon dumbbell. Interestingly, the anticodon and the amino-acid attachment site, two critical loci of tRNAs, are located at the two furthest extremes of the L-shaped structure.

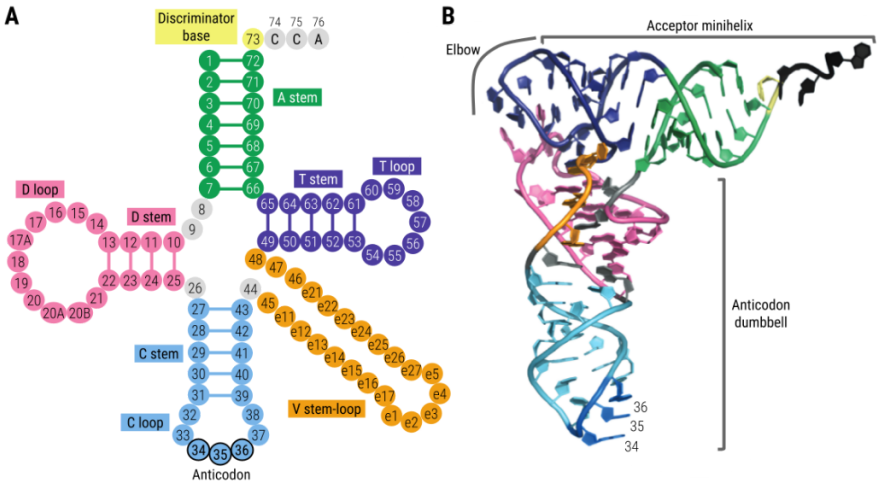


Figure 1.6. Canonical tRNA structure.
(A) Secondary cloverleaf-like structure of tRNAs and its corresponding base numbering. Colors correspond to the acceptor or A stem (green), the dihydrouridine or D stem-loop (pink), the anticodon or C stem-loop (blue), the variable or V stem-loop (orange), and the TΨC or T stem-loop (purple). (B) Three-dimensional L-shaped structure of tRNAs, which consists of the acceptor minihelix and the anticodon dumbbell. Colors of nucleotides are matched between (A) and (B). Figure adapted from Berg et al. (63).

1.3.3. tRNA modification

All tRNAs are chemically modified throughout their biosynthesis, with an average of 13 modifications per molecule (82). However, while some modifications are common in most tRNAs (e.g. dihydrouridine in the D loop and pseudouridine in the T loop), many others are unevenly distributed among specific tRNAs. This tRNA epitranscriptome is orchestrated by tRNA modification enzymes, including both writers and erasers (**Fig. 1.7A-B**) (83).

To date, over 40 types of modifications have been identified in human cytosolic and mitochondrial tRNAs, which include methylations, hydroxylations, acetylations, or deaminations (**Fig. 1.7C**) (84). Of those, more than 20 are located at the first position of the anticodon or wobble position (base 34), which highlights the importance of tRNA modifications to expand or restrict the codon-anticodon recognition during mRNA translation. For instance, while adenine can only pair with uracil, *A-to-I editing* at position 34 by tRNA-dependent adenosine deaminases (ADATs) expands the pairing capacity to U, C and A at the wobble position of codons (**Table 1.1**) (60,85). On the other hand, modifications outside the anticodon loop can also regulate the structural stability and stiffness of tRNAs, their degradation, or their aminoacylation specificity. Furthermore, aberrant tRNA modifications can lead to mitochondrial diseases, neurological disorders or cancer (82,84).

1.3.4. tRNA aminoacylation

Chapeville et al. established in 1962 that ribosomes are blind to the amino acid that is being incorporated during mRNA translation, and hence it is critical that the correct amino acid is coupled to the cognate tRNAs (86). tRNA aminoacylation (a.k.a. tRNA charging) is catalyzed by 20 *aminoacyl-tRNA synthetases* (aaRSs), one enzyme for each amino acid (**Fig. 1.5**) (2). The resulting aminoacylated tRNAs can be then delivered to the ribosomes by elongation factors.

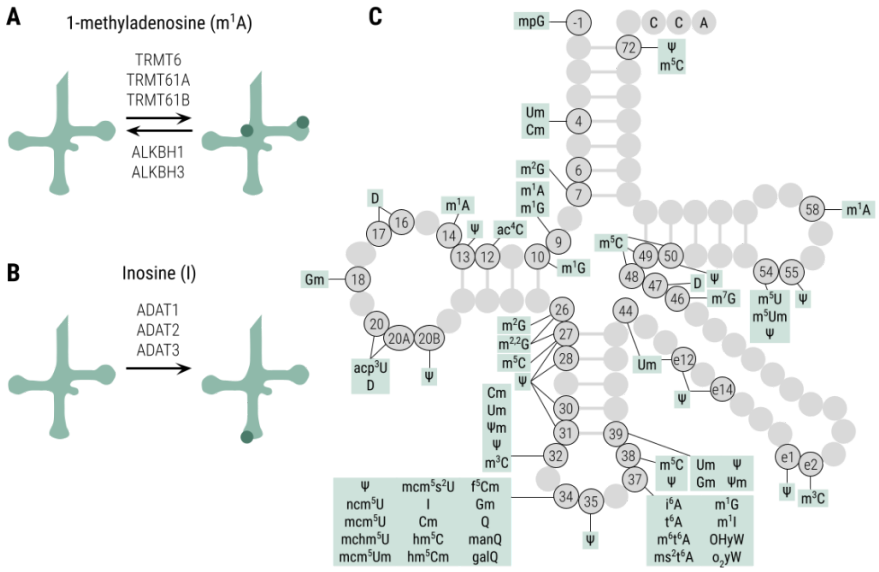


Figure 1.7. tRNA modifications and enzymes.
(A, B) Enzymes writing or erasing m^1A and I modifications on human cytosolic tRNAs. (C) tRNA modifications of human cytosolic tRNAs. Nucleotides are numbered based on the canonical tRNA secondary structure (Fig. 1.6A). Abbreviations correspond to the MODOMICS database (87): ac^4C, *N*^{4}-acetylcytidine; acp^3U, 3-(3-amino-3-carboxypropyl)uridine; Cm, 2'-*O*-methylcytidine; D, dihydrouridine; f^5Cm, 5-formyl-2'-*O*-methylcytidine; galQ, galactosyl-queuosine; Gm, 2'-*O*-methylguanosine; hm^5C, 5-hydroxymethylcytidine; hm^5Cm, 2'-*O*-methyl-5-hydroxymethylcytidine; I, inosine; i^6A, *N*^{6}-isopentenyladenosine; m^1A, 1-methyladenosine; manQ, mannosyl-queuosine; m^3C, 3-methylcytidine; m^5C, 5-methylcytidine; $mchm^5U$, 5-(carboxyhydroxymethyl)uridine methyl ester; mcm^5s^2U, 5-methoxycarbonylmethyl-2-thiouridine; mcm^5U, 5-methoxycarbonylmethyluridine; m^1G, 1-methylguanosine; m^2G, *N*^{2}-methylguanosine; $m^{2,2}G$, *N*^{2},*N*^{2}-dimethylguanosine; m^7G, 7-methylguanosine; m^1I, 1-methylinosine; mpG, 5'-methylphosphoguanosine; ms^2t^6A, 2-methylthio-*N*^{6}-threonylcarbamoyladenosine; m^6t^6A, *N*^{6}-methyl-*N*^{6}-threonylcarbamoyladenosine; m^5U, 5-methyluridine; m^5Um, 2'-*O*-methyl-5-methyluridine; ncm^5U, 5-carbamoylmethyluridine; OHyW, hydroxywybutosine; o_2yW, peroxywybutosine; Q, queuosine; t^6A, *N*^{6}-threonylcarbamoyladenosine; Um, 2'-*O*-methyluridine; Ψ, pseudouridine; Ψm, 2'-*O*-methylpseudouridine. Figure adapted from Suzuki et al. (84).

AaRSs catalyze the esterification between an amino acid and the CCA tail of a tRNA, which requires one ATP molecule (88). The reaction involves two different active sites, a catalytic site and an editing site, which define the "double sieve" model to explain the proofreading activity of aaRSs (89). First, the amino acid is activated by covalently binding AMP, in which amino acids that are large enough are excluded from the active site (90). AaRSs then select the cognate tRNAs by reading both their nucleotide composition (e.g. anticodon, acceptor stem sequence, modified bases) and structural features, and catalyze the reaction between the 3′ adenine and the activated amino acid. However, smaller amino acids than the correct one can slip through this first sieve. Therefore, most aaRSs contain an editing site, which reverses the misactivation of amino acids or the misacylation of tRNAs by hydrolysis (88).

In eukaryotes, some aaRSs are commonly assembled into the *multisynthetase complex*, formed in mammals by nine aaRS (GlnRS, ProRS, GluRS, IleRS, LeuRS, MetRS, LysRS, ArgRS, and AspRS) and three noncatalytic auxiliary proteins (88). The complex has been proposed to help channel the substrates and deliver the product to ribosomes. Furthermore, aaRSs, including their alternatively spliced variants, have been involved in a plethora of other functions outside of mRNA translation, such as signaling, biosynthetic activities, or splicing (81).

1.3.5. tRNA degradation

Mature tRNAs are highly stable with a half-life of 100h due to their abundant modifications and rigid three-dimensional structure (91). However, damaged molecules that are unable to bind to tRNA binding proteins (elongation factors, aaRSs) become unprotected and undergo rapid tRNA decay (14). Apart from the general turnover of tRNAs, cells also dynamically regulate the ready-to-translate tRNA pool under oxidative stress by the angiogenin-induced endonucleolytic cleavage of CCA tails (92). Upon stress relief, CCA termini are rapidly restored by the CCA-adding enzyme.

On the other hand, *tRNA-derived fragments* (tRFs) have been detected among cellular small RNAs (93). Far from being randomly generated debris, tRFs are abundant, have discrete lengths, and originate from specific termini. In particular, six main types of fragments have been described (**Fig. 1.5**): 5′ and 3′ tRNA halves, which originate from cleavage at the anticodon loop; tRF-5 and tRF-3, which are smaller and begin at 5′ or 3′ ends; tRF-1, which comes from the 3′ trailer; and internal tRFs arising from the anticodon loop. While several endonucleases have been involved in tRF generation (93), most studies have focused on the angiogenin-induced cleavage under different types of stress. Given that stress-induced cleavage is <2% of total tRNA (94), the function of tRFs is not likely mediated by depletion of tRNA pools. For instance, tRFs can bind and control translation-related proteins, base-pair with mRNAs and regulate their expression through microRNA-like pathways, or sequester RNA-binding proteins (RBPs). tRFs have been implicated in many processes such as protein synthesis, mRNA degradation, apoptosis, stress granule formation, and epigenetic inheritance (62,63).

1.3.6. Couplers between genes and proteins

The main function of tRNAs is mRNA translation, in which 46 isoacceptors recognize 61 codons to incorporate the correct amino acid into the protein sequence (**Fig. 1.8A**, see section 1.1.2). In consequence, considering the A-to-I editing at the wobble position of ANN anticodons (**Table 1.1**), a high correlation exists between the tRNA gene copy numbers and the genomic codon usage (60). Given these base-pairing rules, all ANN isoacceptors of split codon boxes (Phe/Leu, Tyr/Stop, His/Gln, Asn/Lys, Asp/Glu, Cys/Stop/Trp, Ser/Arg) are absent in the human genome (**Fig. 1.2**), since their A-to-I editing would otherwise lead to mistranslation. In terms of abundance, there exists a correlation between codon usage and tRNA levels, where common codons correspond with enriched tRNA isoacceptors (17). Ribosome profiling data actually indicate that slowly translated codons are generally decoded by low-abundance tRNAs, and vice versa (95).

While tRNAomes and codon usage were long perceived as static, recent reports suggest that a balance exists between the tRNA *supply* and their *demand*—i.e. the consumption of tRNAs in actively translated mRNAs (14). Simultaneously translating ribosomes need to compete for a limited tRNA pool; scarce isoacceptors recognizing rare codons are hence more sensitive to changes in abundance, with higher regulatory potential (**Fig. 1.8B)** (53). In fact, changing the tRNA demand by recoding highly expressed genes alters the translation efficiency of the rest of the proteome, which can then be alleviated by increasing the corresponding supply (96). Furthermore, the supply not only accounts for tRNA abundance, but involves any required process in the formation of the ready-to-translate tRNAome (**Fig. 1.8C**).

In support of this supply-demand model, tRNA abundances, charging and modifications are regulated under different stresses in unicellular organisms, which favor the translation of stress-response transcripts, enriched in rare codons (17,97,98). In multicellular eukaryotes, coordination between tRNAomes and codon usage has been also reported in amino acid deprivation (99,100), cell proliferation (53,101), specific tissues (102–104), and several cancer types (105,106); tRNA isoacceptors appear stable in other cell states (33,68). However, given the interplay with other selection forces over coding sequences (see section 1.2), the existence of an optimization between tRNAomes and codon usage in humans remains under debate (32,107).

Table 1.1. Wobble base pairing rules in eukaryotes.
The recognition rules between codons and anticodons were originally established by Crick (85), and are shaped by the existence of tRNA-dependent adenosine deaminases (ADATs) in eukaryotes (60).

CODON	ANTICODONS
NNU	ANN, GNN, INN
NNC	GNN, INN
NNA	UNN, INN
NNG	CNN, UNN

1.3.6.1. Other functions of tRNAs

Apart from tRF-related mechanisms introduced above, the function of mature tRNAs can also extend beyond protein synthesis (93). Aminoacylated tRNAs can act as amino acid donors, transferring their amino acid to the N-terminus of peptides or to lipids. Uncharged tRNAs can activate stress-response signaling pathways by binding to GCN2. tRNAs are required to prime retrotranscription in retrotransposons and retroviruses such as HIV. Foreign tRNAs can activate the host immune response through recognition of their distinct modification patterns. Clusterted tRNA genes can act as chromatin insulators; they also have higher mutation rates than other parts of the genome which generates variability among individuals.

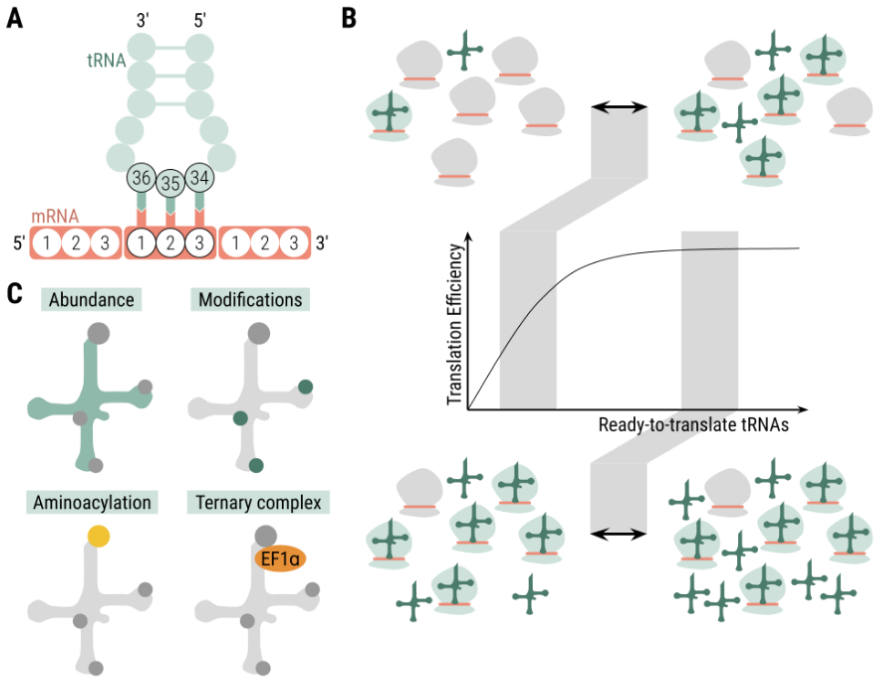


Figure 1.8. Couplers between genes and proteins.
(A) Nucleotides at the anticodon loop of tRNAs base-pair with their cognate codons. (B) Model of constant demand and changing ready-to-translate tRNA levels, adapted from Guimaraes et al. (53). At low concentration ranges, small changes in a limiting tRNA pool can lead to big differences of translation efficiency. High tRNA concentrations over ribosome saturation levels have a small impact on translation. (C) The ready-to-translate tRNA pool is determined by multimodal processes.

Chapter 2

Objectives

The main objective of this PhD thesis is to elucidate the role of tRNAomes in determining changes of mRNA translation across human tissues. As introduced in Chapter 1, coding sequences are under a multitude of selection pressures (codons, mRNA stability, splicing, etc.) and at the same time tRNAs are intricately regulated at multiple levels (expression, modification, aminoacylation). In this thesis, I investigated this complex interface from a systems biology perspective, with the aim of uncovering common and distinct mechanisms of protein synthesis regulation in different human tissues. In the following chapters, I am going to present the results of this work, which I have performed in collaboration with lab colleagues.

In Chapter 3, we analyze small RNA-seq datasets of >8,000 tumor and healthy samples from The Cancer Genome Atlas to determine their tRNA abundance. We find that the variability of the tRNA pool is largely related to the proliferative state across tissues, and the translation efficiency of specific codons appears associated with poor cancer patient survival. To understand how these tRNAs are dysregulated in cancer, we detect that tRNA expression is correlated with tRNA gene copy numbers and anticorrelated with tRNA gene methylation.

Given this dynamic landscape of tRNA pools among tissues and the reliance of viruses on the host translation machinery for productive infection, in Chapter 4 we investigate the translational adaptation of human viruses to the tissues they infect. We observe that viruses infecting different tissues showcase differences in codon usage, which are more pronounced for viral proteins expressed at early infection time points.

Although codon optimization methods are commonly applied in the development of recombinant protein and mRNA-based therapies and vaccines, they currently do not account for tissue-specific aspects of decoding. In Chapter 5, using tissue-wide mRNA-seq and proteomics data, we identify which codons are over or under-represented in specific human tissues. Based on this analysis, we develop CUSTOM, an algorithm that designs optimal coding sequences for protein production in a tissue-specific manner. This data provides the first proof-of-concept evidence that tissue-targeted codon optimization exists.

Finally, in Chapter 6, we uncover an additional layer of tissue-specific tRNA regulation, which was performed during a research stay at Prof. Tao Pan's group (University of Chicago). State-of-the-art tRNA-seq protocols can simultaneously assess the tRNA abundance, some modifications, and aminoacylation. We develop SLAC (SingLe-read Analysis of Crosstalks), which correlates modification and aminoacylation signatures in tRNA at the single-read level to elucidate their crosstalks and their roles in the regulation of mRNA translation. We discover that human tRNAs are dynamically modified under stress, and these modification patterns and tRNA aminoacylation are coordinated to regulate mRNA translation in distinct cellular environments.

Hernandez-Alias X, Benisty H, Schaefer MH, Serrano L. Translational efficiency across healthy and tumor tissues is proliferation-related. Molecular Systems Biology. 2020;16(3):e9275.

Modernist painting of dividing tumor cells

by DALL·E 2

Chapter 3

Translational efficiency across healthy and tumor tissues is proliferation-related

3.1. Abstract

Different tissues express genes with particular codon usage and anticodon tRNA repertoires. However, the codon-anticodon co-adaptation in humans is not completely understood, nor is its effect on tissue-specific protein levels. Here, we first validated the accuracy of small RNA-seq for tRNA quantification across five human cell lines. We then analyzed the tRNA abundance of more than 8000 tumor samples from TCGA, together with their paired mRNA-seq and proteomics data, to determine the Supply-to-Demand Adaptation. We thereby elucidate that the dynamic adaptation of the tRNA pool is largely related to the proliferative state across tissues. The distribution of such tRNA pools over the whole cellular translatome affects the subsequent translational efficiency, which functionally determines a condition-specific expression program both in healthy and tumor states. Furthermore, the aberrant translational efficiency of some codons in cancer, exemplified by ArgAGA, is associated with poor patient survival. The regulation of these tRNA profiles is partly explained by the tRNA gene copy numbers and their promoter DNA methylation.

3.1.1. Synopsis

Quantification of the tRNA expression over thousands of small RNA-seq samples from The Cancer Genome Atlas unveils the existence of tissue-specific translational efficiencies related to proliferation.

- tRNA abundance is tissue- and cancer-type-specific.
- Translational efficiency is globally controlled and the cellular translatome needs to compete for a limiting tRNA pool.
- Proliferation is the major determinant of translational efficiency differences among tissues, and the codon ArgAGA appears particularly favored in cancer.
- Differences at the tRNA pool affect protein translation and subsequently determines specific functional phenotypes.

3.1.2. Additional data access

All supplementary figures and data can be accessed from the original publication through this QR code and link.
Expanded View figures and tables will be referred to as "Figure EV" and "Table EV".

3.2. Introduction

In the light of the genetic code, multiple 3-letter combinations of nucleotides in the mRNA can give rise to the same amino acid, which are known as synonymous codons. However, despite the homology at the protein level, these different codons are recognized distinctly by the transcriptional and translational machineries (17,43), and ultimately cause changes at multiple levels of gene expression. Therefore, the non-uniform abundance of synonymous codons across different tissues and among distinct functional gene sets has been proposed as an adaptive mechanism of gene expression regulation (56), particularly linked to the proliferative state (42). Nevertheless, in human, it is still

under debate whether the efficiency of gene expression is the main selective pressure driving the evolution of genomic codon usage (32).

The 61 amino-acid-coding codons need to be recognized by 46 different tRNA isoacceptors distributed across 428 Pol-III-transcribed tRNA genes (12), thus requiring wobble interactions (non-Watson-Crick base pairing). This complexity of the tRNA repertoire is further enhanced by an average of 11-13 base modifications per tRNA and all possible combinations thereof (81). The underlying mechanisms regulating tRNA gene expression and modification are far from resolved (14,82). However, it has been established that different conditions and tissues showcase distinct tRNA abundances (42,103) and codon usages (56,108).

In order to understand such changes in codon-anticodon co-adaptation, orthogonal datasets of gene expression including tRNA quantification are required, which needs to overcome the challenges of strong secondary structures and abundant chemical modifications. Recent technological developments have paved the way for sensitive high-throughput tRNA sequencing across tissues and conditions (109,110). Aside from these methods and despite the lower coverage, tRNA reads can also be detected from generic small RNA-seq datasets (111–115). In this context, The Cancer Genome Atlas (TCGA) has been recently used to investigate the alteration of tRNA gene expression and translational machinery in cancer, which may play a role in driving aberrant translation (116,117).

To validate the use of small RNA-seq for tRNA quantification, we first compare tRNA levels determined in HEK293 by well-established tRNA sequencing methods (Hydro-tRNAseq and demethylase-tRNA-seq) (109,110,118), with those obtained by small RNA-seq. Then we quantify the tRNA repertoire of five cell lines using Hydro-tRNAseq and perform small RNA-seq in parallel. Comparison of the tRNA measures obtained by both approaches shows that it is possible to accurately estimate relative tRNA abundance of cells and tissues using small RNA-seq. Furthermore, we show that both types of

quantification are informative enough to distinguish between the five analyzed human cell lines covering multiple tissue types. In consequence, we apply a tRNA-specific computational pipeline to re-analyze 8,534 small RNA-seq datasets from TCGA (119). We find that the tissue-specificity of tRNA profiles is largely proliferation-related, even within healthy tissues. The tRNA quantification of TCGA samples enables their comparison with paired and publicly available mRNA-seq, proteomic, DNA methylation and copy number data, which underscores the role of tRNAs in globally controlling a condition-specific translational program. We discover multiple codons, including ArgAGA, whose translational efficiency is compromised and leads to poor prognosis in cancer. Finally, promoter DNA methylation and tRNA gene copy number arise as two regulatory mechanisms controlling tRNA abundances in cancer.

3.3. Results

3.3.1. tRNA quantification and modifications from small RNA-seq data

In order to test how accurately we can extract tRNA abundance information contained in small RNA sequencing data, we re-analyze four publicly-available datasets of the cell line HEK293 (115,120,121). In contrast to previous studies analyzing tRNA expression from small RNA-seq data (116,117), we use a computational pipeline specifically developed for the accurate mapping of tRNA reads (113) in order to quantify all different isoacceptor species (**Fig. 3.1A**, see Methods). To validate the accuracy of these small RNA-seq quantifications, we retrieve four datasets of well-established tRNA sequencing methods (Hydro-tRNAseq and demethylase-tRNA-seq) applied to the same cell type (109,110,118,122), which autocorrelate in the range of 0.75-0.85 among themselves (Table EV1, Fig. EV1A). In comparison, our four HEK293 small RNA-seq quantifications show an average Spearman correlation against these four conventional datasets of 0.73. Compared to the Zhang et al. (116) quantification, which correlate in the range of 0.60-0.77 (Table EV1, Fig. EV1A), our tRNA-specific mapping pipeline

performs slightly better than the previously published protocol. It has been reported that there are tRNA-derived fragments naturally produced and having other functions different from translation (81), which could confound the tRNA quantification. Although we cannot exclude the presence of tRNA-derived fragments in small RNA-seq datasets (123), we found that no differences between reads with or without mismatches are found when compared to tRNA-seq protocols in which tRFs are specifically removed before sequencing.

Further than correlating small RNA-seq data with conventional tRNA-seq datasets, we analyze whether small RNA-seq quantifications are informative enough to distinguish between different human cell lines covering multiple tissue types. We therefore apply both small RNA-seq and Hydro-tRNAseq to HEK293 (kidney), HCT116 (colon), HeLa (cervix), MDA-MB-231 (breast), and BJ fibroblasts. However, given the high variability between replicates of MDA-MB-231 Hydro-tRNAseq quantifications, this cell line was excluded from further analyses (Table EV2). First, the correlations between the two methods of identical samples and computational mapping pipeline range between 0.93 and 0.96 for all cell lines. tRNA quantifications from both protocols are compared and significantly higher Spearman correlations are obtained within matching samples versus mismatching cell lines (**Fig. 3.1B**). In order to assess the amount of tRNA variability coming from either the sequencing method or the cell lines, a Principal Component Analysis of these tRNA quantifications indicates that both factors influence variability to a similar extent (Fig. EV1B, >30% variance each). Furthermore, to validate that small RNA-seq is similarly informative of cell type differences as Hydro-tRNAseq, we show that a Linear Discriminant Analysis of the data is able to discriminate between cell lines regardless of the sequencing protocol (Fig. EV1B).

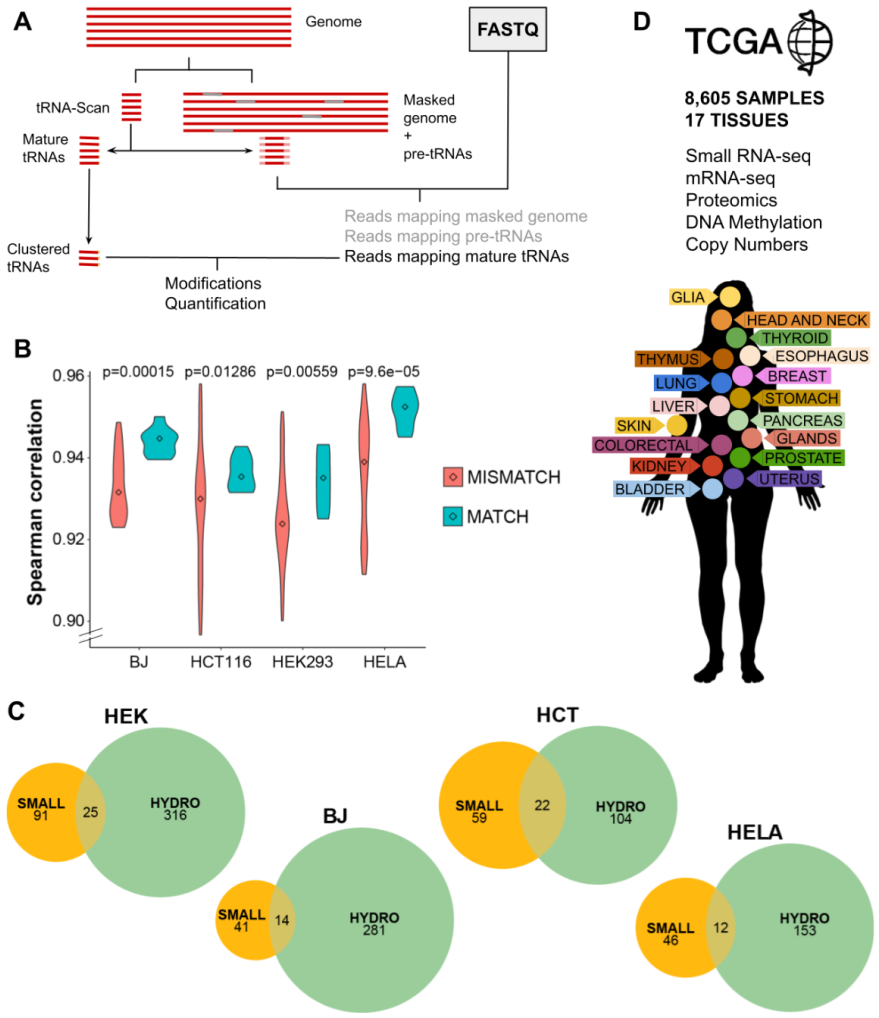


Figure 3.1. tRNA quantification and modifications from small RNA-seq data.
(A) Schematic pipeline for accurate mapping of tRNA reads. (B) Correlations between tRNA quantifications by small RNA-seq and Hydro-tRNAseq of matching (correlations within the same cell line) versus non-matching (different cell lines) samples. Center values represent the median. The p-value corresponds to a one-tailed Wilcoxon rank-sum test, with $n_{matching}$= 9 and $n_{non\text{-}matching}$= 63. (C) Overlap of the detected tRNA modifications upon variant calling by both methods. (D) The TCGA network contains small RNA-seq data alongside mRNA-seq, DNA methylation arrays, non-targeted proteomics, and copy number alteration quantification comprising 17 tissues.

We also detect tRNA base modifications in both protocols by nucleotide variant calling, as described in Hoffmann et al. (113). In all cases, considering the modifications that are detected in all three replicates, Hydro-tRNAseq datasets identify a larger number of modifications than small RNA-seq, as expected by the more uniform and deeper coverage of this method (Fig. EV2, Table EV2). Furthermore, we detect a significant enrichment of the Hydro-tRNAseq modifications in the small RNA-seq data ($p < 1e-16$, Fisher test), indicating that the latter contains also information on tRNA modifications (**Fig. 3.1C**). Although the exact nature of modifications cannot be determined by sequencing, most frequent nucleotide mismatches in both sequencing methods include A-to-G changes at position 34 and 37 (Fig. EV3, Table EV3), which correspond to known modifications such as adenosine-to-inosine editing and 1-methylinosine, respectively (82). Overall, most of the known modification-specific mismatches can be retrieved with both small RNA-seq and Hydro-tRNAseq (Table EV3), while the deeper coverage of the latter improves its sensitivity.

Taken together, these observations demonstrate the applicability of small RNA-seq data for the quantification of tRNAs and their modifications. We therefore apply the same computational pipeline to all healthy and primary tumor small RNA-seq samples from 23 cancer types of The Cancer Genome Atlas (TCGA), which consists of 8,605 samples distributed among 17 different human tissues (**Fig. 3.1D**, number of samples and their abbreviations in Table EV4).

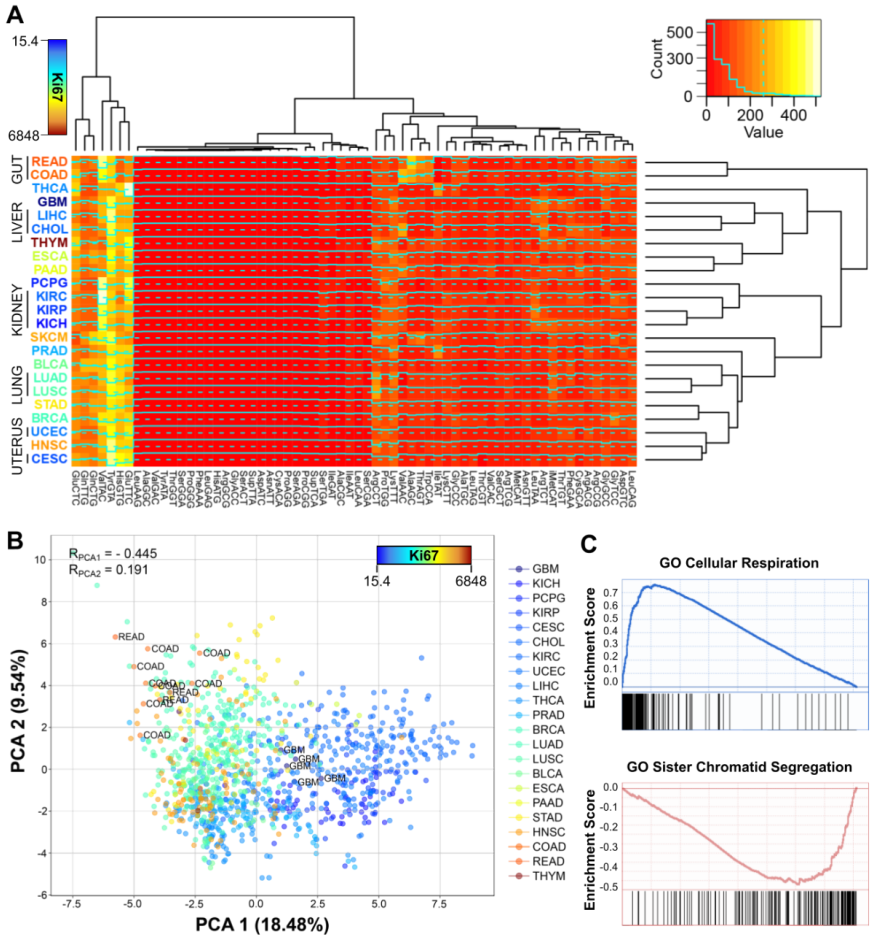


Figure 3.2. Proliferation is the major driver of tissue-specificity in tRNAs.
(A) Medians of square-root-normalized tRNA abundances across all TCGA tissues. The color of the tissue labels correspond to the average Ki67 expression. Refer to Table EV4 for full cancer type names and number of samples. (B) Principal Component Analysis (PCA) of the Relative Anticodon Abundances (RAA, see Methods) of all healthy samples of TCGA, where the color scale corresponds to the mean tissue expression of Ki67. The Spearman correlations of Ki67 with the components are shown, as well as the samples of most extreme tissues. (C) Top positive and negative GO terms upon Gene Set Enrichment Analysis (GSEA) of the correlations of the first PCA component against all genes.

3.3.2. Proliferation is the major driver of tissue-specificity in tRNAs

To determine the tissue-specificity of tRNAs in physiological conditions, the tRNA levels of all 675 healthy samples in TCGA tissues are analyzed from small RNA-seq data. For all 46 annotated anticodons, tRNA abundances have significant differences between tissues, as detected by Kruskal-Wallis test ($q<0.05$, FDR-corrected). Such differences between tissues are also observed by hierarchical clustering of the median abundance between all groups (**Fig. 3.2A**). Furthermore, healthy samples from cancer types originating from the same tissue tend to cluster together: READ and COAD from the gut; KIRC, KIRP and KICH from the kidney; LUAD and LUSC from the lung; UCEC and CESC from the uterus; LIHC and CHOL from the liver (refer to Table EV4 for full cancer names). On the other hand, in terms of anticodon abundances, three main subgroups of tRNAs with low, medium and high levels can be distinguished across all cancer types (**Fig. 3.2A**).

Abbreviations: BLCA (Bladder Urothelial Carcinoma), BRCA (Breast invasive carcinoma), CESC (Cervical squamous cell carcinoma and endocervical adenocarcinoma), CHOL (Cholangiocarcinoma), COAD (Colon adenocarcinoma), ESCA (Esophageal carcinoma), GBM (Glioblastoma multiforme), HNSC (Head and Neck squamous cell carcinoma), KICH (Kidney Chromophobe), KIRC (Kidney renal clear cell carcinoma), KIRP (Kidney renal papillary cell carcinoma), LIHC (Liver hepatocellular carcinoma), LUAD (Lung adenocarcinoma), LUSC (Lung squamous cell carcinoma), PAAD (Pancreatic adenocarcinoma), PCPG (Pheochromocytoma and Paraganglioma), PRAD (Prostate adenocarcinoma), READ (Rectum adenocarcinoma), SKCM (Skin Cutaneous Melanoma), STAD (Stomach adenocarcinoma), THCA (Thyroid carcinoma), THYM (Thymoma), UCEC (Uterine Corpus Endometrial Carcinoma).

Regarding codon usage, a measure of tRNA abundance taking into account the relative contribution of each tRNA anticodon among the set of synonymous codons of a certain amino acid is the Relative Anticodon Abundance (see Methods). Using this measure, a principal component analysis (PCA) of all healthy control samples in TCGA also shows clear differences between tissues (**Fig. 3.2B**). To interrogate the biological functions related to the variability of anticodon abundances between samples, we compute the correlation of the whole mRNA-seq transcriptome against the first PCA component, which explains 18.5% of the variance, and analyze it by Gene Set Enrichment Analysis (GSEA). As a result, the top correlating genes are enriched in proliferation and immune cell activation, while the lowest correlations belong to genes related with oxidative metabolism and respiration (**Fig. 3.2C**, Table EV5). Moreover, our first component correlates positively with the proliferation marker Ki67 ($R_{spearman}$= 0.45) (124). This confirms, as has been previously suggested (42), that there is a proliferative tRNA expression program.

Overall, we observe patterns of tissue-specific tRNA profiles in TCGA healthy samples. Furthermore, based on both the gene set enrichment and the association to a proliferation marker, our analyses identify the proliferative state of tissues as the major biological function driving the variability on tRNA abundances.

3.3.3. tRNA repertoires determine tissue-specific translational efficiency

Given that different tissues express distinct tRNA repertoires, we wondered whether they could have an effect in protein translation elongation. The so-called translational efficiency is defined as the rate of protein production from mRNA, and multiple indices and models can be described to estimate it (125). In this article, and based on previous studies underscoring the global control role of codon usage as a competition for a limited tRNA pool (47,50,96), we define the Supply-to-Demand Adaptation (SDA) as the balance between the supply (i.e. the anticodon tRNA abundances) and demand (i.e. the

weighted codon usage based on the mRNA levels) for each of the 60 codons (excluding methionine and Stop codons). Furthermore, we normalize both the codon and anticodon abundances within each amino acid family (i.e. relative to the most abundant synonymous codon/anticodon), in order to remove the effect of amino acid biases and get a cleaner measure of codon optimality (28).

To validate the suitability of SDA in determining the translational efficiency, we correlate the SDA value of all proteins against the available proteomics data of paired TCGA samples (126,127), which includes breast and colorectal tissues (tumor only, as no healthy samples are available). Although correlations are modest, both the protein abundances and the protein-to-mRNA ratios correlate significantly better with SDA than with the classical tRNA Adaptation Index [tAI] (61,128) or with a relative tAI with normalized weights within each amino acid family [RtAI] (**Fig. 3.3A**, Fig. EV4A-B). In consequence, including the mRNA codon demand into the SDA metric outperforms other tRNA-only metrics of translational efficiency. Furthermore, the correlation of SDA with protein-to-mRNA ratio is slightly but significantly higher than with protein levels alone, which indicates that the first is a better proxy for the process of translation (**Fig. 3.3A**).

Next, we calculate the SDA for the 620 healthy samples for which both tRNA abundances and mRNA levels are available. When analyzing the tissue medians of SDA weights per each codon (SDAw), we observe that most codons are optimally balanced (SDAw =1), while 13.7% and 16.3% of codons are favored (SDAw >2) and disfavored (SDAw <0.5) respectively. The tissue clustering again shows that healthy samples of cancer types from the same tissue have similar SDAw profiles, which separates two major clusters of mostly high-Ki67 and low-Ki67 tissues (Fig. EV4C).

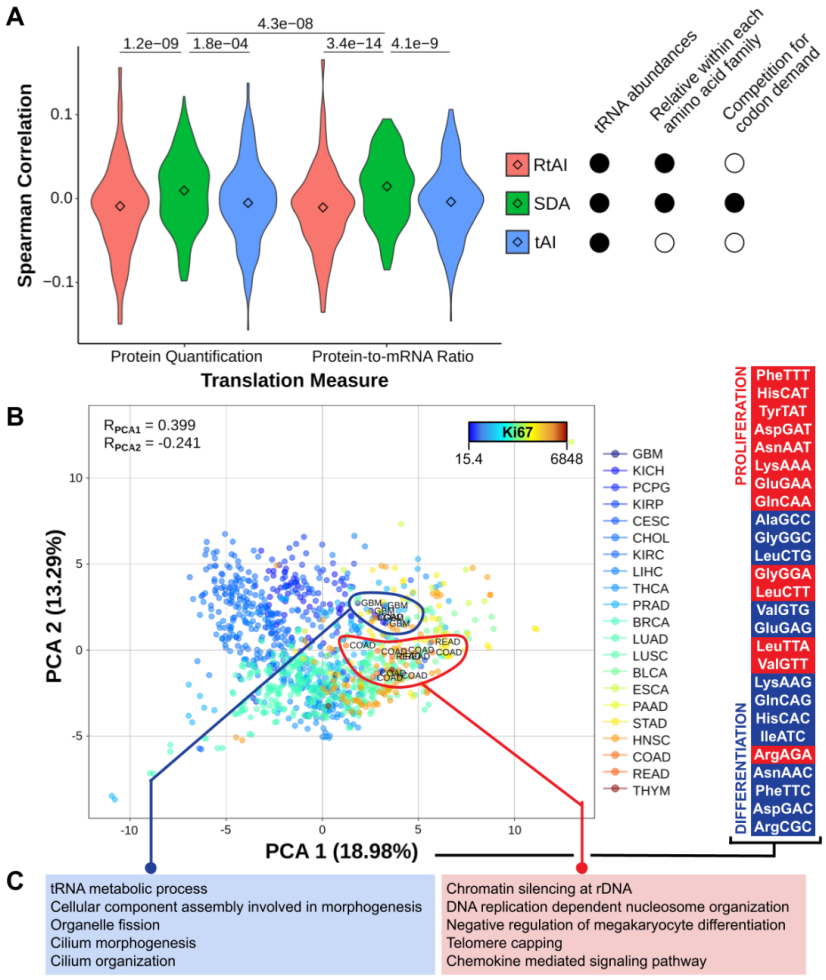


Figure 3.3. tRNA repertoires determine tissue-specific translational efficiency.
(A) Three metrics of translational efficiency (the classical tAI, a relative tAI with normalized weights within each amino acid family, and the Supply-to-Demand Adaptation described in this article) are Spearman correlated against two proxies of translation (protein abundance and protein-to-mRNA ratio) for all samples for which proteomics data is available (BRCA, COAD and READ). Center values represent the median. Statistical differences are determined by sample-paired two-tailed Wilcoxon rank-sum test (n=219). (B) Principal Component Analysis (PCA) of the SDAw of TCGA, where the color scale corresponds to the mean tissue expression of Ki67. The Spearman correlations of Ki67 with the components are shown, as well as the samples of most extreme tissues. On the right, the top and bottom proliferation- and differentiation-related codons, as defined by Gingold et al. (2014), ordered by their contribution to the first PCA component. Refer to Table EV4 for full cancer type names and number of samples. (C) GSEA of the differential SDA between extreme tissues ($\Delta SDA = SDA_{Colorectal} - SDA_{Brain}$), showing five among the top ten GO terms with high (right) and low (left) SDA in colorectal versus glial tissues.

In order to identify the codons contributing most to the differences between tissues, we compute a bidimensional PCA across all samples and SDAw (**Fig. 3.3B**). Both the first and second components significantly correlate with the proliferation marker Ki67 (0.4 and -0.24; see **Fig. 3.3B**). In agreement with the proliferation- and differentiation-related codons of Gingold et al. (42), such proliferative pattern is similarly reproduced by the codons contributing to the first (**Fig. 3.3B**) and second (Table EV6) PCA components. Further, similarly to the tRNA abundances (**Fig. 3.2B**), a GSEA of correlating genes with the first component highlights the link with proliferation-related terms (Table EV6). On the other hand, the first component also clearly separates codons based on the GC content of the third codon base, which has recently been associated with differentiation (high in nnC/G codons) versus self-renewal functions (high in nnA/T) (129), as well as with proliferative transcriptomes (130).

The previous analyses support the idea of proliferation-related tRNAs driving changes in translational efficiencies. In that case, we expect that the two most extreme tissues in terms of proliferation (brain and gut, excluding thymus for its low number of samples) differ in the optimization of proliferation-related proteins. As such, we compute the average SDAw for these two tissues, analyze the subsequent SDA score for each protein, and perform a GSEA of the differential SDA per protein. Consistent with our hypothesis, the results indicate that gut-optimized proteins are enriched in DNA replication, chromatin organization and chemokine signaling, whereas brain-optimized proteins are related to tRNA metabolism and cilium morphogenesis (**Fig. 3.3C**, Table EV7). Taken together, this result confirms that the tRNA-dependent translational efficiency is optimized for the translation of tissue-specific genes, particularly in function of the proliferation state.

3.3.4. Aberrant translational efficiencies drive tumor progression

Given that proliferation is a major determinant of translational efficiency in healthy tissues, its importance could be extrapolated to pathological conditions such as cancer. In fact, aberrant expression of tRNAs and codon usage have been broadly related with tumorigenesis and cancer progression (106,116,117,122). We therefore investigate 22 cancer types from TCGA in order to determine which codons are translationally compromised in disease.

Similar to the analysis performed on the healthy tissues, we quantify all tRNA abundances of TCGA primary tumor samples (Fig. EV5) and determine their corresponding translational efficiencies using the SDA metric. By analyzing the differential SDAw between normal and tumor samples, we observe many significant differences in all 60 codons across the 22 cancer types (**Fig. 3.4A**). Among the most consistent changes, the ArgAGA codon is significantly more favored in tumors for 15 out of 15 cancer types, while the ArgCGG is disfavored in 7 out of 11 cancers (**Fig. 3.4B**). In the case of threonine, ThrACT and ThrACC are better adapted in healthy samples (13/14), whereas tumor mostly favors ThrACG (12/16).

In terms of patient survival, we divide the TCGA patients in two groups based on their low or high tumor SDAw and analyze their survival probability (**Fig. 3.4C**, Table EV8). Among others, and consistent with the previous analysis, high supply-to-demand weights of ArgAGA are associated with poor prognosis in kidney renal clear cell carcinoma and colon adenocarcinoma. Arginine limitation in the kidney cell line HEK293T has been shown to compromise tRNAArg aminoacylation, leading to codon pausing and reduced cell viability (99). In addition, low SDAw of ArgAGG and SerAGT lead to longer survival in kidney renal clear cell carcinoma.

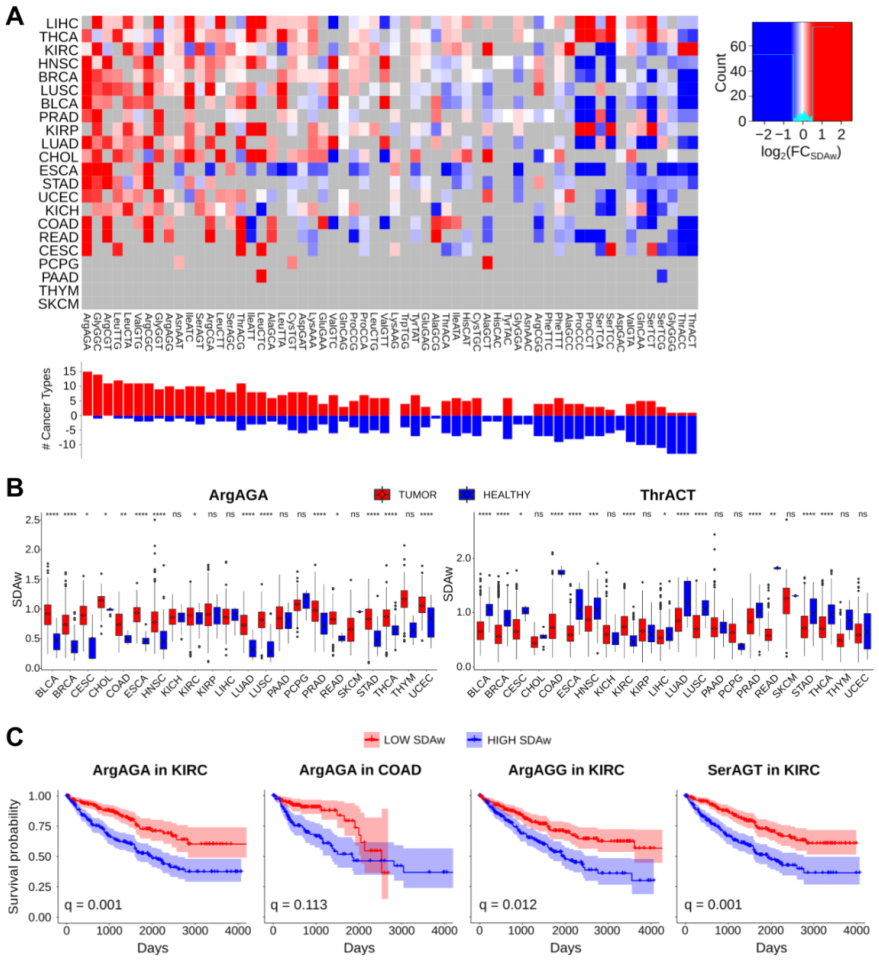


Figure 3.4. Aberrant translational efficiencies drive tumor progression.
(A) Differential SDAw between healthy and tumor samples across 22 cancer types, as measured by $\log_2(SDAw_{Tumor}/SDAw_{Healthy})$. Only significant differences are colored, which are determined using a two-tailed Wilcoxon rank-sum test and corrected for multiple testing by FDR. Refer to Table EV4 for full cancer type names and number of samples. (B) Boxplot of the SDAw of ArgAGA and ThrACT codons across TCGA cancer types. Center values represent the median. Statistical significance: ns ($p > 0.05$), * ($p <= 0.05$), ** ($p <= 0.01$), *** ($p <= 0.001$), **** ($p <= 0.0001$). (C) Survival curves for the previous codons in KIRC and COAD patients. The survival analysis was performed for all codons whose SDAw was significantly different in more than 5 cancer types in the one direction with respect to the other [Abs(UP-DOWN)>5], and correspondingly corrected for multiple comparisons using FDR.

To determine the impact of aberrant translational efficiencies in regulating an oncogenic translation program, we calculate the differential SDA for the whole genome based on the average SDAw of healthy and tumor samples in kidney renal clear cell carcinoma, since it is the cancer type with the most prognostic differences (**Fig. 3.4A**). The GSEA of the resulting ΔSDA score indicates that cancer SDAw should favor the translation of proteins related to DNA replication and gene expression, whereas the healthy kidney samples favor signals transduction and differentiation processes (Table EV9). As the SDAw of the ArgAGA is specifically disturbed in cancer, we also interrogate how this codon is distributed along the genome. We therefore perform a GSEA on the relative codon usage of ArgAGA, which shows that proliferation and immune activation functions lie among the most AGA-enriched genes, while development and differentiation terms are AGA-depleted (Table EV10). Together with the low-proliferative state of kidney (**Fig. 3.2B**), the over-efficiency of a proliferation-related codon in this tissue can thus perturb its cellular SDA.

Overall, we detect differences at the level of SDAw between tumor and healthy tissues, which show a functional relevance to the disease state. Therefore, while the differential expression of tRNAs in TCGA had been already discussed elsewhere (116,117), we could here elucidate their oncogenic effect in translational efficiency. In particular, ArgAGA appears as an interesting codon candidate in favoring tumor progression, which we had also detected in healthy tissues to be associated with proliferation (PCA2 in Table EV6).

3.3.5. Promoter methylation and gene copy number regulate the tRNA abundance

Aberrant translational efficiencies in cancer are partially caused by the differential abundance of tRNA genes (Fig. EV5). To determine the underlying mechanisms driving changes in expression, we retrieve the DNA methylation (typically occurring at CpG dinucleotides) and copy number alteration (CNA) data from TCGA samples, as a possible means for tRNA gene regulation. While CNA information cover 84%

of tRNA genes, the 450K-BeadChip methylation arrays used in TCGA are mostly centered on the coding genome (131) and yield a coverage of only 37%.

In order to make the gene-based data comparable with the measured isoacceptor-based tRNA abundances, we average methylation and CNA levels over all genes within the same isoacceptor family, at the cost of losing resolution. For each isoacceptor and each cancer type, we finally fit a Multiple Linear Regression to determine how are promoter methylation and CNA affecting tRNA expression (**Fig. 3.5A**, Table EV11). Among all models, the significant coefficients for methylation and CNA are significantly negative and positive, respectively. Despite the limited explained variance of the models (average R^2=0.023), such results suggest that promoter methylation could contribute to inhibition of tRNA gene expression, whereas an increase in the gene copy number would enhance tRNA expression.

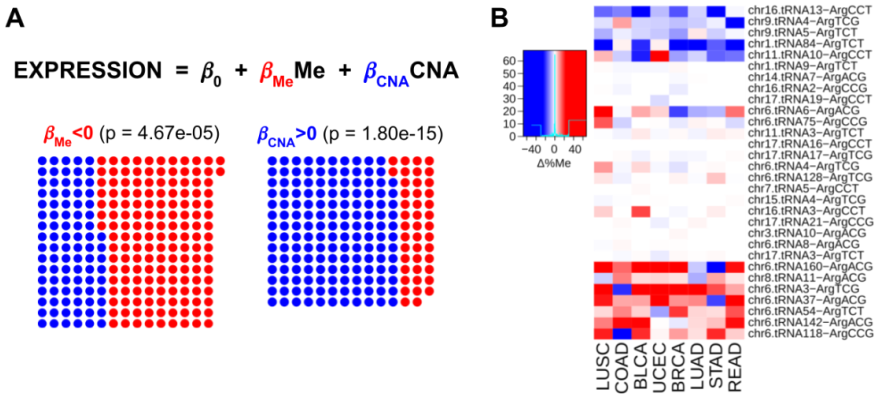


Figure 3.5. Promoter methylation and gene copy number regulate tRNA abundance. (A) A Multiple Linear Regression (MLR) between square-root-normalized tRNA abundance and the average promoter methylation (450K BeadChip array) and gene copy number at the isoacceptor level. Among all MLRs for each isoacceptor and each cancer type separately, the dots show the FDR-normalized significant coefficients based on their corresponding t-statistic p-value, and red/blue show whether they are negative/positive respectively. The p-value corresponds to a two-tailed binomial test between n_{pos} and n_{neg}. (B) Differential promoter methylation (bisulfite sequencing) between healthy and tumor samples of genes expressing arginine tRNAs, as measured by Δ%Me=(%Me_{Tumor}-%$Me_{Healthy}$). Refer to Table EV4 for full cancer type names and number of samples.

Given the association of the codon ArgAGA with cancer prognosis (**Fig. 3.4C**), we explore the abundance pattern of tRNAArg in TCGA. In agreement, the complementary tRNAArg^{TCT} appears overexpressed in 13 out of 15 cancer types (Fig. EV5A), making it a candidate driver of the translational differences. To get a more accurate picture of the tRNA gene methylation levels, we also analyze recently published bisulfite sequencing data (132), which, for 47 samples among nine cancer types, improved the coverage of tRNA genes up to an average of 81%. In total, tRNAArg^{TCT} genes stand among the least methylated arginine isoacceptors in cancer (Fig. EV6A-B), in particular at the chr9.tRNA5 and chr1.tRNA84 genes (**Fig. 3.5B**). Furthermore, tRNAArg^{TCT} gene duplications occur frequently in kidney cancers (Fig. EV6C).

In short, promoter methylation and CNA appear as two possible regulatory mechanisms of tRNA expression in cancer, which suggests that similar mechanisms that control the Pol-II-mediated RNAs might also regulate the expression of Pol-III non-coding transcriptome, such as tRNA genes. However, more accurate and high-throughput data on the methylation and CNA of the non-coding genome together with gene-based tRNA quantifications are needed to make stronger associations.

3.4. Discussion

In this study, we use a systems biology approach to interrogate the multi-omics TCGA dataset under the perspective of translational efficiencies. We therefore first validate the suitability of small RNA-seq data in reproducing conventional tRNA-seq quantifications based on a gold standard set of five tissue-wide human cell lines. In fact, knowing that small RNA-seq datasets have a limited tRNA coverage and tend to be biased towards tRNA fragments and unmodified tRNAs (115,123), we extend and apply a computational pipeline for accurate mapping of tRNA reads (113). As a result, we obtain reproducible and informative quantifications of all isoacceptors in our gold standard cell lines as well as in thousands of samples across 23 cancer types of TCGA, exceeding

the quality of similarly published data (116,117). However, we cannot exclude that tRNA-derived fragments (tRFs) could be interfering with our small RNA-seq quantifications. At the level of nucleotide modifications (82), our tRNA mapping pipeline is also able to detect most of the known mismatch-producing modifications of mature tRNAs. All in all, even though our quantifications from small RNA-seq just give an estimate of the tRNA abundances, the results indicate that they can be rather precise proxies.

From these quantifications, we then elucidate their effect on the translational efficiency by defining the SDA, for Supply-to-Demand Adaptation, which is a balance between the tRNA supply and the codon demand. Although a more accurate SDA would have determined the supply and demand based on the aminoacylated portion of tRNAs (133) and the ribosome-bound mRNAs (134) respectively, we approximate such measures by our tRNA quantifications and the publicly-available mRNA-seq data of TCGA. In agreement with current studies showing that a dynamic codon usage need to compete for a limited tRNA pool (28,96), we demonstrate that SDA is better measure of codon optimality than previously published metrics such as the tAI (61,128). However, far from explaining the translation process, the still low but significant correlations of protein-SDA in human, in contrast to unicellular organisms, suggest that protein expression is also dependent on other layers of regulation, such as transcriptional and post-transcriptional machineries, translation initiation, epigenetic modifications of DNA and RNAs, or protein degradation mechanisms (107).

On the level of translational efficiency, in agreement with previous studies (42,122), we detect that the proliferative state is the major determinant of SDA differences both across healthy tissues and in cancer. Moreover, in contrast to recent work challenging the tissue-specificity of codon-anticodon co-adaptation in human (28,107), our data here support the idea that tissue-specific SDAw have functional implications on the tissue phenotype (e.g. in favoring neural differentiation in brain, or abnormal proliferation in cancer).

Furthermore, we observe a pattern of proliferative nnA/T versus differentiative nnC/G codons. Based on ribosome profiling experiments of pluripotency changes in embryonic stem cells (129), this could be attributed to the slower translation in differentiated cells of codons decoded by tRNAs that require adenosine-to-inosine modification at the wobble-base pairing position. In particular, we detect the ArgAGA codon to be significantly more favored in proliferative cells and leading to poor cancer prognosis in kidney carcinoma, specifically driven by an overexpression of $tRNAArg^{TCT}$ in cancer. Arginine limitation in the kidney cell line HEK293T has indeed been shown to compromise tRNAArg aminoacylation, leading to arginine codon pausing and reduced cell viability (99). Furthermore, in support of our approach for isoacceptor quantification and translational efficiency, similar studies of tRNA levels in TCGA have concordantly claimed a prognostic value for the ArgAGA codon in clear renal cell carcinoma (116,117).

In an effort to elucidate the mechanisms regulating the expression of tRNAs, we observe that the tRNA gene copy number and their DNA methylation state have a positive and inhibitory association with tRNA abundances, respectively. In this context, DNA methylation has previously been linked to the silencing of type II genes (such as tRNAs) of the Pol-III transcriptome (135,136). Here we specifically propose a role for DNA methylation in regulating the overexpression of $tRNAArg^{TCT}$ in cancer, although no direct causal link can yet be established. In terms of the copy number alterations, it is not surprising to detect tRNA gene duplications in tumors, but the functional role in disease of different isodecoder genes that share the same anticodon is still a matter of debate (137). With the advent of more accurate and high-throughput multi-omics datasets, our knowledge on the underlying mechanisms controlling tRNA expression, degradation, and the effect of their modifications will be further expanded (14,82). Recent studies in TCGA have actually observed an upregulation of tRNA-modifying enzymes, as well as proposed a link of tRNA-derived fragments (tRF) to proliferation (116,138).

Overall, this is the first high-throughput study of codon-anticodon translational efficiency over thousands of samples comprising multiple tissues and disease. We therefore demonstrate a functional role for the proliferation-driven tRNA abundance differences in determining a tissue-specific phenotype, both in physiological and pathological conditions. In the future, we expect to validate the effect of such differential translational efficiency by integrating perturbation-based data and including additional gene expression regulatory layers such as tRNA modifications.

3.5. Material and Methods

3.5.1. Reagents and Tools table

Reagent/Resource	Reference or Source	Identifier or Catalog Number
Chemicals, Enzymes and other reagents		
Antarctic phosphatase	New England BioLabs	Cat#M0289
T4 Polynucleotide Kinase	New England BioLabs	Cat#M0201
ProtoScript II Reverse Transcriptase	New England BioLabs	Cat#M0368
miRNeasy Mini kit	Qiagen	Cat#217004
15% TBE-Urea Gels	NOBEX, Invitrogen	Cat#EC6885BOX
RNeasy MinElute Cleanup Kit	Qiagen	Cat#74204
QIAquick PCR Purification Kit	Qiagen	Cat#28106
Experimental Models		
BJ/hTERT	Gift from Anders H. Lund laboratory (Disa Tehler).	N/A
HeLa	ATCC	CCL-2
HEK293	ATCC	CRL-1573
HCT116	ATCC	CCL-247

MDA-MB-231	ATCC	HTB-26
Software		
BBMap [v38.22]	Bushnell B.	https://sourceforge.net/projects/bbmap
FastQC [v0.11.4]	Andrews S.	https://www.bioinformatics.babraham.ac.uk/projects/fastqc
SAMtools [v1.3.1]	(139)	http://samtools.sourceforge.net
tRNAscan-SE [v2.0]	(140)	http://lowelab.ucsc.edu/tRNAscan-SE
BEDtools [v2.27.1]	(141)	https://bedtools.readthedocs.io/en/latest
Segemehl [v0.3.1]	(142)	https://www.bioinf.uni-leipzig.de/Software/segemehl
Picard [v2.18.17]	Broad Institute	https://github.com/broadinstitute/picard
GATK [v3.8]	(143)	https://software.broadinstitute.org/gatk
GSEA [v3.0]	(144)	https://http://software.broadinstitute.org/gsea
BLAST [v2.9.0]	(145)	https://blast.ncbi.nlm.nih.gov

3.5.2. Cell lines

The cell lines included in this study are HeLa, HEK293, HCT116, MDA-MB-231 and fibroblast BJ/hTERT. The sex of each cell line is as follows: HeLa, Female; HEK293, Female; MDA-MB-231, Female; HCT116, Male; BJ fibroblasts, Male. Cells were maintained at 37 °C in a humidified atmosphere at 5% CO2 in DMEM 4.5g/L Glucose with UltraGlutamine media supplemented with 10% of FBS and 1% penicillin/streptomycin.

3.5.3. RNA extraction

Cells were grown in 60mm dishes for 48h. Total RNA from HeLa, HEK293, HCT116, MDA-MB-231 and fibroblast BJ/hTERT was

extracted using the miRNeasy Mini kit. Independent replicates were grown and RNA was extracted on different days. 20 μg of total RNA was treated following either the protocol of Hydro-tRNAseq (109) or generic small RNA-seq.

3.5.4. Hydro-tRNA sequencing

Total RNA was resolved on 15% Novex TBE urea gels and size-selected for 60-100 nt fragments. The recovered material was then alkaline hydrolyzed (10mM sodium carbonate and 10mM sodium bicarbonate) for 10 minutes at 60°C. The resulting RNA was de-phosphorylated with Antarctic Phosphatase (New England Biolabs) at 37ºC for 1 hour. De-phosphorylated RNA was purified with an RNeasy MinElute spin column and re-phosphorylated with Polynucleotide Kinase (NEB). PNK-treated tRNAs were purified with an RNeasy MinElute spin column and, similar to small RNA-seq library preparation, adaptor-ligated, reverse-transcribed and PCR-amplified for 14 cycles. The resulting cDNA was purified using a QIAQuick PCR Purification Kit and sequenced on Illumina HiSeq 2500 platform in 50bp paired-end format. Hydro-tRNAseq data of HCT116, MDA-MB-231 and fibroblast BJ/hTERT has been generated in this study, while sequencing data from HEK293 and HeLa had been previously published (122).

From all five cell lines, the isoacceptor abundances of MDA-MB-231 yielded a median of 3-5 times higher standard deviation than the other Hydro-tRNAseq quantifications (Table EV2), thus suggesting some technical problem with this cell line. In consequence, this cell line was excluded from any further analysis.

3.5.5. Small RNA sequencing

Total RNA was directly adaptor-ligated, reverse-transcribed and PCR-amplified for 12 cycles. The resulting cDNA was then size-selected by gel electrophoresis, and fragments of 145-160 bp were eluted and sequenced on Illumina HiSeq 2500 platform in 50bp single-end format.

3.5.6. The Cancer Genome Atlas multi-omics data

Raw small RNA-sequencing data in BAM format were retrieved from the GDC legacy archive after obtaining the necessary permissions from dbGaP, comprising all healthy samples (NT, solid tissue normal) and their primary tumor (PT) counterparts, which consists of 23 cancer types (BRCA, PRAD, KICH, KIRP, KIRC, LUAD, LUSC, HNSC, UCEC, CESC, LIHC, CHOL, THCA, COAD, READ, ESCA, STAD, BLCA, PAAD, THYM, SKCM, PCPG, GBM). For samples for which more than one BAM was available, all files were downloaded. BAM files were converted to FASTQ using SAMtools [v1.3.1] (139). We retrieved publicly available and pre-processed mRNA-seq gene expression, 450k DNA methylation, bisulfite DNA methylation, and SNP6 segmented copy number alteration (CNA) from firebrowse. As for proteomics, preprocessed protein assembly data and protein relative abundance were obtained from CPTAC for TCGA samples including BRCA, COAD and READ.

3.5.7. tRNA quantification and modification calling

In both Hydro-tRNAseq and small RNA-seq FASTQ files, sequencing adapters were trimmed using BBDuk from the BBMap toolkit [v38.22] (https://sourceforge.net/projects/bbmap): k-mer=10 (allowing 8 at the end of the read), Hamming distance=1, length=10-50bp, Phred>25. Using the human reference genome GRCh38 (Genome Reference Consortium Human Reference 38, GCA_000001405.15), a total of 856 nuclear tRNAs and 21 mitochondrial tRNAs were annotated with tRNAscan-SE [v2.0] (140).

Trimmed FASTQ files were then mapped using a specific pipeline for tRNAs (**Fig. 3.1A**) (113). Summarizing, an artificial genome is first generated by masking all annotated tRNA genes and adding pre-tRNAs (i.e. tRNA genes with 3′ and 5′ genomic flanking regions) as extra chromosomes. Upon mapping to this artificial genome with Segemehl [v0.3.1] (142), reads that map to the tRNA-masked chromosomes or to

the tRNA flanking regions are filtered out in order to remove non-tRNA reads and unmature-tRNA reads respectively.

After this first mapping step, a second library is generated by adding 3′ CCA tails and removing introns from tRNA genes. All 100% identical sequences of these so-called *mature* tRNAs are clustered to avoid redundancy. Next, the subset of filtered reads from the first mapping is aligned against the clustered mature tRNAs using Segemehl [v0.3.1] (142). Mapped reads are then realigned with GATK IndelRealigner [v3.8] (143) to reduce the number of mismatching bases across all reads.

For quantification, isoacceptors were quantified as reads per million (RPM). In order to increase the coverage for anticodon-level quantification, we consider all reads that map unambiguously to a certain isoacceptor, even though they ambiguously map to different isodecoders (i.e. tRNA genes that differ in their sequence but share the same anticodon). Ambiguous reads mapping to genes of different isoacceptors were discarded.

Regarding modification site calling, we only considered gene-level uniquely mapped reads, as described to be optimal in Hoffmann et al. (113). As in their pipeline, in order to distinguish mapping or sequencing errors from true misincorporation sites, we use GATK UnifiedGenotyper [v3.8] (143). Furthermore, given that tRNAs have variable D-loop and V-region, we map the detected modifications to the standard tRNA model to make them comparable. We align our tRNA library to the structurally-annotated human tRNAs from tRNAdb (146) using BLAST [v2.9.0] (145), and fit the secondary structure annotation of the top BLAST hits.

3.5.8. Translational efficiency analysis

3.5.8.1. Relative Codon Usage (RCU) and Relative Anticodon Abundance (RAA)

The RCU/RAA is defined as the contribution of a certain codon/anticodon to the amino acid it belongs to. The RCU of all

synonymous codons and the RAA of all anticodons recognizing synonymous codons therefore sum up to 1.

$$RCU = \frac{x_C}{\sum_{i \in C_{aa}} x_i} \qquad RAA = \frac{x_A}{\sum_{i \in A_{aa}} x_i}$$

where x_C / x_A refers to the abundance of the codon/anticodon C/A, and C_{aa} is the set of all synonymous codons, as well as A_{aa} is the set of all anticodons that decode synonymous codons.

3.5.8.2. tRNA Adaptation Index (tAI)

As described by dos Reis et al. (61,128), the tAI weights every codon based on the wobble-base codon-anticodon interaction rules. Let c be a codon, then the decoding weight is a weighted sum of the square-root-normalized tRNA abundances $tRNA_{cj}$ for all tRNA isoacceptors j that bind with affinity $(1 - s_{cj})$ given the wobble-base pairing rules n_c. However, while dos Reis et al. (61) assumes that highly expressed genes are codon-optimized, here we use the non-optimized s-values to avoid a circularity in our reasoning:

$$s = [0,\ 0,\ 0,\ 0,\ 0.5,\ 0.5,\ 0.75,\ 0.5,\ 0.5]$$

$$w_c = \sum_{j=1}^{n_c} (1 - s_{cj}) tRNA_{cj}$$

And therefore the tAI of a certain protein is the product of weights of each codon i_k at the triplet position k throughout the full gene length l_g, and normalized by the length.

$$tAI = \left(\prod_{k=1}^{l_g} w_{i_k} \right)^{1/l_g}$$

For this and all further analyses, the coding sequences of *Homo sapiens* from RefSeq were downloaded from the Codon/Codon Pair Usage Tables (CoCoPUTs) project release as of February 6, 2019 (30,147).

3.5.8.3. Relative tRNA Adaptation Index (RtAI)

For comparison with the SDA (**Fig. 3.3A**), an amino-acid-normalized tAI measure is defined by dividing each tAI weight by the maximum weight among all codons within each amino acid family.

$$Rw_c = \frac{w_c}{max_{i \in C_{aa}}(w_i)}$$

And therefore the RtAI of a certain protein is the product of weights Rw of each codon i_k at the triplet position k throughout the full gene length l_g, and normalized by the length.

$$RtAI = \left(\prod_{k=1}^{l_g} Rw_{i_k} \right)^{1/l_g}$$

3.5.8.4. Supply-to-Demand Adaptation (SDA)

The SDA aims to consider not only tRNA abundances, but also the codon usage demand. In doing so, it constitutes a global measure of translation control, since the efficiency of a certain codon depends both on its complementary anticodon abundance as well as the demand for such anticodon by other transcripts. This global control has been indeed established to play an important role in defining optimal translation programs (96).

The definition of the SDA is based on similar previously published metrics (47,50), which consists of a ratio between the anticodon supply and demand. On the one hand, the anticodon supply is defined as the relative tAI weights Rw (see previous section). On the other, the anticodon demand is estimated from the codon usage at the

transcriptome level. It is computed as the frequency of each codon in a transcript weighted by the corresponding transcript expression, and finally summing up over all transcripts. Let c be a codon, then the codon usage is a weighted sum of the counts of codon c_i in gene j weighted by the mRNA-seq abundance $mRNA_j$ for all genes in the genome g:

$$CU_c = \sum_{j=1}^{g} c_{ij} mRNA_j$$

Similarly to the supply, the anticodon demand is then normalized within each amino acid family:

$$D_c = \frac{CU_c}{max_{i \in c_{aa}}(CU_i)}$$

Finally, the SDA weights (SDAw) are defined as the ratio between the codon supply S_c and demand D_c:

$$SDAw_c = \frac{S_c}{D_c}$$

And therefore the SDA of a certain protein is the product of weights $SDAw$ of each codon i_k at the triplet position k throughout the full gene length l_g, and normalized by the length.

$$SDA = \left(\prod_{k=1}^{l_g} SDAw_{i_k} \right)^{1/l_g}$$

3.5.9. Gene Set Enrichment Analysis (GSEA)

Gene sets derived from the GO Biological Process Ontology were downloaded from the Molecular Signatures Database [v6.2] (MSigDB)

as a GMT file (144,148). We analyzed the enrichment of gene sets using the GSEA algorithm (144). The score used to generate the ranked list input is specified in the text for each analysis.

3.5.10. Survival Analysis

To analyze how the supply-to-demand ratio of a certain codon (SDAw) can affect the survival probability in cancer, patients of a certain cancer type are divided in two groups of low/high SDAw, which correspond to the patients having the top and bottom 40% SDAw. The Kaplan-Meier curves are then computed to estimate the survival probability of each group along time.

3.5.11. tRNA methylation and copy number

For consistency with the current version of publicly available and pre-processed 450k DNA methylation and SNP6 segmented copy number alteration (CNA) data from firebrowse, we used the human reference genome GRCh37/hg19 (Genome Reference Consortium Human Reference 37, GCA_000001405.1) in this analysis. The coordinates of all nuclear tRNA genes were obtained using tRNAscan-SE [v2.0] (140).

Regarding DNA methylation, we computed the average beta value of each tRNA gene from 1.5kb upstream of the transcription start site (1500TSS) until the end of the gene. For CNA, we retrieved the segmented data of precomputed $log_2(CN) - 1$ from firebrowse and extracted the corresponding value for the genomic coordinates containing the tRNA genes. Whenever the tRNA genes was located between two segments, the weighted average in function of the gene overlap with each segment was computed.

3.5.12. Bisulfite sequencing methylation

As 1500TSS methylation of tRNA genes lead to an average coverage of only 37% genes, we also analyzed the recently published bisulfite

sequencing data of 47 samples across nine cancer types (Table EV4) (132). After retrieving the datasets from the GDC legacy archive, given the higher resolution of bisulfite sequencing data, we restricted the computation of the average promoter methylation of tRNA genes to the GRCh37/hg19 genomic coordinates containing the tRNA genes, since the promoter region of Pol-III-genes is intragenic.

3.5.13. Multiple Linear Regression (MLR)

We fitted a Multiple Linear Regression (MLR) between the square-root-normalized tRNA abundance (dependent variable) and the promoter methylation and gene copy number (independent variables). To make all three layers of information comparable, we considered only samples for which all data was available and performed the regression at the isoacceptor level, thus averaging the methylation and CNA data over all tRNA genes that shared the same anticodon.

$$EXP = \beta_0 + \beta_{Me} Me + \beta_{CNA} CNA$$

We fitted the model parameters for all 64 isoacceptors and 22 cancer types, leading to 22x64=1408 MLRs, among which only significant coefficients (FDR-corrected t-statistic p-value < 0.05) were considered in downstream analyses.

3.5.14. Statistical Analysis

For hypothesis testing, an unpaired two-tailed Wilcoxon rank-sum test was performed, unless stated otherwise. All details of the statistical analyses can be found in the Results section. We used a significance value of 0.05. In differential expression analyses, a False Discovery Rate correction was used to account for multiple testing.

3.6. Data and Software Availability

The datasets and computer code produced in this study are available in the following databases:

- Scripts for analyzing tRNA data of TCGA: GitHub (github.com/hexavier/tRNA_TCGA).
- Scripts for tRNA mapping: GitHub (github.com/hexavier/tRNA_mapping).
- Generated TCGA data (tRNA abundances, SDA, CNA, and DNA methylation): Synapse syn20640275 (www.synapse.org/tRNA_TCGA).
- Hydro-tRNA and small RNA sequencing data of all five cell lines: Gene Expression Omnibus GSE137834 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137834).

3.7. Acknowledgments

We thank Eva Maria Novoa Pardo, Samuel Miravet-Verde, and Marc Weber for stimulating and critical discussions. We thank the CRG Genomics Unit for assistance with RNA sequencing services. The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We acknowledge the support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa', the CERCA Programme / Generalitat de Catalunya, and the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership. The work of X.H. has been supported by a PhD fellowship from the Fundación Ramón Areces.

3.8. Author contributions

Conceptualization, X.H., M.H.S. and L.S.; Methodology, X.H., H.B., M.H.S., L.S.; Software, X.H.; Investigation, H.B., X.H.; Validation, X.H., M.H.S.; Formal analysis, X.H., M.H.S.; Writing-Original Draft, X.H.; Writing-Review & Editing, X.H., H.B., M.H.S., L.S.; Visualisation: X.H., M.H.S.; Funding Acquisition, L.S.; Supervision, M.H.S. and L.S.

Hernandez-Alias X, Benisty H, Schaefer MH, Serrano L.
Translational adaptation of human viruses to the tissues they infect.
Cell Reports. 2021;34(11):108872.

Virus infecting cell in Santiago Rusiñol style

by DALL·E 2

Chapter 4

Translational adaptation of human viruses to the tissues they infect

4.1. Summary

Viruses need to hijack the translational machinery of the host cell for a productive infection to happen. However, given the dynamic landscape of tRNA pools among tissues, it is unclear whether different viruses infecting different tissues have adapted their codon usage toward their tropism. Here, we collect the coding sequences of 502 human-infecting viruses and determine that tropism explains changes in codon usage. Using the tRNA abundances across 23 human tissues from TCGA, we build an *in silico* model of translational efficiency that validates the correspondence of the viral codon usage with the translational machinery of their tropism. For instance, we detect that the coronavirus SARS-CoV-2 is specifically adapted to the upper respiratory tract and alveoli. Furthermore, this correspondence is specifically defined in early viral proteins. The observed tissue-specific translational efficiency could be useful for the development of antiviral therapies and vaccines.

4.1.1. Highlights

- Viruses with distinct tissue tropisms show differences in codon usage.
- Viral tropism defines a unique pattern of translational adaptation to human tissues.
- SARS-CoV-2 is especially favored to the upper respiratory tract and the alveoli.
- Early viral proteins are generally better adapted than late counterparts.

4.1.2. Additional data access

All supplementary figures and data can be accessed from the original publication through this QR code and link.
Supplementary figures and tables will be referred to as "Figure S" and "Table S".

4.2. Introduction

Given the degeneracy of the genetic code, multiple 3-letter combinations of nucleotides can code for the same amino acid. Such synonymous codons are nevertheless not uniformly distributed along the genomes and can significantly deviate between organisms (149). Evolutionary forces that explain the existence of the so-called codon bias include (1) a mutation pressure for a certain GC base composition depending on the species and chromosomal location, and (2) the translational selection for codons corresponding to highly expressed tRNA isoacceptors (24,150,151).

Viruses strongly depend on the translational machinery of the host for the expression of their own proteins and, ultimately, their replication. For instance, given the small size of most viral genomes, no or very few tRNA genes are generally autonomously encoded (152). In terms of

codon usage, it has indeed been shown that bacteriophages are specifically adapted to their microbial hosts (153,154). This information has been applied in the prediction of viral hosts from metagenomics data (155,156). The codon usage of human-infecting viruses is similarly adapted to the host (157,158), and actually the concept of codon deoptimization has been applied in the design of attenuated vaccines (159).

Although translational selection has long been under debate in human (32), recent studies indicate that different tissues and conditions showcase distinct tRNA expression profiles, leading to changes in their respective translational efficiency (42,150). In agreement with this observation, the codon usage of papillomavirus capsid proteins is adapted to the tRNAs of differentiated keratinocytes, where their translation becomes specifically efficient (160,161). In addition, upon HIV-1 infection, the host tRNA pool is reprogrammed to favor translation of late viral genes (162), a phenomenon that is indeed exploited by host antiviral mechanisms (163). Furthermore, some viruses with a specific tissue tropism resemble the codon bias of highly expressed proteins of their respective infecting tissues (164). Nevertheless, despite the few aforementioned studies, a high-throughput analysis of the translational selection of viral genomes to their tissue tropism has been heretofore hindered by the absence of tissue-wide tRNA expression data.

Here, we systematically analyze the relative codon usage landscape of 502 human-infecting viruses together with the recently published tRNA expression profiles of human tissues (150). Among other viral annotated features, including phylogeny and Baltimore classification, their tissue tropism explains more variance in codon usage than the other tested features. In consequence, tropism corresponds with codon optimization patterns that can be associated with tissue-specific profiles of tRNA-based translation efficiencies. Further, by studying the tissue-adaptation among the viral proteome, we also determine that early replication-related proteins are more translationally-adapted than the late structural counterparts. Overall, we observe a tropism-specific

adaptation of the viral proteome to the tRNA profiles of their target tissues, which opens the door to the development of tissue-specific codon-deoptimized vaccines and targeted antiviral therapies.

4.3. Results

4.3.1. Tropism corresponds with differences in Relative Codon Usage of human-infecting viruses

Publicly available genomic data comprised a total of 502 human-infecting viruses, distributed among 35 families and covering all seven Baltimore categories (Table S1). Across this diversity, six main viral tropisms were defined for 228 viruses based on the ViralZone curated database (165): neurons, immune cells, respiratory tract, hepatocytes, intestine, and epithelial cells (**Fig. 4.1A**), while the rest of viruses remained unassigned. Their corresponding coding sequences constituted a total of 6087 viral proteins (Table S1), for which we determined the Relative Codon Usage (RCU, i.e. the contribution of each synonymous codon to the amino acid it encodes, see Methods).

In order to understand the main factors driving differences between viral RCU, we used three internal clustering indexes that assess how similar each virus is to a certain group compared to other groups. Taking the average RCU over each of the 502 viral proteomes, we applied this framework to assess the grouping performance of five different viral features: tropism, type of genetic material (aka Baltimore category), family, genus, and a sequence-based classification by Aiewsakun and Simmonds (2018). In such analysis, the tropism leads the best classification of viral RCUs, followed by the viral genetic type (**Fig. 4.1B**). On the other hand, classical and sequence-based phylogenetic classifications show poor clustering performances.

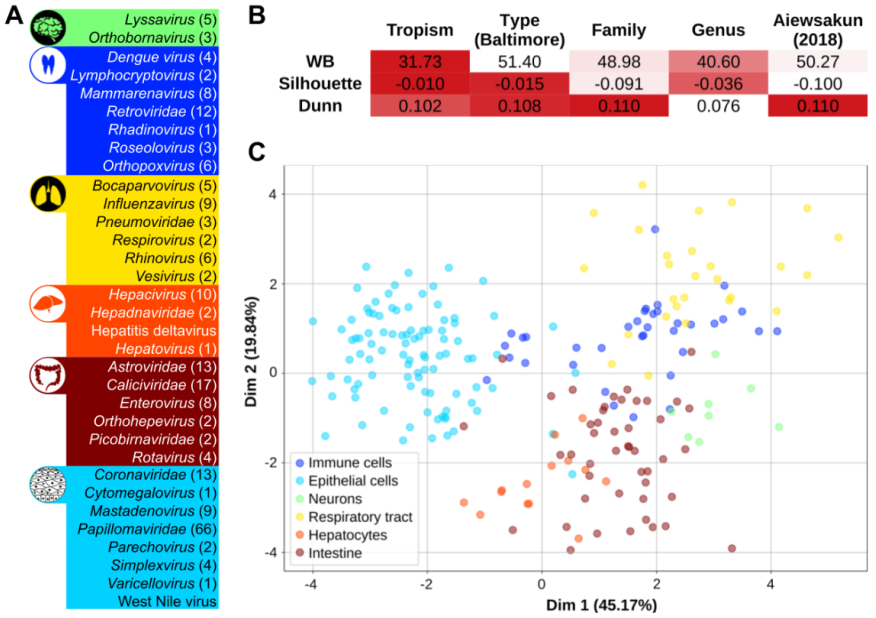


Figure 4.1. Tropism corresponds with differences in relative codon usage of human-infecting viruses.
(A) A total of 502 viruses was distributed among 35 families and covered all seven Baltimore groups. From there, 228 viruses were classified in six general tropisms based on ViralZone annotations (165). (B) Three internal clustering indexes were computed to assess the validity of each viral classification in terms of their RCU. Good cluster performances lead to low WB indexes, but to high Silhouette and Dunn values (as shown in the color code). (C) Linear Discriminant Analysis of the RCU of the 228 tropism-defined viruses. In brackets, the percentage of variance explained by each of the components. See also Fig. S1.

Given the impact of viral tropism on the RCU, we sought to determine the main codon differences between the six defined target tissues. By using a linear discriminant analysis (LDA, see Methods), the 228 tropism-defined viruses were classified in six clear clusters, regardless of other factors such as the phylogenetic lineage (**Fig. 4.1C**). For further validation, by randomizing the set of tropisms, the LDA loses all its discriminating power (Fig. S1A). Supporting the robustness of the clustering, we obtain the same performance using another reduced list of human viruses from ViralZone (165) (Fig. S1B, Table S1).

With the previous results indicating a clear codon usage pattern among tropisms, we then wondered to what extent other factors could in

parallel shape the nucleotide composition of viruses (29). As shown in Fig. S1, we observed that RNA folding, as determined by the minimum free energy, is also non-randomly distributed among tropisms. Other factors such as ribosomal frameshift are not significantly different.

Overall, we observe that specific codon usage profiles are associated with the tissue tropism of human-infecting viruses, together with other determinants such as RNA stability.

4.3.2. Viruses are adapted to the tRNA-based translational efficiencies of their target tissues

Based on the RCU differences between viruses with distinct tropism, we hypothesize that distinct tissues impose selection towards a certain set of translationally-efficient codons. However, a validation for this hypothesis requires the accurate quantification of tissue-specific tRNA profiles, which has been hitherto missing. With the advent of such high-throughput expression data (109,116), here we retrieved the previously-published Supply-to-Demand Adaptation (SDA) estimate for translational efficiency (47,150), which computes the balance between the supply (i.e., the anticodon tRNA abundances) and demand (i.e., the codons expressed in mRNAs) of each codon (see Methods).

Using a total of 620 healthy samples from The Cancer Genome Atlas (TCGA) dataset (150), we first computed the SDA of all viral-protein-coding sequences based on the SDA weights of their constituent codons. Therefore, taking the average of all healthy samples across each of the 23 TCGA cancer types, we determined the estimated translational efficiencies of viral proteins in different human tissues (Table S2).

Next, from the perspective of the translational selection hypothesis, we would expect that viral proteins are translationally adapted to their target tissues. In consequence, we tried to test our hypothesis using a completely blind and unbiased random forest classifier, which applies machine learning in order to predict the tropism of each viral protein

based on the SDA to different tissues (see Methods). The resulting performance of the models, based on the Area Under the Curve (AUC) of their Receiver Operating Characteristic (ROC) curves, ranges between 0.79-0.92 (**Fig. 4.2A**), clearly higher than the no-skill model of 0.5 ($p < 0.01$, permutation test in Table S2). Similar results are also obtained from complementary prediction performance metrics such as Precision-Recall curves (**Fig. 4.2A**). This analysis was replicated with the other list of viruses from ViralZone, leading to comparable predictive outcomes (Fig. S2, $p < 0.01$, permutation test in Table S2). These results indicate that our machine learning model is able to predict the tropism of a viral protein based on its SDA to tissues with high accuracy. In concordance, a linear discriminant analysis of the average SDA of each virus across tissues can similarly separate different clusters of viral tropism based on their translational efficiencies (Fig. S2).

In an attempt to understand which tissues are the most predictive in identifying the viral tropism of proteins, we analyzed the relative feature importance within each random forest classifier, which measures the contribution of each tissue SDA in the decision trees (**Fig. 4.2B**). The main observation is that no single tissue alone is able to discriminate against the specific tropism, since all feature importances lie below 0.09. However, it is also clear that translational adaptation to stomach (STAD, for healthy samples of stomach adenocarcinoma) is a recurrent discriminant feature, while other tissues are specifically important for just one or few tropisms, such as liver (LIHC, for healthy samples of liver hepatocellular carcinoma) in predicting hepatocyte viruses. In any case, the directionality of these features cannot be established.

All these analyses using the TCGA dataset are based on tRNA quantifications derived from generic small RNA sequencing, which we have previously reported to provide consistent measurements compared to other tRNA sequencing techniques such as Hydro-tRNAseq (150). However, to exclude any possible technical bias related to the low tRNA coverage of the technique, we have reproduced the same Random Forest model of viral tropism using an alternative dataset of Hydro-tRNAseq across seven tissue-wide cell lines (HEK293, HCT116, HeLa,

MDA-MB-231, BJ/hTERT, HACAT, and HepG2, see Methods). The results show similar predictive performances compared to TCGA (Fig. S3).

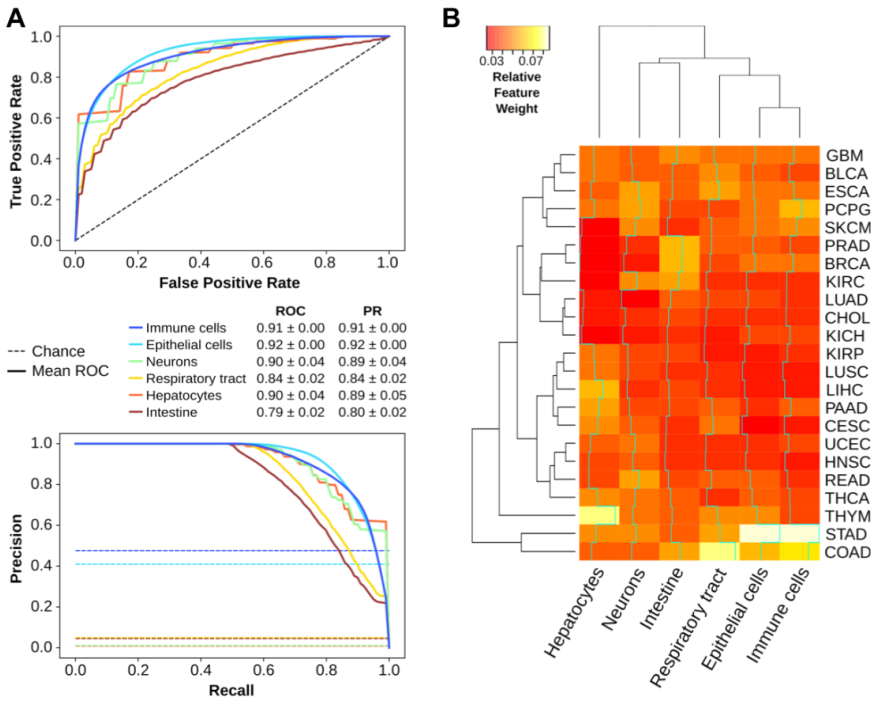


Figure 4.2. Viruses are adapted to the tRNA-based translational efficiencies of their target tissues.

(A) Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves of a Random Forest Classifier, in which the average Supply-to-Demand Adaptation of viral proteins to each of the 23 TCGA tissues is used to predict their corresponding viral tropism of NCBI viruses (see Methods). The area under the curves (AUC) ± SD summarizes the performance of the model. (B) Relative feature weights of each of the 23 TCGA tissues for each of the six tropisms, which measure the contribution of each tissue in the decision trees. The dendrograms show a hierarchical clustering among tissues (left) and among tropisms (top). The cyan lines show the trace of weights along each tropism. Refer to Table S2 for full TCGA cancer type names. See also Fig. S2-5.

In addition, as our systematic analysis suggests that the codon usage of viruses tend to be adapted to the tRNAs of the tissue they infect, we specifically interrogated the translational efficiency of the coronavirus SARS-CoV-2, which is causing the most deadly pandemic of the recent decades (167). As a result, we observe that the coronavirus proteome is especially adapted to the upper respiratory airways and alveoli, but also to other tissues such as the gastrointestinal tract and brain (Fig. S4 and S5, Table S3, see Methods).

Overall, as tropism of viruses can be predicted from their translational adaptation to tissues, these results indicate that viral proteomes are specifically adapted to certain tRNA-based translational efficiencies. In consequence, and complementary to the observations of mutational pressure driving viral codon bias (158,168,169), we describe the basis for a potential tissue-specific translational selection of the viral codon usage.

4.3.3. Early viral proteins are better adapted than late counterparts

Given the tropism-specific adaptation of viral RCU towards the translational machinery of tissues, we wondered whether certain genomic subsets were specifically adapted to the tissue of infection. In particular, we speculated that early replication-related proteins would further benefit from such adaptation than late structural proteins, since once the virus takes control of the cell it could change its tRNA expression program (162,170).

To estimate the adaptation of each protein to the tRNA-based codon efficiencies of each tissue, we computed their SDA (150) (Table S4). For that purpose, we matched each virus to the tRNAs of their tissues of infection (Table S4). In concordance with our hypothesis, based on current viral annotations (VOGdb, vogdb.org), we observed a small but highly significant shift in SDA between structural and replication proteins across most viral tropisms, with the exception of hepatocyte and intestine viruses (**Fig. 4.3A**, paired two-tailed Wilcoxon rank-sum test). Similarly, we performed a Gene Set Enrichment Analysis to

identify which Virus Orthologous Groups (VOGs) were enriched in high-SDA or low-SDA proteins (**Fig. 4.3B**). As determined by current annotations (171), top-VOGs mostly contained replication-related early proteins, whereas bottom-VOGs had structural late functions, with few exceptions to the general trend.

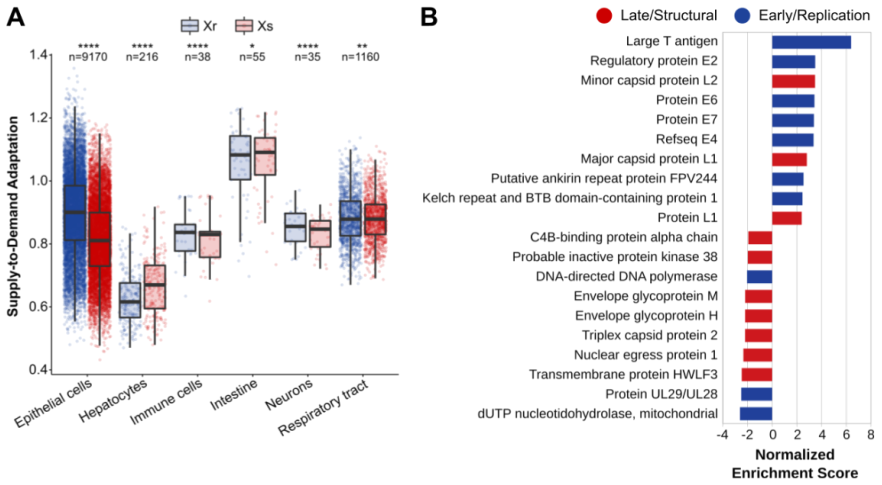


Figure 4.3. Early viral proteins are better adapted than late counterparts.
(A) Average Supply-to-Demand Adaptation of replication (Xr) and structural (Xs) proteins of a total of 104 annotated tropism-specific viruses, respectively matched to 461 samples of their tissues of infection (Table S4). Boxes expand from the first to the third quartile, with the center values indicating the median. The whiskers define a confidence interval of median ± 1.58*IQR/sqrt(n). Statistical significance is determined by paired (structural against replication proteins of each virus) and two-tailed Wilcoxon rank-sum test. (B) Top 10 positive and negative Virus Orthologous Groups upon Gene Set Enrichment Analysis of the SDA of all proteins of tropism-specific viruses (Table S4). Based on their annotations, proteins groups are colored based on their early/replication or late/structural function (171).

Previous studies on the translational adaptation of the human immunodeficiency virus 1 suggested that the host tRNA pool is reprogramed upon viral infection in order to favor the expression of late genes (162). In this direction, we wanted to test whether this tRNA reprogramming is a general adaptive mechanism among viral species. Using three previously published small RNA-sequencing datasets of human cell lines upon viral infection (172–174), we quantified the tRNA abundances at different time points (Table S5). Therefore, in terms of time-course differences, we detected a general decrease in translational efficiency (measured as RtAI, see Methods) upon viral infection, which is relatively more pronounced for late proteins rather than early (**Fig. 4.4A**). At the same time, to compare the absolute translational efficiency of late and early genes, we also compiled previously published proteomic measurements upon infection of these three viruses (175–177). While there are no consistent differences of early versus late protein levels across viruses (**Fig. 4.4B**), we nonetheless observed that most abundant expression classes tend to have higher translation efficiencies (**Fig. 4.4**).

Overall, we determine that the tropism-specific adaptation of viruses is specifically pronounced among early proteins. However, the lower adaption of late viral genes and findings on translation changes upon infection suggest that host cells might be reprogrammed to favor the expression of late viral genes.

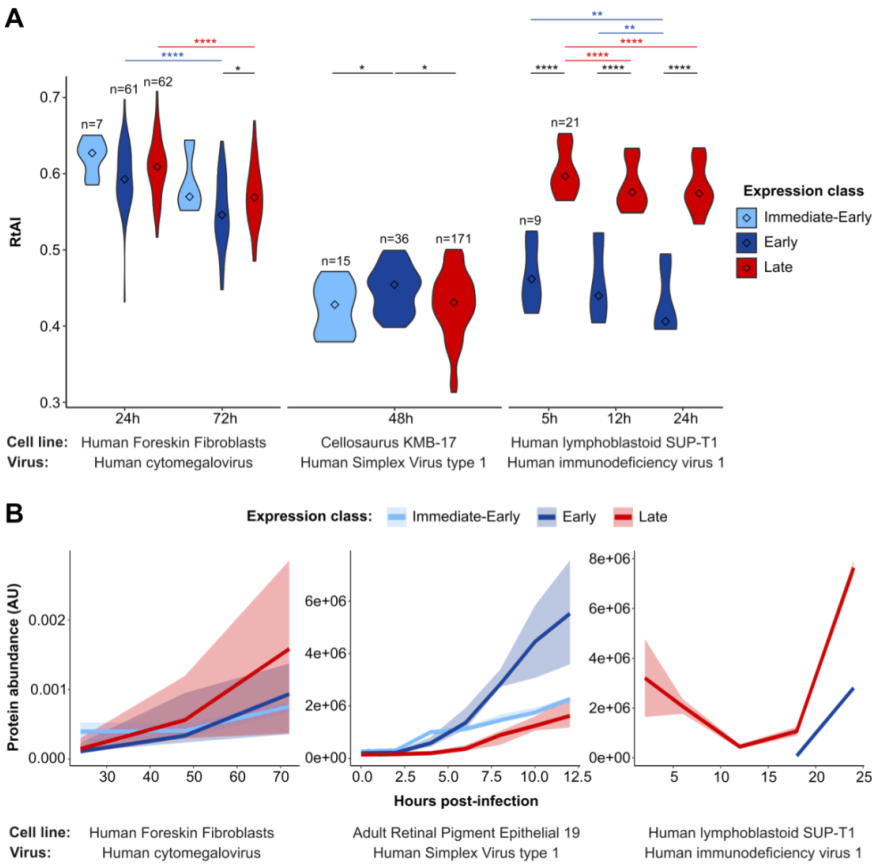


Figure 4.4. Translational adaptation of viral proteins upon infection.
(A) Relative tRNA Adaptation Index (see Methods, Table S5) of viral proteins upon effective viral infections in different cell lines. Proteins are allocated to different time expression classes based on current viral knowledge (171) (Table S5). Center values within the violin plot represent the median. Only significant differences are shown: * (p <= 0.05), ** (p <= 0.01), *** (p <= 0.001), **** (p <= 0.0001). Statistical differences are based on a FDR-corrected two-tailed Wilcoxon rank-sum test, with paired comparisons between time points (written in color) and unpaired comparisons between expression classes. (B) Abundances of viral proteins upon effective viral infections at different time points in different cell lines. Solid lines represent the median of the expression class, surrounded by an uncertainty interval between the 0.4 and 0.6 percentiles.

4.4. Discussion

Tropism is determined by an ensemble of different factors, including the mechanism of viral entry to the host, the immune responses to the infection, or the viral hijacking of the cellular machinery in the interest of replication and propagation. In this article, we study the latter by focusing on the translational adaptation of viral genomes to the host.

There could be a certain controversy to what extent some viruses out of the 502 included in the NCBI database used here are truly adapted to humans, or whether some of them are annotated just because of anecdotal infections. Also, for some viruses, multiple genotypes and variants are represented in the NCBI database, although they actually do showcase differences in codon usage (Table S1, **Fig. 4.1**). To exclude any statistical bias due to the used database, we have duplicated our analysis with the reduced highly curated ViralZone database, obtaining similar results and thus indicating our observations are robust.

While previous studies on the base composition and codon usage of both DNA and RNA viruses (158,169) have attributed most of the codon usage variability to the mutational pressure of viral genomes, our analysis proposes tropism as another potential driving force. By systematically interrogating all human-infecting viruses, we uncover that tissue tropism explains changes in their codon usage more than other viral properties such as type or family. Therefore, as mutational pressure would act more similarly within phylogenetically closer species, such tropism-related differences in codon usage suggests that tissue-specific tRNA expression could be driving a translational selection on viral genomes. However, as suggested by Fig. S1, many other and overlapping forces such as mRNA stability, frameshift motifs, transcriptional regulation, or codon-dependent immune responses are also known to shape the composition of coding sequences (29,163).

Although high-throughput sequencing of tRNAs has been only recently developed, cases of natural selection of codon usage towards the host have been previously proposed. For instance, codon usage of

Parvovirus has been progressively adapted from dogs to cats after the host jump (178). *Influenzaviruses* show a similar adaptation over time of viral isolation, deviating from the codon usage of avian hosts (179,180). However, whether these progressive changes in codon usage over time are directly driven by translational selection has remained elusive. With the advent of tissue-wide datasets of tRNAs and their translational efficiencies (150), we can now compute the Supply-to-Demand Adaptation (SDA) of all viral proteomes in different tissues. From there, we then created a random forest model that predicts with high accuracy the viral tropism of proteins based on their profile of adaptation to human tissues. In consequence, the tRNA-based adaptation profile of a protein is descriptive of their viral tropism, indicating that translational selection could indeed drive tropism differences of codon usage. It is important to remark that viruses could still have a good SDA to non-target tissues with similar tRNA expression patterns that are not infected because they are not exposed to the virus.

In particular, we find that the coronavirus SARS-CoV-2 is highly adapted to the upper respiratory tract and the alveoli (Fig. S4C), which is in agreement with recent single-cell transcriptomic studies reporting the expression of ACE2 in the nasal goblet and ciliated cells as well as the type-2 alveolar epithelial cells (181,182). Apart from the respiratory tract, the gastrointestinal tract emerges as the most translationally adapted tissue, followed by the other epithelial-like tissues and the brain, which concurs with some frequently observed COVID-19 symptoms (183–188). In terms of the evolution of the new coronavirus, given the similarity of SARS-CoV-2 SDA with the phylogenetically closest bat coronavirus (Fig. S5B), it seems that a translational selection to increase SDA would have acted before the putative zoonosis from bats or other intermediate hosts. Furthermore, in agreement with the highest translational potential of SARS-CoV-2 in their target tissues, a recent model of viral tropism suggested that a tradeoff exists between the efficiency of viral translation and the translational load on the host, indicating that an improved codon usage can make the difference between symptomatic and natural hosts (189).

On the other hand, in analyzing differences in codon usage between early and late viral genes, previous studies do not completely agree. While it would be intuitive and some authors claim that late proteins, which often need to be expressed in higher amounts, are better translationally adapted than early counterparts (157), others state otherwise (170,190). Using the tRNA abundances from the TCGA dataset and based on the Supply-to-Demand Adaptation, we therefore validate that early replication-related proteins are generally better adapted to the tissue of infection, despite few exceptions (**Fig. 4.3B**). In agreement with this observation, it is known that host tRNA pools either undergo reprogramming upon HIV-1 infection (162), or get locally channeled to ribosomes in vaccinia and influenza A viruses (191). Upon infection, we propose that translational adaptation could switch in some cases towards the expression of late structural proteins, which has previously been observed in HIV-1 (162).

Overall, this systematic analysis establishes a link between the codon usage of human viruses and the translational efficiency of their tissue of infection. This correspondence is particularly observed in early viral proteins. We therefore envision the development of *ad hoc* gene therapies specifically targeting the tissue of interest.

4.5. STAR Methods

4.5.1. Key Resources Table

Reagent or Resource	**Source**	**Identifier**
Chemicals, Peptides, and Recombinant Proteins		
Antarctic phosphatase	New England BioLabs	Cat#M0289
T4 Polynucleotide Kinase	New England BioLabs	Cat#M0201
ProtoScript II Reverse Transcriptase	New England BioLabs	Cat#M0368
Critical Commercial Assays		
miRNeasy Mini kit	Qiagen	Cat#217004

15% TBE–urea gels	NOVEX, Invitrogen	Cat#EC6885BOX
RNeasy MinElute Cleanup Kit	Qiagen	Cat#74204
QIAquick PCR Purification Kit	Qiagen	Cat#28106
Deposited Data		
Supply-to-Demand Adaptation weights (SDAw) from TCGA samples	(150)	Synapse: syn20640275
SARS-CoV-2 reference genome	(192)	NCBI Reference Sequence: NC_045512.2
Bat coronavirus RaTG13 genome	(193)	GenBank: MN996532.1
Small RNA-seq of HFF infected by HCMV	(174)	GEO: GSE33584
Small RNA-seq of KMB-17 infected by HSV1	(173)	GEO: GSE102470
Small RNA-seq of SUP-T1 infected by HIV1	(172)	GEO: GSE57763
Hydro-tRNAseq of HEK293, HCT116, HeLa, MDA-MB-231, and BJ/hTERT	(150)	GEO: GSE137834
Hydro-tRNAseq of HACAT and HepG2	This study	ArrayExpress: E-MTAB-9905
Experimental Models: Cell Lines		
HACAT	CRG Collection (Center for Genomic Regulation)	RRID: CVCL_0038
HepG2	IMIM Collection (Institut Hospital del	RRID: CVCL_0027

	Mar d'Investigacions Mèdiques)	
Software and Algorithms		
GSEA [v4.0.3]	(144)	http://software.broadinstitute.org/gsea
SciKit Learn [v0.20.1]	(194)	https://scikit-learn.org
Codon Usage tool	(195)	https://www.bioinformatics.org/sms2/codon_usage
ViennaRNA toolkit [v2.4.14]	(162)	https://www.tbi.univie.ac.at/RNA
KnotInFrame	(196)	https://bibiserv.cebitec.uni-bielefeld.de/knotinframe
BBMap [v38.22]	Bushnell B.	https://sourceforge.net/projects/bbmap
FastQC [v0.11.4]	Andrews S.	https://www.bioinformatics.babraham.ac.uk/projects/fastqc
SAMtools [v1.3.1]	(139)	http://samtools.sourceforge.net
tRNAscan-SE [v2.0]	(140)	http://lowelab.ucsc.edu/tRNAscan-SE
BEDtools [v2.27.1]	(141)	https://bedtools.readthedocs.io/en/latest
Segemehl [v0.3.1]	(142)	https://www.bioinf.uni-leipzig.de/Software/segemehl
Picard [v2.18.17]	Broad Institute	https://github.com/broadinstitute/picard
GATK [v3.8]	(143)	https://software.broadinstitute.org/gatk
Other		
Code for tRNA mapping and quantification of Hydro-tRNAseq data	This paper	https://github.com/hexavier/tRNA_mapping

Code for all computational analyses of this report	This paper	https://github.com/hexavier/tRNA_viruses

4.5.2. Resource availability

This study did not generate new unique reagents. The code used in this study is available at GitHub (github.com/hexavier/tRNA_viruses, github.com/hexavier/tRNA_mapping). All generated raw sequencing data can be accessed at ArrayExpress E-MTAB-9905.

4.5.3. Experimental model and subject details

4.5.3.1. Cell lines

The cell lines included in this study are HACAT and HepG2. The sex of each cell line is as follows: HACAT, male; HepG2, male. Cells were maintained at 37°C in a humidified atmosphere at 5% CO2 in DMEM 4.5 g/l Glucose with UltraGlutamine media supplemented with 10% of FBS and 1% penicillin/streptomycin. Cells have been tested negative for mycoplasma.

4.5.4. Method details

4.5.4.1. Biological Assays

RNA extraction

Cells were grown in 60 mm dishes for 48h in triplicates. Total RNA from HACAT and HepG2 was extracted using the miRNeasy Mini kit. 20 μg of total RNA was treated following the protocol of Hydro-tRNAseq (109).

Hydro-tRNA sequencing

Total RNA was resolved on 15% Novex TBE–urea gels and size-selected for 60-100 nt fragments. The recovered material was then alkaline hydrolyzed (10 mM sodium carbonate and 10 mM sodium bicarbonate) for 10 min at 60°C. The resulting RNA was de-phosphorylated with

Antarctic Phosphatase (New England Biolabs) at 37°C for 1 h. De-phosphorylated RNA was purified with an RNeasy MinElute spin column and re-phosphorylated with polynucleotide kinase (NEB). PNK-treated tRNAs were purified with an RNeasy MinElute spin column, adaptor-ligated, reverse-transcribed, and PCR-amplified for 14 cycles. The resulting cDNA was purified using a QIAQuick PCR Purification Kit and sequenced on Illumina HiSeq 2500 platform in 50 bp paired-end format.

4.5.4.2. Data Sources

Viruses and annotations

We included in the analysis all human-infecting viruses from the NCBI Viral Genome Browser, downloaded as of June 9, 2020. Additionally, for its interest, we added *a posteriori* the new SARS-CoV-2 virus. Viral metadata including family, genus, genetic material type and Baltimore category were retrieved either from the ICTV 2018b Master Species List (197) or the ICTV Virus Metadata Resource (talk.ictvonline.org/taxonomy/vmr/). The sequence-based phylogenetic information was obtained from Aiewsakun and Simmonds (166). Tissue and cell type tropism was determined based on the curated database ViralZone (165), and allocated to each of the six main classes based on the main annotation. To exclude any bias due to the source of the list, we also used the list of human-infecting viruses of ViralZone (165). Table S1 contains all human-infecting viruses and their associated metadata.

Coding sequences

The coding sequences of human-infecting viruses from RefSeq were downloaded from the Codon/Codon Pair Usage Tables (CoCoPUTs) project release as of June 9, 2020 (30,147) (Table S1). The SARS-CoV-2 and RaTG13 sequences were directly downloaded from GenBank (Table S3).

Virus Orthologous Groups

Virus Orthologous Groups and their functional annotations (virus structure and replication) were downloaded from VOGdb (vogdb.org, release number vog94). The protein sets of each VOG were formatted to a Gene Matrix Transposed (GMT) file for custom GSEA analyses.

TCGA translational efficiency

The Supply-to-Demand Adaptation (SDA) is the balance between the supply (i.e. the anticodon tRNA abundances) and demand (i.e. the weighted codon usage based on the mRNA levels) for each of the 60 codons (excluding methionine and Stop codons) (150). The SDA weights of all TCGA samples were downloaded from Synapse (www.synapse.org/tRNA_TCGA, syn20640275).

Hydro-tRNAseq of HeLa, HEK293, HCT116, MDA-MB-231, and fibroblast BJ/hTERT

Using the exact same protocol as described above, we have previously generated and published the Hydro-tRNAseq data of five tissue-wide human cell lines: HeLa, HEK293, HCT116, MDA-MB-231, and fibroblast BJ/hTERT (150). The raw data is publicly accessible at the Gene Expression Omnibus (GSE137834).

Small RNA-sequencing datasets upon viral infection

Three small RNA-sequencing datasets were downloaded to analyze the tRNA content of human cell lines upon viral infection. In Stark et al. (2012), samples of Human Foreskin Fibroblasts (HFF) infected with human cytomegalovirus (HCMV) strain Towne at a multiplicity of infection (MOI) of 3, analyzed at 24 and 72 hours post-infection (GSE33584). In Shi et al. (2018), samples of cellosaurus KMB-17 infected with Human Simplex Virus type 1 (HSV1) strain 17 at a MOI of 1, analyzed at 48 hpi (GSE102470). In Chang et al. (2013), samples of lymphoblastoid SUP-T1 cells infected with Human immunodeficiency virus 1 (HIV1) strain LAI at a MOI of 5, at 5, 12 and 24 hpi (GSE57763). The raw FASTQ files were analyzed using the tRNA quantification pipeline below.

Proteomics datasets upon viral infection

Three proteomics datasets were downloaded to analyze the abundances of viral proteins in human cell lines upon infection. In Golumbeanu et al. (175), iBAQ mass spectrometry quantification was used with lymphoblastoid SUP-T1 cells infected with a HIVeGFP-based viral vector, analyzed at 6, 12, 18 and 24 hpi. In Ouwendijk et al. (177), TOP3 MS quantification was used in human retinal pigment epithelial ARPE-19 cells infected with HSV-1 F-strain at a MOI of 1, analyzed at 0, 2, 4, 6, 8, 10 and 12 hpi (label-free absolute measurements of peptides were accessed upon request to the authors). For HCMV, we used two datasets of HFF infected with HCMV strain Merlin at a MOI of 10 with TMT mass spectrometry: (a) WCL3 from Weekes et al. (198), and (b) proteomic series three from Fielding et al. (199). The iBAQ absolute quantifications of these two datasets have been previously published in Nobre et al. (176). Therefore, quantifications at 24, 48 and 72 hpi were determined by distributing the absolute iBAQ quantification among the relative TMT abundances. All proteomic data is accessible in Table S5.

4.5.4.3. Computational Analysis

Relative Codon Usage (RCU)

The RCU is defined as the contribution of a certain codon to the amino acid it belongs to. The RCU of all synonymous codons therefore sum up to 1.

$$RCU = \frac{x_C}{\sum_{i \in C_{aa}} x_i}$$

where x_C refers to the abundance of the codon C, and C_{aa} is the set of all synonymous codons.

tRNA quantification

In both Hydro-tRNAseq and small RNA-seq FASTQ files, sequencing adapters were trimmed using BBDuk from the BBMap toolkit [v38.22]

(https://sourceforge.net/projects/bbmap): k-mer=10 (allowing 8 at the end of the read), Hamming distance=1, length=10-75bp, Phred>25. Using the human reference genome GRCh38, the high confidence set of tRNAs from GtRNAdb (12) was annotated with tRNAscan-SE [v2.0] (140), which includes a total of 432 nuclear tRNAs and 20 mitochondrial tRNAs.

Trimmed FASTQ files were then mapped using a specific pipeline for tRNAs (113). Summarizing, an artificial genome is first generated by masking all annotated tRNA genes and adding pre-tRNAs (i.e. tRNA genes with 3′ and 5′ genomic flanking regions) as extra chromosomes. Upon mapping to this artificial genome with Segemehl [v0.3.1] (142), reads that map to the tRNA-masked chromosomes or to the tRNA flanking regions are filtered out in order to remove non-tRNA reads and unmature-tRNA reads respectively.

After this first mapping step, a second library is generated by adding 3′ CCA tails and removing introns from tRNA genes. All 100% identical sequences of these so-called *mature* tRNAs are clustered to avoid redundancy. Next, the subset of filtered reads from the first mapping is aligned against the clustered mature tRNAs using Segemehl [v0.3.1] (142). Mapped reads are then realigned with GATK IndelRealigner [v3.8] (143) to reduce the number of mismatching bases across all reads.

For quantification, isoacceptors were quantified as reads per million (RPM). In order to increase the coverage for anticodon-level quantification, we consider all reads that map unambiguously to a certain isoacceptor, even though they ambiguously map to different isodecoders (i.e. tRNA genes that differ in their sequence but share the same anticodon). Ambiguous reads mapping to genes of different isoacceptors were discarded.

Relative tRNA Adaptation Index (RtAI)

As described by dos Reis et al. (61,128), the tAI weights every codon based on the wobble-base codon-anticodon interaction rules. Let c be a

codon, then the decoding weight is a weighted sum of the square-root-normalized tRNA abundances $tRNA_{cj}$ for all tRNA isoacceptors j that bind with affinity $(1 - s_{cj})$ given the wobble-base pairing rules n_c. However, while dos Reis et al. (61) assumes that highly expressed genes are codon-optimized, here we use the non-optimized s-values to avoid a circularity in our reasoning:

$$s = [0,\ 0,\ 0,\ 0,\ 0.5,\ 0.5,\ 0.75,\ 0.5,\ 0.5]$$

$$w_c = \sum_{j=1}^{n_c} (1 - s_{cj}) tRNA_{cj}$$

For better comparison with the SDA, an amino-acid-normalized tAI measure is defined by dividing each tAI weight by the maximum weight among all codons within each amino acid family.

$$Rw_c = \frac{w_c}{max_{i \in C_{aa}}(w_i)}$$

And therefore the RtAI of a certain protein is the product of weights Rw of each codon i_k at the triplet position k throughout the full gene length l_g, and normalized by the length.

$$RtAI = \left(\prod_{k=1}^{l_g} Rw_{i_k} \right)^{1/l_g}$$

Supply-to-Demand Adaptation (SDA)

The SDA aims to consider not only tRNA abundances, but also the codon usage demand. In doing so, it constitutes a global measure of translation control, since the efficiency of a certain codon depends both on its complementary anticodon abundance as well as the demand for such anticodon by other transcripts. This global control has been indeed

established to play an important role in defining optimal translation programs (96).

The definition of the SDA is based on similar previously published metrics (47,50,150), which consists of a ratio between the anticodon supply and demand. On the one hand, the anticodon supply is defined as the relative tAI weights Rw (see previous section). On the other, the anticodon demand is estimated from the codon usage at the transcriptome level. It is computed as the frequency of each codon in a transcript weighted by the corresponding transcript expression, and finally summing up over all transcripts. Let c be a codon, then the codon usage is a weighted sum of the counts of codon c_i in gene j weighted by the mRNA-seq abundance $mRNA_j$ for all genes in the genome g:

$$CU_c = \sum_{j=1}^{g} c_{ij} mRNA_j$$

Similarly to the supply, the anticodon demand is then normalized within each amino acid family:

$$D_c = \frac{CU_c}{max_{i \in C_{aa}}(CU_i)}$$

Finally, the SDA weights (SDAw) are defined as the ratio between the codon supply S_c and demand D_c:

$$SDAw_c = \frac{S_c}{D_c}$$

And therefore the SDA of a certain protein is the product of weights $SDAw$ of each codon i_k at the triplet position k throughout the full gene length l_g, and normalized by the length.

$$SDA = \left(\prod_{k=1}^{l_g} SDAw_{i_k} \right)^{1/l_g}$$

Internal clustering validity

Three indexes were used to determine the clustering performance of the RCUs based on different viral features. These are "internal" metrics, since they evaluate the quality of a certain grouping using measures of the dataset itself (homogeneity of clusters, distances within and between clusters, etc.).

- WB index is a ratio of the sum-of-squares (SS) within clusters and the SS between clusters, normalized by the number of clusters (200). Therefore, low values of the WB index are indicative of good clustering.
- Dunn index considers the inter-cluster distance and diameter of the cluster hypersphere (201). A higher Dunn index indicates better clustering.
- Silhouette Coefficient ranges from -1 to +1, and measures how similar an object is to its own cluster (intra-cluster distance) compared to other clusters (nearest-cluster distance) (202). A high value indicates a correct clustering.

Linear Discriminant Analysis of viral RCU

We applied a Linear Discriminant Analysis (LDA) to the viral RCUs, taking for each virus the average RCU of its proteins. We assigned each virus to its corresponding tropism (Table S1) in order to find the linear combination of codon features that maximized differences between viral target tissues. Given the collinear nature of RCUs by definition, the estimated coefficients are impossible to interpret, although it does not hamper the classification performance.

Other determinants of codon usage

To analyze the extent of multiple coding determinants on viral sequences, we computed two metrics associated with the folding of RNAs and the presence of ribosomal frameshift motifs. In both cases, we compared these results to a set of randomized sequences, which code for the exact same protein and have the same codon usage, but their codon composition is shuffled.

- Minimum Free Energy (MFE). RNAs are not simple linear sequences, but rather need to be appropriately folded. As such, we applied the ViennaRNA toolkit (203) to predict the folding of all viral RNA sequences and therefore determine their corresponding Minimum Free Energy (MFE).
- Ribosomal Frameshift prediction. Viruses are known to incorporate ribosomal frameshift events in their genomes in order to induce the expression of downstream coding regions or regulate the expression of protein products (204). As such, we applied the KnotInFrame tool (196) to identify sequences that could induce ribosome frameshift and would therefore be biasing our analysis. The algorithm computes the MFE of the pseudoknot RNA structure, which is known to produce frameshifts, and compares it with the base RNA folding.

Random Forest Classifier

To evaluate the adaptation of the viral proteins to the SDAw of human tissues, we computed their average SDA to each of the 23 TCGA tissues (Table S2). Using the set of 228 tropism-defined viruses, we had a total of 2891 viral proteins. Taking the 23 tissue-specific SDAs as features, we applied a Random Forest (RF) Classifier, populated with 100 decision trees, using the *scikit-learn* package (194). Therefore, for each of the six viral tropisms, we developed a model for predicting the tropism-positive versus tropism-negative proteins based on the translational adaptation across tissues. Given that the size of the tropism-positive and tropism-negative groups were often unbalanced, we iteratively sampled

equal-sized groups, for n=100 iterations. Furthermore, we validated the results with a stratified 5-fold cross-validation.

In order to evaluate the performance of the RF models, we computed the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) and Precision-Recall (PR) plots (**Fig. 4.2A**). We took the average and standard deviation across all iterations. Similarly, we computed the relative feature weights corresponding to each of the 23 TCGA tissues (**Fig. 4.2B**). In addition, we also validate the predictive potential of the model by performing a permutation test over 100 randomizations of the tropism labels of the dataset (Table S2).

For the dataset of Hydro-tRNAseq of human cells, we computed the average viral RtAI to each of the seven cell lines (Table S2). The RtAI is the supply-only version of SDA (see description above), since no codon demand information is available for this dataset. Using the same set of 2891 viral proteins across these seven RtAI features, we performed an identical RF classifier as above.

Linear Discriminant Analysis of tissue-specific SDAs

Similar to the RF classifier, we also computed the average proteome SDA per virus in each of the 23 tissues. We then applied a Linear Discriminant Analysis (LDA) to these averaged SDAs. We assigned each virus to its corresponding tropism (Table S1) in order to find the linear combination of tissue adaptation features that maximized differences between viral target tissues (Fig. S2, Table S2).

Translational adaptation of human coronaviruses

The SARS-CoV-2 coronavirus constitutes the etiologic agent of the biggest pandemic of the 21st century, causing the COVID-19 pneumonia-like disease. As our systematic analysis suggests that the codon usage of viruses tend to be adapted to the tissue they infect, we selected the novel coronavirus SARS-CoV-2 and other related respiratory viruses to further explore their translational adaptation profile over tissues. We initially reconstructed tRNA expression profiles

along the respiratory tract making use of the spatial information associated with healthy TCGA samples from head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) (Table S3). We then computed the SDA of viral proteins from the three pandemic coronaviruses of the last two decades SARS-CoV (205), MERS-CoV (206), and SARS-CoV-2 (193), as well as the common flu-causing influenza A virus (H1N1) along the respiratory tract (Fig. S4A). Apart from the clear viral tropism of SARS-CoV-2 to the respiratory tract, recent studies propose that their tropism can expand to other tissues such as the digestive system or the brain (183,187). For this reason, we also extended our translational analysis to all the 23 tissues of the TCGA dataset (Fig. S5A).

Moreover, given that the tropisms not only depend on the translational adaptation to the host, but also on the expression of the required entry receptors, we measured the respective receptors of each virus (Fig. S4B, Table S3). Influenza A viruses bind to α(2,6)-linked and α(2,3)-linked sialic acids, which are synthesized by the enzymes ST6GAL1 and ST3GAL4, respectively (207). The MERS-CoV uses the parenchyma-specific receptor DPP4 (208). On the other hand, both the SARS-CoV and SARS-CoV-2 strains bind to the ACE2 protein and require the proteolytic priming of the viral spike protein by TMPRSS2 (209), although the receptor BSG/CD147 has also been proposed (210).

In an attempt to elucidate the translational selection that could have benefitted the evolution of the new coronavirus, we also compared the SDA adaptation of SARS-CoV-2 to those of close phylogenetic strains (Fig. S5B): the human-infecting SARS-CoV and the bat coronavirus RatG13, with 79.6% and 96.2% of sequence identity, respectively (193).

Gene Set Enrichment Analysis (GSEA)

We analyzed the enrichment of gene sets of the Virus Orthologous Groups using the GSEA algorithm (144). The score used to generate the ranked list input is specified in the text. For the analysis, all gene sets with at least 10 members appearing in the ranked list were included.

4.5.5. Quantification and statistical analysis

All details of the statistical analyses can be found in the figure legends. For hypothesis testing, a Wilcoxon rank-sum test was performed. We used a significance value of 0.05.

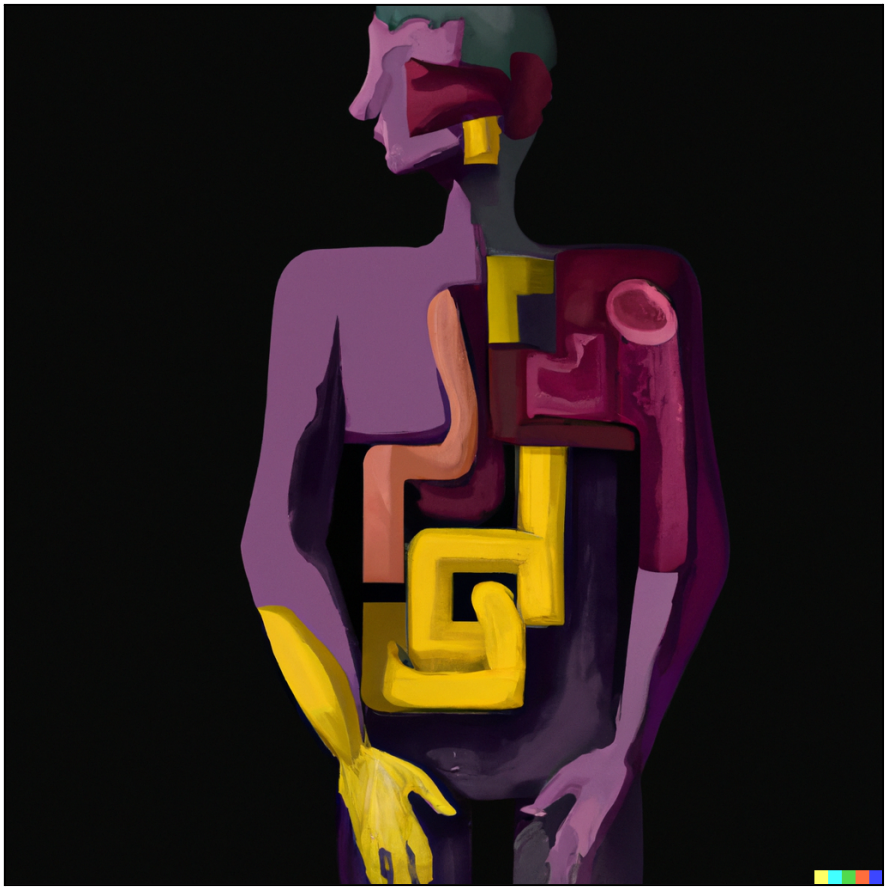
4.6. Acknowledgments

We thank Samuel Miravet-Verde for stimulating and critical discussions. We thank the CRG Genomics Unit for assistance with RNA sequencing services. The results published here are in part based on data generated by the TCGA Research Network: www.cancer.gov/tcga. We acknowledge the support of the Spanish Ministry of Science and Innovation (MICINN) (PGC2018-101271-B-I00 Plan Estatal), 'Centro de Excelencia Severo Ochoa', the CERCA Programme/Generalitat de Catalunya, and the Spanish Ministry of Science and Innovation (MICINN) to the EMBL partnership. The work of X.H. has been supported by a PhD fellowship from the Fundación Ramón Areces.

4.7. Author contributions

Conceptualization, X.H., M.H.S. and L.S.; Methodology, X.H., H.B., M.H.S., L.S.; Software, X.H.; Investigation, X.H., H.B.; Validation, X.H., H.B., M.H.S.; Formal analysis, X.H.; Writing-Original Draft, X.H.; Writing-Review & Editing, X.H., H.B., M.H.S., L.S.; Visualisation: X.H., H.B., M.H.S., L.S.; Funding Acquisition, L.S.; Supervision, M.H.S. and L.S.

Hernandez-Alias X, Benisty H, Serrano L, Schaefer MH. Using protein-per-mRNA differences among human tissues in codon optimization. bioRxiv. 2022;2022.03.22.485268.



Cubist painting of a human body with organs with a dark background

by DALL·E 2

Chapter 5

Using protein-per-mRNA differences among human tissues in codon optimization

5.1. Abstract

Codon usage and nucleotide composition of coding sequences have profound effects on protein expression. However, while it is recognized that different tissues have distinct tRNA profiles and codon usages in their transcriptomes, the effect of tissue-specific codon optimality on protein synthesis remains elusive. Here, we leverage existing state-of-the-art transcriptomics and proteomics datasets from the GTEx project and the Human Protein Atlas to compute the protein-to-mRNA ratios of 36 human tissues. Using this as a proxy of translational efficiency, we build a machine learning model that identifies codons enriched or depleted in specific tissues. In particular, we detect two clusters of tissues with an opposite pattern of codon preferences. We then use the identified patterns for the development of CUSTOM, a codon optimizer algorithm which suggests a synonymous codon design in order to optimize protein production in a tissue-specific manner. In a human cell model, we provide evidence that codon optimization should indeed take into account particularities of the translational machinery of the tissues in which the target proteins are expressed and that our approach can design genes with

tissue-optimized expression profiles. Altogether, CUSTOM could benefit biological and biotechnological research, such as the design of tissue-targeted therapies and vaccines.

5.1.1. Additional data access

All supplementary figures and data can be accessed from the original publication through this QR code and link.
Supplementary figures and tables will be referred to as "ED Figure" and "Sup. Table".

5.2. Introduction

From the advent of synthetic biology, it is widely recognized that gene design needs to be adapted to the expression requirements of the host (211). Within coding sequences, there are manifold overlapping factors that determine translation, mRNA stability, transcription, splicing, methylation, or ribosomal frameshifting, among others (29). Therefore, while the amino acid sequence of proteins is maintained, the usage of synonymous codons can be optimized for heterologous expression.

During the last decades, an extensive number of computational tools have been developed for gene design (36,212). Most commonly, these tools optimize the codon usage in order to resemble that of the host based on the Codon Adaptation Index (CAI) of the genes to be optimized or similar metrics. Other more innovative developments also include neural networks that control translation speed (213) or other machine learning algorithms that optimize mRNA stability (214). Although there is no absolute “best” approach, codon optimization is commonly and successfully applied in gene design. In fact, current knowledge on the effect of synonymous variants on the heterologous expression of the protein GFP shows up to 46-fold expression differences in HeLa cells (27). Similarly, mRNA and protein levels

across thousands of GFP variants strongly correlated with their CAI in *S. cerevisiae* (101).

Nevertheless, codon optimization in multicellular eukaryotes is more intricately determined, since different tissues can showcase differences in codon usage and tRNA expression (54,103,150). The translational efficiency, which constitutes the rate of protein production from mRNA, is therefore dependent on the balance between the codon usage of genes being translated and the abundance of a limited tRNA pool (96,150). In this context, codons translated by highly abundant tRNAs generally correspond to optimal codons in the translatome, as has been reported by ribosome profiling (95). However, detecting differences of translational efficiency between tissues can be challenging, since the larger gene-to-gene variability of protein levels can obscure the actual tissue-to-tissue differences (3).

The advent of high-throughput sequencing has enabled an extensive transcriptome profiling of human tissues (215,216). Based on the mRNA-seq data from the GTEx project, Kames et al. (2020) developed the public resource TissueCoCoPUTs, containing codon and codon pair usage tables of tissue transcriptomes (54). However, current knowledge indicates that tissue-specific variability of gene expression is mostly regulated at the post-transcriptional level and mRNA-seq alone is therefore not able to capture it (21,28). Developments in mass spectrometry have very recently led to the release of deep and quantitative proteome maps of human tissues (217,218).

Using this transcriptomic and proteomic data from the Human Protein Atlas and the GTEx project, we here compute the protein-to-mRNA (PTR) ratios of 36 human tissues as a proxy for translational efficiency. To distinguish high-PTR from low-PTR proteins, we build random forest models that identify which codons are optimal or non-optimal for each tissue. Then we apply these codon preferences to develop a tool, CUSTOM, that optimizes coding sequences for a specific tissue. CUSTOM is publicly available as a Python package (github.com/hexavier/CUSTOM) and as a web interface

(custom.crg.eu). By optimizing eGFP and mCherry proteins to a human cell model of kidney and lung, we provide experimental evidence of how tissue codon optimization could be important e.g. in vaccines or gene therapy.

5.3. Results

5.3.1. Protein-to-mRNA ratios detect differences in translational efficiency among tissues

Translational efficiency (TE) is defined as the rate of protein synthesis from mRNAs, which can be estimated as the protein-to-mRNA (PTR) ratio. To systematically analyze the PTR ratios across a total of 36 human tissues, we retrieved the mRNA-seq and proteomics data from two recent datasets: 29 tissues from the Human Protein Atlas (28,218) (HPA) and 24 tissues from the GTEx project (217) (**Fig. 5.1A-B**, Sup. Table 1). The first study includes one sample per tissue, which are concurrently analyzed by mRNA-seq and label-free iBAQ proteomics. On the latter, a total of 182 matched samples are measured both by mRNA-seq and tandem mass tag 10plex/MS3 mass spectrometry. By correlating the mRNA expression, protein abundance and PTR ratios along the 17 tissues in common, we could ascertain a high correspondence between the two datasets (ED Fig. 1A).

Although to date this data is still relatively rare, a more direct readout of TE is the ratio between ribosome profiling and mRNA abundance. To confirm the validity of using PTR ratios as an estimate of TE, we therefore compared the PTR values to a ribosome profiling dataset of brain, liver, and testis. In all of them we observe a significantly positive correlation across the human genome (22) (**Fig. 5.1C**, Sup. Table 1).

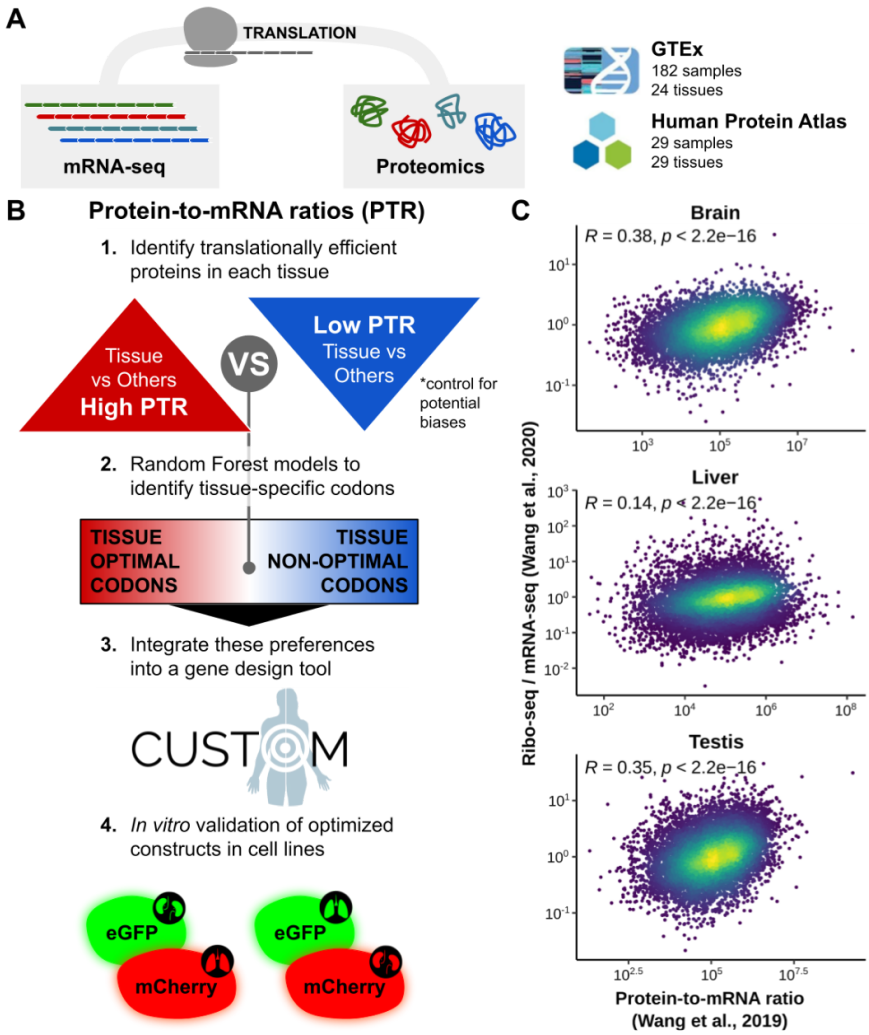


Figure 5.1. Protein-to-mRNA ratios detect differences in translational efficiency among tissues.

(A) Proteomics and mRNA-seq data included in this study contains samples from the GTEx project (217) and Human Protein Atlas (218). (B) Using these datasets, we compute the protein-to-mRNA ratios (PTR) and define tissue-enriched and tissue-depleted sets of proteins for each tissue. By comparing the codon usage of these two sets, we identify the codon optimality pattern of tissues. Using this information, we develop a gene design tool called CUSTOM and validate the method using an *in vitro* cellular model. (C) Spearman correlation between the median translational efficiency (22) (ratio between ribo-seq and mRNA-seq FPKMs) and PTR (218) across genes in brain, liver, and testis. The color code depicts the density of points in the scatter plot.

We next set out to investigate the tissue-to-tissue differences of PTR ratios in the aforementioned datasets. For each tissue, we defined a set of high-PTR and a set of low-PTR genes, described as having a PTR fold change compared to the average of all other tissues larger than 2, and vice versa (Sup. Table 1). We find a significant concordance between the gene sets derived from the HPA and GTEx datasets in most tissues ($p < 0.05$, one-tailed binomial test, Sup. Table 1).

To physiologically interpret the differences between gene sets, we performed an enrichment map among high-PTR and low-PTR sets linking tissues with high overlap of the respective gene sets (ED Fig. 1B). In agreement with their highly tissue-specific function, we detect that tissues group according to their role in the body: eg. nervous tissue (brain and tibial nerve), muscular tissue (skeletal muscle and heart). Moreover, GO analyses of high-PTR genes show significant enrichments for highly tissue-specific biological processes according to the physiological and anatomical function of the tissue ($p <0.05$, Fisher's exact test, ED Fig. 1C).

We next asked if there could be any confounding factors associated with these gene sets, such as protein secretion and degradation, that could bias our analyses. On the one hand, it has been recently reported that constitutively secreted proteins are often detected at the mRNA but not at the protein level (217), which could bias PTR ratios as a measure of TE. While we also observe these differences in our dataset (ED Fig. 2A), the exclusion of secreted proteins from our gene sets does not affect the downstream results (see following section). On the other hand, we analyzed the protein half-life of gene sets based on two recent datasets in five human cell lines (219,220) (Sup. Table 1). The protein half-life is not significantly different between high-PTR and low-PTR gene sets in most of the tissues ($p < 0.05$, two-tailed Wilcoxon rank-sum test), nor is there any trend that one of the groups would be consistently associated with higher or lower half-life (ED Fig. 2B).

Taken together, these observations indicate that PTR ratios can efficiently detect tissue-specific differences in translation. As such, it

constitutes an appropriate dataset to systematically study TE differences across the set of 36 human tissues.

5.3.2. Random Forest models identify two clusters of human tissues with distinct codon signatures

Recent studies show that different tissues can have different tRNA repertoires and codon usage (54,150), which could have an influence on translational efficiency. Therefore, we wondered whether high-PTR and low-PTR sets of genes were specifically enriched or depleted of certain codons. If there is a tissue-specific codon signature, we would expect to be able to predict these differences in PTR.

To that aim, we built a random forest classifier for each tissue that predicts the high-PTR vs low-PTR state of genes based on their codon usage. All 36 resulting models perform with an area under the curve (AUC) of their receiver operating characteristic (ROC) curves higher than the no-skill model of 0.5 (**Fig. 5.2A**, Sup. Table 2). In particular, kidney, breast, lung, rectum and tonsil showcase the highest tissue-specific profiles (**Fig. 5.2A**; all AUC > 0.70). Furthermore, to validate whether these differences in PTR are specifically dependent on codon usage and not from nucleotide composition alone, we compared them with the performance of three control models: +1 and +2 misframed codon usage as well as dinucleotide composition of genes (Sup. Table 2). While these control models also show predictive power, the AUC of the correctly framed codon usage models significantly outperform the controls ($p < 0.05$, one-tailed binomial test).

To examine the tissue-specificity of codons, we next analyzed which particular codons are predictive for high vs low PTR states in each tissue. The relative feature importances of each random forest classifier measure the contribution of codons in the decision trees (ED Fig. 3A). In general, only a few codons (5 to 10) are relevant for each model, but they differ across tissues. A recursive feature elimination of each model similarly substantiates that fewer than 10 codons are sufficient to achieve the maximum AUC performance (ED Fig. 3B).

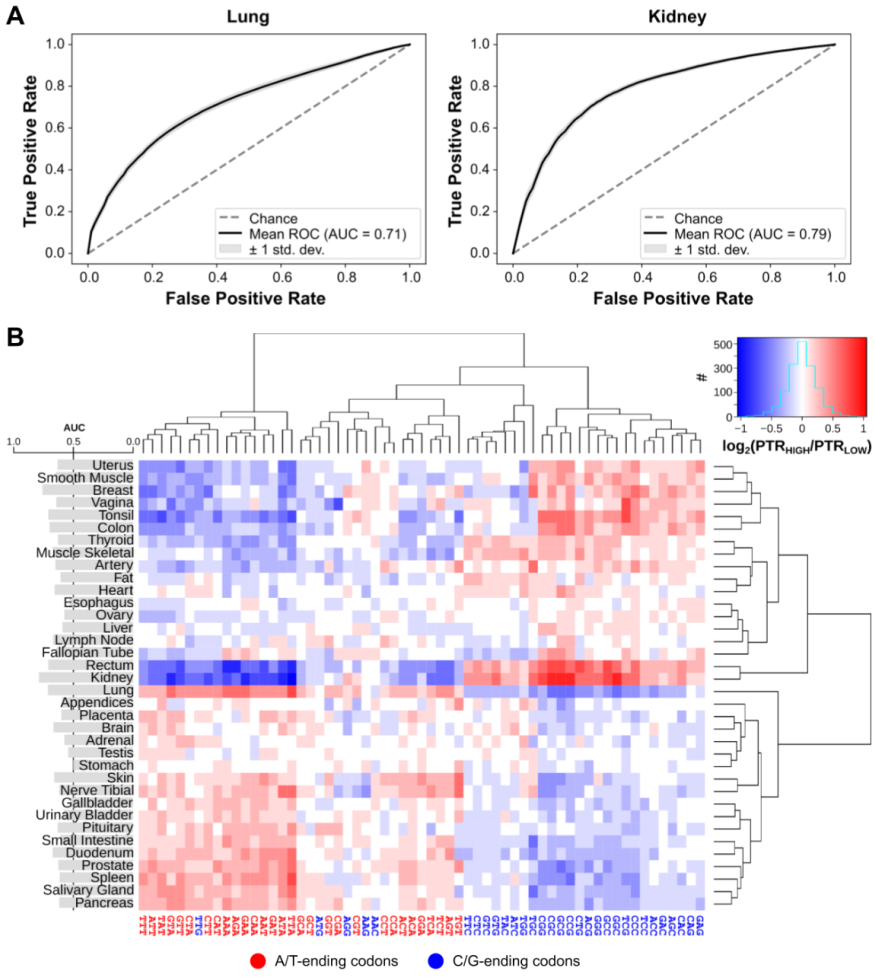


Figure 5.2. Random Forest models identify two clusters of human tissues with distinct codon signatures.

(A) Receiver operating characteristic (ROC) curves of lung and kidney random forest classifiers, in which the codon usage of genes is used to predict whether they are high-PTR or low-PTR in the respective tissue (see Methods). (B) Ratios of the codon usage between high-PTR and low-PTR genes in each tissue. Codons and tissues are hierarchically clustered using euclidean distances and the complete-linkage method. The barplot on the left shows the mean AUC of the ROC curve of the RF model of each tissue.

In addition, by computing the ratio between the codon usage of high-PTR vs low-PTR genes, we observe the enrichment or depletion of codons in specific tissues (**Fig. 5.2B**). There are two main clusters of tissues with opposite codon optimality profiles: the first generally preferring A/T-ending codons while the second favoring C/G-ending ones. Also, as expected, tissues with higher AUC performances showcase more definite codon profile patterns both in terms of their enrichment/depletion (**Fig. 5.2B**) as well as their importance (ED Fig. 3A). As mentioned in the previous section, we also repeated the same analyses with the secretome-excluded sets of genes, which have a highly similar codon optimality profile with all correlations of codon ratios over 0.95 (Sup. Table 2).

Given that some reports highlight the role of codon pair bias in translation (54,221), we similarly analyzed the codon pair usage ratios between high-PTR vs low-PTR genes (Sup. Table 3). A principal component analysis (PCA) of these ratios perfectly separates the exact same two clusters observed above with single codons alone (ED Fig. 4A). To further analyze how much codon pair variance is explained by single codons alone, we compared observed codon pair ratios with their expected values based on their constituent single codons. They relate highly linearly as shown by linear regression models (ED Fig. 4B, Sup. Table 3), which indicates that differences in codon pair ratios can be explained by single codons alone. In fact, codon pairs that deviate the most from linearity just correspond to outliers with very low counts within gene sets (ED Fig. 4C).

Overall, our random forest classifiers can predict the PTR of genes in a certain tissue based on their codon usage. As such, the observed differences in codon preference or avoidance across tissues can be exploited to optimize tissue-specific gene design.

5.3.3. CUSTOM generates fluorescent variants with desired tissue-specific expression

To translate differences in tissue-specific PTR into a codon optimizer tool, we developed CUSTOM as a probabilistic approach (see Methods, custom.crg.eu). Given a certain amino acid sequence and a target tissue, codons are selected with a probability proportional to their tissue importance in the model (ED Fig. 3A). Then, based on the ratio of the selected codon (**Fig. 5.2B**), it is either added or avoided in the generated sequence. This process is performed along the whole sequence, and repeated iteratively to generate a pool of hundreds of optimized sequences. Among this pool of sequences, given that tissue-specific codon usage is not the only factor influencing coding sequences (29), the top scoring ones can be selected based on other commonly used parameters of codon bias or mRNA stability (36) (Codon Adaptation Index, Codon Pair Bias, Minimum Free Energy, Effective Number of Codons, see Methods).

To validate the predictor, we chose the proteins eGFP and mCherry, and optimized them with CUSTOM to either kidney or lung (Sup. Table 4). Taking eight among the top optimized sequences (**Fig. 5.3A**, 2x $eGFP_{Kidney}$, 2x $eGFP_{Lung}$, 2x $mCherry_{Kidney}$, 2x $mCherry_{Lung}$), we then designed four constructs, placing in each of them one eGFP and one mCherry optimized each one for a different tissue and under an inducible bidirectional promoter (**Fig. 5.3B**). These constructs were then simultaneously expressed in the lung and kidney cell lines A549 and HEK293T, respectively. Based on available proteomics data of these cell lines (222), the proteome of A549 clearly resembles that of lung, while HEK293 is a closer model to kidney (ED Fig. 5A).

We then analyzed the eGFP and mCherry fluorescence of each construct in each cell line. For all cases, we observe that the eGFP/mCherry ratio is significantly higher in the tissue for which eGFP is optimized (**Fig. 5.3C**, $p < 0.05$, two-tailed Wilcoxon rank-sum test, ED Fig. 5B), which validates our tissue-specificity hypothesis. We further observe that (1) the two constructs with $eGFP_{Lung}$ have generally lower eGFP/mCherry

ratios compared to the ones with $eGFP_{Kidney}$, and (2) the differences in eGFP/mCherry ratios between constructs are more variable in HEK293 than A549 cells. Altogether, these observations suggest that A/T-ending codons are generally lower expressed than C/G-ending counterparts, but tissues like lung tolerate them better.

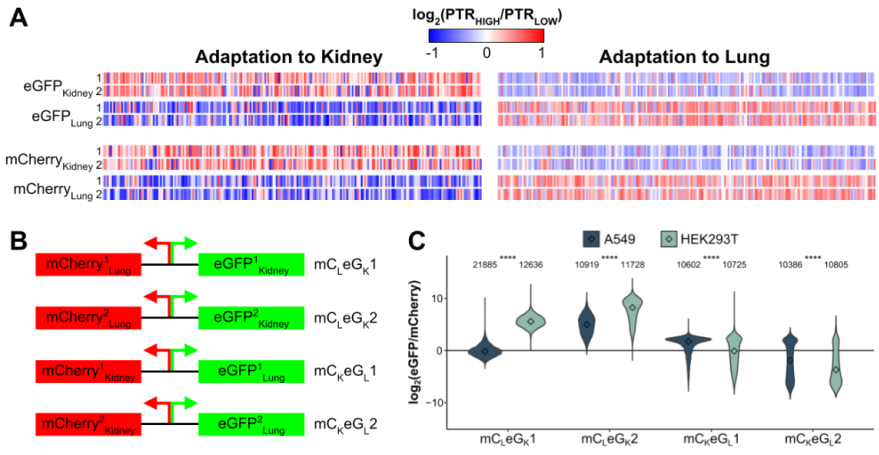


Figure 5.3. CUSTOM generates fluorescent variants with desired tissue-specific expression.

(A) Selected eGFP and mCherry sequences optimized to lung and kidney using CUSTOM. The color code corresponds to the optimality ratios of Fig. 5.2B. (B) Using these sequences, we designed four of constructs by placing a mCherry and an eGFP with opposite tissue-specificity under an inducible bidirectional promoter. (C) Ratios of eGFP and mCherry for each of the four constructs detected by flow cytometry. The number of cells within each group is specified. Center values represent the median. Statistical differences were determined by two-tailed Wilcoxon rank-sum test, and are denoted as follows: *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001, ****p ≤ 0.0001. Two additional replicates are shown in ED Fig. 5A.

5.4. Discussion

Current analyses of the mRNA and protein levels among human tissues distinguish between across-gene and within-gene (i.e. across-tissue) variability (3). In fact, the coefficient of variation of mRNA and protein levels across genes highly exceeds that of across tissues. In consequence, studies of codon usage on human transcriptomes and PTR ratios so far were dominated by the across-gene variability, and thus overlooked the smaller across-tissue differences (28,54). The approach taken here puts the focus on the across-tissue variability of PTR ratios rather than the overall genome, which is actually the major source of post-transcriptional regulation (21,28). In fact, we provide evidence that high-PTR gene sets of tissues are particularly enriched for tissue-specific functions.

Given the high GC content of the human genome as a whole, G/C-ending codons are generally more abundant (i.e. higher CAI), and relate to higher mRNA and protein expression levels (27,223,224). But again, moving away from this across-gene perspective of human codon usage to look at the across-tissue variation, we here report that distinct tissues showcase different codon preferences. All in all, as also determined experimentally, we observe that the expression of a certain protein is dependent on two axes: (1) the across-gene axis with G/C-ending codons favoring higher absolute expression, and (2) the tissue-specific axis with the codon preferences observed in Fig. 5.2B. Moreover, we also report that some tissues have a more definite codon profile than others, where this second axis is less evident. In agreement with our observed tissue-specific axis, Allen et al. (2022) recently reported that testis and brain (in contrast to other tissues such as ovary) better tolerate the translation of rare A/T-ending codons in *Drosophila melanogaster* (225).

The codon optimization tool CUSTOM is able to exploit these codon preferences for the design of tissue-targeted genes. In fact, all four designed constructs expressed in a kidney and lung cell line showed the predicted tissue-specificity. To make CUSTOM readily available to the

community, we developed it completely open source and made it accessible through a web server.

Human tissues are ensembles of heterogeneous cell types, and therefore observed differences in codon optimality are actually a composition of the constituent cell types. However, single-cell technologies of mRNA and protein measurements fall still far from complete cellular atlases (226). Instead, we used the most up-to-date and complete tissue-wide maps of the human transcriptome and proteome, which have been generated by cutting-edge mass spectrometry and mRNA sequencing techniques (215,217,218).

Finally, the results presented here constitute a proof-of-concept that tissue-specific codon usage exists and can be applied to gene design. In particular, this tool could be used in the development of optimized gene therapies or mRNA vaccines with more targeted tissue targets and therefore potentially less side effects. Nevertheless, factors other than codon usage also play a role in gene expression (29), and therefore changes in synonymous codons can as well interfere with other processes such as mRNA folding and stability, mRNA modifications, protein folding, or translational kinetics (227,228). As such, tissue-specific codon usage will constitute one additional instrument in the gene design tool set.

5.5. Methods

5.5.1. Codon optimizer for tissue-specific expression

CUSTOM is implemented in Python (version >= 3.7) and available on GitHub (github.com/hexavier/CUSTOM) and as a web interface (custom.crg.eu). The landscape of possible synonymous sequences is vast and manifold factors overlap in defining the code. Therefore, we follow a simple probabilistic approach with two steps: (1) translate tissue-specific codon preferences into a pool of optimal sequences, and (2) select the desired sequence based on other parameters of relevance.

5.5.1.1. Create a pool of tissue-optimized sequences

The algorithm requires two main input data: the amino acid sequence to be optimized (or DNA sequence) and the target tissue. For each iteration of the optimization, the sequence is optimized taking two factors into account: how important the codon is in defining tissue-specificity (relative feature weights in ED Fig. 3A) and whether it is enriched or depleted in the tissue (codon ratios in Fig. 5.2B). Therefore, for each amino acid, a certain codon is selected with a probability proportional to the first. If the selected codon is enriched in the tissue, it is incorporated into the sequence. If it is depleted, the codon is excluded and another codon is selected based on the same probabilities as before. This process is repeated along the full sequence, and for as many iterations as desired. Furthermore, given that 5-10 top codons are often sufficient to achieve the full AUC prediction (ED Fig. 3B), users can also control whether optimizing all codons or only the top ones.

5.5.1.2. Selecting the top scoring candidates

Once a pool of optimized sequences has been generated, the best-ranked ones can be selected as the user desires. Given that no ground truth is known, the default *select_best* method of the package measures a list of standard metrics frequently used in gene design and computes an average to select the top scoring sequences. The following factors can be included:

- Minimum Free Energy (MFE): a measure of mRNA stability from the ViennaRNA package (203). CUSTOM distinguishes between the first 40 nucleotides (whose weak secondary structure leads to increased translation initiation) and the rest of the sequence (whose strong secondary structure relates to longer mRNA half-lives) (36).
- Codon Adaptation Index (CAI): a measure of similarity between the codon usage of the sequence and that of the human genome (229).
- Codon Pair Bias (CPB): a measure of similarity between the codon pair usage of the sequence and that of the human genome (221).

- Effective Number of Codons (ENC): a measure of codon evenness. A value of 20 means that all 100% codons are biased towards the most common codon, while 61 corresponds to no bias at all (230).
- GC content: a measure of similarity between the sequence GC content and a desired target value of GC.
- Homopolymers: filters out sequences with homopolymers of a certain length, which can lead to worse expression.
- Motifs: filters out sequences containing certain motifs.

5.5.2. Experimental model and protocol

5.5.2.1. Human cell models

The cell lines included in this study are HEK293T and A549. The sex of each cell line is as follows: HEK293T, female; A549, male. Cells were maintained at 37°C in a humidified atmosphere at 5% CO2 in DMEM 4.5 g/l Glucose with UltraGlutamine media supplemented with 10% of FBS and 1% penicillin/streptomycin.

5.5.2.2. Expression vectors design

We applied CUSTOM to the protein sequences of eGFP and mCherry (Uniprot ID: C5MKY7, X5DSL3). Sequences were optimized to either lung or kidney, generating a total of *n_pool* = 1000. Sequences with homopolymers equal or larger than 7 were filtered out and scored with:

opt.select_best(by={"MFE":"min", "MFEini":"max", "CAI":"max", "CPB":"max", "ENC":"min"}, homopolymers=7, top=10)

Among the top 10 scoring candidates of each optimization, we selected 2x $eGFP_{Kidney}$, 2x $eGFP_{Lung}$, 2x $mCherry_{Kidney}$, and 2x $mCherry_{Lung}$ (Sup. Table 4).

For gene overexpression experiments, the two selected eGFP and and mCherry were cloned into a modified version of the XLone-GFP vector (Addgene#96930). The modification consisted of replacing the promoter of XLone-GFP with a bidirectional TRE3G promoter

(Clontech), which allows the simultaneous expression of both genes. The four constructs consisted in a combination of $eGFP_{Lung}$ + $mCherry_{Kidney}$ and $eGFP_{Kidney}$ + $mCherry_{Lung}$.

5.5.2.3. Flow cytometry

HEK293T and A549 cells were seeded in 6-well plates. Gene expression was induced with 500 ng/mL of doxycycline during 48h. To measure the expression of the fluorescent proteins, cells were trypsinized and resuspended with 500 µL of media. Samples were applied on a FACS Fortessa analyser. Approximately 10^4 live single-cell events were collected per sample. BD FACSDiva software was used for gating and analysis. The fluorescence intensity for each population in the FITC channel and PE–Texas Red channel was obtained.

5.5.3. Data sources

5.5.3.1. Protein-to-mRNA ratios

The PTR ratios of the HPA were directly retrieved from the Table EV3 of Eraslan et al. (2019). In this dataset, protein levels are determined as absolute abundances based on their iBAQ quantification. As for the GTEx data, we retrieved protein and mRNA levels from Table S2 and Table S3 of Jiang et al. (2020), respectively. In this case, the proteomics measurements are relative quantifications from a tandem mass tag (TMT) 10plex/MS3 mass spectrometry strategy. To compute their PTR ratios, we followed the same pipeline as in the HPA: (1) proteins with an abundance of 0 were considered as missing values (NA); (2) protein quantifications were adjusted to have in each tissue the same median than the overall median; (3) genes with a TPM lower than 10 were taken as non-transcribed (NA). With that, comparable PTR values between HPA and GTEx are obtained (ED Fig. 1A).

5.5.3.2. Codon and codon pair usage tables

The codon usage and codon pair usage tables of *Homo sapiens* from RefSeq were downloaded from the Codon/Codon Pair Usage Tables (CoCoPUTs) project release as of June 9th, 2020 (30). Regarding the

codon usage of misframed coding sequences and their dinucleotide composition, we computed them from the latest release of the CCDS database of human sequences (release 22) (231).

5.5.3.3. Translational efficiencies

The processed data of matched ribosome profiling and mRNA-seq samples from brain, liver and testis was retrieved from ArrayExpress (E-MTAB-7247) (22). Translational efficiencies were then computed as the ratio $FPKM_{Ribo\text{-}seq}/FPKM_{mRNA\text{-}seq}$.

5.5.3.4. Protein half-life

The log-10-transformed protein half-lives for B cells, NK cells, hepatocytes, monocytes, and HeLa cells were downloaded from Eraslan et al. (2019) (219,220). Given the concordance of half-lives among the five cell types (Sup. Table 1), we used their average for the analysis in this work (ED Fig. 2B).

5.5.3.5. Blood secretome

Using the predictions by the HPA (216), there are 2641 secretome genes, 729 of which are secreted to blood. Given that we were concerned on proteins that are not detected at the protein levels because of their systemic rather than local secretion, we focused our analysis on the latter (Sup. Table 1).

5.5.4. Computational analysis

5.5.4.1. High-PTR and low-PTR gene sets

As PTR values from GTEx were computed from relative TMT proteomics in contrast to the absolute iBAQ quantification of HPA, they were not directly comparable and thus we defined the high-PTR and low-PTR gene sets for each dataset separately. On the one hand, high-PTR genes fulfilled three conditions: (1) genes having a PTR fold change compared to the average of all other tissues larger than 2, (2) genes with the highest PTR among all tissues, (3) genes detected in at

least 3 tissues in the dataset. On the other hand, low-PTR genes were defined as: (1) genes having a PTR fold change compared to the average of all other tissues smaller than 0.5, (2) genes with the lowest PTR among all tissues, (3) genes detected in at least 3 tissues in the dataset. As a result, we defined one high-PTR and one low-PTR gene set for each tissue in each dataset. For those 17 tissues in common between both HPA and GTEx datasets, the union between both datasets was taken except for genes with contradictory labels, which were excluded.

5.5.4.2. Random Forest classifiers

To identify the most important codons determining high-PTR vs low-PTR genes, we computed their codon usage normalized by length, so that all 61 amino-acid-encoding codons sum up to 1. Taking this table of normalized codon usage as features, we applied a Random Forest (RF) classifier, populated with 100 decision trees, using the scikit-learn package (194). Therefore, for each of the 36 tissues, we developed a model for predicting the high-PTR vs low-PTR genes based on their codon usage. To control for size differences between high-PTR and low-PTR groups, we iteratively sampled equal-sized groups, for n = 100 iterations. Furthermore, we validated the results with a stratified 5-fold cross-validation. In order to evaluate the performance of the RF models, we computed the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) plots (Fig. 5.2A). We took the average and standard deviation across all iterations. Similarly, we computed the relative feature weights corresponding to each of the 61 codons (Fig. 5.2B).

To validate that the predictive potential of RF classifiers were codon-specific, we similarly computed the length-normalized codon usage of +1 and +2 misframed coding sequences as well as dinucleotide usage. By running the exact same pipeline as above, we determined the average AUC of these three control RF classifiers (Sup. Table 2). We used a one-tailed binomial test to analyze whether the AUCs of controls were lower than the original model more often than expected by chance ($p = 1/2$).

While the relative feature weights determine the importance of each codon in distinguishing high-PTR vs low-PTR genes, they do not provide any directionality. To analyze whether codons are enriched or depleted in high-PTR vs low-PTR genes, we computed the ratios between the average length-normalized codon usage of high-PTR and low-PTR genes. Similarly, codon pair ratios were computed in the same way.

Among the total of amino-acid-encoding 61 codons, we also analyzed how many of them were actually informative in the models using a Recursive Feature Elimination (RFE). Therefore, for each tissue, we started by building a full model with all 61 codons and then recursively removed the least important one, as determined by the relative feature weights, until only one was left. At each step, we computed the AUC of the ROC curve of the model as explained above (ED Fig. 3B).

5.5.4.3. Enrichment map

For this analysis, in order to allow an overlap between tissue gene sets, we used a slightly less stringent tissue-specificity definition. High-PTR sets were defined as (1) genes having a PTR fold change compared to the average of all other tissues larger than 2 and (2) genes detected in at least 3 tissues in the dataset, and vice versa for low-PTR sets.

To analyze the overlap between tissue gene sets, we used the EnrichmentMap app from Cytoscape (232). We defined a generic input of high-PTR and low-PTR sets of proteins per tissue. Similarity was computed as the overlap coefficient ([size of (A intersect B)] / [size of (minimum(A ,B))]).

5.5.4.4. Gene Ontology enrichment analysis

Gene Ontology (GO) categories of Biological Processes were analyzed for enrichment as of May 27th, 2021 (233). Enrichment analyses were performed by PANTHER using the Fisher's exact test and Bonferroni correction for multiple testing (234).

5.5.4.5. Principal Component Analysis of codon pairs

We applied Principal Component Analysis to the codon pair ratios of each tissue in order to explore the main variability among tissues along the 4096 codon pair ratios.

5.5.4.6. Linear regression of codon pairs

We fitted a linear regression model between the observed codon ratios (dependent variable) and the expected ratios based on single codons alone (independent variable). The expected values were computed as the product of the ratios of the two codons that constitute the pair. For each model, we computed the R squared, the Residual Standard Error (RSE), and the model p-value (Sup. Table 3).

5.5.4.7. Statistical analysis

All details of the statistical analyses can be found in the Results section and the figure legends. We used a significance value of 0.05.

5.6. Acknowledgments

We acknowledge the support of the Spanish Ministry of Science and Innovation (MICINN) (PGC2018-101271-B-I00 Plan Estatal), 'Centro de Excelencia Severo Ochoa', the CERCA Programme/Generalitat de Catalunya. The work of X.H. has been supported by a PhD fellowship from the Fundación Ramón Areces.

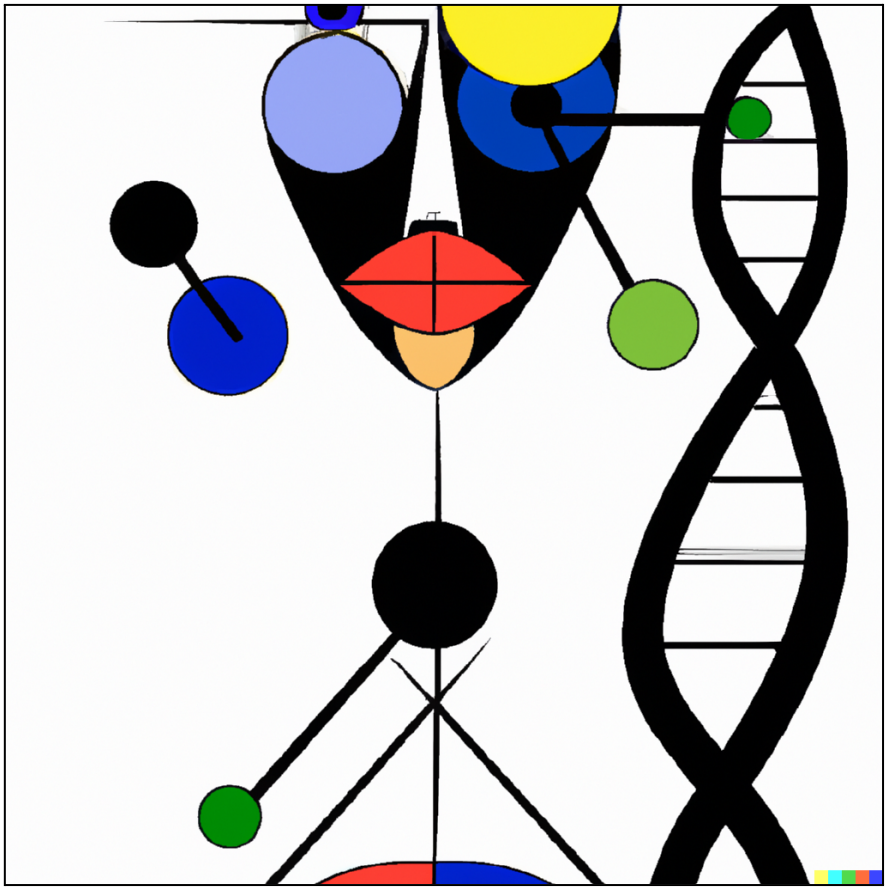
5.7. Author contributions

Conceptualization, X.H., H.B., M.H.S. and L.S.; Methodology, X.H., H.B., M.H.S., L.S.; Software, X.H.; Investigation, X.H., H.B.; Validation, X.H., H.B., M.H.S.; Formal analysis, X.H., H.B.; Writing-Original Draft, X.H., H.B.; Writing-Review & Editing, X.H., H.B., M.H.S., L.S.; Visualisation: X.H., H.B., M.H.S., L.S.; Funding Acquisition, L.S.; Supervision, M.H.S. and L.S.

5.8. Data and Code Availability

The code used in this study is available at GitHub (github.com/hexavier/codon_optimization), and the CUSTOM software is available as a Python package (github.com/hexavier/CUSTOM) and a web interface (custom.crg.eu). The published article includes all datasets generated or analyzed during this study.

Hernandez-Alias X, Katanski CD, Zhang W, Assari M, Watkins CP, Schaefer MH, Serrano L, Pan T. Single-molecule tRNA-seq analysis reveals coordination of tRNA modification and charging and fragmentation. Nucleic Acids Research. 2022; in press.



Nucleotide in Picasso style

by DALL·E 2

Chapter 6

Single-read tRNA-seq analysis reveals coordination of tRNA modification and aminoacylation and fragmentation

6.1. Abstract

Transfer RNA (tRNA) utilizes multiple properties of abundance, modification, and aminoacylation in translational regulation. These properties were typically studied one-by-one; however, recent advance in high throughput tRNA sequencing enables their simultaneous assessment in the same sequencing data. How these properties are coordinated at the transcriptome level is an open question. Here, we develop a single-read tRNA analysis pipeline that takes advantage of the pseudo single-molecule nature of tRNA sequencing in NGS libraries. tRNAs are short enough that a single NGS read can represent one tRNA molecule, and can simultaneously report on the status of multiple modifications, aminoacylation, and fragmentation of each molecule. We find correlations among modification-modification, modification-aminoacylation, and modification-fragmentation. We identify interdependencies among one of the most common tRNA modifications, m^1A58, as coordinators of tissue-specific gene expression. Our method, **S**ing**L**e-read **A**nalysis of **C**rosstalks (SLAC), reveals tRNAome-wide networks of modifications, aminoacylation, and fragmentation. We observe changes of these networks under different

stresses, and assign a function for tRNA modification in translational regulation and fragment biogenesis. SLAC leverages the richness of the tRNA-seq data and provides new insights on the coordination of tRNA properties.

6.1.1. Additional data access

All supplementary figures and data can be accessed from the original publication through this QR code and link.
Supplementary figures and tables will be referred to as "Figure S" and "Table S".

6.2. Introduction

Transfer RNAs are highly abundant non-coding RNAs of 65-94 nucleotides, which form a rigid secondary structure. Up to 20% of all residues in an eukaryotic tRNA are modified (84). tRNAs are aminoacylated (charged) with an amino acid at the 3' end. Charged tRNAs are required for protein synthesis by recognizing their cognate codons in the elongating ribosomes (81). The fine-tuned regulation of tRNA abundance, modification and charging determines the structure and function of tRNA in translation, and their alterations can lead to human diseases including neurological disorders and cancer (84,105,106).

All three properties of the cellular tRNAome are highly dynamic and respond to development and environmental cues (14,82). tRNA abundance is tissue-specific in humans and reprogrammed in cancers that adapt to the codon usage changes of the tissues and cancer phenotypes (106,150). tRNA charging levels are sensitive to stress conditions; its response to amino acid starvation regulates selective translation of stress-response genes and tunes global translation activity through phosphorylation of eIF2α by the GCN2 kinase (99,105). Dynamic response of tRNA modifications occurs at many modification

types and sites which fine-tune translation of stress response genes in a codon-dependent manner, as well as the biogenesis of tRNA fragments that are functional small RNAs in the regulation of mRNA stability and cell-cell communication (81,84).

The advent of high throughput tRNA sequencing (67,110,133,235) allows for the examination of tRNA abundance, modification and charging in the same sequencing library. NGS sequencing requires cDNA synthesis of the RNA by reverse transcription. Certain tRNA modifications that perturb the Watson-Crick base pairing such as N1-methyladenosine (m^1A) do not always stop reverse transcription, rather, some reverse transcriptases can readthrough these modifications while leaving behind a "mutation" signature relative to the tRNA reference sequence in data analysis (67,236,237). All mature tRNAs end with 3'CCA. The aminoacylated tRNA levels can be measured upon incorporating a periodate oxidation and beta-elimination step in the sequencing library preparation; periodate only oxidizes uncharged tRNAs and the oxidized 3' A residue is removed upon beta-elimination. Therefore, uncharged tRNA ends with 3'CC and charged tRNA ends with 3'CCA, and the aminoacylation levels for each tRNA can be measured by the ratio of their 3'CCA and 3'CC ending reads (67,133).

A human tRNA contains on average 13 modifications per molecule, how these modifications are linked to each other in their installation and function is currently under intense investigation (238). For example, the installation of multiple anticodon loop modifications in yeast $tRNA^{Phe}$ exhibits a specific order (239). Q and m^5C modifications in *S. pombe* $tRNA^{Asp}$ depend on one another (240). Such studies so far have mostly focused on the multiple anticodon loop modifications that are all located in the same hairpin loop (241–243). How modifications that are distal in the tRNA secondary structure are dependent on each other remains an open question. In addition, very little is known about the potential crosstalks between tRNA modification-modification and tRNA charging-modification (244).

Here, we implement SLAC (**S**ing**L**e-read **A**nalysis of **C**rosstalks), a computational pipeline to investigate the crosstalk between tRNA modifications, aminoacylation and fragmentation transcriptome-wide. SLAC is a single-read analysis pipeline of tRNA-sequencing data that takes advantage of recent advances in read-length, charging, modification, and fragment detection in tRNA. We show that correlating modification-induced mutation signatures with aminoacylation measurements at the single-read level confirms known crosstalks in yeast, and identifies new crosstalks in human tRNAs. We identify tissue-specific crosstalk signatures in mice. We observe stress-response changes in modification and charging that are consistent with these crosstalks. We show that tRNA modification and fragmentation are associated in a tRNA-type and cleavage-site dependent manner. Our results support the notion that tRNA modification and charging crosstalks may play a previously under-appreciated role in translational regulation.

6.3. Materials and Methods

6.3.1. Data sources

6.3.1.1. tRNA-seq datasets

The raw FASTQ files of mim-tRNA-seq of *S. cerevisiae* and of HEK293T cells were retrieved from Gene Expression Omnibus (GEO): GSE152621 (67). In this dataset, all HEK293T samples, three WT yeast samples ("WT" in Table S2) and the *Trm7Δ* mutant are periodate oxidized, and therefore both aminoacylation and modification levels are detectable. However, three WT yeast samples ("WTox0" in Table S2), and *Trm1Δ* and *Trm10Δ* mutants are not periodate oxidized and only modification levels are detectable.

The raw FASTQ files of QuantM-tRNA-seq of mouse tissues and of HEK293T cells were retrieved from GEO: GSE141436 (237). The raw FASTQ files of Charged DM-tRNA-seq of HEK293T cells were retrieved from GEO: GSE97259 (133). The raw FASTQ files of total

and polysome MSR-seq data of HEK293T in control and stress were retrieved from GEO: GSE198441 (235).

6.3.1.2. Coding sequences and expression

The mRNA-seq gene expression data of total and polysome HEK293T in control and stress were directly downloaded from Watkins et al. (235). The coding sequences of all human transcripts were computed from the same reference transcriptome.

6.3.2. Computational analysis

6.3.2.1. Read preprocessing

For QuantM-tRNA-seq data, raw FASTQ reads were first trimmed using Seqtk (github.com/lh3/seqtk, v1.3) and then 3' and 5' adaptors removed with BBDuk (sourceforge.net/projects/bbmap, v38.79), as described in their original protocol (237). For MSR-seq data, raw FASTQ reads were adaptor-removed using BBMerge (245) (sourceforge.net/projects/bbmap, v38.79) with: mininsert=20, mininsert0=20, trimpolya=t, usequality=f, forcemerge=t, entropy=f, adapter1=GATCGTCGGACTGTAGAA, adapter2=AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC. Then 3' and 5' ends were trimmed to remove the barcode and library-added nucleotides using Seqtk -b 7 -e 6 (github.com/lh3/seqtk, v1.3). For DM-tRNA-seq data, raw FASTQ reads were adaptor-removed using BBMerge (245) (sourceforge.net/projects/bbmap, v38.79) with: mininsert=20, mininsert0=20, trimpolya=t, usequality=f, forcemerge=t, entropy=f, adapters=AGATCGGAAGAGCACACGTCTGAACTCCAGTCA. Then RT adapters were trimmed from 3'-ends with BBDuk (sourceforge.net/projects/bbmap, v38.79): literal=CTTTGAGCCTAATGCCTGAAAGATCGGAAGAGCAC ACGTCTAGTTCTACAGTCCGACGATC, mink=8, ktrim=r, k=10, hdist=1, minlength=10.

6.3.2.2. tRNA quantitation and analysis

Quantitation and analysis of tRNA expression, charging and modifications were performed using the mim-tRNA-seq computational pipeline (*67*) (github.com/nedialkova-lab/mim-tRNAseq, v0.3.4.5). The following parameters were used:

S. cerevisiae: --species Scer --cluster-id 0.90 --min-cov 0.0005 --max-mismatches 0.1 --control-condition WT --max-multi 4 --remap --remap-mismatches 0.075

M. musculus: --species Mmus --cluster-id 0.95 --min-cov 0.0005 --max-mismatches 0.1 --control-condition Liver --max-multi 4 --remap --remap-mismatches 0.075

H. sapiens: --species Hsap --cluster-id 0.95 --min-cov 0.0005 --max-mismatches 0.1 --control-condition control_total --max-multi 4 --remap --remap-mismatches 0.075

6.3.2.3. Single-read analysis

After read deconvolution of the mim-tRNA-seq computational pipeline (*67*), all mismatches and charging of each read were recorded in a tab-delimited format. For all positions with at least 5% mismatch rate, the single-read analysis was performed. For each pair of either two modified positions or modification and charging, the algorithm computed the odds ratio (OR).

$$OR = \frac{(\#reads\ with\ both\ A\ and\ B)\times(\#reads\ without\ A\ nor\ B)}{(\#reads\ with\ A,\ but\ not\ B)\times(\#reads\ with\ B,\ but\ not\ A)}$$

The significance of the crosstalk was determined by Fisher's exact test. Finally, p-values were FDR-corrected for multiple comparisons with Benjamini & Hochberg.

6.3.2.4. tRNA fragmentation crosstalks

For fragmentation analysis, only 5′ tRFs were considered, since 3′ tRFs cannot be distinguished from RT stops. tRNA reads were classified in four different classes based on their 3′ end: tRFs terminating at positions 30-39 (C-loop), tRFs at 40-49 (V-loop), tRFs at 50-59 (T-loop), and full-length tRNAs longer than position 60. Similar to the single-read analysis above, we computed the odds ratio between all positions with at least 5% mismatch rate and each of the three tRF classes versus full length tRNAs.

$$OR = \frac{(\#tRX\ reads\ with\ modification)\times(\#Full\ tRNA\ reads\ without\ modification)}{(\#tRX\ reads\ without\ modification)\times(\#Full\ tRNA\ reads\ with\ modification)}$$

The significance of the crosstalk was determined by Fisher's exact test. Finally, p-values were FDR-corrected for multiple comparisons with Benjamini & Hochberg.

6.3.2.5. Simulated data

To validate the method against an artificially generated dataset, we generated samples containing 20,000 full-length yeast $tRNA^{Phe}$(GAA)-2 reads each. Mismatches were incorporated into positions 26 and 58 of the sequence with a certain probability (p_A, p_B). Given a certain OR, the probability of having both modifications in the same read was determined as:

$$p_{AB} = \frac{1+(p_A+p_B)(OR-1)-\sqrt{\left(1+(p_A+p_B)(OR-1)\right)^2+4\cdot OR(1-OR)\cdot p_A\cdot p_B}}{2(OR-1)}$$

A total of 700 samples were generated with all combinations of p_A,p_B = [0.075, 0.175, 0.275, 0.375, 0.475, 0.575, 0.675, 0.775, 0.875, 0.975] and $\log_2$(OR) = [-1.0, -0.5, -0.25, 0.0, 0.25, 0.5, 1.0]. They were finally aligned and analyzed by SLAC.

6.3.2.6. Differential modification and aminoacylation analysis

To test the significance of changes in modification between conditions, we first computed the counts of reads with or without mismatch at modified positions (>5% absolute mismatch rate). For changes in aminoacylation in periodate-treated samples, we counted the reads ending with CCA or CC. Next, a contingency table was built with the counts of modified/unmodified or charged/uncharged reads in condition A and condition B, and the odds ratios were calculated as:

$$OR = \frac{(\#reads\ modified\ or\ charged\ in\ A)/(\#reads\ unmodified\ or\ uncharged\ in\ A)}{(\#reads\ modified\ or\ charged\ in\ B)/(\#reads\ unmodified\ or\ uncharged\ in\ B)}$$

The significance of each OR was determined by chi-square tests. Finally, p-values were FDR-corrected for multiple comparisons with Benjamini & Hochberg.

6.3.2.7. Translational efficiency analysis

Using the human mRNA-seq data of control and stressed HEK293T cells, we computed the Translational Efficiency (TE) as the ratio between polysome TPMs versus total RNA TPMs, taking only genes with at least 10 TPM in both datasets.

6.3.2.8. Relative Synonymous Codon Usage (RSCU)

The RSCU is defined as the ratio of the observed frequency of a certain codon to the expected frequency given that all the synonymous codons for the same amino acid were used equally. The RSCU is therefore a real value between 0 and the number of synonymous codons for that amino acid, with values below 1 indicating a lower observed usage than expected, and vice versa.

$$RSCU = \frac{x_C}{\frac{\sum_{i \in C_{aa}} x_i}{n_{aa}}} \times n_{aa}$$

where x_C refers to the abundance of the codon C, C_{aa} is the set of all synonymous codons, and n_{aa} is the number of synonymous codons.

To determine the RSCU of a specific condition, we computed the average of the RSCU of all genes weighted by their standardized $\log_2$(TE).

6.3.2.9. Statistical analysis

All details of the statistical analyses can be found in the text and figure legends. We used a significance level of 0.05 for all analyses. Significant differences are abbreviated as follows: ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), *** ($p \leq 0.001$), and **** ($p \leq 0.0001$).

6.4. Results

6.4.1. Single-read tRNA-seq analysis reveals known and new crosstalks in yeast $tRNA^{Phe}$

Because of its size, tRNA-seq produces reads that can cover the entire length of the tRNA; tRNA-seq also captures certain tRNA modifications as "mutations" relative to the reference tRNA sequence, and depending on the library construction protocol the charging status by the heterogeneous 3' ends in the sequencing data (67,133,235). On average, one of every four modifications leave a mismatch signature upon reverse transcription, which makes them detectable by tRNA-seq (Table S1). However, the analysis of modifications and charging is generally reduced to a simple pileup percentage, which does not take into account the single-read nature of the data. In fact, previous reports using Sanger sequencing have provided proof-of-concept evidence that modification signatures can be quantified and correlated at the single-read level to detect crosstalks (246). In SLAC, we consider all pairwise combinations of tRNA modified positions (i.e. positions with at least 5% reads containing "mutated" reads) and charging, as well as all pairwise combinations of two modified positions that are detectable through "mutations" (**Fig. 6.1A**). For charging and modification

crosstalks, we determine the number of reads for: (i) tRNA is charged and modified, (ii) tRNA is charged but not modified, (iii) tRNA is not charged but modified, and (iv) tRNA is not charged and not modified. For any two modification crosstalks, we determine the number of reads for: (i) both sites are modified, (ii) site 1 is, site 2 is not modified, (iii) site 1 is not, site 2 is modified, and (iv) both sites are not modified.

Our analysis produces an odds ratio (OR) that informs whether tRNA charging and modification or any pair of modifications tend to appear together in the same read (OR>1, stimulatory crosstalk) or tend to be exclusive of one another (OR<1, inhibitory crosstalk), as well as calculates the significance of this interdependence using Fisher's exact test. Given that the tRNA-seq data in general has very high coverage for each tRNA species, hundreds to thousands of pairs for a specific tRNA can be analyzed simultaneously and the resulting p-values are FDR-corrected for multiple comparisons. We implemented our method within the open-source mim-tRNA-seq computational pipeline (67) (github.com/nedialkova-lab/mim-tRNAseq).

To validate our method, we analyzed yeast mim-tRNA-seq data from Behrens et al. (67), since most prior knowledge on tRNA crosstalks was centered on yeast $tRNA^{Phe}$ (238) (Table S2). Among all modifications detectable by tRNA-seq (**Fig. 6.1B**, Fig. S1A), there were two known crosstalks against which our method could be tested: the interdependence between wybutosine (yW) at position 37 (all residue numbering are according to the standard tRNA nomenclature, 37 corresponds to the 3' immediate nucleotide after the anticodon) and charging (244) and the association between N2,2-dimethyl-G (m^2_2G) at position 26 (in between the D and anticodon stems) and m^1A at position 58 (in T loop) (239). Our single-read pipeline identified both crosstalks in all three biological replicates (**Fig. 6.1C-D**). Furthermore, we identified interdependencies (yW37-m^1A58, m^2_2G26-Charging) that were previously unknown. Apart from $tRNA^{Phe}$, we detected crosstalks expanding to other tRNA species in wild-type yeast cells (Fig. S1B, Table S2).

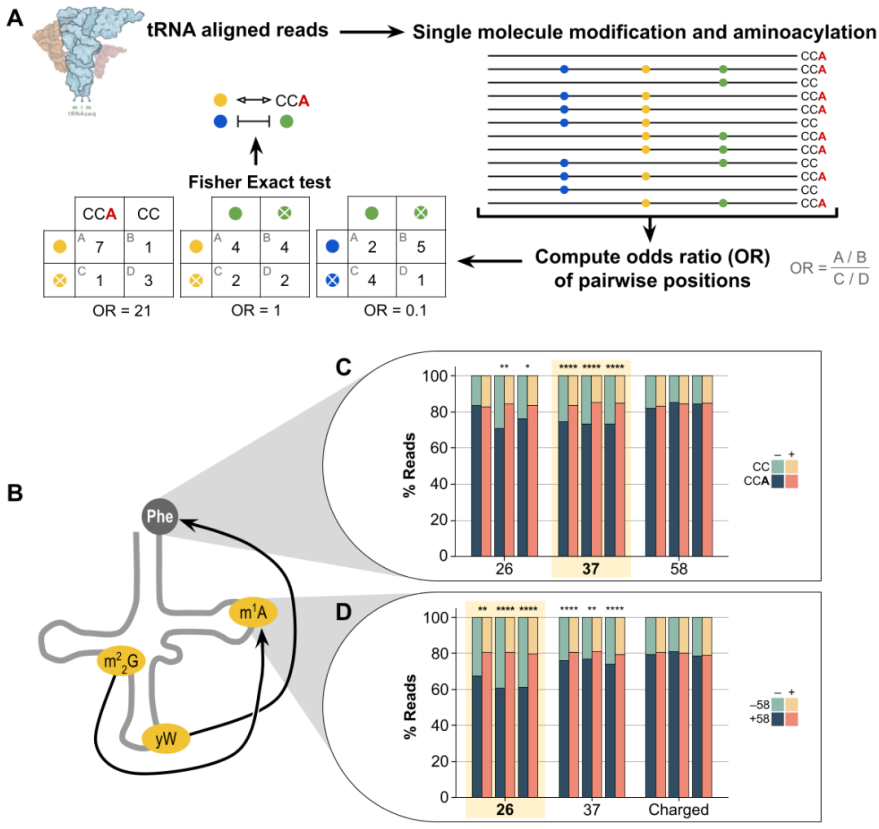


Figure 6.1. Single-read tRNA-seq analysis reveals known and new crosstalks in yeast $tRNA^{Phe}$.

(A) Schematic pipeline for the single-read analysis of tRNA reads. OR: odds ratio. (B) Detected known crosstalks of $tRNA^{Phe}$ of yW37-Charging and m^2_2G26-m^1A58 that are highlighted in (C) and (D). (C) Percentage of charged tRNA reads with possible crosstalks with the m^2_2G26, yW37, and m^1A58 modifications. (D) Percentage of m^1A58-modified reads with possible crosstalks with the m^2_2G26, yW37, and charging. In (C) and (D) the biological triplicates are plotted separately. Significant crosstalks are indicated with * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 10^{-3}$), **** ($p < 10^{-4}$). Significance is determined by Fisher's exact test, and FDR-corrected for multiple comparisons with Benjamini & Hochberg.

Although SLAC detects statistical association between two modifications or modification-charging, the method inherently lacks causality or directionality. Some of the detected crosstalks may be causal, i.e. they represent coordinated actions of tRNA acting enzymes, whereas others may be derived from indirect or independent events. Evidence of causality may be established through tRNA acting gene

knockout experiments. We analyzed mim-tRNA-seq data of yeast strains deficient of modification enzyme *Trm7* (lacking yW37 in $tRNA^{Phe}$(GAA), Fig. S1C, left panel) (67). $tRNA^{Phe}$ has known crosstalks between 37 and charging (**Fig. 6.1C**). Indeed, *Trm7Δ* led to a significant decrease in charging, consistent with yW37 modification enhancing $tRNA^{Phe}$ charging.

We also analyzed data from yeast strains with deletions of *Trm1* (lacking m^2_2G26, Fig. S1C, middle panel) and *Trm10* (lacking m^1G9, Fig. S1C, right panel) (67). In contrast to *Trm7Δ*, the *Trm1Δ* and *Trm10Δ* samples were not treated with periodate in library construction and hence only modification crosstalks could be examined. We established that the gene knockouts led to large decreases in the mutation fractions of the corresponding positions in the known tRNA substrates as expected, 37 for *Trm7Δ*, 26 for *Trm10Δ*, and 9 for *Trm10Δ* (Fig. S1C, Table S2). Among the previously identified crosstalks (Fig. S1B), *Trm1Δ* increased the m^1G9 level of $tRNA^{Lys}$(CTT) and decreased m^1A58 level of $tRNA^{Lys}$(CTT), $tRNA^{Phe}$(GAA), and $tRNA^{Thr}$(TGT). These results suggest that m^2_2G26 partially inhibits m^1G9 modification as well as enhances m^1A58 modification in these tRNAs.

We also found some modification crosstalks present in the wild-type samples that did not respond or responded differently to the modification enzyme knockouts, or new crosstalks only present in the enzyme knockout samples. The former includes 26-37, 37-58 in *Trm7Δ*; 26-32, 26-37, several 26-58 in *Trm1Δ*; and 9-26, 9-32, and 9-58 in *Trm10Δ*. These may represent crosstalks that are either unidirectional, e.g. m^2_2G26-m^1G9 crosstalk of $tRNA^{Lys}$(CTT) in *Trm1Δ* but not in *Trm10Δ*, or derived from multiple or independent events rather than coordinated action of two specific modification enzymes. The latter, such as three 9-26 in *Trm1Δ* and two 9-58 in *Trm10Δ*, may represent "synthetic" crosstalks that become only prevalent upon the knockout of a specific modification enzyme, akin to the synthetic phenotypes in genetics that become observable only upon the deletion of a specific gene. Therefore, while all known crosstalks in the literature are recovered by SLAC (**Fig. 6.1C-D**), not all detected crosstalks by SLAC

change in the same way in every tRNA upon modification enzyme perturbation.

We further validated the sensitivity of the method by generating simulated reads with different odds ratios (see Methods) and analyzing them with SLAC. We found that we can detect with confidence odds ratios as low as 1.20 (i.e. $\log_2(OR) = 0.25$), especially for mutation fractions in the sequencing data ranging between 5-95% (Fig. S1D). These results indicate that tRNA-seq data can be used at single-read level to retrieve interdependence information between charging and modification, and between two modified positions.

6.4.2. m^1A58-related crosstalks are abundant in the human tRNAome

We next analyzed the DM-tRNA-seq dataset of human HEK293T cells which also measured charging (133) (Table S3). Among all detected interdependencies, we identified m^1A58 crosstalks with positions 26 and 37 among the most frequent, present in 22 and 19 different tRNA isodecoders, respectively (**Fig. 6.2A**). By analyzing the OR of the 37-58 pair (**Fig. 6.2B**), we found that the m^1A58 modification is generally positively correlated with m^1G37/m^1I37 (OR>1), except in $tRNA^{Ala}$ isodecoders where the m^1A58 modification has an inhibitory effect on m^1I37 (OR<1). On the other hand, $m^2{}_2G26$ appears always positively correlated with m^1A58. Moreover, most crosstalks do not occur alone, rather they coordinate with other pairs within the same tRNA molecule. Therefore, specific tRNAs form intricate interdependent networks of multiple modified positions and charging. For instance, the isodecoders $tRNA^{Ala}$(AGC)-4/8 exhibit a network among three modification sites (**Fig. 6.2C**) involving $m^2{}_2G26$, m^1I37, and m^1A58. In $tRNA^{Leu}$(CAG)-1/2 (**Fig. 6.2D**), the network is even more complex with a total of eight significant crosstalks among $m^2{}_2G26$, m^1G37, m^3C-e2 (in the loop of the variable hairpin loop in type II tRNA), m^1A58, and charging. Other isodecoders show simpler networks with just one or few detected crosstalks (Fig. S2A).

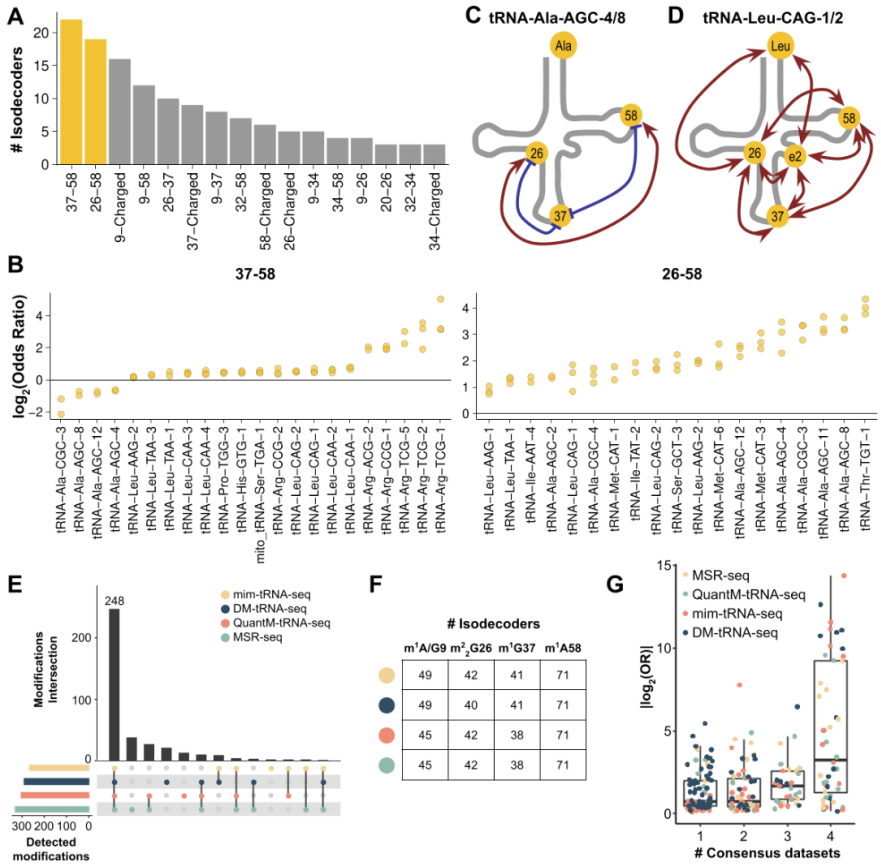


Figure 6.2. m^{1}A58-related crosstalks are abundant in the human tRNAome.
(A) Histogram of all significant crosstalks in at least two of the 3 biological replicates of human tRNA in HEK293T cells. Only crosstalks detected in three or more tRNA isodecoders are shown. Known modifications that generate high levels of mutation signatures in this sequencing experiment are m^{1}G9, acp^{3}U20, m$^2{}_2$G26, m^{3}C32, I34, m^{1}I37/m^{1}G37, and m^{1}A58. (B) Distribution of OR of detected 37-58 and 26-58 pairs, with each dot corresponding to individual replicates. (C-D) Significant crosstalks detected in at least two of the 3 biological replicates of tRNA^{Ala}(AGC)-4/8 and tRNA^{Leu}(CAG)-1/2. Position e2 is a m^{3}C modification. (E) Upset plot of detected modification sites (>5% mismatch in read alignment) in all replicates of HEK293T cells by mim-tRNA-seq, DM-tRNA-seq, QuantM-tRNA-seq, and MSR-seq. (F) Number of isodecoders with detected m^{1}A/G9, m$^2{}_2$G26, m^{1}G37, and m^{1}A58 modifications in all replicates of HEK293T cells by mim-tRNA-seq, DM-tRNA-seq, QuantM-tRNA-seq, and MSR-seq. (G) Absolute log$_2$-transformed odds ratios of crosstalks detected in HEK293T cells by 1, 2, 3 or 4 methods. The residue numbers for each tRNA is according to the tRNA nomenclature, e.g. the wobble anticodon nucleotide is always 34. tRNA transcript label is according to the genomic tRNA database (12).

For pairwise correlation of modifications we benchmarked SLAC against published QuantM-tRNA-seq, mim-tRNA-seq, DM-tRNA-seq and MSR-seq methods where datasets of HEK293T cells were available (67,133,235,237) (Table S3). In the detection of the mutation signatures, MSR-seq and QuantM-tRNA-seq used the SuperScript IV (SSIV) reverse transcriptase while mim-tRNA-seq and DM-tRNA-seq employed TGIRT that have similar tendencies to generate mutation signature in the tRNA-seq data. We detected a good degree of concordance, with a total of 248 consensus modifications identified by mutation signatures among all 380 mutated positions identified (>65%, **Fig. 6.2E-F**). However, at the quantitative level, modification signatures of TGIRT-based methods are clearly different from SSIV-based methods (Fig. S2B-C). As a result of these protocol differences and the distinct depth of reads coverage, we observed many common crosstalks as well as others that were dependent on the tRNA-seq dataset, which expanded our initial set of detected crosstalks (Fig. S2E, Table S3). A more detailed analysis actually showed that the consensus detected crosstalks between datasets are generally pairs with more extreme OR (**Fig. 6.2G**). Altogether, our exploration of the HEK293T human tRNAome shows a high interconnectivity among modifications and between modifications and charging, often involving m^1A58.

6.4.3. Tissue-specificity of m^1A58 and crosstalks across mouse tissues

We next characterized the relevance of tRNA modifications and their crosstalks beyond cell cultures using the publicly available QuantM-tRNA-seq data of seven mouse tissues (237) (Table S4). The m^1A58 modification is among the most widespread tRNA modifications in mammals, is present in the T loop of almost all cytosolic tRNAs, and can be detected at high sensitivity through mutation signatures in tRNA-seq (110,247). As previously reported, the mutation fraction derived from m^1A58 has a highly tissue-specific pattern (Fig. S3). A principal component analysis (PCA) of all modifications can clearly discriminate four main clusters of samples corresponding to brain (containing cortex, spinal cord, medulla

oblongata, and cerebellum; without clear differences among them), liver, tibialis, and heart tissues (**Fig. 6.3A**). The first component, which explains almost 70% of the variance, is predominantly determined by the m^1A58 fraction of several isodecoders among $tRNA^{Ala}$, $tRNA^{Asp}$, and $tRNA^{Glu}$ (**Fig. 6.3A**). On the other hand, the second component, which explains less than 10% of the variance, is more related to the technical variability of mismatches, which is particularly evident for one biological replicate of the cortex. We also observe a significant tissue-specificity of the interdependence between m^1A58 and m^1G9 in $tRNA^{Glu}$(TTC)-1 (ANOVA, $p<0.05$) (**Fig. 6.3B**), with liver and heart tissues showing a higher concurrence of both modifications. Overall, the mouse tissue data indicates a tissue-specific nature of the m^1A58 modification and its related crosstalks.

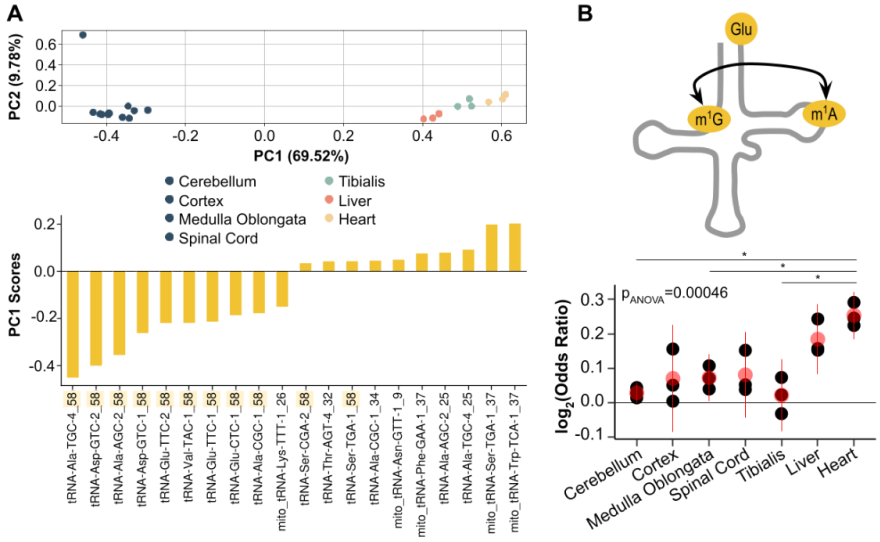


Figure 6.3. Tissue-specificity of m^1A58 and crosstalks across mouse tissues.
(A) Principal Component Analysis (PCA) of all tRNA modifications across seven mouse tissues (top), and the respective top 10 positive and negative contributors of the first component (bottom). (B) Changes in OR among tissues of $tRNA^{Glu}$(TTC)-1. Black dots are individual replicates, red dots and lines indicate the mean ± SD. Significance is determined with ANOVA and post-hoc Student's t-Test, FDR-corrected for multiple comparisons with Benjamini & Hochberg.

6.4.4. Crosstalks recapitulate modification and charging changes upon stress to potentially regulate translation

To investigate the regulatory potential of tRNA modification and charging crosstalks in stress response, we analyzed the MSR-seq data of HEK293T cells upon three different stresses: heat shock, hydrogen peroxide, and arsenite, which have been shown to alter tRNA properties that regulate translation (235). The MSR-seq data contains tRNA abundance, modification and charging measurements and mRNA measurements for both total RNA and polysome associated RNA, so that the effect on translational response under stress can be readily examined. For X-Y crosstalks with OR>1, we hypothesize a perturbation that causes an increase in X would also induce an increase in Y, and vice versa. In contrast, for X-Z crosstalks with OR<1, an increase in X would produce a decrease in Z, and vice versa. We first identified all changes of >3% in modification or charging upon each stress as well as their statistical significance, and asked whether their interdependent positions were also changing as expected (a.k.a. "TRUE", Table S5). We observed that the detected crosstalks significantly explain changes in modification and charging in all three stresses (**Fig. 6.4A**, Fig. S4A). This analysis was generally robust to changes in the selected percentage threshold of 3%, with TRUE cases always exceeding FALSE ones (Table S5). However, similar to the yeast mutants, a minority of the SLAC crosstalks did not change as expected, which reinforces the idea that detected crosstalks do not necessarily imply causality. A more detailed analysis also indicated that crosstalks with more extreme OR have generally higher percentage of TRUE cases (Fig. S4B). Moreover, TRUE cases are often dependent on the type of interconnected pair and on the amino acid family they belong to (Fig. S4C-D).

Given the near universal presence of m^1A58 in all cytosolic tRNAs, we analyzed differences of OR between m^1A58-Charging in tRNA isodecoders and their changes upon stress. For seven out of eight 4-codon-box amino acid families (Gly, Arg, Leu, Pro, Ser, Val, Thr; Ala is the exception), we detected significantly different ORs between tRNAs with A/T versus C/G at wobble anticodon position 34 in many

conditions (**Fig. 6.4B**, Fig. S5A). However, directionality was dependent on the amino acid family. Because the wobble nucleotides read the third nucleotide of codons, these differences can lead to differential translational regulation of synonymous codons. As GC-ending codons appear to be more efficiently translated than AT-ending codons (Fig. S5B) and this difference is enhanced under stress (Fig. S5C) (235), these differences in m^1A58-Charging crosstalks suggest a distinct regulation of the m^1A58 modification and charging that unexpectedly depend on the identity of their wobble nucleotides.

To further characterize the role of m^1A58 in translation, we compared the level of this modification between total tRNA and the polysome-associated tRNA. Surprisingly, we found that the m^1A58 fraction is overall lower in polysome-associated tRNA than those in the total tRNA. In-depth analysis revealed, however, that this anti-selection by polysomes only occurs for tRNAs with OR>1, but not for tRNAs with OR<1 between m^1A58 and charging (**Fig. 6.4C**). These results are consistent with polysome selectively enriching m^1A58-hypomodified tRNAs that are also not charged. On the other hand, the relative charging levels between polysome-associated and total RNA are about the same for tRNAs with OR>1 and OR<1 (Fig. S5D), which is consistent with m^1A58-hypomodified tRNAs losing their charging (i.e. synthesizing the peptide bond) more slowly in translation. This selective enrichment may be useful to temporarily pause the polysome at specific codons (OR>1 codons in **Fig. 6.4B**), which sensitizes the polysome to rapid changes in stress.

In short, detected crosstalks provide a snapshot of the changes in modification and charging upon stress. Specifically, m^1A58 modification appears anti-selected in translation, revealing m^1A58-Charging crosstalk as a potential parameter of translational regulation.

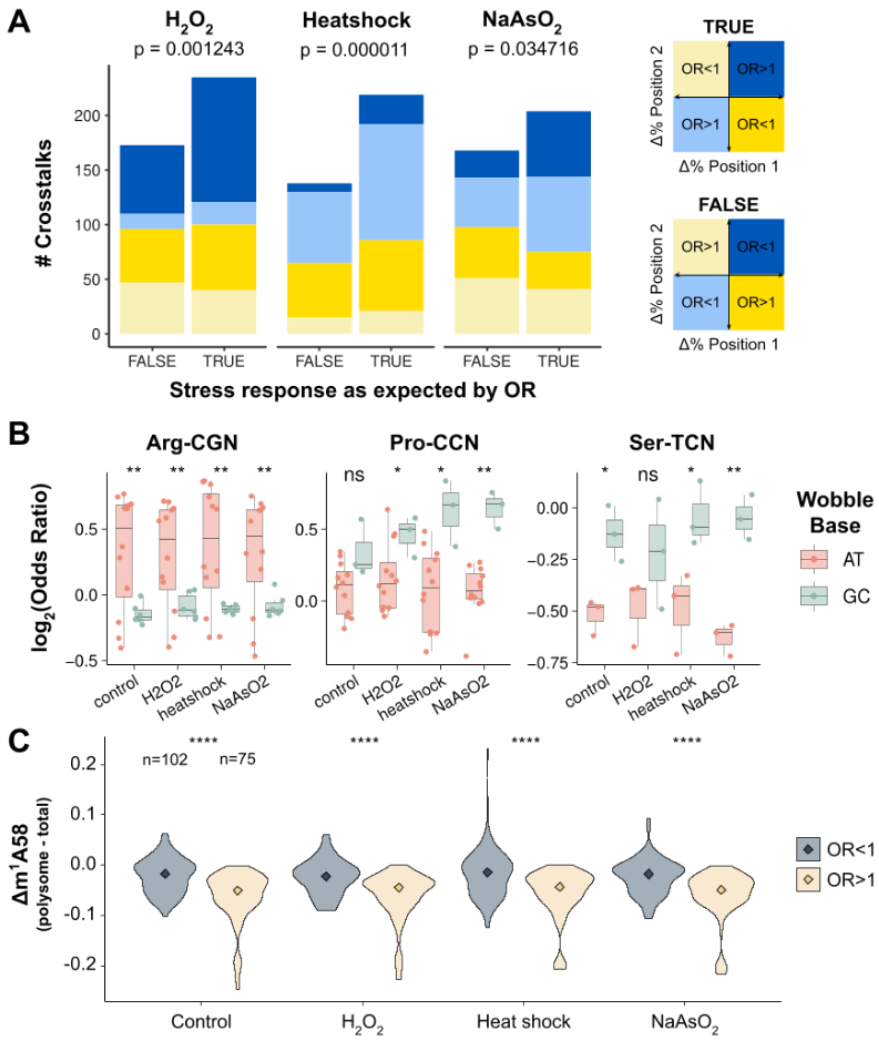


Figure 6.4. Crosstalks recapitulate modification and charging changes upon stress to potentially regulate translation.

(A) Number of position pairs changing in stress as expected by SLAC crosstalks (TRUE) or not (FALSE), related to Fig. S4A. All significant crosstalks of >3% changes in mismatch in read alignment are included. A one-sided binomial test is used to determine whether the observed frequency of successes is significantly higher than the null model p=0.5. (B) Differences in OR of m^1A58-Charged between isodecoders having AT vs GC at the wobble position for 4-codon-box readers of $tRNA^{Arg}$, $tRNA^{Pro}$, and $tRNA^{Ser}$. Significance is determined by a two-sided Student's t-test: * ($p<0.05$), ** ($p<0.01$), *** ($p<10^{-3}$), **** ($p<10^{-4}$). (C) Differences in m^1A58 modification between polysome versus total tRNA among isodecoders with OR>1 and OR<1. Changes between tRNAs with OR>1 and OR<1 are detected by two-tailed Wilcoxon rank-sum test (**** $p<10^{-4}$).

6.4.5. Modification crosstalks with tRNA fragmentation patterns

tRNA fragments (tRF) are small non-coding RNAs originated through enzymatic cleavage of tRNAs, which are implicated in many cellular processes such as cell proliferation, RNA silencing, or translational regulation (81). In the biogenesis of tRF, the modification state of tRNA molecules has been observed to play an important role (248–250). However, systematic analyses of existing crosstalks between tRNA modification and fragmentation have not been performed. We took advantage of MSR-seq that simultaneously sequences both full-length and fragmented tRNAs (235) and applied SLAC to uncover crosstalks. In particular, we binned tRNA fragments depending on their 3′ end, classifying them into three groups: 30 to 39 nt (terminating at anticodon loop or C-loop), 40 to 49 nt (V-loop), and 50 to 59 nt (T-loop). Using SLAC, we then computed the odds ratios of certain modifications being detected more or less frequently in the fragmented versus the full-length tRNA reads.

Among the most abundant crosstalks in control HEK293T cells (Table S5), we found that m^2_2G26 is associated with fragmentation at any of the three fragment size groups (**Fig. 6.5A**). $tRNA^{Leu}$ isodecoders are frequently detected to establish m^2_2G26 crosstalks with multiple loops fragments (**Fig. 6.5B**). In particular, for 30-39 fragmentation we observe a negative odds ratio for $tRNA^{Leu}$(CAA)-4 which corresponds to reads containing m^2_2G26 being less fragmented, whereas $tRNA^{Leu}$(TAA)-1 and $tRNA^{Leu}$(AAG)-2 show positive odds ratios which corresponds to m^2_2G26 increasing fragmentation in these two tRNAs. For 40-49 fragmentation we observe a negative odds ratio for $tRNA^{Leu}$(AAG)-2, but positive ratios for $tRNA^{Leu}$(CAA)-4 and $tRNA^{Leu}$(CAA)-1. For 50-59 fragmentation the odds ratios are negative for $tRNA^{Leu}$(TAA)-1 and $tRNA^{Leu}$(AAG)-2, but positive for $tRNA^{Leu}$(CAA)-1. These results indicate that a single tRNA modification can have both stimulatory and inhibitory effects on tRNA fragmentation; whether stimulatory or inhibitory depends on the specific tRNA and the location of cleavage.

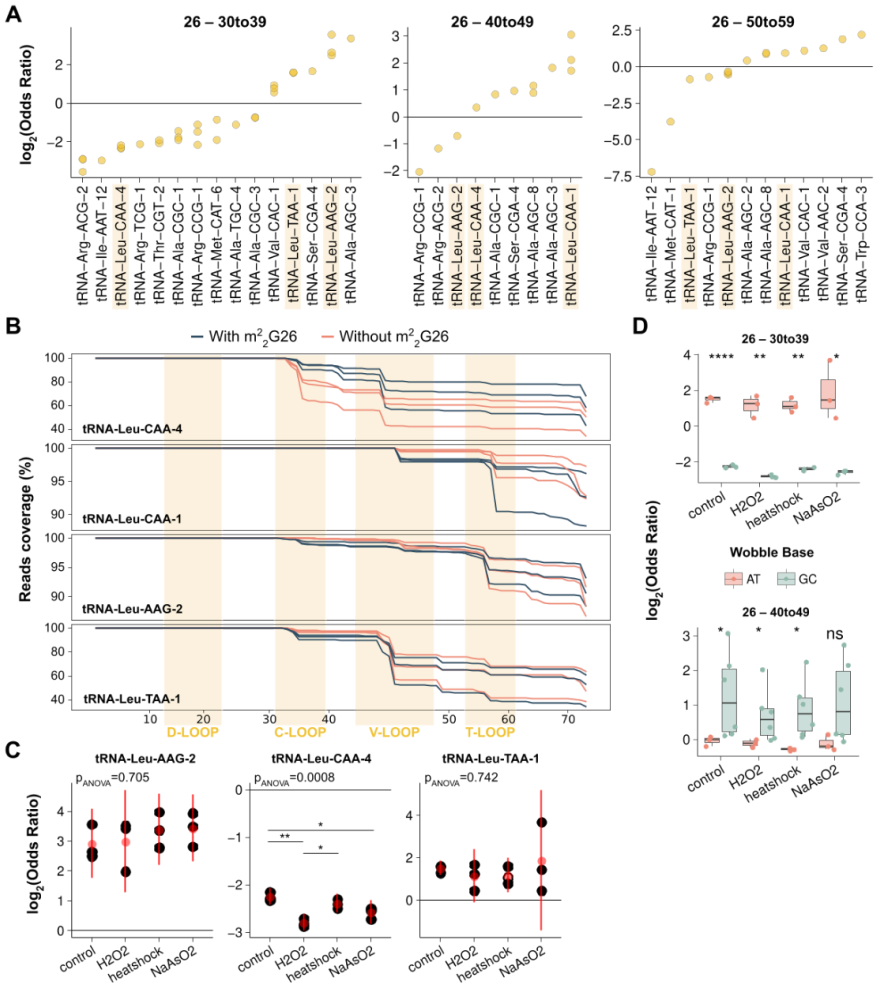


Figure 6.5. Modifications establish crosstalks with tRNA fragmentation patterns.
(A) Distribution of OR of all significant crosstalks of m^2_2G26 with tRNA fragmentation at positions 30-39, 40-49, and 50-59. Each dot corresponds to individual biological replicates of the sequencing data. (B) Coverage of reads with and without mismatch at position 26, mapping to $tRNA^{Leu}$(CAA)-4, $tRNA^{Leu}$(CAA)-1, $tRNA^{Leu}$(AAG)-2, and $tRNA^{Leu}$(TAA)-1. The residue numbers for each tRNA is according to the tRNA nomenclature, e.g. the wobble anticodon nucleotide is always 34. tRNA transcript label is according to the genomic tRNA database (12). (C) Changes in OR among conditions of the 26-30to39 crosstalk in $tRNA^{Leu}$(AAG)-2, $tRNA^{Leu}$(CAA)-4, and $tRNA^{Leu}$(TAA)-1. Black dots are individual replicates, red dots and lines indicate the mean ± SD. Significance is determined with ANOVA and post-hoc Student's t-Test, FDR-corrected for multiple comparisons with Benjamini & Hochberg. (D) Differences in OR of 26-30to39 and 26-40to49 between $tRNA^{Leu}$ isodecoders having AT vs GC at the wobble position, cognate of Leu-TTN codons. Significance is determined by a two-sided Student's t-test.

As tRNA fragments are known to play a role in stress response (250), we next interrogated whether these detected crosstalks were condition-dependent, focusing our analysis on the m^2_2G26-30to39 crosstalks. We observed that, while these crosstalks in some tRNAs are maintained in the unstressed and three stress conditions, for $tRNA^{Leu}$(CAA)-4 it becomes stronger under stress (**Fig. 6.5C**). Finally, we analyzed the differences between $tRNA^{Leu}$ isodecoders, and observed that isodecoders with a GC-wobble nucleotide had negative m^2_2G26-30to39 crosstalks, while the crosstalks were positive in isodecoders with AT-wobble nucleotide (**Fig. 6.5D**). In contrast, m^2_2G26-40to49 crosstalks show an opposite trend (**Fig. 6.5D**), suggesting that the m^2_2G26 modification in $tRNA^{Leu}$(CAA) is protective against anticodon-loop fragmentation, but induces V-loop fragmentation. Given these differences between isodecoders decoding AT-ending versus GC-ending leucine codons (Fig. S4C-D), the m^2_2G26 modification may also play a role in translational regulation in a way that depends on tRNA fragmentation.

Altogether, we provide evidence that single-read analysis can be extended to systematically detect crosstalks between modifications and tRNA fragmentation patterns in tRF biogenesis.

6.5. Discussion

tRNA-seq data is inherently multi-modal. While mRNA-seq is often used for quantifying transcript abundance only, tRNA-seq quantifies abundance, modification, charging and fragmentation (67,133,235,237). This work adds another aspect of tRNA-seq data which identifies networks of crosstalks between tRNA modifications, tRNA charging, and tRNA fragmentation. To make SLAC readily accessible to the community, we developed it within the open-source mim-tRNA-seq computational pipeline (github.com/nedialkova-lab/mim-tRNAseq) (67). The main limitation of our method requires that modifications elicit a misincorporation during the reverse transcription to generate a mutation signature. This restricts the types of modifications assessable by tRNA-seq (67,235),

which are to a certain extent dependent on the tRNA-seq method of choice. However, with continuous improvements in both the experimental and computational techniques of tRNA-seq, more modification types will become accessible by including specific chemical treatments that detect otherwise silent modifications such as pseudouridine or 5-methyl-cytosine. In addition, tRNA-seq datasets intrinsically suffer from a 5′ end coverage drop-off due to RT stops at some modified sites or rigid secondary structures, which leads to SLAC having a higher statistical power towards the 3′ ends. Incorporation of modification-specific insertions, deletions and RT stops into the pipeline could further expand our ability to access more modification types.

So far, the study of crosstalks has been mostly limited to *ad hoc* time-course or depletion/overexpression setups followed by sequencing, mass spectrometry or NMR (46,251,252). For instance, m^3C32 depended on i^6A37 upon *Tit1* deletion (253), NMR and LC-MS of yeast $tRNA^{Phe}$ following time-course maturation or enzyme deletions revealed modification dependencies (239), tRNA methylation was affected upon queuine depletion (254), or NSun2-mediated m^5C methylation was protective against 5′ tRNA fragmentation (249). With the development of SLAC, we can now leverage single-read information to accurately capture tRNAome-wide associations from unperturbed and physiologically-relevant tRNA-seq datasets.

SLAC does not establish a causality in the detected crosstalks, as the existence of invariant pairs in yeast strains with modification defects suggest. Our results indicate that crosstalks identified by SLAC can be derived from multiple origins, some causal and others indirect or independent (255). Our simulation experiment indicates that SLAC can also suffer from false negatives in modification sites with >95% mismatch rate, which is sometimes the case of tRNAs. The potential for causality or indirectness is idiosyncratic among specific tRNAs and modification sites. Even though we could not readily tell which crosstalks are causal at this time, SLAC is still useful in first identifying modification-modification, modification-charging, and

modification-fragmentation crosstalks that otherwise elude our current methods of detecting them.

Our systematic analysis of human tRNA crosstalks reveals that modifications and charging can be interconnected, potentially wiring a complex regulatory network. By analyzing the observed changes upon three different stresses, we support the physiological relevance of tRNA crosstalk networks in recapitulating the changes of modifications and charging. Translation has been widely studied in the context of stress response, which has been related to defects in protein homeostasis and human disease (256–258). For instance, the differential expression of tRNAs recognizing A- vs G-ending codons in polysomes, together with changes of specific tRNA modifications, leads to the selective enhancement of stress-dependent transcripts with a biased codon usage towards G/C-ending-codons (235). Here we provide evidence that tRNA crosstalks play a role in the stress response, and that m^1A58-Charging crosstalk can be differently wired between isodecoders decoding AT- vs GC-ending codons.

The tRNA m^1A58 is among the most interconnected modifications with m^1A58-m^2_2G26 and m^1A58-m^1G37 as abundant crosstalks in the human tRNAome, and also constitutes the most tissue-specific modification in mice. These results indicate that m^1A58 showcases a previously underappreciated regulatory potential. In this context, we observe that m^1A58-modification levels for some tRNAs are lower in polysomes compared to total tRNA, however, this result is only for tRNAs with a positive m^1A58-Charging crosstalk. This result is consistent with ribosome subtlely enriching tRNAs lacking m^1A58 modification which are also not charged. Presumably, this type of selection by the polysome may recapitulate an enrichment of uncharged and m^1A58-hypomodified tRNA in the A-site for translational pausing.

Beyond translation, tRNAs can be enzymatically cleaved and give rise to tRNA fragments which are involved in many processes such as stress response, cancer, aging, or development (81,259). By extending our single-read tRNA analysis to tRFs, we identify specific tRNA

modifications that either correlate or anti-correlate with fragment biogenesis, which recapitulates previously observed associations (235). In particular, we detect m^2_2G26 as a crucial parameter for $tRNA^{Leu}$ fragmentation; this modification can be either protective or inductive for fragmentation depending on the specific isodecoder and the location of the cleavage site. The magnitude of these crosstalks can also be modulated under stress.

The networks of crosstalks that we identified are diverse at the tRNA isodecoder level. These networks are dynamic in stress response and distinct in mouse tissues. As such, the tRNA crosstalk networks are, at first approach, resistant to satisfying and simple rules - a theme in tRNA biology harkening back to deciphering protein-RNA interactions between aminoacyl-synthetases and specific tRNA substrates (260). There is likely a complex grammar that we have just started to decipher, that introduces many new questions. What mechanisms give rise to crosstalks: tRNA biogenesis, tRNA synthetases, writer and eraser enzymes, nuclear export and re-import? How do these crosstalks affect translation in each tRNA context? On what timescales are these networks wired and rewired; does rewiring require tRNA turnover? Answering these questions will benefit from the characteristically high sequencing depth of many tRNA-seq datasets and their potential to detect for the first time tRNA crosstalks *in vivo*, such as the reported m^1A58-m^1G9 crosstalk of $tRNA^{Glu}$(TTC)-1 detected in liver and heart but not in other mouse tissues. Our work opens a new avenue to study crosstalks in physiologically-relevant conditions or diseases such as cancer that are known to alter tRNAomes and ultimately translation (84,150,261).

6.6. Data Availability

The code used in this study is available at GitHub (github.com/hexavier/tRNA_crosstalks), and the single-read analysis software is available within the within the open-source mim-tRNA-seq computational pipeline (github.com/hexavier/mim-tRNAseq).

6.7. Funding

This study was performed during a research stay, financially supported by EMBO Fellowships (Scientific Exchange Grant number 9170 to X.H.). The work of X.H. has been supported by a PhD fellowship from the Fundación Ramón Areces. Funding was also provided by the NIH (RM1HG008935 to T.P.). We acknowledge support of the Spanish Ministry of Science and Innovation, the Centro de Excelencia Severo Ochoa and the CERCA Programme / Generalitat de Catalunya.

6.8. Author contributions

Conceptualization, X.H., C.D.K., T.P.; Methodology, X.H., C.D.K., W.Z., C.P.W., T.P.; Software, X.H.; Investigation, X.H., C.D.K., W.Z., M.A., C.P.W.; Validation, X.H., C.D.K., T.P.; Formal analysis, X.H.; Writing-Original Draft, X.H., C.D.K., T.P.; Writing-Review & Editing, X.H., C.D.K., T.P., M.H.S., L.S.; Visualisation: X.H., C.D.K., T.P.; Funding Acquisition, X.H., T.P., L.S.; Supervision, T.P., M.H.S., L.S.

Chapter 7

Discussion

Along the genetic information flow from the genotype to the phenotype, proteins rather than transcripts are a more direct readout of gene expression (3). However, with the higher resolution of sequencing methods over mass-spectrometry, expression studies have so far mostly focused on the analysis of transcriptomes. Moreover, away from the long perceived static nature of mRNA translation, current evidence indicates that ribosome composition, tRNA pools and their post transcriptional modifications are highly dynamic and important regulators of mRNA translation (14,262).

In fact, alteration of tRNA repertoires has often been related to human diseases (263): mRNA translation rate is limited by scarce $tRNA^{Pro}$ aminoacylation in kidney cancer (105), $tRNA^{Glu}$(TTC) and $tRNA^{Arg}$(CCG) expression promote breast cancer metastasis (106), a mutation in $tRNA^{Arg}$(TCT) causes tRNA maturation defects leading to neurodegeneration (102), $tRNA^{Leu}$ aminoacylation by LARS is repressed in breast cancer (264), $tRNA^{Gly}$ sequestration by a mutated GARS leads to Charcot-Marie-Tooth disease (265), or aberrant tRNA-modifying enzymes can cause mitochondrial and neurological disorders and cancer, known as "tRNA modopathies" (266). In

consequence, the homeostasis of the cellular tRNAome needs to be globally controlled at multiple levels to maintain physiological levels of protein synthesis.

Stemming from a systems biology concept of translation regulation, in this thesis, we leverage and analyze available multi-omic datasets using statistical and machine learning algorithms. In particular, we analyze tissue-wide tRNA abundances and their role in cancer (Chapter 3) and viral infection (Chapter 4); detect and apply tissue-specific preferences of codon usage (Chapter 5); and study tRNA crosstalks as a coordination mechanism of tRNAomes (Chapter 6).

7.1. Translation regulation in cellular proliferation

tRNAs are among the most abundant RNA species in the cell, together with rRNAs. Given their strong secondary structure and abundant modifications, the coverage of tRNA reads in standard small RNA-seq datasets is limited and heterogeneous. Therefore, sequencing of tRNA molecules generally requires *ad hoc* protocols—e.g. using highly processive retrotranscriptases (67,110,235). However, in order to leverage publicly available standard small RNA-seq datasets, in Chapter 3, we benchmarked tRNA quantifications obtained either by standard small RNA-seq or by Hydro-tRNA-seq protocols (109). As a result, we provided proof-of-concept evidence that tRNA expression among several human cell lines can be quantified from small RNA-seq data. In fact, several computational pipelines have been designed to analyze this type of data (113,267), and the tRNA abundances of many human high-throughput datasets are currently available in databases (116,268).

With the analysis of more than 8,000 tumor and healthy small RNA-seq samples of the TCGA dataset, we identified tissue-specific differences in tRNA expression, which had only been previously reported by tRNA microarrays of limited sample size (103). Consistent with our results, a recent mouse tRNA-seq dataset similarly highlighted the upregulation of $tRNA^{Arg}$(TCT) and $tRNA^{Ala}$(TGC) in brain tissues (237).

Moreover, we observed that cell proliferation is the major factor explaining the tRNA level variability among human tissues. Previous studies indicate that A/T-cognate tRNAs, in contrast to G/C-cognate tRNAs, are overexpressed in cancer samples and models, which relates to the opposite codon usage patterns observed in proliferation- vs differentiation-related genes, respectively (42). In fact, CRISPR-targeting of tRNA genes showed that "proliferative" tRNAs were essential in rapidly dividing cell lines (269). This diverging tRNA program between proliferation and differentiation was also recapitulated in tumor vs healthy tissues across 23 cancer types. For instance, we reported the downregulation of $tRNA^{Leu}$(CAG) expression in nine different cancer types (including BRCA), which is actually the most repressed $tRNA^{Leu}$ upon depletion of the tumor suppressor LARS in breast cancer (264). Similarly, we found $tRNA^{Arg}$(TCT) upregulated in 13 cancer types, whose increased stability upon m^7G46 modification by METTL1 has been observed to induce malignant transformation (270).

Our work also revealed coordination between tRNA supply and demand (see Chapter 1), showing that the translation efficiency estimate Supply-to-Demand Adaptation (SDA) positively correlated with protein abundances. This positive correlation is nevertheless modest, as reported in other similar settings (95), which suggests that translation elongation in multicellular eukaryotes is also determined by factors other than tRNAs (18). Using SDAs, we therefore identified two sets of proliferation and differentiation codons, generally corresponding to A/T-ending and G/C-ending codons, respectively. In agreement, current models indicate that the decoding of rare A/T-ending codons by scarce tRNAs is the major translation bottleneck for rapid proliferation (53). In consequence, synonymous mutations from A/T to G/C-ending codons that alleviate this constraint are frequently observed in cancer-related genes (271,272), while upregulation of the limiting A/T-cognate tRNAs decreases ribosomal pausing and leads to tumorigenesis (99,270). Altogether, we hypothesize that the regulated expression of specific tRNAs rewires the supply/demand balance of rate-limiting codons, which enhances the mRNA translation elongation of A/T-rich oncogenes and represses G/C-rich tumor suppressors (122).

7.2. Translation interplay between host and virus

Viruses need to hijack the translational machinery of the host, which involves the interplay between the codon usage of viruses and the tRNA expression, modification and aminoacylation of hosts (273). For instance, it has been reported that a positive correlation exists between the codon usage of viruses and hosts, although relatively higher in prokaryotes than eukaryotes (157,274). Given the observed differences in tRNA expression among human tissues, in Chapter 4, we provided the first systematic study of the codon-anticodon interface of human viruses. We observed differences in codon usage related to the tissue tropism of viruses, which could be predicted based on their SDA adaptation.

In support of this translational adaptation, chikungunya infection for example induces translation in the host of virus-rich GAA, AAA, CAA, AGA and GGA codons by increasing the KIAA1456-mediated tRNA modification $mcm^5$34 (275). Therefore, viruses generally evolve to resemble the host codon usage and thus hijack its translation machinery, but an excessive adaptation can also be deleterious, since disruption of host translation impedes further replication (189,276). In fact, more virulent/symptomatic viruses are generally associated with more similar virus-host codon usages (189,277). However, the coding sequences of viruses are actually dependent on many other factors than translational adaptation alone, such as efficient transcription, mRNA export, or immune evasion (278). For instance, codon deoptimization of HIV-1 based on codon usage alone does not always lead to virus attenuation, and vice versa (279).

In the context of the COVID-19-causing SARS-CoV-2, we reported a high translational adaptation to the upper respiratory tract and the alveoli. The similarity between the codon usage of the virus and the human lung also supports this high adaptation to lung tissues (280,281). Moreover, given the stable codon usage since the first zoonosis, reports suggest that translational adaptation of SARS-CoV-2 happened prior to the human infection (282). In fact, since the first

sequences in December 2019, synonymous mutations appearing in the viral genome have slightly increased the similarity to human codon usage (283,284). However, the overall adaptation including both synonymous and nonsynonymous substitutions has globally decreased, probably driven by other CUB-unrelated evolutionary forces (285–287). This global codon dissimilarity to humans has also been related to the decrease in pathogenicity over time (277,287). On the other hand, the observed host promiscuity of coronaviruses has been associated with a generalist codon usage (282,288). As a side project of this thesis, we actually observed that humans, compared to other species with lower infectivity, showcased both a higher translational adaptation and an optimal interaction with the entry receptor ACE2 (289).

7.3. Tissue-specific codon usage in biotechnology

As different tissues exhibit different tRNA abundances, in Chapter 5, we measured the effect of codon usage on tissue-specific protein synthesis. To estimate translation efficiencies, we leveraged tissue-wide transcriptomics and proteomics datasets to compute their PTRs, and controlled for possible translation-independent biases. A more direct readout of translation efficiencies will come from ribosome profiling datasets, but such data is still scarce (22,290) and the experimental protocol is challenging and prone to many biases (213,291). Interestingly, detected high-PTR and low-PTR proteins of each tissue were highly dependent on their constituent cell types. For instance, skeletal and heart muscle shared many low-PTR proteins compared to other tissues, or nervous system tissues presented similar high-PTR proteins, which were actually enriched in neural-related functionalities. However, the cell-type contribution to detected bulk PTRs cannot be fully established without single-cell measurements of translation efficiency, which are still lacking (52). Approximations using scRNA-seq and scATACseq to measure codon usage and tRNA expression, respectively, also highlighted particularities of translation efficiency of neurons and muscles (104).

Codon usage studies have hitherto focused on the variability across genes rather than across tissues (28,54). With the computed PTRs, we therefore developed a random forest ensemble model that identified tissue-specific codon preferences, showing two clearly distinct sets of tissues with opposite codon patterns. Consistent with our results, testis and brain tissues in *Drosophila* present tissue-specific translation efficiency of A/T-ending-rich GFP constructs *in vivo* (292).

A wide variety of codon optimization algorithms exist, which have been commonly applied to increase the protein yield for biotechnological and therapeutic purposes (36,293,294). However, codon optimization is generally directed to resemble the genomic codon usage, which does not take into account the existing differences of gene expression among cell types (54). To translate the detected codon patterns into a tissue-targeted codon optimization tool, we then developed the open-source package and web server CUSTOM. By optimizing reporter proteins to lung and kidney, we demonstrated their tissue-specific expression in a human cell model. This was therefore the first proof-of-concept evidence and application of codon optimization to specific human tissues. Nevertheless, as introduced in Chapter 1, there are factors other than codon usage that shape coding sequences, which can alter their binding partners, transcript stability, translation, cotranslational folding, etc. (25,29). In consequence, codon optimization algorithms, alongside increased protein production (27), can sometimes also have other unpredictable effects (293,295). Altogether, tissue-specific codon usage constitutes an additional optimization factor within the multi-objective landscape of synthetic gene design.

7.4. Regulatory mechanisms of dynamic tRNAomes

While Chapter 3 reported that tRNAs were variable across tissues and cancer types, the regulatory mechanisms coordinating these changes in tRNAomes are still largely unknown (see Chapter 1). By integrating tRNA abundances with paired DNA methylation and copy number data, we revealed that tRNA gene expression is generally associated with

low DNA methylation and high copy numbers across cancer types. In fact, DNA methylation of the detected cancer-related $tRNA^{Arg}$(TCT) was recently reported to prevent binding of the transcriptional machinery and inhibit endometrial cancer cell growth and migration (77). Copy numbers of tRNA genes can also often vary among the human population (296). With the availability of this and other small RNA-seq datasets such as ENCODE or GTEx (297,298), the integration of tRNAomes with other multi-omic data types will shed light on the regulatory mechanisms of tRNA expression.

To investigate further the systems-wide coordination of tRNAomes, we developed SLAC in Chapter 6, which leverages the single-read nature of tRNA-seq data to identify crosstalks between tRNA modifications, aminoacylation and fragmentation. We validated known crosstalks of yeast $tRNA^{Phe}$(GAA), as well as explored tRNAome-wide crosstalks in yeast, human cells, and mouse tissues. In fact, tRNA biogenesis involves many tRNA-modifying enzymes or tRNA synthetases that often form multimeric complexes and requires sequential tRNA processing steps (88,255,299), which can lead to the widespread detection of interdependent modification sites and tRNA aminoacylation. While detected correlations do not imply a direct causal relationship, SLAC provided the first high-throughput exploration of crosstalks across multiple eukaryotic species. Studies of modification and aminoacylation crosstalks had hitherto been limited to targeted or time-course assays (238,239), and mostly focused at the tRNA anticodon loop (300,301).

The modification and charging state of tRNAs is dynamic and can regulate mRNA translation in stress response, development, or disease (129,258,265,302). We therefore reported that crosstalks can mediate the fine-tuning of tRNAomes across several cellular stresses. In fact, by leveraging polysome-associated data, tRNA species correlatively lacking both m^1A58 and charging were linked to slower translation. Furthermore, certain tRNA modifications can also protect or induce tRNA cleavage in response to stress (303); for instance, NSUN2- and DNMT2-mediated m^5C modification is protective against cleavage by angiogenin (249,250). Using SLAC, we also observed that specific

tRNA modifications such as m^2_2G26 can either protect or induce cleavage of human tRNAomes. Altogether, we provided a method to detect tRNAome-wide crosstalks in physiologically-relevant conditions, which captures snapshots in the study of complex and dynamic tRNAomes.

7.5. Further perspectives

7.5.1. Evolutionary forces on codon usage bias

Under the neutral theory of molecular evolution, selection on synonymous sites has long been considered too weak to exert an effect on fitness (304,305). However, as the results presented in this thesis indicate, the synonymous codon usage of genes can regulate gene expression and therefore have a phenotypic impact. As such, the assumption of neutrality of synonymous substitutions is in clear dispute (44,59). In contrast, evidence also indicates that the nucleotide composition of surrounding non-coding regions can explain most of the variability in codon and amino acid usage of genes (306,307). Disentangling the relative contribution of mutational pressure versus natural selection on codon usage bias will benefit from advances in experimental technologies such as CRISPR-Cas9-based mutant libraries (26,308), as well as the development of codon-based models of evolution (309).

Conservation at synonymous sites has been detected to be associated with mRNA translation, mRNA structure, and splicing signals (309). Furthermore, in a side project of this thesis, we have observed that A/T-ending codons show higher codon conservation in mammals compared to G/C-ending codons (57). Genes rich in these highly conserved A/T-ending codons were coordinately expressed together across human tissues and more likely to form protein complexes. These observations are in line with "rare" A/T-ending codons being commonly rate-limiting and hence showing higher regulatory potential (see Chapter 1) in contexts such as proliferation or specific tissues.

7.5.2. Dynamic regulation of mRNA translation

This thesis has provided evidence that tRNAomes are dynamic across tissues, between disease states, and upon cellular stresses. However, the exact mechanisms that regulate these dynamic changes in tRNAomes remain largely unaddressed. Recent developments in tRNA-seq technologies now allow the accurate and simultaneous determination of tRNA abundances, aminoacylation, and some modifications (67,235). Further techniques to capture tRNAome dynamics also include mass spectrometry for measuring tRNA modifications (310) or direct tRNA sequencing by Oxford Nanopore (311). Altogether, the advent of multi-omic datasets across diverse cell states and perturbations will enable the exploration of the regulatory mechanisms involved (64,297). For instance, in Chapter 3 and Chapter 6, we showed the potential of multi-omic data integration for elucidating the regulation of tRNAomes.

To what extent these changes in tRNAomes subsequently control mRNA translation elongation is also under debate. In Chapter 3, we observed that tRNA-based translation estimates correlated only modestly with protein abundances. Similarly modest correlations have also been reported between codon efficiency from ribosome profiling data and tRNA abundances (33,95). One possible explanation for this lack of strong correlations is that mRNA translation elongation is rate-limiting only for a subset of codons/tRNAs, as the results in Chapter 5 suggested. As high-resolution ribosome profiling and tRNA-seq datasets become available, the rate-limiting codons/tRNAs in specific cell states will be assessable.

7.5.3. tRNA-based therapeutics

Not only does mRNA translation explain the most variability of gene expression across tissues (21,28), but it is actually deregulated in many diseases such as cancer, metabolic disorders, neuropathies, or viral infections (62,263). Therefore, several approaches of modulation of tRNAomes have been proposed for therapeutic intervention (263).

First, tRNA overexpression can restore defective tRNAs, which has been effective to treat Charcot-Marie-Tooth (CMT) peripheral neuropathy in *Drosophila* and mouse models, caused by the $tRNA^{Gly}$ sequestration by a mutant glycyl-tRNA synthetase (265). Second, as reported in Chapter 3, several cancer types are associated with aberrant upregulation of specific tRNAs, such as $tRNA^{Arg}$(TCT) (270), whose expression can be therapeutically repressed by small molecules or RNA interference (106,312,313). Finally, nonsense mutations—single nucleotide mutations that convert an amino-acid-encoding codon into a premature stop codon—result in loss of protein expression, which account for 10-15% of genetic diseases (314). Engineering suppressor tRNAs that can read-through these premature stop codons emerges as a promising therapeutic intervention (315,316), which has been successfully applied in a mouse model of nonsense mutation through AAV-delivered gene therapy (317).

7.6. Concluding remarks

The typical human body is composed of dozens of different tissues and hundreds of cell types. While they all share a common genotype, their gene expression needs to be finely regulated at many levels to showcase distinct phenotypes and functions. In this thesis, we have applied systems biology approaches to uncover the contribution of dynamic codon usage and tRNAomes on tissue-specific mRNA translation.

From the tRNA perspective, we repurposed available tissue-wide small RNA-seq datasets for quantification of tRNA abundances, whose expression patterns were closely linked to the proliferative state of tissues. In the aberrant case of cancer, specific tRNA isoacceptors showcased altered expression and prognostic value, which related to changes in mRNA translation. These tissue-specific tRNA repertoires similarly reported the adaptation between the codon usage of human viruses to their tissue tropism. On the methods side, we measured and exploited tissue-specific differences in codon usage to develop a codon optimization tool for tissue-targeted gene expression. Furthermore, we designed a pipeline for the exploration of tRNAome-wide crosstalks

between modifications, aminoacylation and fragmentation, which highlighted the intricacy of tRNA dynamics.

Therefore, while we have only started to grasp the full complexity of mRNA translation control, the resources and findings of this thesis will be directly applicable to the development of tRNA-based interventions or tissue-targeted gene therapies and vaccines.

Bibliography

1. Crick F. Central Dogma of Molecular Biology. Nature. 1970;227(5258):561–3.
2. Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Molecular Biology of the Cell. Sixth Edition. Garland Science; 2015.
3. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 2020;21(10):630–44.
4. Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, Goodman B, et al. mRNA structure regulates protein expression through changes in functional half-life. Proc Natl Acad Sci. 2019;116(48):24075–83.
5. Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol. 2010;6(2):e1000664.
6. Tuller T, Zur H. Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res. 2015;43(1):13–28.
7. Lodish H, Berk A, Kaiser CA, Krieger M, Bretscher A, Ploegh H, et al. Molecular Cell Biology. Eighth Edition. W. H. Freeman and Company; 2016.
8. Maquat LE. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol. 2004;5(2):89–99.
9. Doma MK, Parker R. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. Nature. 2006;440(7083):561–4.
10. van Hoof A, Frischmeyer PA, Dietz HC, Parker R. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. Science. 2002;295(5563):2262–4.
11. Wilson RC, Doudna JA. Molecular mechanisms of RNA interference. Annu Rev Biophys. 2013;42:217–39.
12. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016;44(D1):D184-189.
13. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290(5806):457–65.
14. Rak R, Dahan O, Pilpel Y. Repertoires of tRNAs: The Couplers of Genomics and Proteomics. Annu Rev Cell Dev Biol. 2018;34:239–64.
15. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147(4):789–802.
16. Wu B, Eliscovich C, Yoon YJ, Singer RH. Translation dynamics of single mRNAs in live cells and neurons. Science. 2016;352(6292):1430–5.
17. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol. 2017;19:20–30.

18. Neelagandan N, Lamberti I, Carvalho HJF, Gobet C, Naef F. What determines eukaryotic translation elongation: recent molecular and quantitative analyses of protein synthesis. Open Biol. 2020;10(12):200292.
19. Schwartz AL, Ciechanover A. Targeting proteins for destruction by the ubiquitin system: implications for human pathobiology. Annu Rev Pharmacol Toxicol. 2009;49:73–96.
20. Pohl C, Dikic I. Cellular quality control by the ubiquitin-proteasome system and autophagy. Science. 2019;366(6467):818–22.
21. Franks A, Airoldi E, Slavov N. Post-transcriptional regulation across human tissues. PLOS Comput Biol. 2017;13(5):e1005535.
22. Wang Z-Y, Leushkin E, Liechti A, Ovchinnikova S, Mößinger K, Brüning T, et al. Transcriptome and translatome co-evolution in mammals. Nature. 2020;588(7839):642–7.
23. Schwanhäusser B, Wolf J, Selbach M, Busse D. Synthesis and degradation jointly determine the responsiveness of the cellular proteome. BioEssays News Rev Mol Cell Dev Biol. 2013;35(7):597–601.
24. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12:32–42.
25. Liu Y, Yang Q, Zhao F. Synonymous but not Silent: The Codon Usage Code for Gene Expression and Protein Folding. Annu Rev Biochem. 2021;90:375–401.
26. Shen X, Song S, Li C, Zhang J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature. 2022;606:725–31.
27. Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, et al. Codon Usage and Splicing Jointly Influence mRNA Localization. Cell Syst. 2020;10(4):351-362.e8.
28. Eraslan B, Wang D, Gusic M, Prokisch H, Hallström BM, Uhlén M, et al. Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. Mol Syst Biol. 2019;15(2):e8513.
29. Bergman S, Tuller T. Widespread non-modular overlapping codes in the coding regions. Phys Biol. 2020;17(3):031002.
30. Alexaki A, Kames J, Holcomb DD, Athey J, Santana-Quintero LV, Lam PVN, et al. Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design. J Mol Biol. 2019;431(13):2434–41.
31. Duret L, Galtier N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311.
32. Pouyet F, Mouchiroud D, Duret L, Sémon M. Recombination, meiotic expression and human codon usage. eLife. 2017;6:e27344.
33. Gobet C, Weger BD, Marquis J, Martin E, Neelagandan N, Gachon F, et al. Robust landscapes of ribosome dwell times and aminoacyl-tRNAs in response to nutrient stress in liver. Proc Natl Acad Sci. 2020;117(17):9630–41.
34. Translation elongation can control translation initiation on eukaryotic mRNAs. EMBO J. 2014;33:21–34.

35. Yang Q, Yu C-H, Zhao F, Dang Y, Wu C, Xie P, et al. eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. Nucleic Acids Res. 2019;47(17):9243–58.
36. Gould N, Hendy O, Papamichail D. Computational tools and algorithms for designing customized synthetic genes. Front Bioeng Biotechnol. 2014;2:41.
37. Mordret E, Dahan O, Asraf O, Rak R, Yehonadav A, Barnabas GD, et al. Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. Mol Cell. 2019;75(3):427-441.e5.
38. Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, Kline N, et al. Codon Optimality Is a Major Determinant of mRNA Stability. Cell. 2015;160(6):1111–24.
39. Wu Q, Medina SG, Kushawah G, DeVore ML, Castellano LA, Hand JM, et al. Translation affects mRNA stability in a codon-dependent manner in human cells. Sonenberg N, Struhl K, Weissman JS, editors. eLife. 2019;8:e45396.
40. Bae H, Coller J. Codon optimality-mediated mRNA degradation: Linking translational elongation to mRNA stability. Mol Cell. 2022;82(8):1467–76.
41. Radhakrishnan A, Chen Y-H, Martin S, Alhusaini N, Green R, Coller J. The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. Cell. 2016;167(1):122-132.e9.
42. Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, et al. A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. Cell. 2014;158(6):1281–92.
43. Supek F. The Code of Silence: Widespread Associations Between Synonymous Codon Biases and Gene Function. J Mol Evol. 2016;82(1):65–73.
44. Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 2006;7(2):98–108.
45. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell. 2010;141(2):344–54.
46. Torres AG, Rodríguez-Escribà M, Marcet-Houben M, Santos Vieira HG, Camacho N, Catena H, et al. Human tRNAs with inosine 34 are essential to efficiently translate eukarya-specific low-complexity proteins. Nucleic Acids Res. 2021;49(12):7011–34.
47. Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol. 2013;20(2):237–43.
48. Chaney JL, Steele A, Carmichael R, Rodriguez A, Specht AT, Ngo K, et al. Widespread position-specific conservation of synonymous rare codons within coding sequences. Wilke CO, editor. PLOS Comput Biol. 2017;13(5):e1005531.
49. Rosenberg AA, Marx A, Bronstein AM. Codon-specific Ramachandran plots show amino acid backbone conformation depends on identity of the translated codon. Nat Commun. 2022;13(1):2815.

50. Gingold H, Dahan O, Pilpel Y. Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. Nucleic Acids Res. 2012;40(20):10053–63.
51. Man O, Pilpel Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet. 2007;39(3):415–21.
52. VanInsberghe M, van den Berg J, Andersson-Rolf A, Clevers H, van Oudenaarden A. Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. Nature. 2021;597:561–5.
53. Guimaraes JC, Mittal N, Gnann A, Jedlinski D, Riba A, Buczak K, et al. A rare codon-based translational program of cell proliferation. Genome Biol. 2020;21(1):44.
54. Kames J, Alexaki A, Holcomb DD, Santana-Quintero LV, Athey JC, Hamasaki-Katagiri N, et al. TissueCoCoPUTs: Novel Human Tissue-Specific Codon and Codon-Pair Usage Tables Based on Differential Tissue Gene Expression. J Mol Biol. 2020;432(11):3369–78.
55. Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci. 2004;101(34):12588–91.
56. Najafabadi HS, Goodarzi H, Salavati R. Universal function-specificity of codon usage. Nucleic Acids Res. 2009;37(21):7014–23.
57. Benisty H, Hernandez-Alias X, Weber M, Anglada-Girotto M, Mantica F, Radusky L, et al. Evolutionary conservation of A/T-ending codons reflects co-regulation of expression and complex formation. bioRxiv. 2022. p. 2022.01.17.475622.
58. Hershberg R, Petrov DA. Selection on Codon Bias. Annu Rev Genet. 2008;42(1):287–99.
59. Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, et al. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Mol Biol Evol. 2018;35(5):1092–103.
60. Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. A Role for tRNA Modifications in Genome Structure and Codon Usage. Cell. 2012;149(1):202–13.
61. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004;32(17):5036–44.
62. Kirchner S, Ignatova Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nat Rev Genet. 2015;16(2):98–112.
63. Berg MD, Brandl CJ. Transfer RNAs: diversity in form and function. RNA Biol. 2021;18(3):316–39.
64. Behrens A, Gao L, Rodschinka G, Wani S, Strasser K, Nedialkova D. Dynamics of human tRNA repertoires as a function of cell identity. Poster session presented at: EMBL Conference: Protein Synthesis and Translational Control. Online; 2021.

65. Thornlow BP, Hough J, Roger JM, Gong H, Lowe TM, Corbett-Detig RB. Transfer RNA genes experience exceptionally elevated mutation rates. Proc Natl Acad Sci U S A. 2018;115(36):8996–9001.
66. Torres AG. Enjoy the Silence: Nearly Half of Human tRNA Genes Are Silent. Bioinforma Biol Insights. 2019;13:1177932219868454.
67. Behrens A, Rodschinka G, Nedialkova DD. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. Mol Cell. 2021;81(8):1802-1815.e7.
68. Schmitt BM, Rudolph KLM, Karagianni P, Fonseca NA, White RJ, Talianidis I, et al. High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA–tRNA interface. Genome Res. 2014;24(11):1797–807.
69. Michels AA, Robitaille AM, Buczynski-Ruchonnet D, Hodroj W, Reina JH, Hall MN, et al. mTORC1 directly phosphorylates and regulates human MAF1. Mol Cell Biol. 2010;30(15):3749–57.
70. Orioli A, Praz V, Lhôte P, Hernandez N. Human MAF1 targets and represses active RNA polymerase III genes by preventing recruitment rather than inducing long-term transcriptional arrest. Genome Res. 2016;26(5):624–35.
71. Campbell KJ, White RJ. MYC regulation of cell growth through control of transcription by RNA polymerases I and III. Cold Spring Harb Perspect Med. 2014;4(5):a018408.
72. Chu WM, Wang Z, Roeder RG, Schmid CW. RNA polymerase III transcription repressed by Rb through its interactions with TFIIIB and TFIIIC2. J Biol Chem. 1997;272(23):14755–61.
73. Cairns CA, White RJ. p53 is a general repressor of RNA polymerase III transcription. EMBO J. 1998;17(11):3112–23.
74. Yang J, Smith DK, Ni H, Wu K, Huang D, Pan S, et al. SOX4-mediated repression of specific tRNAs inhibits proliferation of human glioblastoma cells. Proc Natl Acad Sci U S A. 2020;117(11):5782–90.
75. Gerber A, Ito K, Chu C-S, Roeder RG. Gene-Specific Control of tRNA Expression by RNA Polymerase II. Mol Cell. 2020;78(4):765-778.e7.
76. Rosselló-Tortella M, Bueno-Costa A, Martínez-Verbo L, Villanueva L, Esteller M. DNA methylation-associated dysregulation of transfer RNA expression in human cancer. Mol Cancer. 2022;21:48.
77. Acton RJ, Yuan W, Gao F, Xia Y, Bourne E, Wozniak E, et al. The genomic loci of specific human tRNA genes exhibit ageing-related DNA hypermethylation. Nat Commun. 2021;12:2655.
78. Van Bortle K, Phanstiel DH, Snyder MP. Topological organization and dynamic regulation of human tRNA genes during macrophage differentiation. Genome Biol. 2017;18:180.
79. Sekulovski S, Trowitzsch S. Transfer RNA processing - from a structural and disease perspective. Biol Chem. 2022;403(8–9):749–63.
80. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, et al.

Structure of a Ribonucleic Acid. Science. 1965;147(3664):1462–5.
81. Schimmel P. The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. Nat Rev Mol Cell Biol. 2018;19:45–58.
82. Pan T. Modifications and functional genomics of human transfer RNA. Cell Res. 2018;28(4):395–404.
83. de Crécy-Lagard V, Boccaletto P, Mangleburg CG, Sharma P, Lowe TM, Leidel SA, et al. Matching tRNA modifications in humans to their known and predicted enzymes. Nucleic Acids Res. 2019;47(5):2143–59.
84. Suzuki T. The expanding world of tRNA modifications and their disease relevance. Nat Rev Mol Cell Biol. 2021;22(6):375–92.
85. Crick FH. Codon--anticodon pairing: the wobble hypothesis. J Mol Biol. 1966;19(2):548–55.
86. Chapeville F, Lipmann F, Von Ehrenstein G, Weisblum B, Ray WJ, Benzer S. On the role of soluble ribonucleic acid in coding for amino acids. Proc Natl Acad Sci U S A. 1962;48:1086–92.
87. Dunin-Horkawicz S, Czerwoniec A, Gajda MJ, Feder M, Grosjean H, Bujnicki JM. MODOMICS: a database of RNA modification pathways. Nucleic Acids Res. 2006;34(Database issue):D145-149.
88. Rubio Gomez MA, Ibba M. Aminoacyl-tRNA synthetases. RNA N Y N. 2020;26(8):910–36.
89. Fersht AR, Dingwall C. Evidence for the double-sieve editing mechanism in protein synthesis. Steric exclusion of isoleucine by valyl-tRNA synthetases. Biochemistry. 1979;18(12):2627–31.
90. Rajendran V, Kalita P, Shukla H, Kumar A, Tripathi T. Aminoacyl-tRNA synthetases: Structure, function, and drug discovery. Int J Biol Macromol. 2018;111:400–14.
91. Choe BK, Taylor MW. Kinetics of synthesis and characterization of transfer RNA precursors in mammalian cells. Biochim Biophys Acta BBA - Nucleic Acids Protein Synth. 1972;272(2):275–87.
92. Czech A, Wende S, Mörl M, Pan T, Ignatova Z. Reversible and rapid transfer-RNA deactivation as a mechanism of translational repression in stress. PLoS Genet. 2013;9(8):e1003767.
93. Su Z, Wilson B, Kumar P, Dutta A. Noncanonical Roles of tRNAs: tRNA Fragments and Beyond. Annu Rev Genet. 2020;54:47–69.
94. Saikia M, Krokowski D, Guan B-J, Ivanov P, Parisien M, Hu G, et al. Genome-wide Identification and Quantitative Analysis of Cleaved tRNA Fragments Induced by Cellular Stress. J Biol Chem. 2012;287(51):42708–25.
95. Wu CC-C, Zinshteyn B, Wehner KA, Green R. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. Mol Cell. 2019;73(5):959-970.e5.
96. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc Natl Acad Sci U S A. 2018;115(21):E4940–9.

97. Elf J, Nilsson D, Tenson T, Ehrenberg M. Selective Charging of tRNA Isoacceptors Explains Patterns of Codon Usage. Science. 2003;300(5626):1718–22.
98. Torrent M, Chalancon G, de Groot NS, Wuster A, Madan Babu M. Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. Sci Signal. 2018;11(546):eaat6409.
99. Darnell AM, Subramaniam AR, O'Shea EK. Translational Control through Differential Ribosome Pausing during Amino Acid Limitation in Mammalian Cells. Mol Cell. 2018;71(2):229-243.e11.
100. Saikia M, Wang X, Mao Y, Wan J, Pan T, Qian S-B. Codon optimality controls differential mRNA translation during amino acid starvation. RNA. 2016;22(11):1719–27.
101. Chen S, Li K, Cao W, Wang J, Zhao T, Huan Q, et al. Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level. Mol Biol Evol. 2017;34(11):2944–58.
102. Ishimura R, Nagy G, Dotu I, Zhou H, Yang X-L, Schimmel P, et al. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. Science. 2014;345(6195):455–9.
103. Dittmar KA, Goodenbour JM, Pan T. Tissue-Specific Differences in Human Transfer RNA Expression. PLOS Genet. 2006;2(12):e221.
104. Gao W, Gallardo-Dodd CJ, Kutter C. Cell type-specific analysis by single-cell profiling identifies a stable mammalian tRNA-mRNA interface and increased translation efficiency in neurons. Genome Res. 2022;32:97–110.
105. Loayza-Puch F, Rooijers K, Buil LCM, Zijlstra J, F. Oude Vrielink J, Lopes R, et al. Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. Nature. 2016;530(7591):490–4.
106. Goodarzi H, Nguyen HCB, Zhang S, Dill BD, Molina H, Tavazoie SF. Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. Cell. 2016;165(6):1416–27.
107. Rudolph KLM, Schmitt BM, Villar D, White RJ, Marioni JC, Kutter C, et al. Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. Galtier N, editor. PLOS Genet. 2016;12(5):e1006024.
108. Waldman YY, Tuller T, Shlomi T, Sharan R, Ruppin E. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. Nucleic Acids Res. 2010;38(9):2964–74.
109. Gogakos T, Brown M, Garzia A, Meyer C, Hafner M, Tuschl T. Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. Cell Rep. 2017;20(6):1463–75.
110. Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, et al. Efficient and quantitative high-throughput tRNA sequencing. Nat Methods. 2015;12(9):835–7.
111. Guo Y, Bosompem A, Mohan S, Erdogan B, Ye F, Vickers KC, et al. Transfer RNA detection by small RNA deep sequencing and disease association with myelodysplastic syndromes. BMC Genomics. 2015;16:727.

112. Guo Y, Xiong Y, Sheng Q, Zhao S, Wattacheril J, Flynn CR. A micro-RNA expression-signature for human NAFLD progression. J Gastroenterol. 2016;51(10):1022–30.
113. Hoffmann A, Fallmann J, Vilardo E, Mörl M, Stadler PF, Amman F. Accurate mapping of tRNA reads. Bioinformatics. 2018;34(7):1116–24.
114. Pundhir S, Gorodkin J. Differential and coherent processing patterns from small RNAs. Sci Rep. 2015;5:12062.
115. Torres AG, Piñeyro D, Rodríguez-Escribà M, Camacho N, Reina O, Saint-Léger A, et al. Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. Nucleic Acids Res. 2015;43(10):5145–57.
116. Zhang Z, Ye Y, Gong J, Ruan H, Liu C-J, Xiang Y, et al. Global analysis of tRNA and translation factor expression reveals a dynamic landscape of translational regulation in human cancers. Commun Biol. 2018;1:234.
117. Zhang Z, Ruan H, Liu C-J, Ye Y, Gong J, Diao L, et al. tRic: a user-friendly data portal to explore the expression landscape of tRNAs in human cancers. RNA Biol. 2019;17(11):1674–9.
118. Mattijssen S, Arimbasseri AG, Iben JR, Gaidamakov S, Lee J, Hafner M, et al. LARP4 mRNA codon-tRNA match contributes to LARP4 activity for ribosomal protein mRNA poly(A) tail length protection. eLife. 2017;6:e28889.
119. Chu A, Robertson G, Brooks D, Mungall AJ, Birol I, Coope R, et al. Large-scale profiling of microRNAs for The Cancer Genome Atlas. Nucleic Acids Res. 2016;44(1):e3.
120. Flores O, Kennedy EM, Skalsky RL, Cullen BR. Differential RISC association of endogenous human microRNAs predicts their inhibitory potential. Nucleic Acids Res. 2014;42(7):4629–39.
121. Mefferd AL, Kornepati AVR, Bogerd HP, Kennedy EM, Cullen BR. Expression of CRISPR/Cas single guide RNAs using small tRNA promoters. RNA N Y N. 2015;21(9):1683–9.
122. Benisty H, Weber M, Hernandez-Alias X, Schaefer MH, Serrano L. Mutation bias within oncogene families is related to proliferation-specific codon usage. Proc Natl Acad Sci U S A. 2020;117(48):30848–56.
123. Torres AG, Reina O, Attolini CS-O, Pouplana LR de. Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. Proc Natl Acad Sci. 2019;116(17):8451–6.
124. Scholzen T, Gerdes J. The Ki-67 protein: From the known and the unknown. J Cell Physiol. 2000;182(3):311–22.
125. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. Mol Syst Biol. 2011;7:481.
126. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016;534(7605):55–62.
127. Slebos RJC, Wang X, Wang X, Zhang B, Tabb DL, Liebler DC. Proteomic analysis of colon and rectal carcinoma using standard and customized databases.

Sci Data. 2015;2:150022.
128. dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res. 2003;31(23):6976–85.
129. Bornelöv S, Selmi T, Flad S, Dietmann S, Frye M. Codon usage optimization in pluripotent embryonic stem cells. Genome Biol. 2019;20:119.
130. Fornasiero EF, Rizzoli SO. Pathological changes are associated with shifts in the employment of synonymous codons at the transcriptome level. BMC Genomics. 2019;20:566.
131. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95.
132. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018;50(4):591–602.
133. Evans ME, Clark WC, Zheng G, Pan T. Determination of tRNA aminoacylation levels by high-throughput sequencing. Nucleic Acids Res. 2017;45(14):e133.
134. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science. 2009;324(5924):218–23.
135. Park J-L, Lee Y-S, Kunkeaw N, Kim S-Y, Kim I-H, Lee YS. Epigenetic regulation of noncoding RNA transcription by mammalian RNA polymerase III. Epigenomics. 2017;9(2):171–87.
136. Besser D, Götz F, Schulze-Forster K, Wagner H, Kröger H, Simon D. DNA methylation inhibits transcription by RNA polymerase III of a tRNA gene, but not of a 5S rRNA gene. FEBS Lett. 1990;269(2):358–62.
137. Lant JT, Berg MD, Heinemann IU, Brandl CJ, O'Donoghue P. Pathways to disease from natural variations in human cytoplasmic tRNAs. J Biol Chem. 2019;294(14):5294–308.
138. Telonis AG, Loher P, Magee R, Pliatsika V, Londin E, Kirino Y, et al. tRNA Fragments Show Intertwining with mRNAs of Specific Repeat Content and Have Links to Disparities. Cancer Res. 2019;79(12):3034–49.
139. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009;25(16):2078–9.
140. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol Biol Clifton NJ. 2019;1962:1–14.
141. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl. 2010;26(6):841–2.
142. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol. 2009;5(9):e1000502.
143. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al.

The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
144. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
145. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
146. Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 2009;37(Database issue):D159–62.
147. Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, et al. A new and updated resource for codon usage tables. BMC Bioinformatics. 2017;18:391.
148. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015;1(6):417–25.
149. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 1981;9(1):r43-74.
150. Hernandez-Alias X, Benisty H, Schaefer MH, Serrano L. Translational efficiency across healthy and tumor tissues is proliferation-related. Mol Syst Biol. 2020;16(3):e9275.
151. Sharp PM, Stenico M, Peden JF, Lloyd AT. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans. 1993;21(4):835–41.
152. Morgado S, Vicente AC. Global In-Silico Scenario of tRNA Genes and Their Organization in Virus Genomes. Viruses-Basel. 2019;11(2):180.
153. Lucks JB, Nelson DR, Kudla GR, Plotkin JB. Genome Landscapes and Bacteriophage Codon Usage. Plos Comput Biol. 2008;4(2):e1000001.
154. Carbone A. Codon Bias is a Major Factor Explaining Phage Evolution in Translationally Biased Hosts. J Mol Evol. 2008;66(3):210–23.
155. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. 2017;45(1):39–53.
156. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017;5:69.
157. Bahir I, Fromer M, Prat Y, Linial M. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol Syst Biol. 2009;5:311.
158. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 2003;92(1):1–7.
159. Lauring AS, Jones JO, Andino R. Rationalizing the development of live

attenuated virus vaccines. Nat Biotechnol. 2010;28(6):573–9.
160. Zhao K-N, Gu W, Fang NX, Saunders NA, Frazer IH. Gene codon composition determines differentiation-dependent expression of a viral capsid gene in keratinocytes in vitro and in vivo. Mol Cell Biol. 2005;25(19):8643–55.
161. Zhou J, Liu WJ, Peng SW, Sun XY, Frazer I. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J Virol. 1999;73(6):4972–82.
162. van Weringh A, Ragonnet-Cronin M, Pranckeviciene E, Pavon-Eternod M, Kleiman L, Xia X. HIV-1 modulates the tRNA pool to improve translation efficiency. Mol Biol Evol. 2011;28(6):1827–34.
163. Li M, Kao E, Gao X, Sandig H, Limmer K, Pavon-Eternod M, et al. Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. Nature. 2012;491(7422):125–8.
164. B. Miller J, Hippen AA, M. Wright S, Morris C, G. Ridge P. Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect. Biomed Genet Genomics. 2017;2(2):1–5.
165. Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, et al. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res. 2011;39(Database issue):D576-582.
166. Aiewsakun P, Simmonds P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome. 2018;6:38.
167. World Health Organization. Novel Coronavirus (2019-nCoV) situation reports [Internet]. [cited 2020]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
168. Belalov IS, Lukashev AN. Causes and Implications of Codon Usage Bias in RNA Viruses. Digard P, editor. PLoS ONE. 2013;8(2):e56642.
169. Shackelton LA, Parrish CR, Holmes EC. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J Mol Evol. 2006;62(5):551–63.
170. Goz E, Mioduser O, Diament A, Tuller T. Evidence of translation efficiency adaptation of the coding regions of the bacteriophage lambda. DNA Res Int J Rapid Publ Rep Genes Genomes. 2017;24(4):333–42.
171. Knipe DM, Howley PM. Fields Virology. Sixth Edition. Lippincott Williams and Wilkins; 2013.
172. Chang ST, Thomas MJ, Sova P, Green RR, Palermo RE, Katze MG. Next-generation sequencing of small RNAs from HIV-infected cells identifies phased microrna expression patterns and candidate novel microRNAs differentially expressed upon infection. mBio. 2013;4(1):e00549-12.
173. Shi J, Hu N, Mo L, Zeng Z, Sun J, Hu Y. Deep RNA Sequencing Reveals a Repertoire of Human Fibroblast Circular RNAs Associated with Cellular Responses to Herpes Simplex Virus 1 Infection. Cell Physiol Biochem Int J Exp

Cell Physiol Biochem Pharmacol. 2018;47(5):2031–45.
174. Stark TJ, Arnold JD, Spector DH, Yeo GW. High-resolution profiling and analysis of viral and host small RNAs during human cytomegalovirus infection. J Virol. 2012;86(1):226–35.
175. Golumbeanu M, Desfarges S, Hernandez C, Quadroni M, Rato S, Mohammadi P, et al. Proteo-Transcriptomic Dynamics of Cellular Response to HIV-1 Infection. Sci Rep. 2019;9:213.
176. Nobre LV, Nightingale K, Ravenhill BJ, Antrobus R, Soday L, Nichols J, et al. Human cytomegalovirus interactome analysis identifies degradation hubs, domain associations and viral protein functions. Garrett WS, Maes P, Maes P, Szpara ML, editors. eLife. 2019;8:e49894.
177. Ouwendijk WJD, Dekker LJM, van den Ham H-J, Lenac Rovis T, Haefner ES, Jonjic S, et al. Analysis of Virus and Host Proteomes During Productive HSV-1 and VZV Infection in Human Epithelial Cells. Front Microbiol. 2020;11:1179.
178. Franzo G, Tucciarone CM, Cecchinato M, Drigo M. Canine parvovirus type 2 (CPV-2) and Feline panleukopenia virus (FPV) codon bias analysis reveals a progressive adaptation to the new niche after the host jump. Mol Phylogenet Evol. 2017;114:82–92.
179. Luo W, Tian L, Gan Y, Chen E, Shen X, Pan J, et al. The fit of codon usage of human-isolated avian influenza A viruses to human. Infect Genet Evol. 2020;81:104181.
180. Wong EHM, Smith DK, Rabadan R, Peiris M, Poon LLM. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. BMC Evol Biol. 2010;10:253.
181. Sungnak W, Huang N, Bécavin C, Berg M, Queen R, Litvinukova M, et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat Med. 2020;26(5):681–7.
182. Ziegler CGK, Allon SJ, Nyquist SK, Mbano IM, Miao VN, Tzouanas CN, et al. SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. Cell. 2020;181(5):1016-1035.e19.
183. Baig AM, Khaleeq A, Ali U, Syeda H. Evidence of the COVID-19 Virus Targeting the CNS: Tissue Distribution, Host-Virus Interaction, and Proposed Neurotropic Mechanisms. ACS Chem Neurosci. 2020;11(7):995–8.
184. Li Y-C, Bai W-Z, Hashikawa T. The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. J Med Virol. 2020;92(6):552–5.
185. Mao L, Jin H, Wang M, Hu Y, Chen S, He Q, et al. Neurologic Manifestations of Hospitalized Patients With Coronavirus Disease 2019 in Wuhan, China. JAMA Neurol. 2020;77(6):683–90.
186. Pan L, Mu M, Yang P, Sun Y, Wang R, Yan J, et al. Clinical Characteristics of COVID-19 Patients With Digestive Symptoms in Hubei, China: A Descriptive, Cross-Sectional, Multicenter Study. Am J Gastroenterol. 2020;115(5):766–73.

187. Zhang H, Kang Z, Gong H, Xu D, Wang J, Li Z, et al. Digestive system is a potential route of COVID-19: an analysis of single-cell coexpression pattern of key proteins in viral entry process. Gut. 2020;69(6):1010–8.
188. Zhou Z, Zhao N, Shu Y, Han S, Chen B, Shu X. Effect of gastrointestinal symptoms on patients infected with COVID-19. Gastroenterology. 2020;158(8):2294–7.
189. Chen F, Wu P, Deng S, Zhang H, Hou Y, Hu Z, et al. Dissimilation of synonymous codon usage bias in virus–host coevolution due to translational selection. Nat Ecol Evol. 2020;4(4):589–600.
190. Mioduser O, Goz E, Tuller T. Significant differences in terms of codon usage bias between bacteriophage early and late genes: a comparative genomics analysis. BMC Genomics. 2017;18:866.
191. Pavon-Eternod M, David A, Dittmar K, Berglund P, Pan T, Bennink JR, et al. Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation. Nucleic Acids Res. 2013;41(3):1914–21.
192. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.
193. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–3.
194. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–2830.
195. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. BioTechniques. 2000;28(6):1102, 1104.
196. Theis C, Reeder J, Giegerich R. KnotInFrame: prediction of −1 ribosomal frameshift events. Nucleic Acids Res. 2008;36(18):6013–20.
197. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Dempsey DM, Dutilh BE, et al. Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). Arch Virol. 2019;164(9):2417–29.
198. Weekes MP, Tomasec P, Huttlin EL, Fielding CA, Nusinow D, Stanton RJ, et al. Quantitative Temporal Viromics: An Approach to Investigate Host-Pathogen Interaction. Cell. 2014;157(6):1460–72.
199. Fielding CA, Weekes MP, Nobre LV, Ruckova E, Wilkie GS, Paulo JA, et al. Control of immune ligands by members of a cytomegalovirus gene expansion suppresses natural killer cell activation. Yokoyama WM, editor. eLife. 2017;6:e22206.
200. Zhao Q, Fränti P. WB-index: A sum-of-squares based index for cluster validity. Data Knowl Eng. 2014;92:77–89.
201. Dunn JC. Well-Separated Clusters and Optimal Fuzzy Partitions. J Cybern.

1974;4(1):95–104.
202. Al- Zoubi MB, Raw M. An Efficient Approach for Computing Silhouette Coefficients. J Comput Sci. 2008;4(3):252–5.
203. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.
204. Bekaert M, Firth AE, Zhang Y, Gladyshev VN, Atkins JF, Baranov PV. Recode-2: new design, new search tools, and many more genes. Nucleic Acids Res. 2010;38:D69–74.
205. Drosten C, Günther S, Preiser W, van der Werf S, Brodt H-R, Becker S, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med. 2003;348(20):1967–76.
206. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012;367(19):1814–20.
207. Broszeit F, Tzarum N, Zhu X, Nemanichvili N, Eggink D, Leenders T, et al. N-Glycolylneuraminic Acid as a Receptor for Influenza A Viruses. Cell Rep. 2019;27(11):3284-3294.e6.
208. Raj VS, Mou H, Smits SL, Dekkers DHW, Müller MA, Dijkman R, et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. Nature. 2013;495(7440):251–4.
209. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181(2):271-280.e8.
210. Wang K, Chen W, Zhou Y-S, Lian J-Q, Zhang Z, Du P, et al. SARS-CoV-2 invades host cells via a novel route: CD147-spike protein. bioRxiv. 2020;2020.03.14.988345.
211. Ranaghan MJ, Li JJ, Laprise DM, Garvie CW. Assessing optimal: inequalities in codon optimization algorithms. BMC Biol. 2021;19(1):36.
212. Watts A, Sankaranarayanan S, Watts A, Raipuria RK. Optimizing protein expression in heterologous system: Strategies and tools. Meta Gene. 2021;29:100899.
213. Tunney R, McGlincy NJ, Graham ME, Naddaf N, Pachter L, Lareau LF. Accurate design of translational output by a neural network model of ribosome distribution. Nat Struct Mol Biol. 2018;25(7):577–82.
214. Medina-Muñoz SG, Diez M, Castellano LA, Pescador G da S, Wu Q, Bazzini AA. iCodon: ideal codon design for customized gene expression. bioRxiv. 2021;2021.05.06.442969.
215. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
216. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.
217. Jiang L, Wang M, Lin S, Jian R, Li X, Chan J, et al. A Quantitative Proteome Map of the Human Body. Cell. 2020;183(1):269-283.e19.

218. Wang D, Eraslan B, Wieland T, Hallström B, Hopf T, Zolg DP, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol Syst Biol. 2019;15(2):e8503.
219. Mathieson T, Franken H, Kosinski J, Kurzawa N, Zinn N, Sweetman G, et al. Systematic analysis of protein turnover in primary cells. Nat Commun. 2018;9(1):689.
220. Zecha J, Meng C, Zolg DP, Samaras P, Wilhelm M, Kuster B. Peptide Level Turnover Measurements Enable the Study of Proteoform Dynamics *. Mol Cell Proteomics. 2018;17(5):974–92.
221. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. Science. 2008;320(5884):1784–7.
222. Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics MCP. 2012;11(3):M111.014050.
223. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. PLOS Biol. 2006;4(6):e180.
224. Hia F, Yang SF, Shichino Y, Yoshinaga M, Murakawa Y, Vandenbon A, et al. Codon bias confers stability to human mRNAs. EMBO Rep. 2019;20(11):e48220.
225. Allen SR, Stewart RK, Rogers M, Ruiz IJ, Cohen E, Laederach A, et al. Distinct responses to rare codons in select Drosophila tissues. bioRxiv. 2022;2022.01.06.475284.
226. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):31.
227. Mauro VP. Codon Optimization in the Production of Recombinant Biotherapeutics: Potential Risks and Considerations. BioDrugs. 2018;32(1):69–81.
228. Alexaki A, Hettiarachchi GK, Athey JC, Katneni UK, Simhadri V, Hamasaki-Katagiri N, et al. Effects of codon optimization on coagulation factor IX translation and structure: Implications for protein and gene therapies. Sci Rep. 2019;9(1):15449.
229. Sharp PM, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281–95.
230. Wright F. The 'effective number of codons' used in a gene. Gene. 1990;87(1):23–9.
231. Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. Nucleic Acids Res. 2018;46(D1):D221–8.

232. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PloS One. 2010;5(11):e13984.
233. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
234. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47(D1):D419–26.
235. Watkins CP, Zhang W, Wylder A, Katanski CD, Pan T. A multiplex platform for small RNA sequencing elucidates multifaceted tRNA stress response and translational regulation. Nat Commun. 2022;13:2491.
236. Clark WC, Evans ME, Dominissini D, Zheng G, Pan T. tRNA base methylation identification and quantification via high-throughput sequencing. RNA. 2016;22(11):1771–84.
237. Pinkard O, McFarland S, Sweet T, Coller J. Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. Nat Commun. 2020;11(1):4104.
238. Li J, Zhu W-Y, Yang W-Q, Li C-T, Liu R-J. The occurrence order and cross-talk of different tRNA modifications. Sci China Life Sci. 2021;64(9):1423–36.
239. Barraud P, Gato A, Heiss M, Catala M, Kellner S, Tisné C. Time-resolved NMR monitoring of tRNA maturation. Nat Commun. 2019;10(1):3373.
240. Johannsson S, Neumann P, Wulf A, Welp LM, Gerber H-D, Krull M, et al. Structural insights into the stimulation of S. pombe Dnmt2 catalytic efficiency by the tRNA nucleoside queuosine. Sci Rep. 2018;8(1):8880.
241. Huang Z-X, Li J, Xiong Q-P, Li H, Wang E-D, Liu R-J. Position 34 of tRNA is a discriminative element for m5C38 modification by human DNMT2. Nucleic Acids Res. 2021;49(22):13045–61.
242. Guy MP, Shaw M, Weiner CL, Hobson L, Stark Z, Rose K, et al. Defects in tRNA Anticodon Loop 2'-O-Methylation Are Implicated in Nonsyndromic X-Linked Intellectual Disability due to Mutations in FTSJ1. Hum Mutat. 2015;36(12):1176–87.
243. Li J, Wang Y-N, Xu B-S, Liu Y-P, Zhou M, Long T, et al. Intellectual disability-associated gene ftsj1 is responsible for 2'-O-methylation of specific tRNAs. EMBO Rep. 2020;21(8):e50095.
244. Han L, Guy MP, Kon Y, Phizicky EM. Lack of 2'-O-methylation in the tRNA anticodon loop of two phylogenetically distant yeast species activates the general amino acid control pathway. PLOS Genet. 2018;14(3):e1007288.
245. Bushnell B, Rood J, Singer E. BBMerge – Accurate paired shotgun read merging via overlap. PLOS ONE. 2017;12(10):e0185056.
246. Rubio MAT, Ragone FL, Gaston KW, Ibba M, Alfonzo JD. C to U Editing Stimulates A to I Editing in the Anticodon Loop of a Cytoplasmic Threonyl tRNA in Trypanosoma brucei*. J Biol Chem. 2006;281(1):115–20.
247. Liu F, Clark W, Luo G, Wang X, Fu Y, Wei J, et al. ALKBH1-Mediated tRNA

Demethylation Regulates Translation. Cell. 2016;167(3):816-828.e16.
248. Cozen AE, Quartley E, Holmes AD, Hrabeta-Robinson E, Phizicky EM, Lowe TM. ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat Methods. 2015;12(9):879–84.
249. Blanco S, Dietmann S, Flores JV, Hussain S, Kutter C, Humphreys P, et al. Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. EMBO J. 2014;33(18):2020–39.
250. Schaefer M, Pollex T, Hanna K, Tuorto F, Meusburger M, Helm M, et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes Dev. 2010;24(15):1590–5.
251. Su D, Chan CTY, Gu C, Lim KS, Chionh YH, McBee ME, et al. Quantitative analysis of ribonucleoside modifications in tRNA by HPLC-coupled mass spectrometry. Nat Protoc. 2014;9(4):828–41.
252. Helm M, Motorin Y. Detecting RNA modifications in the epitranscriptome: predict and validate. Nat Rev Genet. 2017;18(5):275–91.
253. Arimbasseri AG, Iben J, Wei F-Y, Rijal K, Tomizawa K, Hafner M, et al. Evolving specificity of tRNA 3-methyl-cytidine-32 (m3C32) modification: a subset of tRNAsSer requires N6-isopentenylation of A37. RNA. 2016;22(9):1400–10.
254. Müller M, Hartmann M, Schuster I, Bender S, Thüring KL, Helm M, et al. Dynamic modulation of Dnmt2-dependent tRNA methylation by the micronutrient queuine. Nucleic Acids Res. 2015;43(22):10952–62.
255. Dixit S, Henderson JC, Alfonzo JD. Multi-Substrate Specificity and the Evolutionary Basis for Interdependence in tRNA Editing and Methylation Enzymes. Front Genet. 2019;10:104.
256. Endres L, Dedon PC, Begley TJ. Codon-biased translation can be regulated by wobble-base tRNA modification systems during cellular stress responses. RNA Biol. 2015;12(6):603–14.
257. Tahmasebi S, Khoutorsky A, Mathews MB, Sonenberg N. Translation deregulation in human disease. Nat Rev Mol Cell Biol. 2018;19(12):791–807.
258. Nedialkova DD, Leidel SA. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. Cell. 2015;161(7):1606–18.
259. Pandey KK, Madhry D, Ravi Kumar YS, Malvankar S, Sapra L, Srivastava RK, et al. Regulatory roles of tRNA-derived RNA fragments in human pathophysiology. Mol Ther Nucleic Acids. 2021;26:161–73.
260. Saks ME, Sampson JR, Abelson JN. The transfer RNA identity problem: a search for rules. Science. 1994;263(5144):191–7.
261. Begik O, Lucas MC, Liu H, Ramirez JM, Mattick JS, Novoa EM. Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. Genome Biol. 2020;21(1):97.
262. Genuth NR, Barna M. The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. Mol Cell. 2018;71(3):364–74.
263. Orellana EA, Siegal E, Gregory RI. tRNA dysregulation and disease. Nat Rev

Genet. 2022;23:651–64.
264. Passarelli MC, Pinzaru AM, Asgharian H, Liberti MV, Heissel S, Molina H, et al. Leucyl-tRNA synthetase is a tumour suppressor in breast cancer and regulates codon-dependent translation dynamics. Nat Cell Biol. 2022;24(3):307–15.
265. Zuko A, Mallik M, Thompson R, Spaulding EL, Wienand AR, Been M, et al. tRNA overexpression rescues peripheral neuropathy caused by mutations in tRNA synthetase. Science. 2021;373(6559):1161–6.
266. Chujo T, Tomizawa K. Human transfer RNA modopathies: diseases caused by aberrations in transfer RNA modifications. FEBS J. 2021;288(24):7096–122.
267. Lee J-O, Chu J, Jang G, Lee M, Chung Y-J. tReasure: R-based GUI package analyzing tRNA expression profiles from small RNA sequencing data. BMC Bioinformatics. 2022;23(1):155.
268. Lee J-O, Lee M, Chung Y-J. DBtRend: A Web-Server of tRNA Expression Profiles from Small RNA Sequencing Data in Humans. Genes. 2021;12(10):1576.
269. Aharon-Hefetz N, Frumkin I, Mayshar Y, Dahan O, Pilpel Y, Rak R. Manipulation of the human tRNA pool reveals distinct tRNA sets that act in cellular proliferation or cell cycle arrest. Topisirovic I, Cole PA, Topisirovic I, Serrano L, editors. eLife. 2020;9:e58461.
270. Orellana EA, Liu Q, Yankova E, Pirouz M, De Braekeleer E, Zhang W, et al. METTL1-mediated m7G modification of Arg-TCT tRNA drives oncogenic transformation. Mol Cell. 2021;81(16):3323-3338.e14.
271. Ran X, Xiao J, Cheng F, Wang T, Teng H, Sun Z. Pan-cancer analyses of synonymous mutations based on tissue-specific codon optimality. Comput Struct Biotechnol J. 2022;20:3567–80.
272. Li Q, Li J, Yu C, Chang S, Xie L, Wang S. Synonymous mutations that regulate translation speed might play a non-negligible role in liver cancer development. BMC Cancer. 2021;21(1):388.
273. Nunes A, Ribeiro DR, Marques M, Santos MAS, Ribeiro D, Soares AR. Emerging Roles of tRNAs in RNA Virus Infections. Trends Biochem Sci. 2020;45(9):794–805.
274. Simón D, Cristina J, Musto H. Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts. Front Microbiol. 2021;12:646300.
275. Jungfleisch J, Böttcher R, Talló-Parra M, Pérez-Vilaró G, Merits A, Novoa EM, et al. CHIKV infection reprograms codon optimality to favor viral RNA translation by altering the tRNA epitranscriptome. Nat Commun. 2022;13(1):4725.
276. Alonso AM, Diambra L. SARS-CoV-2 Codon Usage Bias Downregulates Host Expressed Genes With Similar Codon Usage. Front Cell Dev Biol. 2020;8:831.
277. Chen F, Yang J-R. Distinct codon usage bias evolutionary patterns between weakly and strongly virulent respiratory viruses. iScience. 2021;25(1):103682.
278. Mordstein C, Cano L, Morales AC, Young B, Ho AT, Rice AM, et al. Transcription, mRNA Export, and Immune Evasion Shape the Codon Usage of

Viruses. Genome Biol Evol. 2021;13(9):evab106.
279. Jordan-Paiz A, Franco S, Martinez MA. Synonymous Codon Pair Recoding of the HIV-1 env Gene Affects Virus Replication Capacity. Cells. 2021;10(7):1636.
280. Zhang Y, Jin X, Wang H, Miao Y, Yang X, Jiang W, et al. Compelling Evidence Suggesting the Codon Usage of SARS-CoV-2 Adapts to Human After the Split From RaTG13. Evol Bioinforma. 2021;17:11769343211052012.
281. Yu C, Li J, Li Q, Chang S, Cao Y, Jiang H, et al. Hepatitis B virus (HBV) codon adapts well to the gene expression profile of liver cancer: an evolutionary explanation for HBV's oncogenic role. J Microbiol Seoul Korea. 2022;60(11):1106–12.
282. MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, et al. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLOS Biol. 2021;19(3):e3001115.
283. Ramazzotti D, Angaroni F, Maspero D, Mauri M, D'Aliberti D, Fontana D, et al. Large-Scale Analysis of SARS-CoV-2 Synonymous Mutations Reveals the Adaptation to the Human Codon Usage During the Virus Evolution. Virus Evol. 2022;8(1):veac026.
284. Bai H, Ata G, Sun Q, Rahman SU, Tao S. Natural selection pressure exerted on "Silent" mutations during the evolution of SARS-CoV-2: Evidence from codon usage and RNA structure. Virus Res. 2022;323:198966.
285. Mogro EG, Bottero D, Lozano MJ. Analysis of SARS-CoV-2 synonymous codon usage evolution throughout the COVID-19 pandemic. Virology. 2022;568:56–71.
286. Hussain S, Rasool ST, Pottathil S. The Evolution of Severe Acute Respiratory Syndrome Coronavirus-2 during Pandemic and Adaptation to the Host. J Mol Evol. 2021;89:341–56.
287. Huang W, Guo Y, Li N, Feng Y, Xiao L. Codon usage analysis of zoonotic coronaviruses reveals lower adaptation to humans by SARS-CoV-2. Infect Genet Evol. 2021;89:104736.
288. Carmi G, Gorohovski A, Mukherjee S, Frenkel-Morgenstern M. Non-optimal codon usage preferences of coronaviruses determine their promiscuity for infecting multiple hosts. FEBS J. 2021;288(17):5201–23.
289. Delgado Blanco J, Hernandez-Alias X, Cianferoni D, Serrano L. In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. PLoS Comput Biol. 2020;16(12):e1008450.
290. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The Translational Landscape of the Human Heart. Cell. 2019;178(1):242-260.e29.
291. MGlincy NJ, Ingolia NT. Transcriptome-wide measurement of translation by ribosome profiling. Methods. 2017;126:112–29.
292. Allen SR, Stewart RK, Rogers M, Ruiz IJ, Cohen E, Laederach A, et al. Distinct responses to rare codons in select Drosophila tissues. eLife. 2022;11:e76893.

293. Mauro VP. Codon Optimization: Codon Optimization of Therapeutic Proteins: Suggested Criteria for Increased Efficacy and Safety. In: Sauna ZE, Kimchi-Sarfaty C, editors. Single Nucleotide Polymorphisms. Cham: Springer; 2022. p. 197–224.
294. Lin BC, Kaissarian NM, Kimchi-Sarfaty C. Implementing computational methods in tandem with synonymous gene recoding for therapeutic development. Trends Pharmacol Sci. 2022;S0165-6147(22)00204-8.
295. Picard MAL, Leblay F, Cassan C, Willemsen A, Daron J, Bauffe F, et al. The multi-level phenotypic impact of synonymous substitutions: heterologous gene expression in human cells. bioRxiv. 2022;2022.01.07.475042.
296. Iben JR, Maraia RJ. tRNA gene copy number variation in humans. Gene. 2014;536(2):376–84.
297. Ardlie K. A portal and integrative collaborative analysis platform for GTEx [Internet]. [cited 2022]. Available from: https://reporter.nih.gov/search/FSa0vivtb0uWpsyy7wgS8g/project-details/10405210
298. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818):699–710.
299. Maraia RJ, Arimbasseri AG. Factors That Shape Eukaryotic tRNAomes: Processing, Modification and Anticodon–Codon Use. Biomolecules. 2017;7(1):26.
300. Han L, Phizicky EM. A rationale for tRNA modification circuits in the anticodon loop. RNA. 2018;24(10):1277–84.
301. Kleiber N, Lemus-Diaz N, Stiller C, Heinrichs M, Mai MM-Q, Hackert P, et al. The RNA methyltransferase METTL8 installs m3C32 in mitochondrial tRNAsThr/Ser(UCN) to optimise tRNA structure and mitochondrial translation. Nat Commun. 2022;13(1):209.
302. Martin S, Allan KC, Pinkard O, Sweet T, Tesar PJ, Coller J. Oligodendrocyte differentiation alters tRNA modifications and codon optimality-mediated mRNA decay. Nat Commun. 2022;13(1):5003.
303. Lyons SM, Fay MM, Ivanov P. The role of RNA modifications in the regulation of tRNA cleavage. FEBS Lett. 2018;592(17):2828–44.
304. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968;217(5129):624–6.
305. King JL, Jukes TH. Non-Darwinian evolution. Science. 1969;164(3881):788–98.
306. Eyre-Walker A, Hurst LD. The evolution of isochores. Nat Rev Genet. 2001;2(7):549–55.
307. Brbić M, Warnecke T, Kriško A, Supek F. Global Shifts in Genome and Proteome Composition Are Very Tightly Coupled. Genome Biol Evol. 2015;7(6):1519–32.
308. Kruglyak L, Beyer A, Bloom JS, Grossbach J, Lieberman TD, Mancuso CP, et al.

No evidence that synonymous mutations in yeast genes are mostly deleterious. bioRxiv. 2022;2022.07.14.500130.
309. Cannarozzi GM, Schneider A. Codon Evolution: Mechanisms and Models. Oxford University Press; 2012.
310. Wetzel C, Limbach PA. Mass spectrometry of modified RNAs: recent developments. The Analyst. 2016;141(1):16–23.
311. Thomas NK, Poodari VC, Jain M, Olsen HE, Akeson M, Abu-Shumays RL. Direct Nanopore Sequencing of Individual Full Length tRNA Strands. ACS Nano. 2021;15(10):16642–53.
312. Yankova E, Blackaby W, Albertella M, Rak J, De Braekeleer E, Tsagkogeorga G, et al. Small-molecule inhibition of METTL3 as a strategy against myeloid leukaemia. Nature. 2021;593(7860):597–601.
313. Girstmair H, Saffert P, Rode S, Czech A, Holland G, Bannert N, et al. Depletion of cognate charged transfer RNA causes translational frameshifting within the expanded CAG stretch in huntingtin. Cell Rep. 2013;3(1):148–59.
314. Porter JJ, Heil CS, Lueck JD. Therapeutic promise of engineered nonsense suppressor tRNAs. Wiley Interdiscip Rev RNA. 2021;12(4):e1641.
315. Lueck JD, Yoon JS, Perales-Puchalt A, Mackey AL, Infield DT, Behlke MA, et al. Engineered transfer RNAs for suppression of premature termination codons. Nat Commun. 2019;10(1):822.
316. Ignatova Z, Torda A, Matthies M. Synthetic transfer RNA with extended anticodon loop. United States patent US 11,434,485 B2, 2022.
317. Wang J, Zhang Y, Mendonca CA, Yukselen O, Muneeruddin K, Ren L, et al. AAV-delivered suppressor tRNA overcomes a nonsense mutation in mice. Nature. 2022;604:343–8.