UNIVERSITAT DE BARCELONA

Exploring the human tRNAome:
From transcriptional and processing dynamics to genomic organization and somatic mutagenesis

Marina Murillo Recio

Exploring the human tRNAome: From transcriptional and processing dynamics to genomic organization and somatic mutagenesis

Marina Murillo Recio

Universitat de Barcelona
Facultat de Química

Programa de doctorat en Biomedicina

Institut de Recerca Biomèdica de Barcelona

Exploring the human tRNAome: From transcriptional and processing dynamics to genomic organization and somatic mutagenesis

Director
Lluís Ribas de Pouplana

Director
Adrian Gabriel Torres

Tutor
Modesto Orozco

PhD candidate
Marina Murillo Recio

2025

Universitat de Barcelona
Facultat de Química

Programa de doctorat en Biomedicina

Institut de Recerca Biomèdica de Barcelona

Explorant el tRNAoma humà: Des de la dinàmica transcripcional i de processament fins a l'organització genòmica i la mutagènesi somàtica

Doctoranda
Marina Murillo Recio

“What we know is a drop, what we don’t know is an *ocean”*

Isaac Newton

A mis padres, por ser mi *calma*.
Y a todos los que me acompañaron y guiaron en este camino.

AGRADECIMIENTOS

Y aquí llega el final de una etapa, un camino lleno de aprendizaje, superación y también de experiencias inolvidables. Sin duda, recorrer este camino ha sido mucho más fácil gracias a todas las personas que han estado a mi lado y han aportado su granito de arena.

En primer lugar, quiero agradecer a mis padres. Siento que las palabras nunca serán suficientes para agradeceros todo lo que habéis hecho por mí. Gracias por ser siempre mi calma, mi refugio y mi apoyo. Gracias por acompañarme en todas mis etapas, por estar siempre, por recordarme que “el saber no ocupa lugar” y por mantener viva mi curiosidad. Por enseñarme a ver siempre el lado positivo y por hacer que todo sea más ameno y bonito con vuestro sentido del humor. Gracias también por animarme a seguir adelante y por celebrar conmigo cada logro, por pequeño que sea. Sin vosotros, nada de lo que soy ni de lo que he conseguido sería posible.

Vull agrair a en Lluís per donar-me l’oportunitat de formar part del Ribas Lab. Per guiar-me, però al mateix temps donar-me la llibertat necessària per aprendre i créixer. Per compartir les teves idees i deixar-me explorar. Gràcies també per acostar-me a un equip increïble, la meva família científica, i als meus grans “descobriments”: l’Adrian, en Lluís, la Noelia i l’Alba. Em sento molt afortunada i agraïda d’haver compartit tot això amb vosaltres. Gràcies per recolzar-me, per fer-me sentir estimada i per la vostra paciència (que sé que aquests últims mesos ha estat molta, jajaja). No us imagineu com d’agraïda estic de tenir-vos.

Gracias, Adrian, por ser un pilar fundamental durante esta etapa. Me diste todas las herramientas necesarias desde el inicio para poder crecer como científica y afrontar los retos de la tesis con seguridad. Por darme la confianza que necesitaba en mí misma. Por todas esas charlas científicas que me recordaban el porqué de esta tesis y por compartir conmigo tu pasión por la ciencia. Por enseñarme a mantener la calma ante un “quilombo”. Por estar siempre disponible y ser un supervisor en todos los sentidos, desde psicólogo hasta guía gastronómico, y a quien ahora considero un gran amigo. Gracias también a Patri, por estar en esas quedadas que tantas veces fueron terapia.

Luis, qué agradecida estoy de que te cruzaras en mi camino y de haber podido recorrer gran parte de el junto a ti. Qué importante ha sido para mí tenerte a mi lado, tanto fuera como dentro del lab, y qué bonito ha sido crecer juntos, casi como un hermano. Mi *drama queen*, gracias por sacarme siempre una sonrisa. Es una maravilla poder rodearse de personas tan buenas y bonitas como tú. *Time flies when I’m with you!* Por más aventuras juntos y más atardeceres en el Jardim do Morro, en tu querido Porto. Muito obrigada, adoro-te, amigo.

A Noelia, la luz del lab, ese rayito de sol que se ve al entrar al laboratorio. Gracias por las mini charlas matutinas que me llenan de energía y por hacerme reír cuando más lo necesitaba. He aprendido tanto de ti… Gracias por acompañarme en este “viaje”, pasar tiempo contigo siempre es un “regalo”. ¡Gracias por formar parte de mi “Isla”!

A Alba, por estar siempre dispuesta a ayudarme. Por todos tus consejos que me han ayudado a mantener la calma. Por estar desde los inicios y por todas esas quedadas que me recargaban las pilas. Per continuar trobant-nos a plaça d'Osca, recreant moments èpics de Paquita Salas.

También quiero agradecer a todas las personas que formaron parte de Ribas Lab, y que, en algún momento, recorrieron este camino conmigo, en especial a Neus, Lina, Aure, Ash, Despina, Georgios, Aina, Jan, Khadija, Sundar y Nerea. Gracias por crear en todo momento un espacio donde se puede ser uno mismo, rodeada de personas maravillosas que me motivan y enseñan día a día.

A todas aquellas personas que hacen que estar en el IRB sea como estar en casa. A la facility de Biostatistics/Bioinformatics: Óscar y Camille gracias por darme la oportunidad de rodearme de vuestro conocimiento y experiencia. A los miembros del TAC, Toni, Fran y Eulalia, por enseñarme a ver el proyecto desde diferentes perspectivas. Gracias a Fran y Marina S por darme la oportunidad de colaborar con ellos y sumergirme en el mundo de la genómica. A todas esas personas que compartieron momentos conmigo tanto dentro como fuera del IRB, Niko, Karen (I'll miss your sweetnes and your amazing cakes!) y al grupo de los "Gabaldones" por acogerme como si fuera una más, en especial a Olfat, Vlad y Islam.

Gràcies, Olfat, per ser el meu far i la meva guia, per estar sempre al meu costat. Que agraïda estic d'haver trobat una companya d'aventures com tu, sobretot en aquesta aventura que ja fa més de deu anys que va començar. Quins records més bonics hem creat, ens hem descobert l'una a l'altra i també a la nostra passió per la bioinformàtica. Gràcies, com sempre et dic, per ser una inspiració, no només per a mi sinó per a tothom qui t'envolta. Gràcies per motivar-me i per animar-me a emprendre camins i viure experiències que per a mi eren un gran repte. Per compartir amb mi la teva passió per comunicar la ciencia. Per seguir sent "chichoses" fins a l'eternitat!

També vull agrair a la Cris L., l'Irene, la Queralt, la Cris O, el Salva, en Carles i en Brian, perquè vau ser essencials en etapes anteriors i, d'una manera o altra, també m'heu acompanyat al llarg d'aquesta.

A mi familia, por acompañarme siempre y por enseñarme el verdadero significado de equipo. ¡Os quiero mucho a todos! A mi hermana, mi referente. Admiro tu forma de ver la vida siempre en positivo y con esa energía que se contagia. Gracias, Javi, por cuidarnos. A mis sobrinas, Ainara y Leire, porque cada momento con vosotras siempre implica diversión. A Laura, por ser más que mi familia, una amiga desde siempre y para siempre. Aunque nos separe un trocito de mar, gracias por hacerme sentir que siempre estas aquí conmigo. Gracias por cuidarme y por enseñarme tanto. Te admiro muchísimo, amiga. A Maria, que para mí ya es familia. Porque en todos los momentos importantes siempre has estado ahí. Gracias por tu amistad, por ser hogar para mí y por ser la mejor organizadora de planes, esos que siempre me ayudan a desconectar.

Gracias por todo!

SUMMARY

Transfer RNAs (tRNAs) play a fundamental role in protein synthesis. They mediate the decoding of messenger RNA into specific amino acids through codon-anticodon interactions, according to the genetic code. The cellular tRNA pool is mainly determined by the transcription rate of tRNA genes (tDNAs), which is governed by many regulatory layers. Besides being transcribed, to be fully active, tRNAs undergo several post-transcriptional processing steps that include the addition of chemical modifications. Alterations in any of those steps have been implicated in diseases such as cancer and neurological disorders. Nevertheless, due to the complexity of tRNA biology, characterizing the cellular tRNA pool is quite complex and is often accompanied by methodological challenges. Therefore, many aspects of tRNA biology remain unknown and require further research to understand their full impact on diseases. For this reason, in this thesis, we focus on the application and development of bioinformatic strategies to decipher different aspects of tRNA biology.

Currently, one of the most used techniques for analyzing the tRNA pool is small RNA sequencing (tRNA-Seq). However, the analysis of tRNA-Seq data requires the adaptation of the computational workflow from standard approaches. For this reason, we first developed tRNAstudio, an integrative pipeline that includes a mapping strategy that allows the characterization of tRNA processing and modification landscape, as well as the implementation of tRNA differential expression analysis. All pipeline components were integrated into a graphical user interface (GUI) that eases the analysis of tRNA-Seq data for non-computational users.

Next, we focused on the functional relevance of the modification of adenine (A) to inosine (I) at position 34 of tRNA anticodons (I34-tRNAs), a modification crucial for the expansion of the decoding capacity of tRNAs. In Eukarya, this modification is catalyzed by Adenosine Deaminase Acting on tRNA (ADAT), composed of ADAT2 and ADAT3. Mutations in genes encoding ADAT have been associated with neurological disorders. To comprehend the impact of changes in the levels of I34-tRNAs, a knockdown model of ADAT2 (KD) was generated by our group. Using tRNAstudio, we analyzed the impact of ADAT2 KD on the tRNA pool. The results validated the model by verifying a reduction of I34-tRNAs in the context of ADAT2 KD. Moreover, we observed that the cell is unable to compensate for its loss by upregulating the expression of alternative tRNAs, highlighting the critical role of I34-tRNAs.

Lastly, given that many factors can regulate tDNA transcription and that tDNAs are not randomly distributed within the genome, we wanted to assess whether their organization within the genome may impact their transcriptional activity. To investigate this, we first characterized the localization of tDNAs in the latest human genome assembly (T2T-CHM13), identifying patterns of tDNA clustering. We then determined that these clusters are associated with increased transcriptional activity. Building on these findings, we explored whether the transcriptional activity of tDNAs also influences their susceptibility to somatic mutagenesis. Our analysis revealed that tDNAs are exceptionally prone to accumulate somatic mutations, with mutation rates up to nine-fold higher than those of protein-coding genes. Moreover, mutation rate at tDNAs increased with transcriptional activity, and mutational loads were tumor-type and age-dependent. The analysis of mutational signatures identified APOBEC3 activity as the main

contributor to tDNA somatic mutagenesis. Notably, Mutations at structurally conserved tRNA positions appear to be under negative selection, preventing mutation accumulation at these critical sites. By contrast, other positions were hypermutated, which could disrupt tRNA biogenesis and impair tRNA function, potentially contributing to cellular dysfunction. We propose that the accumulation of somatic mutagenesis at tDNAs could cause proteome heterogeneity and compromise proteostasis, factors that could contribute both to tissue aging and to the development or progression of cancer.

Keywords: tRNAs, biogenesis, genomics, somatic mutations, mistranslation

RESUM

Els ARN de transferència (tRNAs) tenen un paper fonamental en la síntesi de proteïnes. Descodifiquen l'ARN missatger en aminoàcids mitjançant les interaccions codó-anticodó, en base el codi genètic. El conjunt de tRNAs cel·lulars estan determinats principalment per la transcripció dels gens de tRNA (tDNAs), controlada per diversos mecanismes reguladors. A més de ser transcrits, per ser completament actius, els tRNA experimenten un processament post-transcripcional, incloent l'addició de modificacions químiques. Alteracions en la biogènesi dels tRNAs s'han associat amb malalties com el càncer i trastorns neurològics. Degut a la complexitat de la seva biologia, la caracterització del repertori cel·lular de tRNAs sovint comporta reptes metodològics. En conseqüència, la biologia dels tRNA presenta encara múltiples incògnites que cal resoldre mitjançant estudis més detallats per entendre la seva participació en els mecanismes moleculars de les malalties. Per aquest motiu, en aquesta tesi, ens centrem en l'aplicació i el desenvolupament d'estratègies bioinformàtiques per desxifrar diferents aspectes de la biologia dels tRNAs.

Actualment, una de les tècniques més utilitzades per l'anàlisis de tRNAs es la seqüenciació d'ARN petit (tRNA-Seq). Tanmateix, l'anàlisi de dades de tRNA-Seq requereix l'adaptació de procediments computacionals estàndard. Per aquest motiu, primer vam desenvolupar tRNAstudio, una *pipeline* integrativa que inclou una estratègia de mapatge que permet la caracterització del processament i la identificació de modificacions de tRNA, així com la implementació de l'anàlisi d'expressió diferencial de tRNAs. Tots els components es van integrar en una interfície gràfica d'usuari (GUI) que facilita l'anàlisi de dades de tRNA-Seq per a usuaris no computacionals.

A continuació, ens vam centrar en la rellevància funcional de la modificació d'adenina (A) a inosina (I) a la posició 34 dels anticodons de tRNA (I34-tRNAs). Una modificació crucial per a l'expansió de la capacitat de descodificació dels tRNAs. En Eukarya, aquesta modificació està catalitzada per l'adenosina desaminasa específica per tRNA (ADAT), composta per ADAT2 i ADAT3. Les mutacions en els gens que codifiquen per ADAT s'han relacionat amb trastorns neurològics. Per comprendre l'impacte dels canvis en els nivells d'I34-tRNAs, el nostre grup va generar un model de d'ADAT2 *knockdown* (KD). Utilitzant tRNAstudio, vam analitzar l'impacte de l'ADAT2 KD en el conjunt de tRNAs. Els resultats van validar el model verificant una reducció dels I34-tRNAs. A més, vam observar que la cèl·lula no pot compensar la pèrdua regulant l'expressió de tRNAs alternatius, destacant el paper crític dels I34-tRNAs.

Finalment, considerant que molts factors poden regular la transcripció dels tDNAs i que els tDNAs no es distribueixen aleatòriament dins del genoma, vam voler avaluar si la seva organització dins del genoma pot afectar la seva activitat transcripcional. Per examinar aquesta qüestió, primer vam caracteritzar la localització dels tDNAs en el darrer assemblatge del genoma humà (T2T-CHM13), identificant clústers de tDNAs. Després vam determinar que aquests clústers estan associats amb una major activitat transcripcional. A partir d'aquestes troballes, vam explorar si l'activitat transcripcional dels tDNAs també influeix en la seva susceptibilitat a la mutagènesi somàtica. Els resultats van revelar que els tRNAs són excepcionalment propensos a acumular mutacions somàtiques, amb taxes de mutació fins a nou vegades superiors a les dels gens que codifiquen per a proteïnes. A més, les taxes de mutació

dels tRNAs augmenten amb l'activitat transcripcional, i els nivells de mutació depenen del tipus tumoral i de l'edat. L'anàlisi de les signatures mutacionals va identificar l'activitat d'APOBEC3 com el principal contribuent a la mutagènesi somàtica dels tDNAs. Cal destacar que les mutacions en posicions estructuralment conservades semblen estar sota selecció negativa. En canvi, altres posicions estan hipermutades, cosa que podria afectar la funció dels tRNAs, contribuint potencialment a alteracions en la síntesi de proteïnes. Proposem que l'acumulació de mutagènesi somàtica als tDNAs podria causar heterogeneïtat del proteoma i comprometre la proteòstasi, factors que podrien contribuir tant a l'envelliment dels teixits com al desenvolupament o la progressió del càncer.

Paraules clau: tRNAs, biogènesi, genòmica, mutacions somàtiques, proteòstasi

ABBREVIATIONS

A	Adenine
A3A	APOBEC3A (Apolipoprotein B mRNA Editing Catalytic Polypeptide-like 3A)
A3B	APOBEC3B (Apolipoprotein B mRNA Editing Catalytic Polypeptide-like 3B)
A3G	APOBEC3G (Apolipoprotein B mRNA Editing Catalytic Polypeptide-like 3G)
A3H	APOBEC3H (Apolipoprotein B mRNA Editing Catalytic Polypeptide-like 3H)
A34	Adenosine at position 34
aa	Amino acid
aa-AMP	Aminoacyl-adenylate
aaRS	Aminoacyl-tRNA Synthetase
aa-tRNA	Aminoacyl-tRNA
ADAT	Adenosine Deaminase Acting on tRNA
ADAT1	Adenosine Deaminase tRNA specific 1
ADAT2	Adenosine Deaminase tRNA specific 2
ADAT3	Adenosine Deaminase tRNA specific 3
ADP	Adenosine Diphosphate
Ago	Argonaute
ANG	Angiogenin
APOBEC	Apolipoprotein B mRNA editing catalytic polypeptide-like
ATP	Adenosine Triphosphate
BER	Base excision DNA repair
BLCA	Bladder Cancer (Bladder Urothelial Carcinoma)
C	Cytosine
CF	Cystic Fibrosis
Chr	Chromosome
cDNA	Complementary DNA
COPD	Chronic Obstructive Pulmonary Disease
DNA	Deoxyribonucleic Acid
ECM	Extracellular matrix
eEFs	Eukaryotic Elongation Factors
eIFs	Eukaryotic Initiation Factors
eRFs	Eukaryotic Release Factors
E-site	Exit site (ribosome)
ELAC2	ElaC Ribonuclease Z 2
G	Guanine
GtRNAdb	Genomic tRNA Database

GUI	Graphical User Interface
hg19	Human Genome version 19
hg39	Human Genome version 39
I	Inosine
I34	Inosine at position 34 of tRNA
I34-tRNAs	tRNAs with inosine at position 34
iMet	Initiator Methionine
kb	Kilobase
MMR	DNA Mismatch Repair
MSS	Microsatellite Stability
MSI	Microsatellite Instability
mRNA	Messenger RNA
mt-tDNAs	Mitochondrial tRNA genes
mt-tRNAs	Mitochondrial tRNAs
NMF	Non-negative Matrix Factorization
POLE	DNA Polymerase epsilon
POLRMT	Mitochondrial RNA polymerase
Pol II	RNA Polymerase II
Pol III	RNA Polymerase III
pre-tRNA	Precursor tRNA
P-site	Peptidyl site (ribosome)
rDNA	Ribosomal DNA
RNases	Ribonucleases
RNase P	Ribonuclease P
RNase Z	Ribonuclease Z
RNUs	Small nuclear RNA genes
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
RT	Reverse Transcriptase
SHOT-RNAs	Sex Hormone-dependent tRNA-derived RNAs
SNP	Single Nucleotide Polymorphism
ssDNA	Single-stranded DNA
SVs	Structural variants
T	Thymine
T2T	Telomere-to-Telomere
TADs	Topologically Associating Domains
TAM	Transcription Associated Mutagenesis

TCR	Transcription-coupled repair
tgCNV	tRNA gene Copy Number Variation
tiRNAs	tRNA halves
tRFs	tRNA-derived fragments
tDNAs	tRNA genes
tRNAs	Transfer RNAs
tRNA-Seq	tRNA Sequencing
tsRNAs	tRNA-derived small RNAs
U	Uracil
UNG	Uracil-DNA glycosylase
VNTRs	Variable Number Tandem Repeats
WGS	Whole Genome Sequencing
TSEN	tRNA splicing endonuclease complex
miscRNA	Miscellaneous RNA

Amino acids

Three-Letter Code	One-Letter Code	Amino Acid Name
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Sec	U	Selenocysteine

TABLE OF CONTENTS

1 GENERAL INTRODUCTION

The central dogma of molecular biology describes the flow of genetic information. It involves two major processes: transcription and translation (F. Crick, 1970) **(Fig. 1)**. Both processes are used by the cell in order to obtain proteins from the information contained in the genome (DNA). During transcription, the information in DNA is transferred into an RNA molecule. This is followed by translation, where the resulting messenger RNA (mRNA) from protein-coding genes is delivered to the ribosome for decoding. During this decoding process, transfer RNAs (tRNAs) are used as adaptor molecules that, following the genetic code, translate the mRNA sequence into amino acids for protein synthesis.

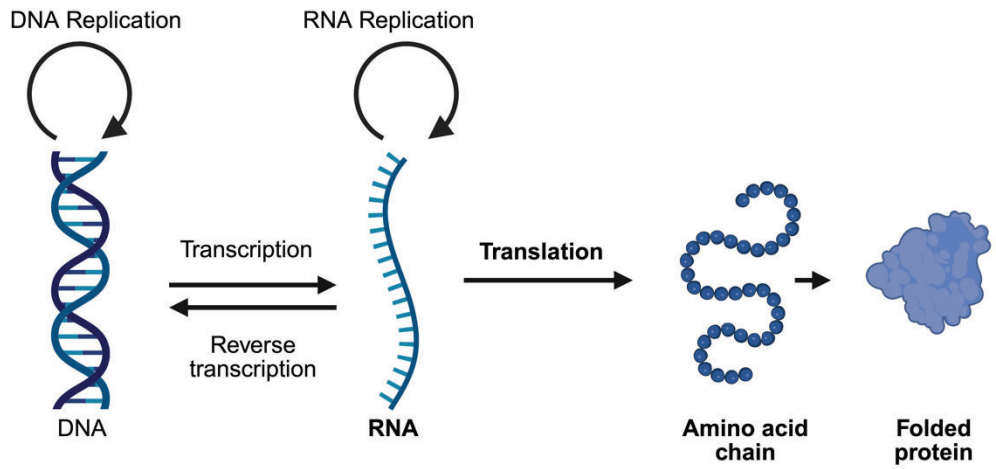


Figure 1. The central dogma of molecular biology. The central dogma describes the process by which genetic information flows from DNA to RNA through transcription and from RNA to amino acids through translation (F. Crick, 1970). Created with BioRender.

The genetic code **(Fig. 2)** is the bridge between nucleic acids and amino acids, providing the set of instructions that enables cells to translate mRNA sequences into functional proteins. This code is nearly universal among all living organisms, reflecting a deep evolutionary conservation with only rare and minor exceptions (Vetsigian et al., 2006). The genetic code is based on codons, that are specific sequences of three consecutive nucleotides in mRNA, each of which corresponds to a particular amino acid (F. H. Crick et al., 1961; Nirenberg & Matthaei, 1961). In DNA, these nucleotides are, guanine (G), cytosine (C), thymine (T), and adenine (A). In RNA, uracil (U) replaces thymine, which results in the nucleotides A, C, U, and G. One of the main characteristics of the genetic code is its unambiguous nature, as each codon defines only one amino acid in order to ensure precise translation (e.g., the codon CUA always codes for leucine). However, there are a total of 61 codons that code for the 20 amino acids, denoting that the code is also redundant or degenerate. This means that multiple codons can code for the same amino acid. For instance, leucine can be coded by six codons: UUA, UUG, CUU, CUC, CUA, and CUG. Furthermore, the genetic code includes specific start and stop signals. The codon

AUG serves as both the code for methionine and the start signal for protein synthesis, whereas 3 specific stop codons (UAA, UAG, UGA) are used to trigger the end of protein synthesis (Alberts et al., 2002).

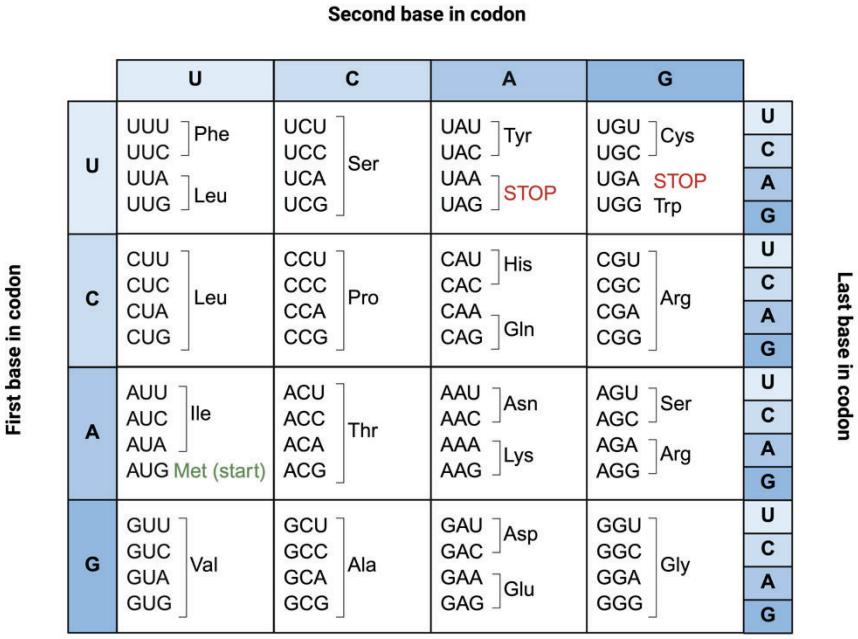


Figure 2. The standard genetic code. Universal set of rules represented as a comprehensive dictionary. Illustrates the translation of each of the 64 possible codons into their corresponding amino acid or a stop signal. Created with BioRender.

The process of translating these codons into their corresponding amino acids depends on tRNAs. Therefore, tRNAs play a fundamental role in protein synthesis, working as genetic code decoders, recognized as ancient, universal molecules essential in all domains of life (Eigen et al., 1989). In addition to their role in translation, tRNAs are involved in other cellular processes, highlighting the relevance and complexity of their biology (Su et al., 2020).

In the following sections, we will describe key concepts of tRNA biology and the theoretical background on which this thesis is based.

1.1 tRNA sequence and structure

tRNAs play a fundamental role in protein synthesis, and both their sequence and structure are essential for this function **(Fig. 3)**. Each tRNA molecule carries its specific cognate amino acid and contains the three-base anticodon sequence that recognizes the appropriate codon on the mRNA. To ensure the correct amino acid is incorporated into the growing polypeptide chain, the anticodon of the tRNA pairs with the mRNA codon via canonical Watson-Crick base-pairing (A:U and G:C) (F. H. Crick, 1958). Furthermore, non-canonical Watson-Crick base-pairing can occur at the first position of the tRNA anticodon (position 34), to allow the expansion of their decoding capacity (F. H. Crick, 1966).

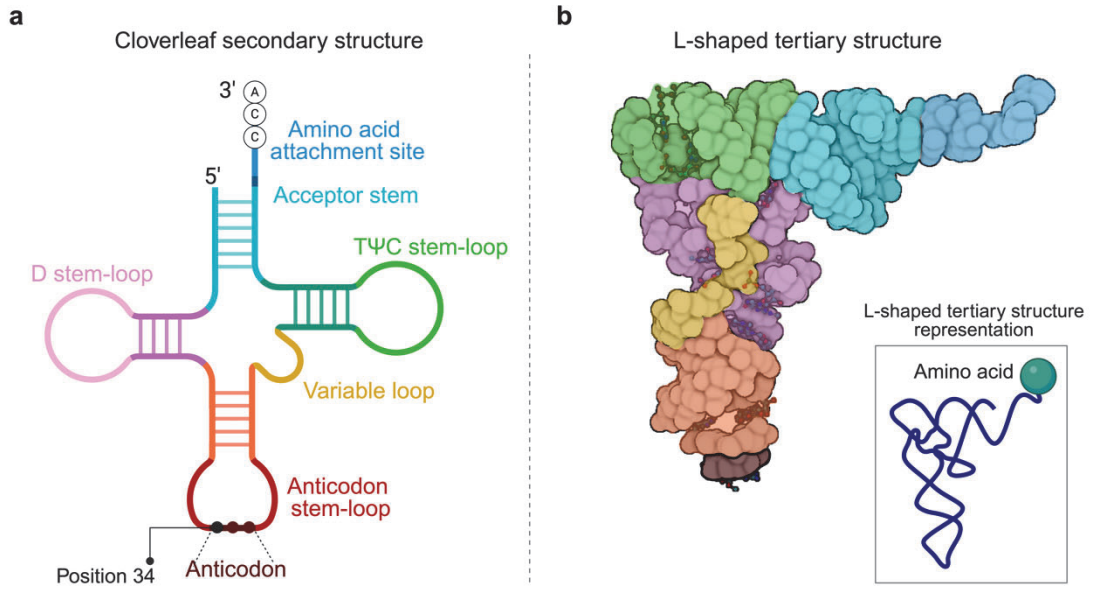


Figure 3. Schematic representation of the tRNA structure. (a) Consensus cloverleaf-shaped representation, with the key distinctive regions indicated in the figure. **(b)** L-shaped 3D structure. Created with BioRender.

The primary structure of tRNA is characterized by a single-stranded RNA sequence of around 76 nucleotides (with variations observed between 75-95 nucleotides) (Sprinzl et al., 1998). The secondary tRNA structure is defined by a hydrogen-bonded stem-loop conformation, which is commonly represented by the well-known cloverleaf shape **(Fig. 3a)**. The stem-loops define five distinct regions with different functions (Berg & Brandl, 2021; Giegé et al., 2012; Holley et al., 1965). First, from the base-pairing between the tRNA 5' and 3' terminal regions, results the acceptor stem that serves as the attachment site for the amino acid as it incorporates the essential 3'-CCA sequence required for the reaction of aminoacylation. Following is the dihydrouridine (D) stem-loop, named for its dihydrouridine base that plays a crucial role in stabilizing the tRNA tertiary structure. Next, on the opposite end of the acceptor stem is the anticodon stem-loop containing the anticodon sequence (position 34-36) that base-pairs with the cognate mRNA codon. Finally, the thymidine-pseudouridine-cytidine (TΨC) stem-loop is

essential for the interaction between tRNA and the ribosome, among other components of the translational machinery (Feinberg & Joseph, 2001). Between the anticodon stem and the TΨC stem-loop larger eukaryotic tRNAs for serine (Ser), selenocysteine (Sec), and leucine (Leu) have a variable loop with diverse sizes and structures (Giegé et al., 2012). By stabilized intramolecular interactions primarily between the TΨC- and D-arms the secondary configuration folds into an L-shaped tertiary structure **(Fig. 3b)** (Kim et al., 1974).

Furthermore, along the tRNA sequence, there are several nucleotide positions that are highly conserved, highlighting their functional and structural importance. For instance, U8, A14, G18, G19, A21, U33, G53, T54, U55, C56, A58, C61, C74, C75, and A76 are the most conserved nucleotides within tRNAs and Y11, R15, R24, Y32, R37, Y48, R57, and Y60 (R = purine; Y = pyrimidine) are also identified as semiconserved nucleotides present in numerous tRNA sequences (Biela et al., 2023; Giegé et al., 2012).

tRNAs can be further classified into groups of isoacceptors and isodecoders according to their sequence similarities (Goodenbour & Pan, 2006). Isoacceptors are tRNAs charged with the same amino acid but with different anticodon sequences (e.g., the isoacceptor group for valine (Val) include: tRNA-Val-AAC, tRNA-Val-GAC, tRNA-Val-CAC, and tRNA-Val-TAC). Whereas, isodecoders have the same anticodon but have differences in other regions of the sequence (e.g., tRNA-Val-AAC-1, tRNA-Val-AAC-2) (Goodenbour & Pan, 2006).

1.2 Transfer RNA genes (tDNAs)

tRNAs genes (tDNAs) are encoded in both the nucleus and mitochondrial genomes (P. P. Chan et al., 2021; Ojala et al., 1981). While nuclear tRNAs participate in cytoplasmic protein synthesis, mitochondrial tRNAs (mt-tRNAs) participate specifically in mitochondrial translation (Ojala et al., 1981). In the human mitochondrial genome, each tRNA isoacceptor is represented by a single-copy gene, with a total of 22 mt-tRNAs genes (mt-tDNAs) (S. Anderson et al., 1981). In contrast, the nuclear genome contains multiple copies, and some of them with identical sequences that can be grouped into families based on 100% sequence similarity. For instance, five tRNA-Val-AAC-1 genes share identical sequences. However, other tDNAs have unique sequences and do not belong to a family but still contribute to the diversity of the tRNAome (P. P. Chan & Lowe, 2016).

The number of nuclear-encoded tDNAs varies widely among species, even between eukaryotes (Bermudez-Santana et al., 2010; Goodenbour & Pan, 2006). In humans, tDNA detection tools like tRNAscan-SE 2.0 have predicted around 600 nuclear-encoded tDNAs (P. P. Chan & Lowe, 2016; Geslain & Pan, 2010).

1.2.1 Genomic organization of tDNAs

Nuclear-encoded tDNAs are dispersed throughout the human genome across multiple loci and are found in all chromosomes except chromosome Y (P. P. Chan & Lowe, 2016). Although tDNAs are scattered throughout the genome, studies on the genomic arrangement of the set of tDNAs indicate an unexpected prevalence of proximate tDNAs and that most are linearly grouped in tDNA clusters (Bermudez-Santana et al., 2010; Dieci et al., 2013; Iben & Maraia, 2014; Mungall et al., 2003; Van Bortle & Corces, 2012). tDNA clusters are prevalent across various domains of life, having evolved independently in each domain and are more abundant in eukaryotes (Bermudez-Santana et al., 2010; Morgado & Vicente, 2019). Previous studies in humans using the genome reference assemblies GRCh38/hg38 and GRCh37/hg19 identified many tDNA clusters localized in different chromosomes. For example, a cluster located on chromosome 6 is composed of 157 tDNAs, including almost all tDNA species except for asparagine (tDNA-Asn) and cysteine (tDNA-Cys) (Acton et al., 2021; Mungall et al., 2003). Additionally, other tDNA clusters have been found on chromosomes 1, 5, 7, 14, 16, and 17 (Gao et al., 2024; Iben & Maraia, 2014; Mungall et al., 2003). This non-random organization of tDNAs suggests that the genomic context may play an important role in their function, regulation and evolution (Bermudez-Santana et al., 2010; Van Bortle et al., 2017).

1.2.2 Genetic variability and mutagenesis in tDNAs

tDNAs are subject to genetic variability between individuals. In this context, tDNAs can exhibit copy number variations (tgCNVs) that refer to structural variations in the genome, where sections of DNA are duplicated or deleted, resulting in differences in the number of tDNAs copies (Iben & Maraia, 2012, 2014). tgCNV is of great interest because of its potential impact on tRNA availability, which can affect the translation efficiency of specific mRNAs (Iben & Maraia, 2014; Novoa et al., 2012). For example, the loss of only one gene for tRNA phenylalanine (tRNA-Phe) has been previously described to affect neuronal function in mice (Hughes et al., 2023). Also, in mice, the lack of tRNA lysine (Lys-UUU) in chromosome 7 was associated with type 2 diabetes, but in humans, it did not have an associated phenotype linked with diseases (Lant et al., 2019; P. Yang et al., 2019).

tgCNV is not limited to single genes but also affects larger tDNAs clusters. For instance, a cluster located on chromosome 1 (Chr1:161,413,094–161,440,995; hg19) exhibits tgCNV across individuals (Iben & Maraia, 2014). This locus is described as a tandem repeat of four units, each unit containing a set of five tDNAs for glutamine (tRNA-Glu-CTC), glycine (tRNA-Gly-TCC and tRNA-Gly-GCC), asparagine (tRNA-Asp-GTC), and leucine (tRNA-Leu-CAG) (Gao et al., 2024; Iben & Maraia, 2014). In some cases, the repeats or portions of them have undergone varying degrees of duplication across different individuals. This leads to variable number tandem repeats that produce differences in the number of tDNAs copies within this cluster, as observed in both humans and mice (Darrow & Chadwick, 2014; Iben & Maraia, 2014). However, tgCNVs are still overlooked, and their characterization and association with specific disease phenotypes are still lacking.

Genetic variability among individuals at the nucleotide level has also been observed in both nuclear and mitochondrial tDNAs (Berg et al., 2019; Lant et al., 2019; Orellana et al., 2022; Parisien et al., 2013; Seplyarskiy et al., 2023; Suzuki et al., 2011). Approximately 200 mt-tDNA single nucleotide polymorphisms (SNPs) have been linked to various pathologies related to mitochondrial dysfunction (Richter et al., 2021; Suzuki et al., 2011). The fact that mt-tRNAs are encoded by a single gene copy for each isoacceptor facilitates the study of these genes and their association with diseases. In contrast, the multicopy nature of nuclear-encoded tDNAs makes such research more challenging. Although nuclear-encoded tDNAs are under strong purifying selection in human germline cells, studies have demonstrated that these genes exhibit mutation rates 7 to 10 times higher than the genome average, resulting in variation between individuals (Berg et al., 2019; Lant et al., 2019; Seplyarskiy et al., 2023; Thornlow et al., 2018). This high conservation rate in germinal cells suggests that mutations in nuclear-encoded tDNAs can have deleterious effects. In fact, some studies have associated SNPs in nuclear tDNAs with disease. A SNP (T>C; rs46447118) in one gene for tRNA arginine induces ribosome stalling, resulting in a neurological phenotype that leads to neurodegeneration in mice (Ishimura et al., 2014). Another case is an SNP identified in the human gene for selenocysteine, which has been associated with symptoms such as abdominal pain and fatigue (Schoenmakers et al., 2016).

Beyond germline variation, tDNAs have also been identified as hotspots for somatic mutagenesis in yeast, bacteria, and human (Saini et al., 2017; Sakhtemani et al., 2019; Sui et al., 2020). The mechanism behind tDNA mutagenesis has been linked with the activity of Apolipoprotein B mRNA Editing Catalytic Polypeptide-like 3 (APOBEC3) enzymes (Saini et al., 2017; Sakhtemani et al., 2019; Thornlow et al., 2018). APOBEC3 enzymes are cytidine deaminases that mutate C-to-U in single-stranded DNA (ssDNA), leading to C-to-T or C-to-G mutations when repaired **(Fig. 4)** (Butler & Banday, 2023). They act on TCN context (where N is any nucleotide), and have a preference for TCW motifs (where W is A or T) (Petljak et al., 2022; Roberts et al., 2013). The main function of APOBEC3 members is to defend the organism against viral infections by inducing mutations in viral genomes. However, APOBEC3 enzymes can be off-target and produce mutations in the host genome (Roberts et al., 2013). From the APOBEC3 members, APOBEC3A (A3A) and APOBEC3B (A3B) have been repeatedly related with cancer mutagenesis (Behjati et al., 2014; Burns et al., 2013b; Petljak et al., 2022; Roberts et al., 2013; Supek & Lehner, 2017).

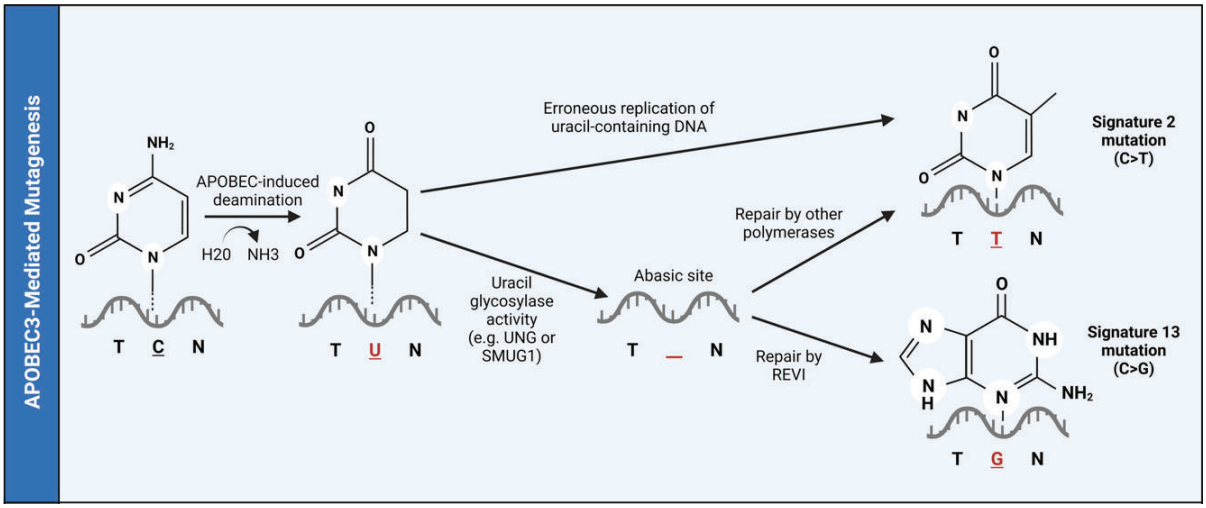


Figure 4. Mechanism of APOBEC3-mediated mutagenesis. APOBEC enzymes catalyze the deamination of C-to-U in ssDNA. In most cases replication occurs leading to C-to-T mutations. Alternatively, when uracil-DNA glycosylase (UNG) recognizes and removes uracil, it creates an abasic site that, when repaired leads to C-to-G mutations. As a result, APOBEC activity is associated with two patterns of DNA mutations (mutational signatures): SBS2 (C-to-T) and SBS13 (C-to-G), both occurring in a TCN sequence context (where N is any nucleotide) with a preference for TCW motifs (where W is A or T) (Alexandrov et al., 2020). Adapted from (Butler & Banday, 2023).

APOBEC3 mutagenesis increases with the presence of ssDNA secondary structures **(Fig. 5)**. For instance, during transcription, the RNA transcript that is being produced can bind to the DNA transcribed strand, producing DNA:RNA hybrids, also known as R-loops. This secondary DNA structure increases the exposure of the non-transcribed strand, making it more susceptible to APOBEC3 mutagenesis (McCann et al., 2023). Moreover, ssDNA regions contain inverted repeat sequences that can fold into DNA hairpins (stem-loop structures). The loops in these DNA hairpins can contain TCN motifs, providing high-affinity sites for the APOBEC3 enzymes (Buisson et al., 2019; Butt et al., 2024; Langenbucher et

al., 2021; Sakhtemani et al., 2019). In bacteria and yeast, the overexpression of A3A and A3B has been identified as a major cause of DNA damage and somatic mutagenesis in tDNAs (Saini et al., 2017; Sakhtemani et al., 2019). As previously mentioned, tDNAs have been reported to be hotspots of somatic mutagenesis in human samples. However, a deeper characterization of their mutagenic profile, the somatic mutational mechanisms acting upon tDNAs, and their tissue-specific impact is still lacking.

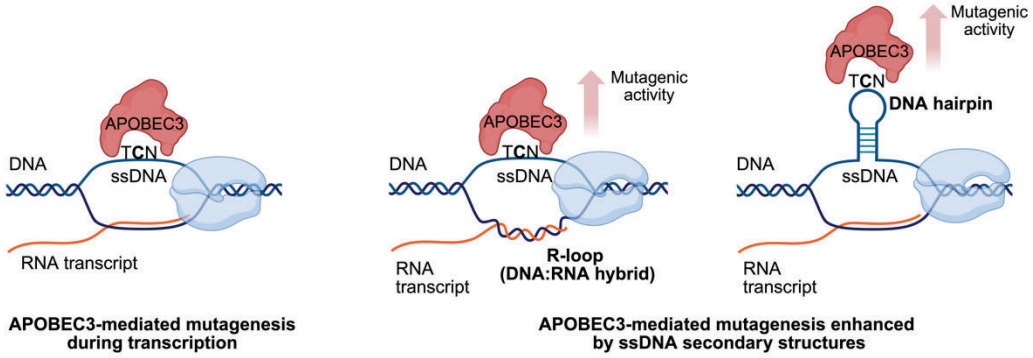


Figure 5. ssDNA secondary structures that enhance APOBEC3 mutagenesis. Schematic representation of the ssDNA secondary structures produced during transcription that can increase the mutagenic activity of APOBEC3.

1.2.3 Computational identification of tDNAs

Identifying genomic regions that encode tDNAs presents significant challenges. Some of these challenges involve the fact that tDNAs are typically very short and are found in highly repetitive regions (Lowe & Eddy, 1997). Additionally, there are sequences that reassemble tDNA genes, which include short interspersed nuclear elements (SINEs) that derive from tDNAs (P. P. Chan et al., 2021; Kramerov & Vassetzky, 2011; Lowe & Eddy, 1997), and other nuclear and mitochondrial "tDNA-lookalike" sequences found in the human genome (Telonis et al., 2014). In most cases, those sequences conserve functional motifs common of tDNAs, such as the internal promoter regions, but often have mutations that do not allow them to produce a functional secondary structure (P. P. Chan et al., 2021). Therefore, simply identifying a genomic sequence that matches a potential tDNA is not sufficient and can introduce false positives.

In this regard, different computational approaches have been developed to predict and identify *bona fide* tDNA loci within a genome. Among all of them the tDNA predictor most widely used is tRNAscan-SE (P. P. Chan et al., 2021). The original algorithm of tRNAscan-SE was first developed in 1991 by Fichant and Burks (Fichant & Burks, 1991), followed by the release of tRNAscan-SE in 1997 by Lowe and Eddy (Lowe & Eddy, 1997), which integrated multiple search algorithms that improved accuracy. Between 2019 and 2021, the last version of tRNAscan-SE incorporated significant upgrades in sensitivity and false-positive reduction (P. P. Chan et al., 2021; P. P. Chan & Lowe, 2019).

The default search of tRNAscan-SE identifies and classifies tDNAs from nuclear and mitochondrial genomic sequences by focusing on conservation and structural tRNA properties. To improve structural alignment accuracy, tRNAscan-SE 2 uses Infernal (INFERence of RNA ALignment) (Nawrocki & Eddy, 2013), a covariance model based approach that enhances tDNA detection. It evaluates the tDNA sequence conservation and ensures that it is able to produce a tRNA with the characteristic cloverleaf structure, including the formation of the corresponding stems and loops. This step allows for the filtering of potential tRNA-lookalike sequences. In addition, the predicted genes are aligned against isodecoder-specific covariance models for better classifying their functions in relation to the consensus tRNA isodecoder group.

tRNAscan-SE applies specific models for Eukarya, Bacteria, and Archaea to ensure an accurate identification of tDNAs unique for each domain. Furthermore, tRNAscan-SE 2.0 distinguishes between functional tDNAs and those that are non-functional or of uncertain function. tDNAs classified as functional are termed high-confident tDNAs, that correspond to predictions that exhibit sequence and structural features that align with known active tDNA sequences (P. P. Chan et al., 2021). In contrast, tRNA pseudogenes are sequences that resemble tDNAs sequences but have lost their ability to produce a functional tRNA. This is because they may arise from gene duplication events and contain sequence defects that prevent them from being transcribed or properly folded into functional tRNAs. Additionally, tDNAs reported as uncertain-function, correspond to sequences whose identity is not clearly established. They may not produce canonical tRNA structures or have sequences that deviate from known functional tRNAs, making their role in protein synthesis ambiguous (P. P. Chan et al., 2021). The final output of tRNAscan-SE is a full description of each of the tDNAs that have been identified, including genomic localization, genomic sequence, secondary structure prediction and in some cases the intronic regions (P. P. Chan et al., 2021). Therefore, tools like tRNAscan-SE that allow the prediction of tDNAs are essential for understanding tRNA biology.

Moreover, tRNAscan-SE tDNAs predictions are available for over 4,000 genomes in the Genomic tRNA Database (GtRNAdb), including human tDNA predictions for GRCh37/hg19 and GRCh38/hg38 genome assemblies (P. P. Chan & Lowe, 2016). All the tDNAs annotations are based on a consensus nomenclature, with the aim of providing a universal standardized naming system **(Fig. 6)**.

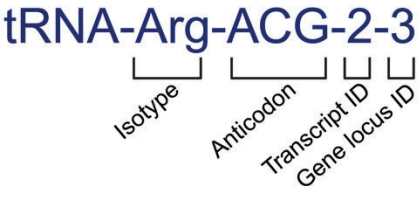


Figure 6. tDNA consensus nomenclature of GtRNAdb. The naming system uses identifiers such as tRNA-Arg-ACG-2-3. Isotype: Corresponds to a three-letter abbreviation for the isoacceptor based on the identified anticodon in the predicted gene sequence; in this example, arginine (Arg). Anticodon: Indicates the anticodon sequence in the predicted gene sequences. Transcript ID: Numeric ID of a unique tDNA transcript used to distinguish between different tDNA copies. All tDNAs with that number will have the exact same sequence and will be considered to be from the same gene family. Gene locus ID: For tDNAs that belong to the same family, this ID determines each particular gene copy in the genome. Adapted from (P. P. Chan & Lowe, 2016).

1.3 tRNA Biogenesis

tRNA biogenesis **(Fig. 7)** starts with the transcription of tDNA to produce a precursor tRNA (pre-tRNA). Then, the pre-tRNA is processed, post-transcriptionally modified and aminoacylated to become a mature and functional tRNA able to perform its main role in translation. Additionally, both pre-tRNA and mature tRNA molecules can be cleaved into tRNA-derived small RNAs (tsRNAs) that are involved in non-canonical functions beyond protein synthesis. The following sections will describe the different processes involved in human tRNA biogenesis.

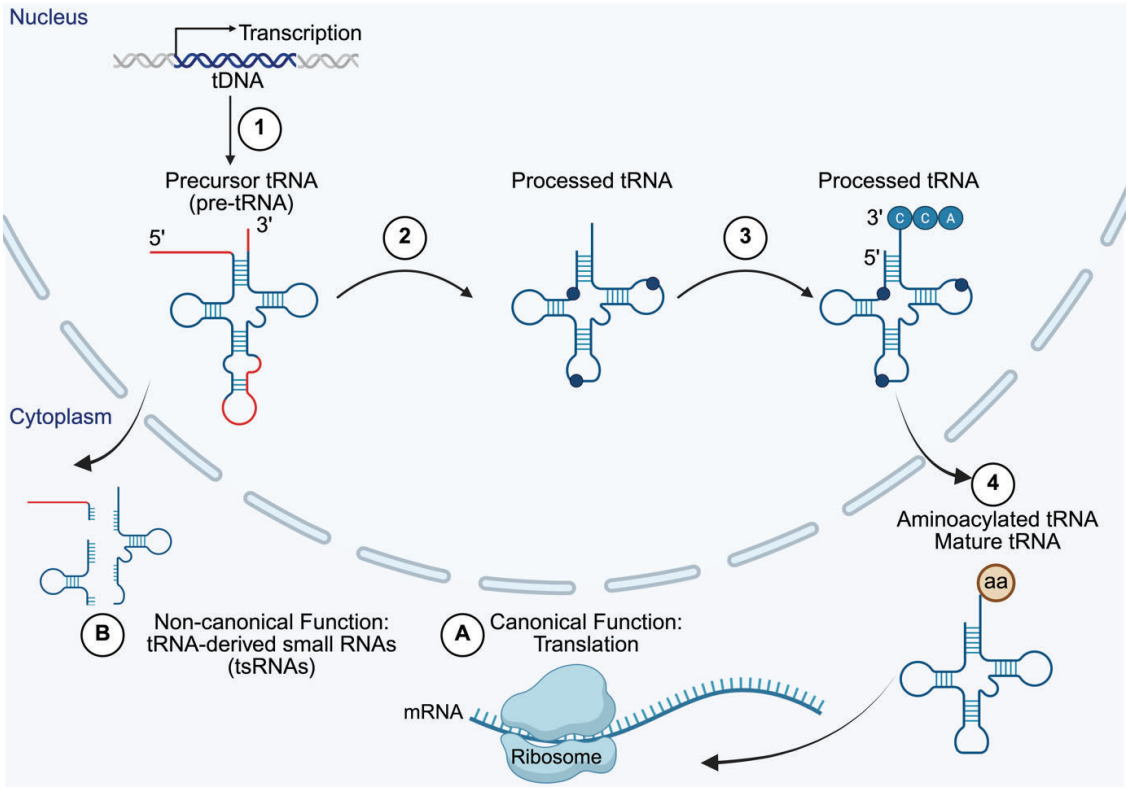


Figure 7. Schematic representation of tRNA processing. Transfer RNA (tRNA) biogenesis involves several key steps: **(1)** tDNA transcription by RNA polymerase III (Pol III) to produce a precursor tRNA (pre-tRNA) sequence. **(2)** Cleavage of the 5' leader, 3' trailer, and intron sequences from the pre-tRNA. **(3)** Addition of chemical modifications and the 3'-CCA tail. **(4)** Attachment of the corresponding amino acid (aa) to the tRNA by aminoacyl-tRNA synthetases (aaRS) to produce a fully mature tRNA. **(A)** The mature tRNA then goes to the ribosome to perform its main function in protein synthesis. **(B)** Both pre-tRNA and processed/mature tRNA sequences can be cleaved into tRNA-derived small RNAs (tsRNAs), which have many functions beyond protein synthesis. Created with BioRender.

1.3.1 tDNA transcription

The biogenesis of cytoplasmic tRNAs begins with the transcription of nuclear-encoded tDNAs by RNA polymerase III (Pol III) **(Fig. 8)**. Pol III is not only responsible for the transcription of tDNAs but also for the transcription of other essential small non-coding RNAs, including 5S ribosomal RNA genes (5S rDNA) and small nuclear RNA genes (RNUs) (Dieci et al., 2013; Turowski & Tollervey, 2016). Pol III-mediated transcription in tDNAs first requires the recognition of two internal conserved promoter sequences that are located at positions 8-19 (box A) and 52-62 (box B) within the tDNA sequence (Goodenbour & Pan, 2006). Both sequences are first recognized by the transcription factor TFIIIC (a six-subunit complex), which then recruits TFIIIB (composed of BRF1 and BDP1). The formation of this pre-initiation complex facilitates the recruitment of Pol III, allowing it to initiate transcription and produce a pre-tRNA (Male et al., 2015; Schramm & Hernandez, 2002). Another well-conserved regulatory feature is the poly-T tract, found downstream of the tDNA sequences, that consist of a stretch of thymine residues involved in pre-tRNA transcription termination and stabilization (Hummel et al., 2019).

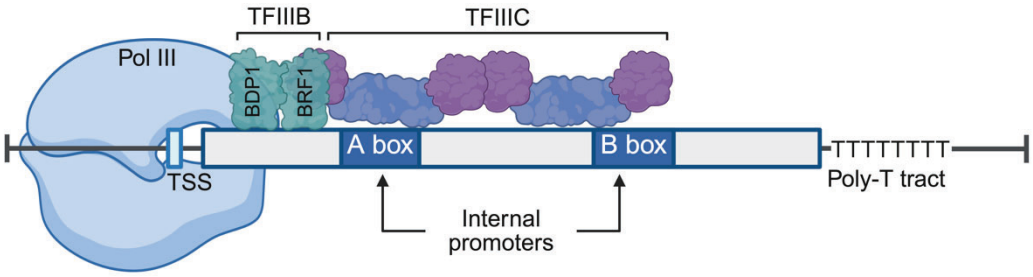


Figure 8. tDNA transcription. This figure illustrates tDNA transcription, highlighting the genomic regions and enzymes essential for this process. Note that tDNAs refer to DNA sequences that encode mature tRNA. Created with BioRender.

Unlike nuclear-encoded tDNAs that are transcribed individually, the 22 human mt-tRNAs are transcribed as a part of two polycistronic transcripts that also contain ribosomal coding sequences. The transcription is performed by the mitochondrial RNA polymerase (POLRMT) using just two bidirectional promoters (HSP and LSP). The long polycistronic transcripts from both strands are processed to generate separate rRNAs, mt-mRNAs, and tRNAs, to then undergo maturation (Ojala et al., 1981; Suzuki et al., 2011).

Regulation of tDNA transcription

The cell has the ability to optimize their tRNA content to match protein synthesis demands or to produce specific tsRNAs involved in cellular processes beyond protein synthesis (Dittmar et al., 2006; Novoa & Ribas de Pouplana, 2012; Torres et al., 2019).

In this regard, nuclear tDNA transcription is modulated under different environmental conditions such as cellular proliferation (Goodarzi et al., 2016), differentiation (Gingold et al., 2014; Van Bortle et al., 2017),

stress or in a disease context like cancer (Pavon-Eternod et al., 2009; Torrent et al., 2018). Moreover, the transcriptional activity of tDNAs is cell type- and tissue-specific (Dittmar et al., 2006; Gogakos et al., 2017; Torres et al., 2019). Interestingly, half of human nuclear-encoded tDNAs were identified as genes constitutively inactive (Gogakos et al., 2017; Thornlow et al., 2018; Torres, 2019; Torres et al., 2019). This indicates that even if a tDNA has all the constitutive elements to be transcribed, it will not necessarily be active (Thornlow et al., 2020). Altogether, these findings underscore the relevance of understanding the regulatory mechanisms behind tDNA transcription.

Some of the regulatory mechanisms controlling nuclear-encoded tDNA transcription will be explained below:

Pol III-dependent regulation

Besides direct regulatory mechanisms such as the interaction between Pol III, transcription factors and promoter regions, other indirect factors can also control Pol III-mediated transcription. For example, it has been described how MAF1 represses Pol III transcription in response to a variety of stresses, including nutrient deficiency (Gao et al., 2024; Hummel et al., 2019). Furthermore, oncogenic signaling pathways like AKT-mTOR, RAS-MAPK, NOTCH1, Ras, and MYC can promote Pol III transcription. By contrast, the tumor suppressors p53 and Rb are known to negatively regulate Pol III transcription. In most cases, this positive and negative regulation of Pol III is TFIIIB-mediated (Orellana et al., 2022). In addition, some subunits of Pol III can be post-transcriptional modified to regulate its transcriptional activity (Hummel et al., 2019).

Genomic context

The position effect describes how besides the genomic sequence of a gene its expression can be influenced by its genomic location (Hummel et al., 2019). In this regard, the expression levels of tDNAs are strongly correlated with CpG (C immediately followed by a G) content in their neighboring regions (Thornlow et al., 2020). tDNAs located in CG-rich areas are more likely to be expressed, since these regions often correspond to open chromatin states that facilitate Pol III binding (Thornlow et al., 2020). By contrast, inactive tDNAs can contain poly-T stretches in the upstream regions that act as constitutive Pol III termination signals inhibiting transcription (Gao et al., 2024). Altogether, this evidence suggests that the surrounding context is an important factor in tDNA transcription and can even be used to predict tDNA activity (Thornlow et al., 2020). Furthermore, recent studies have suggested that as human tDNAs can be linearly organized into clusters, their proximity could favor tDNA expression through the formation of ‘transcription factories’ (Gao et al., 2024). This may occur because such proximity facilitates the recruitment of Pol III and its corresponding transcription factors to multiple tDNAs simultaneously (Gao et al., 2024). However, further analyses are needed in order to characterize tDNA clustering and understand the link between tDNA proximity and tDNA transcription.

Epigenetic context

Epigenetic mechanism influence gene activity without changing the DNA sequence. These mechanisms include processes such as chromatin remodeling and DNA methylation.

Chromatin consists of nucleosomes that are segments of DNA wrapped around histone proteins. These nucleosomes are organized into chromatin fibers, which fold into local short-range interacting structures named Topologically Associating Domains (TADs) (Dixon et al., 2012). TADs are self-interacting genomic units that help organize the genome into functional compartments. Furthermore, regions that are more distant from each other can interact through long-range interactions shaping the 3D organization of the genome (Rowley & Corces, 2018). These physical contacts can influence many cellular processes, such as DNA replication, transcription and repair (Supek & Lehner, 2015). For instance, according to the chromatin configuration the genome is divided into replication domains. Each of these domains duplicates the DNA at a specific point during the S phase (Synthesis phase) of the cell cycle. This temporal regulation of DNA replication is tightly linked to transcriptional activity, as studies have consistently shown that early-replicating regions are often associated with active gene expression (Maric & Prioleau, 2010; Müller & Nieduszynski, 2017; Rhind & Gilbert, 2013). These regions are characterized by open chromatin (euchromatin), which is less condensed to be more accessible for replication and transcription machinery. In contrast, late-replicating regions are frequently linked to repressed gene expression with closed chromatin (heterochromatin) structures. This relationship also shows how replication timing can be used as a proxy for gene activity (Supek & Lehner, 2015).

Evidence supports that the transcription of tDNAs can be regulated in different ways based on their three-dimensional chromatin organization. For instance, in *Saccharomyces cerevisiae*, although tDNAs are dispersed around the genome, the tDNAs appear clustered at the nucleolus together with other Pol III-transcribed genes to promote tDNA transcription (Good et al., 2013; Thompson et al., 2003). The topological association of tDNAs has been observed in humans, particularly in a study on human macrophages differentiation (Van Bortle et al., 2017). They reported that the organization of tDNAs into TADs contributes to the regulation of tDNA transcription during macrophage differentiation. Furthermore, tDNAs undergo long-range interaction, and changes in their 3D organization influence their transcriptional levels, highlighting the role of chromatin architecture in adjusting tRNA pools to meet protein synthesis needs (Van Bortle et al., 2017).

DNA methylation is a gene silencing mechanism that represses the transcription of genes through the addition of a methyl group, typically at cytosine bases in CpG (Moore et al., 2013). In cattle, alterations in the methylation patterns of tDNA clusters alter tRNA abundance, which affects protein synthesis and correlates with large offspring syndrome (Goldkamp et al., 2022). In plants, tDNA clusters expression is silenced via CpG methylation (Hummel et al., 2019). In humans, changes in tDNA expression controlled by methylation processes are related to cancer progression and aging. For example, it has been reported that hypermethylation at tDNA loci increases with age in humans, which leads to transcriptional repression of specific tDNAs (Acton et al., 2021). As well, shifts in tDNAs methylation levels contribute to changes in tRNA abundances observed in cancer cells (Rosselló-Tortella et al., 2022).

In summary, the regulation of nuclear-encoded tDNA transcription is a complex, multi-layered process that go from the regulation of individual genes to short-range interactions and long-range coordination of dynamic tDNA expression (Gao et al., 2024; Torres et al., 2019; Van Bortle et al., 2017).

1.3.2 Pre-tRNA processing

After transcription the obtained pre-tRNA molecules are processed to obtain a mature and active tRNA **(Fig. 9)**. The processing of pre-tRNAs begins in the nucleus with the removal of non-functional sequences for mature tRNA. These sequences are the 5' leader and the 3' trailer, which consist of sequences with an average length of 10 and 6 nucleotides, respectively. In some tRNAs the trailer lengths show a broader distribution that can extend to 40 nucleotides (Gogakos et al., 2017). First, the 5’ leader sequence is cleaved by the enzyme Ribonuclease P (RNase P). Then, ElaC Ribonuclease Z 2 (ELAC2), the human homolog of Ribonuclease Z (RNase Z), trims the 3' trailer sequence (Xue et al., 2024). Additionally, in humans, approximately 30 tRNAs have introns that need to be removed (C. A. Schmidt & Matera, 2020; Yoshihisa, 2014). Introns vary in length from 6 to 133 nucleotides and always start at the canonical site between nucleotides 37 and 38 (P. P. Chan & Lowe, 2016; Yoshihisa, 2014). The removal of the intron is carried out by the tRNA splicing endonuclease complex (TSEN) (Yuan et al., 2023). Additionally, a guanosine is added to the 5′ end only in tRNAs for histidine (tRNA-His), which is a modification essential the aminoacylation of those tRNAs (Rudinger et al., 1994). Finally, a CCA sequence is added to the 3' end by tRNA nucleotidyltransferase, commonly known as the CCA-adding enzyme. This modification is crucial for amino acid attachment to tRNA (Betat & Mörl, 2015).

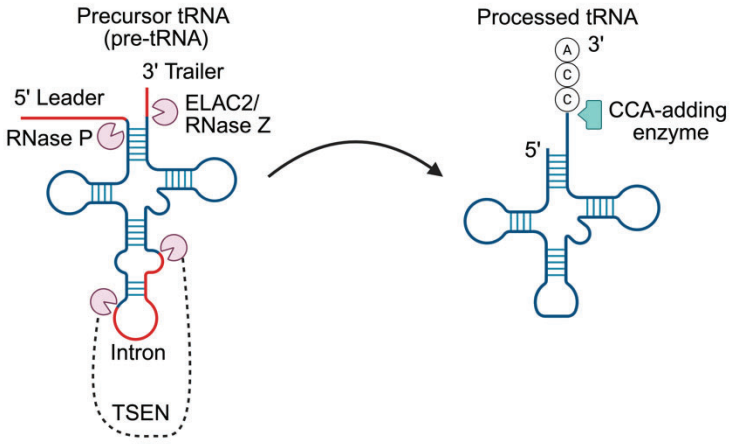


Figure 9. Processing of the precursor tRNA (pre-tRNA) sequence. The illustration shows the processing of pre-tRNA sequence. Processing includes the cleavage of the 5’ leader and 3’ trailer, as well as the removal of intron sequences. Created with BioRender.

1.3.3 tRNA chemical modifications

tRNAs are known as the most chemically modified RNAs in the cell, with over 100 different types of chemical modifications identified across all domains of life (Pan, 2018). While some modifications are conserved and found universally in Bacteria, Archaea, and Eukarya, others are domain-specific, highlighting their diverse functional roles (de Crécy-Lagard & Jaroch, 2021; Machnicka et al., 2014). In human cells, nuclear tRNAs contain an average of 13 modifications **(Fig. 10)**, whereas mitochondrial tRNAs are less modified, with an average of 5 modifications per molecule (Pan, 2018). The functional impact of tRNA modifications varies depending on their location. Modifications in the anticodon loop are essential for codon recognition and translation efficiency, while those in the tRNA body contribute to tRNA structure, stability, localization, interaction with the ribosome and efficient aminoacylation (Pan, 2018; Suzuki, 2021; Torres et al., 2014a).

The complexity of tRNA modifications ranges from simple to complex and large base hypermodifications, whose synthesis often requires a whole cascade of enzymatic reactions. The set of modifications that can be found in a tRNA molecule include methylation, acetylation, deamination, isomerization, glycosylation, thiolation and pseudouridylation (Suzuki, 2021), each of them produced by a specific enzyme (Boccaletto et al., 2022; de Crécy-Lagard et al., 2019).

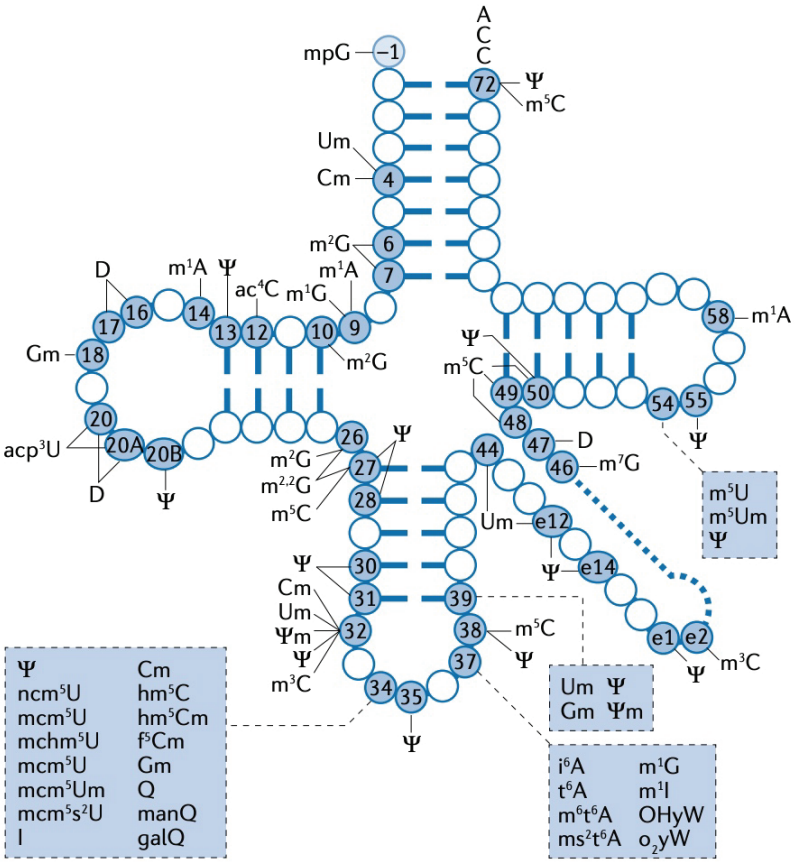


Figure 10. Post-transcriptional chemical modifications in human tRNAs. The positions reported to be modified are numbered, and the corresponding modifications are indicated with the corresponding

symbol. The modification at position -1 is specific of tRNA-His. Modification abbreviations are: ac^4C (N4-acetylcytidine), acp^3U (3-amino-3-carboxypropyluridine), 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 (N6-isopentenyladenosine), m^1A (1-methyladenosine), manQ (mannosyl-queuosine), m^3C (3-methylcytidine), m^5C (5-methylcytidine), $mchm^5U$ ((carboxyhydroxymethyl)uridine methyl ester), mcm^5s^2U (methoxycarbonylmethyl-2-thiouridine), mcm^5U (methoxycarbonylmethyluridine), m^1G (1-methylguanosine), m^2G (N2-methylguanosine), $m^{2,2}G$ (N2,N2-dimethylguanosine), m^7G (7-methylguanosine), m^1I (1-methylinosine), mpG (5′-methylphosphoguanosine), ms^2i^6A (2-methylthio-N6-isopentenyladenosine), ms^2t^6A (2-methylthio-N6-threonylcarbamoyladenosine), m^6t^6A (N6-methyl-N6-threonylcarbamoyladenosine), m^5U (5-methyluridine), m^5Um (2′-O-methyl-5-methyluridine), ncm^5U (carbamoylmethyluridine), OHyW (hydroxywybutosine), o^2yW (peroxywybutosine), Q (queuosine), t^6A (N6-threonylcarbamoyladenosine), Um (2′-O-methyluridine), Ψ (pseudouridine), Ψm (2′-O-methylpseudouridine). Adapted from (Suzuki, 2021).

I34 and decoding expansion

As aforementioned, modifications in the anticodon of tRNAs play a crucial role in translation fidelity and efficiency. One of the most important modifications in tRNAs is the deamination of adenosine (A34) to inosine (I34) at the first position of the anticodon (Gerber & Keller, 1999; Srinivasan et al., 2021), referred to as the "wobble position" (Crick, 1966). This A-to-I editing expands the decoding capacity of tRNAs by enabling non-Watson-Crick codon-anticodon pairing. Specifically, tRNAs with A34 (A34-tRNAs) can only decode U-ended codons, whereas tRNAs with I34 (I34-tRNAs) can decode U-, A- and C-ended codons, allowing a single tRNA to be able to pair with multiple mRNA codons (F. H. Crick, 1966).

In Bacteria, I34 is commonly found in tRNA-Arg-ACG and less frequently, in tRNA-Leu-AAG as observed in *Oenococcus oeni* and *Streptococcus pyogenes* (Rafels-Ybern et al., 2019, 2018; Srinivasan et al., 2021; Wulff et al., 2024). In contrast, in Eukarya I34 is found in eight tRNA isoacceptors, including threonine (T), alanine (A), proline (P), serine (S), leucine (L), isoleucine (I), valine (V), and arginine (R), referred as ‘TAPSLIVR’. Specifically in the isodecoders tRNA-Thr-AGT(IGU), tRNA-Ala-AGC(IGC), tRNA-Pro-AGG(IGG), tRNA-Ser-AGA(IGA), tRNA-Leu-AAG(IAG), tRNA-Ile-AAT(IGU), tRNA-Val-AAC(IAC), and tRNA-Arg-ACG(ICG) (Rafels-Ybern et al., 2015). The enzyme responsible for A-to-I editing in Bacteria is a homodimeric tRNA-specific deaminase, known as TadA. In eukaryotes, A-to-I editing is catalyzed by the enzyme adenosine deaminase acting on tRNA (ADAT). ADAT is a heterodimer that emerged early in eukaryotic evolution from the bacterial TadA through a gene duplication event, resulting in the heterodimeric ADAT composed by ADAT2 and ADAT3 (Gerber & Keller, 1999; Torres et al., 2014b; Wolf et al., 2002). While ADAT2, is considered the catalytic subunit, ADAT3 provides binding and structural support (Srinivasan et al., 2021).

Bacteria commonly use G34-tRNAs to decode TAPSLIVR codons, while Eukarya primarily use I34-tRNAs (Maraia & Arimbasseri, 2017; Rafels-Ybern et al., 2015, 2019, 2018). The widespread use of I34 in eukaryotes influenced genome evolution, particularly in the context of codon usage and tDNA

abundance. Because this modification enables a single tRNA to recognize multiple codons, Eukaryotic genomes favor codons that correspond to A34, contributing to codon bias (Novoa et al., 2012).

Inosine can also be found at other positions in tRNA. For example, m^1I37 (1-methylinosine at position 37) has been identified exclusively in eukaryotic tRNA-Ala. Although m^1I37 is not located in the anticodon itself, it is situated within the anticodon loop and plays a role in ensuring accurate decoding by stabilizing codon-anticodon pairing (Han & Phizicky, 2018; Suzuki, 2021). This modification is catalyzed by the homodimeric enzyme ADAT1 (Torres et al., 2014b).

1.3.4 tRNA aminoacylation

Aminoacylation refers to the attachment of the cognate amino acid at the 3' end of the tRNA to produce aminoacyl-tRNA (aa-tRNA). This reaction is produced by aminoacyl-tRNA synthetases (aaRS) and takes place in the cytoplasm after the export of processed tRNA from the nucleus. The aminoacylation of the tRNA involves two major steps (Alberts et al., 2002) **(Fig. 11)**.

First, the amino acid is attached to the aaRS through an ATP-dependent reaction that allows the formation of aminoacyl-adenylate (aa-AMP). The aa-AMP works as an intermediate molecule for aaRS to attach the amino acid to the corresponding tRNA. For a limited number of aaRS, tRNA binding is required prior to aa-AMP formation (Rubio Gomez & Ibba, 2020). The recognition of the tRNA by aaRS is guided by identity elements that can be located in different regions of the tRNA such as the acceptor stem, the anticodon loop, or the D-loop. This implies that each of the 20 amino acids has a specific synthetase that recognizes both the cognate amino acid and its corresponding tRNA to ensure an accurate and specific pairing (Giegé & Eriani, 2023; Ibba & Söll, 2000). Whereas the anticodon bases are the primary identity element for many tRNAs, the nucleotide in the acceptor stem at position 73 is considered a "discriminator base" that also determines the recognition by aaRS (Giege et al., 1998).

Second, after the recognition of the specific tRNA by the aminoacyl-tRNA synthetase and thanks to the reactivity properties of the 3'-CCA tail, the amino acid is transferred from the aa-AMP to the tRNA producing an aminoacyl-tRNA (aa-tRNA). As a result, AMP is released from aaRS and the aa-tRNA dissociates from the aaRS, ready to be delivered to the ribosome for protein synthesis (Alberts et al., 2002).

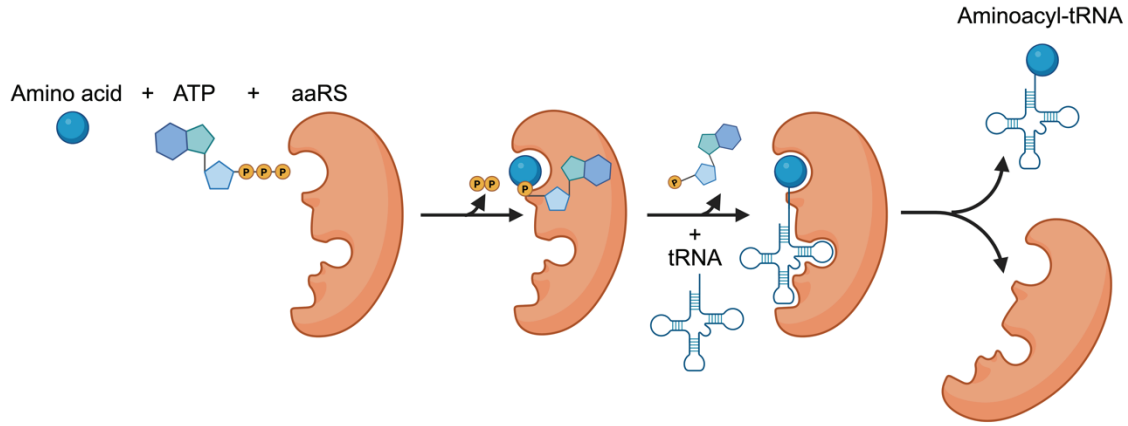


Figure 11. Aminoacylation reaction. Schematic representation of the aminoacylation of the tRNA, in which the tRNA is paired with its cognate amino acid by aminoacyl tRNA synthetase (aaRS). Created with BioRender.

1.3.5 tsRNAs processing

tRNAs have another layer of complexity, as they can be further processed into tsRNAs. tsRNAs can be divided into two main subgroups: tRNA-derived fragments (tRFs) and tRNA halves (also known as stress-induced RNAs or tiRNAs) (P. Anderson & Ivanov, 2014; Y. S. Lee et al., 2009; Oberbauer & Schaefer, 2018; Telonis et al., 2015) **(Fig. 12)**.

tRFs (~12-30 nt) can be categorized based on their origin within the precursor and mature tRNA molecule (P. Anderson & Ivanov, 2014; Y. S. Lee et al., 2009; Oberbauer & Schaefer, 2018; Telonis et al., 2015). These categories include tRF-1 (or 3' U-tRFs), derived from precursor tRNA trailer sequences; 5' leader-exon tRFs, which contain both leader and exon sequences from pre-tRNA; tRF-2, produced by cleavage of the anticodon loop; internal tRFs (i-tRFs), derived from sequences within the mature tRNA, specifically containing part of the sequence of the anticodon loop; and finally, tRF-3 and tRF-5, which originate from the 3' and 5' ends of mature tRNAs. As well, tiRNAs (~30-50 nt) can be classified into two primary types based on their origin from the mature tRNA: 5' halves (5' tiRNAs), which derive from the 5' end of the tRNA up to the anticodon loop, and 3' halves (3' tiRNAs), which originate from the anticodon loop to the 3' end. Furthermore, a specialized type, SHOT-RNAs (Sex HOrmone-dependent TRNA-derived RNAs) , is recognized as a subset of tiRNAs whose production is specifically regulated by sex hormones (Honda et al., 2015).

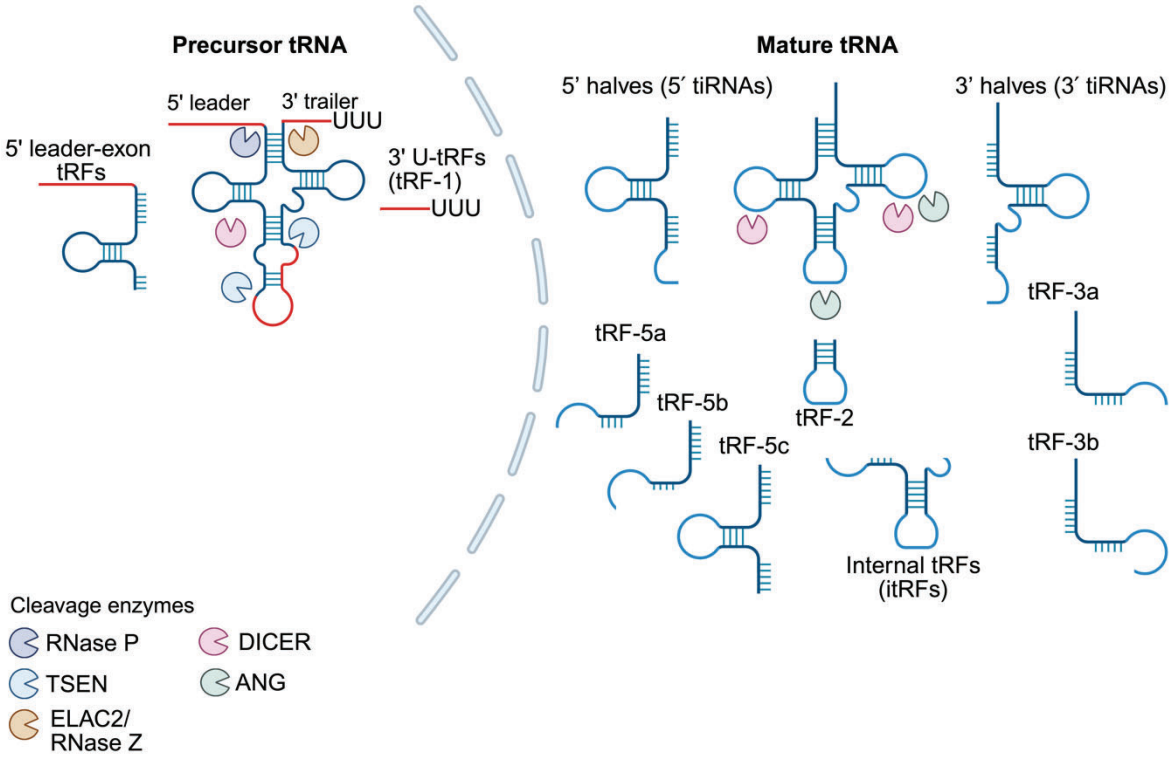


Figure 12. Biogenesis and classification of tRNA-derived small RNAs (tsRNAs). The schema describes the classification of tsRNAs from both pre-tRNA and mature sequences. Enzymes involved in tRNA molecule cleavage are also included. The enzymes that participate in pre-tRNA cleavage and the production of tsRNAs containing precursor sequences are RNase P, TSEN, and ELAC2 (also named RNase Z). The enzymes that participate in the production of tsRNAs derived from mature sequences are Dicer and angiogenin (ANG). Created with BioRender.

The production of tsRNAs is orchestrated by different ribonucleases (RNases), including RNase Z/ELAC2 (Y. S. Lee et al., 2009), Dicer, and angiogenin (ANG) (Ivanov et al., 2011; Kumar et al., 2014, 2016; Su et al., 2019) **(Fig. 12)**. In humans tRF-1 is produced during the processing of pre-tRNA sequences by ELAC2 (Y. S. Lee et al., 2009). Then, the fragments derived from mature sequences (i-tRF, tRF-2, tRF-3, and tRF-5) result from the cleavage produced by ANG and Dicer. tiRNAs are divided into two major subtypes, 5′ tiRNA and 3′ tiRNA, which are derived from mature tRNAs cleaved by ANG at the anticodon loop (P. Anderson & Ivanov, 2014). Although tsRNAs seem to be extensively described, in some cases the enzyme or mechanism responsible for the cleavage of specific tsRNAs is still unknown.

tRNA Modifications can influence the cleavage of tsRNAs. In most cases, modifications prevent the formation of tRFs, offering protection from cleavage. For example, m^5C38 modification, prevents the cleavage produced by ANG under oxidative damage (Schaefer et al., 2010). In that sense, hypo-modified tRNAs can become more vulnerable to degradation or fragmentation by various ribonucleases (Su et al., 2020).

1.4 Canonical function of tRNAs: Translation

After achieving its mature and active form through several processing steps, tRNA is ready to perform its main role in translation. Translation occurs in three main steps: initiation, elongation, and termination (Alberts et al., 2002).

Translation initiation is mediated by eukaryotic translation initiation factors (eIFs) that promote the recognition of the start codon (AUG) on the mRNA by a specialized initiator methionine tRNA (tRNA-iMet). This signals the recruitment of the ribosomal subunits (60s and 40s) to produce a functional ribosome (80s) ready to begin protein synthesis (Sonenberg & Hinnebusch, 2009).

Next, elongation begins with the recruitment of eukaryotic translation elongation factors (eEFs) (Rodnina & Wintermeyer, 2016). The polypeptide chain elongates as the ribosome moves along the mRNA one codon at a time. In this process, tRNA anticodon base-pair with the cognate mRNA codon and enter the ribosomal acceptor aminoacyl site (A-site) **(Fig. 13)**. The amino acid carried by the tRNA is then transferred to the growing polypeptide chain located at the peptidyl site (P-site), a reaction catalyzed by the ribosome. The now uncharged tRNA shifts to the exit site (E-site) and leaves the ribosome (Beringer & Rodnina, 2007). The ribosome then moves to the next codon and the A-site is empty and available for the next aa-tRNA.

The final step, termination, occurs when the ribosome encounters one of the stop codons (UAA, UAG, or UGA) in the mRNA. Release factors (eRFs) bind to the stop codon, signaling the ribosome to stop translation. This promotes the hydrolysis of the bond between the polypeptide chain and tRNA in the P site, leading to the release of the synthesized protein (Schuller & Green, 2018).

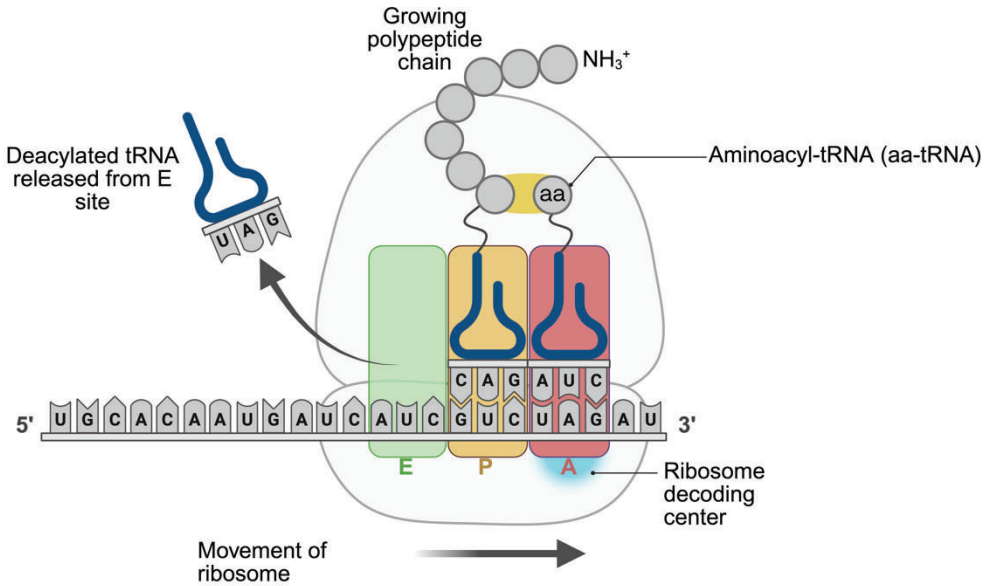


Figure 13. Translation machinery. Schematic representation of the decoding process of mRNA produced in the ribosome. The representation includes the sites for tRNA binding and release, including

the aminoacyl site (A-site), where the aminoacyl-tRNA is located; the peptidyl site (P-site), which contains the growing polypeptide chain; and the exit site (E-site), where the deacylated tRNA leaves. Created with Biorender.

Codon usage bias, tRNAs, translation efficiency and fidelity

Codon usage bias is the phenomenon where genomes, genes, and even localized mRNA regions exhibit a non-random preference for specific synonymous codons (i.e., codons coding for the same amino acid) (Ikemura, 1985; Parvathy et al., 2022). The preference for particular codons, directly influences the speed, efficiency and accuracy of protein synthesis (Novoa & Ribas de Pouplana, 2012; Quax et al., 2015; Weinberg et al., 2016). The core of this influence stems from the interplay with tRNA availability, as the specific codons within a gene sequence dictate which tRNAs are going to be needed. Although a strong correlation between tRNA abundance, tRNA gene copy number, and codon usage is well established in eubacteria, it is still uncertain how much this relationship holds true in eukaryotic genomes (Bermudez-Santana et al., 2010; Ikemura, 1985; Sabi & Tuller, 2014; Spencer et al., 2012). The widespread use and functional relevance of inosine in eukaryotes helps explain why the correlation between tDNA copy number and codon usage is often weaker. This is because inosine at the wobble position expands the decoding capacity, which enables a single tRNA to pair with multiple codons, so fewer distinct tRNA genes are required to decode the entire set of codons (Novoa et al., 2012; Rafels-Ybern et al., 2018).

Furthermore, cells often increase the supply of tRNAs that match frequently used, or optimal codons, leading to faster and more efficient protein synthesis, especially in highly expressed genes (Novoa & Ribas de Pouplana, 2012). To speed translation some regions of the gene present the same codon in close proximity to allow tRNA recycling (Ingolia et al., 2009). Interestingly, it has been observed that proteins with similar functions or related to the same biological pathway can be enriched in the same type of codons. This highlights how codon bias can be used as a strategy to optimize the translation of a specific set of mRNAs (Novoa & Ribas de Pouplana, 2012; Presnyak et al., 2015). In contrast, non-optimal codons, characterized by lower tRNA availability, can slowdown translation. In some scenarios, the slowdown of translation is needed to facilitate co-translational folding by allowing nascent proteins more time to form correct structures in order to prevent misfolding (Novoa & Ribas de Pouplana, 2012). In other instances, non-optimal codons can decrease translation fidelity leading to the erroneous incorporation of an amino acid (i.e., mistranslation) (Akashi, 1994; Drummond & Wilke, 2008; Hanson & Coller, 2018). In this scenario, mistranslation will occur because if the cognate tRNA is not available, the translation machinery will be forced to use a non-cognate tRNA, adding to the polypeptide an incorrect amino acid (Akashi, 1994).

There are proteins with a highly biased or repetitive composition of amino acids in specific regions, that are known as ‘low-complexity’ domains (Mier et al., 2020). The mRNA region from low-complexity domains corresponds to codon stretches with a strong codon bias that can impair translation, acting as non-optimal codons. When specific regions of an mRNA are enriched in particular codons, the

corresponding cognate tRNA may not be available at the necessary speed, which leads to ribosomal stalling and the consequently slowdown of in translation (Frugier et al., 2010; Lassak et al., 2016). Low-complexity domains can also produce mistranslation (Davies & Rubinsztein, 2006).

Altogether, these observations underscore that, especially in Eukarya, ensuring an efficient translation of all codons to promote stable protein synthesis depends on maintaining the interplay between tDNA copy number, tDNA transcription, codon usage bias, and decoding expansion through base modifications (Novoa & Ribas de Pouplana, 2012; Quax et al., 2015; Rafels-Ybern et al., 2019, 2018; Torres et al., 2019). As previously noted, changes in any of these factors can result in mistranslation. However, mistranslation can occur through multiple mechanisms, including the reduced availability of cognate tRNAs, mis-aminoacylation of tRNAs by aaRSs or ribosome-induced mRNA misreading (Schuntermann et al., 2024).

Interestingly, tRNA-mediated mistranslation can be genetically encoded. While both the acceptor stem and the anticodon are an identity element in most tRNAs, for tRNA-Ala, tRNA-Leu, tRNA-Ser and tRNA-Tyr, their respective aaRS recognize the acceptor stem rather than the anticodon (Giegé & Eriani, 2023; Ibba & Söll, 2000). If mutations occur in the anticodon of these isoacceptors, they can still be aminoacylated with the corresponding amino acid, producing a ‘chimeric tRNA’ (i.e., tRNA that carries a specific amino acid but recognizes a different mRNA codon) (Geslain et al., 2010). Therefore, chimeric tRNAs can produce amino acid substitutions across the proteome, resulting in widespread mistranslation (Geslain et al., 2010; M. Santos et al., 2018). Surprisingly, tDNA anticodon variants exist within the population, and among these, variants for tRNA-Ala, tRNA-Ser, and tRNA-Leu produce nonsynonymous anticodons, which results in the presence of genetically encoded chimeric tRNAs (Lant et al., 2019; Schuntermann et al., 2024). An analysis of 1000 Genomes Project data revealed 14 unique nonsynonymous anticodon variants within these tRNAs. While most of these nonsynonymous anticodon variants were rare (present in less than 1% of the individual), one tRNA-Ala variant, which contains a mutation that leads to a Gly anticodon, was found in more than 6% of 1000 individuals. However, evidence linking these chimeric tRNAs with protein synthesis alterations is still lacking (Lant et al., 2019).

The biological impact of mistranslation depends on the specific codon that is mistranslated and the amino acid that is misincorporated. Moreover, tDNA copy number can “buffer” the effects of chimeric tRNAs and mitigate mistranslation (Lant et al., 2019). Additionally, mistranslation tolerance can be determined by the inherent characteristics of the species in which it takes place (Schuntermann et al., 2024). For example, in bacteria, Pro-to-Thr mistranslation, is significantly more detrimental to cellular function than either Pro-to-Ala or Pro-to-Asn (Schuntermann et al., 2023). Whereas, in yeast, the degree of toxicity relies on the specific Ala codon that undergoes mistranslation (Cozma et al., 2023).

Mistranslation events can lead to deleterious alterations. For instance, in mice, Ala-to-Ser mistranslation produces the accumulation of misfolded proteins that lead to neurodegeneration and cardioproteinopathies (J. W. Lee et al., 2006; Y. Liu et al., 2014). In zebrafish, chimeric tRNAs that misincorporate Ser at non-cognate codon sites were shown to alter protein synthesis. Those specific tRNAs were proposed to be drivers of progressive cellular degeneration and disease through

mechanisms that involve protein aggregation, mitochondrial dysfunction, and genome instability (Reverendo et al., 2014).

In other instances, mistranslation may be used as a mechanism to overcome unfavorable environmental or physiological changes (Correia et al., 2024; Miranda et al., 2007; Mohler & Ibba, 2017; Netzer et al., 2009; Pan, 2013; Ribas de Pouplana et al., 2014; Schuntermann et al., 2024). Certain organisms can tolerate high degrees of mistranslation and use mistranslation as a source of proteome diversity. In this regard, mistranslation can be used under cellular stresses, such as nutrient deprivation, exposure to antibiotics, or chemical and immune-related stressors, by organisms to adapt to new cellular demands (Ribas de Pouplana et al., 2014). For example, in mammalian cells, exposure to oxidative stress can induce mistranslation by promoting the incorporation of Met at non-cognate sites. Given that Met has antioxidant properties, it could be used to protect proteins from oxidative damage (Netzer et al., 2009). A different type of adaptive strategy is found in *Candida albicans*, where a unique form of controlled mistranslation serves as an adaptive strategy. In most organisms, the CUG codon is translated to the amino acid Leu. However, in *C. albicans*, the CUG codon is translated as a mix of Ser and Leu (M. A. S. Santos & Tuite, 1995). This generates diverse protein forms from a single gene, providing a mechanism for proteome plasticity (Miranda et al., 2013). This ability to create protein variety allows *C. albicans* to quickly adapt to environmental stresses, including the development of drug resistance, such as fluconazole (Bezerra et al., 2013; Weil et al., 2017). Another example of mistranslation has been described for chimeric tRNAs that produce Ser-to-Ala misincorporation, which increases cell proliferation, producing faster-growing tumors in mice, suggesting that mistranslation can be advantageous in the context of cancer (M. Santos et al., 2018). In this regard, aberrant peptides produced by mistranslation may enable cancer cells to acquire new properties that support tumor progression (M. Santos et al., 2019; Weller et al., 2025).

1.5 Non-canonical functions of tRNAs: Beyond translation

Beyond their role as genetic code decoders, tRNAs are implicated in multiple biological processes. Some of these processes can be attributed to full-length tRNAs, including the control of apoptosis pathways, modification of proteins through arginylation or acting as sensors for amino acid starvation (Avcilar-Kucukgoze & Kashina, 2020; R. Gupta & Laxman, 2020). tRNAs can be processed into tsRNAs that are also involved in a wide range of functions, including gene silencing, translational regulation, cell signaling, cell survival, apoptosis, and amino acid metabolism (P. Anderson & Ivanov, 2014; Raina & Ibba, 2014; Schimmel, 2018; Sheppard et al., 2008; Soares & Santos, 2017; Su et al., 2020). Furthermore, tDNAs are involved in other cellular processes related to genome organization (Raab et al., 2012).

1.5.1 The roles of tsRNAs

As mentioned previously, both pre-tRNA and mature tRNAs can be cleaved into tsRNAs (P. Anderson & Ivanov, 2014; Y. S. Lee et al., 2009; Pandey et al., 2021). Those tRNA-derived products have non-related mRNA decoding functions and are implicated in different biological contexts (Ivanov et al., 2011; Raina & Ibba, 2014; Su et al., 2020).

For instance, tsRNAs have emerged as potential regulators of gene silencing, mimicking the functions of microRNAs (Kumar et al., 2014; Kuscu et al., 2018). Like microRNAs counterparts, tsRNAs are able to associate with argonaute (Ago) proteins or PIWI proteins, to produce the core of gene-silencing complexes (Raina & Ibba, 2014). In those gene-silencing complexes, tsRNAs act as guides to identify specific mRNAs based on sequence complementarity. This targeting can effectively silence or repress mRNA, avoiding the production of the corresponding protein. This mechanism is particularly used by the cell under stress conditions, where tsRNAs can contribute to reprogramming gene expression, which can help cells to deal with new protein demands. Besides silencing mRNAs, tsRNAs can also regulate translation initiation and inhibit ribosome biogenesis (Ivanov et al., 2011; Prehn & Jirström, 2020).

Beyond their role in translation regulation, tsRNAs are involved in a variety of cellular processes, including cell cycle control, apoptosis inhibition, and cell-cell communication, mediated by tsRNAs found within extracellular vesicles (Kumar et al., 2016; Oberbauer & Schaefer, 2018; Raina & Ibba, 2014; Su et al., 2020). Furthermore, many tRFs have been described to be involved in disease (P. Anderson & Ivanov, 2014). To date, new functions continue to be attributed to tsRNAs.

1.5.2 tDNAs in genome organization

As described previously, numerous tDNAs are conserved within the genome despite remaining untranscribed and functionally inactive (Torres, 2019). Altogether, these observations suggest that nuclear-encoded tDNAs (both transcriptionally active and silent ones) may be playing other roles. Indeed, multiple studies have demonstrated that tDNAs play important roles in genomic organization, regulation, and stability (Guimarães et al., 2021; Hamdani et al., 2019; Iwasaki et al., 2020; McFarlane & Whitehall, 2009; Raab et al., 2012; Sizer et al., 2022; Van Bortle & Corces, 2012; Van Bortle et al., 2017).

Regions containing tDNAs are described as conserved genomic boundary elements, which separate chromatin into domains, preventing one domain from influencing its neighbors (McFarlane & Whitehall, 2009; Raab et al., 2012; Sizer et al., 2022). tDNAs can behave as chromatin insulators, also known as heterochromatin barriers (McFarlane & Whitehall, 2009; Raab et al., 2012; Sizer et al., 2022). For example, in tDNAs the Pol III transcription machinery can recruit protein complexes that organize and condense DNA (e.g. cohesin and condensin), blocking the spread of heterochromatin in yeast (Guimarães et al., 2021). Additionally, tDNAs participate in the spatial organization of the genome. In many organisms, despite being linearly dispersed, tDNAs can be found clustered together in the three-dimensional space and can present local and long-range interactions (Dixon et al., 2012; Hamdani et al., 2019; Iwasaki et al., 2020; Van Bortle et al., 2017).

Under specific conditions, tRNAs may influence recombination events and contribute to genomic instability. As tDNAs are multi-copy genes spread through the genome, they can generate homologous regions that can act as substrates for recombination, acting similarly as SINEs, which are derived from tDNAs (Kramerov & Vassetzky, 2011; McFarlane & Whitehall, 2009). Furthermore, tDNA transcription by Pol III can interfere with the progression of DNA replication forks, contributing to genome instability and the formation of genomic fragile sites (Bermudez-Santana et al., 2010; McFarlane & Whitehall, 2009; Yeung & Smith, 2020).

1.6 tRNAs in disease

Alterations in tRNAs, occurring at any point of tRNA biogenesis, including transcription, pre-tRNA processing, aminoacylation, chemical modifications, and tsRNA production have been linked to many diseases including diabetes, cancer and neurological disorders (P. Anderson & Ivanov, 2014; Orellana et al., 2022; Torres et al., 2014a). In this thesis, we mainly focus on the following diseases:

1.6.1 I34 in human disease

As previously mentioned, inosine at position 34 (I34) is an essential modification that enables the decoding expansion of tRNAs. Alterations in the subunits of ADAT (Gerber & Keller, 1999; Torres et al., 2014b), the enzyme responsible for this modification, have been linked to neurological disorders (Torres et al., 2014a). For instance, a mutation in ADAT3 that produces a change between valine to methionine at position 144 (V144M), was previously associated with intellectual disability and strabismus (Alazami et al., 2013; Salehi Chaleshtori et al., 2018; Thomas et al., 2019). Another ADAT3 mutation consists of an 8-nucleotide duplication that leads to altered protein structure and is associated with intellectual disability and hyperactivity (Salehi Chaleshtori et al. 2018). Furthermore, mutations affecting the deaminase domain of ADAT3 were detected in two siblings who exhibited developmental delay, abnormal brain structure, and intellectual disability, among other disorders (Thomas et al., 2019). Alterations in any of the subunits of ADAT could impair A-to-I editing, reducing the availability of I34-tRNAs within the cell.

Given that genes exhibit specific codon usage bias, a deficiency of I34-tRNAs could selectively impact proteins enriched in codons that are preferentially decoded by I34-tRNAs, referred to as ADAT-sensitive codons or TAPSLIVR codons (Rafels-Ybern et al., 2015). Those proteins could be linked to specific pathways that when altered produce phenotypes related to neurological diseases. However, the full functional relevance of I34-tRNAs in human cells is still not completely understood.

1.6.2 tRNAs in cancer

Cancer is a multifaceted disease characterized by uncontrolled cell proliferation. In order to support their rapid division, cancer cells require a constant supply of proteins. Consequently, cancer cells frequently upregulate the expression of tDNAs to ensure an efficient translation of the mRNAs required for continued cell growth and cell division (Goodarzi et al., 2016; T. Gupta et al., 2022; Pavon-Eternod et al., 2009; Pinzaru & Tavazoie, 2023; Z. Zhang et al., 2018). As mentioned previously, overexpression of tRNA levels via Pol III regulation is an important mechanism in the control of tumorigenesis, as several

well-known oncoproteins and tumor suppressors can modulate Pol III activity (Hummel et al., 2019; Orellana et al., 2022).

Changes in the biogenesis of tRNAs are also related with tumor development and progression. For instance, alterations in tRNA modifications, in most cases produced by aberrant expression of modification enzymes have been linked to cancer (Torres et al., 2014a). Specific tRNA isoacceptors and post-transcriptional base modifications regulate cancer cell survival and influence metastatic potential by controlling the translation of genes related with proliferation (Earnest-Noble et al., 2022; García-Vílchez et al., 2023). Furthermore, tsRNAs have emerged as significant players in almost all major cancers (P. Anderson & Ivanov, 2014; Pekarsky et al., 2023; Soares & Santos, 2017). As an example, a well-known 3' U-tRF (identified as tRF-1001), which is processed from pre-tRNA-Ser, is prevalent in various cancer cells and appears to play a significant role in promoting proliferation in prostate cancer (Y. S. Lee et al., 2009). This underscores that the overexpression of tRNAs in cancer can also be related to non-canonical tRNA functions (Cabrelle et al., 2024; Huang et al., 2018).

As mentioned previously, cancer cells can benefit from mistranslation (Kochavi et al., 2024; M. Santos et al., 2018). This phenomenon allows cancer cells to acquire aberrant proteins, increasing proteome diversity produced by non-genetic alterations of protein-coding genes (Weller et al., 2025). Therefore, mistranslation could enable the production of oncoproteins (e.g., modifying the polypeptide sequence of signaling or regulatory proteins). Mistranslation can occur due to a dysregulation of the tRNA pool, which can increases the provability to introduce non-cognate amino acids during translation (M. Santos et al., 2019; Weller et al., 2025). For example, metabolic and therapeutic stresses, such as chemotherapy, induce aberrant protein production through altered tRNA pool (Kochavi et al., 2024; Wernaart et al., 2024; C. Yang et al., 2024). Therefore, alterations in the tRNA pool could be used by cancer cells as a mechanism for adaptive mistranslation to reshape the cancer proteome and support tumor adaptation and therapy resistance.

1.7 Experimental and computational challenges in tRNA analysis

To fully understand tRNA biology and its role in disease, several tools have been developed to analyze the tRNA pool. However, due to the inherent complexity of tRNA biology characterizing the tRNA pool is quite complex and is often followed by numerous experimental and computational challenges.

Nowadays, the detection of tRNA modifications and the quantification of tRNAs can be achieved through different methodologies. For instance, tRNA modifications can be identified and quantified using liquid chromatography coupled with mass spectrometry (LC/MS) (C. T. Y. Chan et al., 2010), whereas tRNA quantification can be achieved through microarrays (Dittmar et al., 2005; Pavon-Eternod et al., 2010). However, the gold standard methodology used to characterize the tRNA pool is high-throughput sequencing of small RNA sequences (tRNA-Seq) (Cozen et al., 2015; Gogakos et al., 2017; Hu et al., 2021; Shigematsu et al., 2017; Zheng et al., 2015). This is because tRNA-Seq can quantify tRNA abundance and detect tRNA modifications simultaneously. However, several experimental caveats need to be addressed when working with tRNA-Seq data. Those caveats come from the tRNA biology itself. For instance, post-transcriptional modifications in tRNAs can hinder the reverse transcription step, which is crucial to obtain complementary DNA (cDNA) from tRNA for sequencing **(Fig. 14)**. These modifications can cause the reverse transcriptase (RT) enzyme to stall or stop prematurely, leading to "RT-blocks", which results in truncated cDNA fragments that produce short sequencing reads that do not represent full length tRNAs (Torres et al., 2019). In other cases, instead of pausing RT, some modifications can cause the reverse transcriptase to insert an incorrect nucleotide, leading to misincorporations (W. Zhang et al., 2022) **(Fig. 14)**. Furthermore, the stable L-shaped tertiary structure of tRNAs can act as a physical roadblock during RT, which results in sequencing bias and reduces tRNA reads content (Torres et al., 2019).

To overcome these caveats, different methods have focused on the development of specific experimental protocol variations (Cabrelle et al., 2024; W. Zhang et al., 2022). Some strategies include the circularization of cDNA previous to DNA amplification, to allow the detection of those short sequences produced due to RT-blocks (e.g., CircRNA-Seq) (Zheng et al., 2015). The hydrolysis or fragmentation of tRNAs to avoid issues coming from their secondary structure (e.g., Hydro-tRNA-Seq (Gogakos et al., 2017)). The removal of specific post-transcriptional modifications like methylations (e.g., DM-tRNA-Seq (Zheng et al., 2015), ARM-seq (Cozen et al., 2015)) or other modifications (e.g., RNA bisulfite sequencing (Schaefer et al., 2009), DM-Ψ-seq (Song et al., 2020), HAC-seq (Cui et al., 2021) or PANDORA-Seq (J. Shi et al., 2021)). Other approaches involve the use of highly processive reverse transcriptase enzymes (e.g., mim-tRNA-Seq (Behrens et al., 2021)), or the optimization of library construction via the ligation of specific adapters (e.g., YAMAT-seq (Shigematsu et al., 2017), QuantM-tRNAseq (Pinkard et al., 2021), AQRNA-seq (Hu et al., 2021), LOTTE-seq (Erber et al., 2020), ALL-tRNAseq (Scheepbouwer et al., 2023)).

However, those strategies still face significant caveats. These methods still introduce biases during library preparation, which leads to inaccurate representations of tRNA populations. Consequently, Nanopore sequencing emerges as a promising alternative, as it directly sequences native RNA overcoming the problematic cDNA synthesis step, allowing the unbiased detection of tRNA abundance and tRNA modifications (Lucas et al., 2024). Nevertheless, Nanopore sequencing is not yet widely adopted for tRNA analysis. This is because the adaptation of standard Nanopore protocols is needed to efficiently handle tRNAs. For example, specific protocols are needed to handle RNA molecules shorter than 200 nt, such as tRNAs, and to accurately interpret complex basecalling errors as genuine modification signals (Lucas et al., 2024). Such specialized experimental and computational approaches are not yet available in all laboratories interested in tRNA analysis. For that reason, many laboratories still use small RNA-Seq to analyze the tRNA pool. As well, many datasets of small RNA-seq data are available in different repositories which can be reused to analyze the tRNA pool once new bioinformatic strategies and tools to study tRNAs are developed.

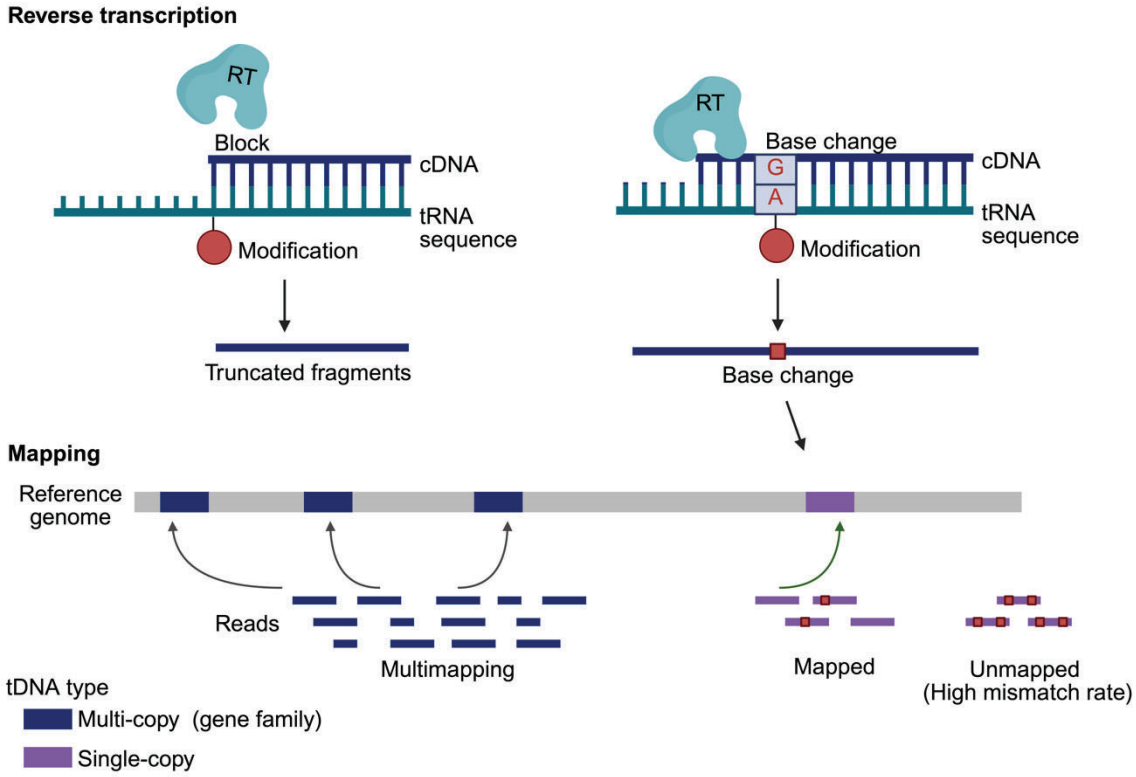


Figure 14. Experimental and computational challenges associated with tRNA-Seq analysis. First, during reverse transcription, if the tRNA sequence contains modifications, reverse transcriptase (RT) can stop the synthesis of complementary DNA (cDNA), producing RT-blocks and leading to truncated reads. If synthesis continues, it can produce base changes, leading to reads containing mismatches. Such truncated fragments do not represent the full-length tRNA sequence, and if specific experimental approaches are not used, such as the circularization of DNA, these truncated reads can be lost. During the mapping of tRNA-derived reads, reads containing a high rate of base modifications can be unmapped because of a high mismatch rate. Moreover, sequence similarity between tDNAs (multicopy tDNAs) can produce ambiguous read assignments, leading to multimapping. Created with Biorender.

The computational analysis of tRNA-Seq data presents several challenges that compromise the alignment of tRNA-related reads and complicates the interpretation of the data (A. Hoffmann et al., 2018; Padhiar et al., 2024). These challenges include the previously mentioned RT-blocks and RT misincorporations caused by the chemical modifications of tRNAs. By analyzing each of the signatures of RT-blocks (indicated by a drop in coverage immediately after a known modification site) or RT misincorporations (mismatches in the alignment at specific positions), it is possible to detect the presence of certain tRNA modifications (Motorin & Helm, 2019, 2024; Ryvkin et al., 2013). Examples of modifications that can produce these signatures include methylations such as m^3C, m^1A, m^1I, $m^{2,2}G$, and m^1G. For instance, inosine frequently changes to guanosine during RT, leading to characteristic A-to-G changes in the resulting reads (W. Zhang et al., 2022). Consequently, standard mapping strategies can classify *bona fide* tRNA reads as unmapped, as tRNA-derived reads can contain base changes that lead to several mismatches within only 76 nucleotides, frequently exceeding the typical thresholds for mapping accuracy.

The complexity of tRNA biology further complicates the accurate assignment of the biological origin of the reads. The sequencing biases produced due to the experimental procedures can complicate the differentiation between pre-tRNAs, mature tRNAs and tsRNAs (Torres et al., 2019). Moreover, as mentioned, tDNAs are highly conserved, often with multiple identical sequences, this leads to multimapping as tRNA reads can map to multiple tDNA locations **(Fig. 14)**. As a result, specific approaches are required to identify and control for multimapping (A. Hoffmann et al., 2018; Torres et al., 2019). Additionally, the short length of the resulting reads, the presence of introns and the post-transcriptional addition of the 3'-CCA tail further contributes to making tRNA-related reads a particular difficult case for read mapping algorithms. Beyond alignment, there is also the essential need to further computationally process the data to classify tRNA-related reads according to their nature, such as distinguishing reads that come from mitochondrially encoded or nuclear-encoded tDNAs, as well as pre-tRNAs and mature/processed tRNAs sequences.

As a consequence, specific mapping strategies are required to analyze expression levels and chemical modifications produced in tRNAs. Different approaches have been tested in order to achieve those goals, such as mapping against native genome further extended with mature tRNA sequences (Cozen et al., 2015; A. Hoffmann et al., 2018), or directly mapping against custom genomes with only tRNA sequences (Behrens et al., 2021; Clark et al., 2016; Hauenschild et al., 2015; Selitsky & Sethupathy, 2015). Most of the pipelines do not include pre-tRNA data and only focus on the analysis of mature or processed sequences (Behrens et al., 2021; P. P. Chan et al., 2025; Clark et al., 2016; Hauenschild et al., 2015; Holmes et al., 2022; Selitsky & Sethupathy, 2015). Other strategies focus on the detection of tsRNAs (P. P. Chan et al., 2025; Donovan et al., 2021; La Ferlita et al., 2025; Loher et al., 2017; Zahra et al., 2023). While others focus on the differential expression analysis of tRNAs (Holmes et al., 2022; J.-O. Lee et al., 2022; Scheepbouwer et al., 2023; Torres et al., 2019). Additionally, there are other approaches developed in order to detect more effectively modifications in tRNA sequences (Behrens et al., 2021; Cozen et al., 2015; A. Hoffmann et al., 2018; Selitsky & Sethupathy, 2015).

Each of the described strategies for tRNA-Seq analyses focuses on a specific outcome (e.g., analyzing only mature sequences). However, phenotypes related to tRNA alterations can arise from disruptions at any step during tRNA biogenesis. Additionally, the analysis of tRNA-Seq data requires strong bioinformatics and programming skills. Therefore, developing integrated pipelines that enable the characterization of the tRNA pool from multiple perspectives, alongside user-friendly tools that facilitate tRNA-Seq analysis for a broader range of researchers without computational expertise, will facilitate the analysis of tRNA-Seq data.

2 OBJECTIVES

The main goal of this thesis is to provide new insights into different aspects of tRNA biology by developing and applying diverse bioinformatic approaches. To address this objective, the following aims were pursued:

1. To develop an integrated bioinformatics pipeline specifically designed to tackle challenges in tRNA-seq data analysis and extract information on tRNA gene expression, processing and modification patterns.

2. To apply the tool developed in Objective 1 for the detailed study of a highly relevant process in tRNA biology such as the importance of I34-tRNAs.

3. To characterize the genomic localization and organization of tDNAs across the human genome and investigate how their genomic organization may influence tDNA transcription.

4. To analyze the somatic mutational landscape of tDNAs in order to elucidate the mechanisms and potential effects of tDNAs mutagenesis.

3 RESULTS

Report on Marina Murillo Recio's Research Contributions

As supervisors of Marina's Ph.D. thesis, we are pleased to report that Marina has made very significant contributions across several key projects in our lab. First, she constructed tRNAstudio, a pipeline to study the presence of tRNA and tsRNA reads in sequencing datasets that now we use routinely to analyze the dynamics of tRNA populations. At the same time, she has helped in interpreting the tRNA-seq data from several projects in our group and those of collaborators. Once she had produced tRNAstudio she focused on the mapping of tRNA genes in the human genome, which resulted in a high-resolution map that has become the basis for many future projects in the group. She continued to use this data to study the distribution of somatic mutations at human tRNA genes, an exercise that has produced a very exciting result that will be the basis for our future research in the years to come. So, to summarize, Marina's scientific contribution is outstanding, and has and will have a lasting impact in the understanding of the dynamics of the populations of human tRNAs in cancer and aging.

Published and submitted articles related to this thesis:

3.1. tRNAstudio: facilitating the study of human mature tRNAs from deep sequencing datasets.

Analyzing tRNA-seq data is challenging due to the inherent complexity of tRNA biology. To address this, we developed tRNAstudio, a pipeline designed to provide the scientific community with a comprehensive tool for extensive analysis of the tRNA pool within a specific biological context. Marina was the core developer of this pipeline, including the crucial task of writing code for both the bioinformatics and statistical analyses, as well as generating all associated plots and figures. Furthermore, she contributed significantly to the writing of the manuscript describing tRNAstudio.

*Note that the supplementary tables generated in this work are available online. Due to their large size, they are not included in this thesis.

Authors: Murillo-Recio, M., Martínez de Lejarza Samper, I. M., Tuñí I Domínguez, C., Ribas de Pouplana, L., & Torres, A. G.
Publication date: 7 April 2022
Journal: *Bioinformatics*
Impact factor: *4.4 (2023) Quartile 1 (Q1)*
Publisher: Oxford Academic **DOI:** https://doi.org/10.1093/bioinformatics/btac198

3.2. Human tRNAs with inosine 34 are essential to efficiently translate eukarya-specific low-complexity proteins.

tRNAs harbor numerous modifications, with Inosine at position 34 (I34) being essential for tRNAs decoding expansion. This study focused on elucidating the biological relevance of I34-tRNAs. Marina conducted several key bioinformatics analyses for this project. She applied tRNAstudio pipeline to determine tRNA modification levels (Fig. S2B) and to run differential expression analyses of tRNAs (Fig. 2C and Fig. S2C). On top of that, she contributed on generating the plots for analyzing how I34-tRNAs connect with eukaryotic systems (Fig. 10C, Fig. 10E, Fig. S7). The experimental work underlying this Project was not conducted by Marina. As a bioinformatician, the contribution of Marina was centered on the design, implementation, and interpretation of computational analyses, which complement the experimental results.

*Note that the supplementary tables generated in this work are available online. Due to their large size, they are not included in this thesis.

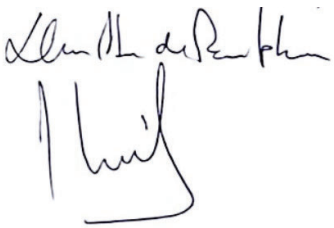
Publication date: 14 June 2021
Authors: Torres, A. G., Rodríguez-Escribà, M., Marcet-Houben, M., Santos Vieira, H. G., Camacho, N., Catena, H., Murillo-Recio, M., Rafels-Ybern, À., Reina, O., Torres, F. M., Pardo-Saganta, A., Gabaldón, T., Novoa, E. M., & Ribas de Pouplana, L.
Journal: *Nucleic Acids Research*
Impact Factor: *16.6 (2023) Quartile 1 (Q1)*
Publisher: Oxford Academic **DOI:** https://doi.org/10.1093/nar/gkab461

Project 3.3: Genomic Organization, Transcription, and Somatic Mutagenesis of tRNA Genes: Implications for Proteome Integrity and Disease

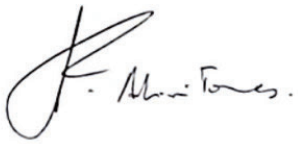
This study aimed to understand several fundamental aspects related to the genomics of tRNA genes (tDNAs). These included: (i) The characterization of the genomic distribution of tDNAs and its correlation with their expression levels. (ii) The analysis of the mutational landscape of tDNAs. For the first part of the study, Marina developed all the necessary code for the bioinformatics and statistical analyses and generated all associated plots and figures. For the second part, which involved the analysis of somatic mutations, we collaborated with Fran Supek's group (Genome Data Science - IRB Barcelona) and she worked under the supervision of Marina Salvadores. In this context, she applied the previously developed code and made the necessary adaptations for tDNA analysis, performing comprehensive statistical analyses and generating the relevant plots and figures. Beyond these substantial technical contributions, Marina also had major contributions to the conceptualization, interpretation, and writing of the manuscript for this study. Note that in this thesis this is included as a chapter that consist on an expanded version of a paper currently submitted for publication.

Publication date: Submitted to Genome Research – 2025
Authors: Murillo-Recio, M., Salvadores, M., Vaquer-Picó, A., Tsapanou, L., Torres, A. G., Supek, F., & Ribas de Pouplana, L.

Dr. Lluís Ribas de Pouplana



Dr. Adrian Gabriel Torres



Gene Translation Laboratory
Institute for Research in Biomedicine (IRB)

3.1

tRNAstudio: facilitating the study of human mature tRNAs from deep sequencing datasets

Bioinformatics, 38(10), 2022, 2934–2936
https://doi.org/10.1093/bioinformatics/btac198
Advance Access Publication Date: 7 April 2022
Applications Note

Sequence analysis

tRNAstudio: facilitating the study of human mature tRNAs from deep sequencing datasets

Marina Murillo-Recio[1], Ignacio Miguel Martínez de Lejarza Samper[1], Cristina Tuñí i Domínguez[1], Lluís Ribas de Pouplana [1,2,*] and Adrian Gabriel Torres [1,*]

[1]Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Catalonia 08028, Spain and [2]Catalan Institution for Research and Advanced Studies, Barcelona, Catalonia 08010, Spain

*To whom correspondence should be addressed.

Associate Editor: Christina Kendziorski

Received on September 10, 2021; revised on March 17, 2022; editorial decision on March 29, 2022

Abstract

Summary: High-throughput sequencing of transfer RNAs (tRNA-Seq) is a powerful approach to characterize the cellular tRNA pool. Currently, however, analyzing tRNA-Seq datasets requires strong bioinformatics and programming skills. tRNAstudio facilitates the analysis of tRNA-Seq datasets and extracts information on tRNA gene expression, post-transcriptional tRNA modification levels, and tRNA processing steps. Users need only running a few simple bash commands to activate a graphical user interface that allows the easy processing of tRNA-Seq datasets in local mode. Output files include extensive graphical representations and associated numerical tables, and an interactive html summary report to help interpret the data. We have validated tRNAstudio using datasets generated by different experimental methods and derived from human cell lines and tissues that present distinct patterns of tRNA expression, modification and processing.
Availability and implementation: Freely available at https://github.com/GeneTranslationLab-IRB/tRNAstudio under an open-source GNU GPL v3.0 license.
Contact: adriangabriel.torres@irbbarcelona.org or lluis.ribas@irbbarcelona.org.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transfer RNAs (tRNAs) are small non-coding RNAs that bring amino acids to the ribosome for protein synthesis. They are transcribed as longer precursor tRNAs (pre-tRNAs) that need to be processed and chemically modified in order to become fully active. Mature tRNAs can also be further processed into tRNA-derived fragments (tRFs) that perform a wide range of non-canonical tRNA functions (Su *et al.*, 2020).

High-throughput sequencing of tRNAs is a powerful approach to study tRNA biology (Torres *et al.*, 2015, 2019). Several methods have been developed to sequence tRNAs, ranging from standard small RNA-Seq to specialized methods such as DM-tRNA-Seq (Zheng *et al.*, 2015), Arm-Seq (Cozen *et al.*, 2015), YAMAT-Seq (Shigematsu *et al.*, 2017), mim-tRNA-Seq (Behrens *et al.*, 2021), AQRNA-Seq (Hu *et al.*, 2021), among others (we herein refer to any deep sequencing method that can detect tRNA reads as 'tRNA-Seq'). However, analyzing tRNA-Seq datasets is computationally challenging and requires specialized bioinformatics and programming skills (Hoffmann *et al.*, 2018).

Here, we present tRNAstudio, an integrative pipeline to analyze human tRNA-Seq datasets that is packaged into a user-friendly graphical user interface (GUI) implemented in local mode. Using publicly available datasets, we show that tRNAstudio can extract information on tRNA expression, processing and post-transcriptional modification status. The pipeline output includes an interactive html summary report, extensive graphical data representations, and spreadsheets useful for custom analyses.

2 Description and implementation

tRNAstudio is implemented as a GUI (Fig. 1), built with the Python library Tkinter, prepared to run in Mac (OS X El Capitan or higher) and Linux-based platforms, and designed for the analysis of human tRNAs using, as input, tRNA-Seq datasets generated by single- or paired-end sequencing. The code was primarily written in Python3 and R. tRNAstudio uses a Conda environment that includes the installation of R package, python modules and all the requirements and dependencies needed to perform tRNA analyses (Bowtie2, Samtools, Bedtools, Pysam, Pysamstats and Picard). To run tRNAstudio, the user will need to be familiar with a command-line interface and simple bash commands. The installation of Conda and the creation of the environment is executed by running the requirements script ('bash requirements.sh'). To activate the Conda environment and to launch the GUI the user needs to run only two

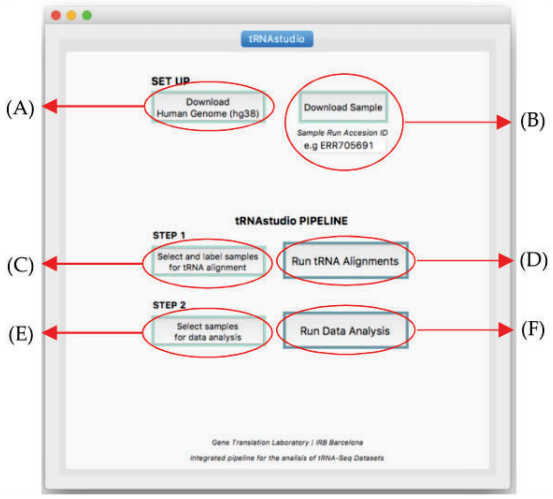


Fig. 1. tRNAstudio GUI visualized in macOS systems

commands: 'conda activate tRNAstudioEnv' and 'python3 tRNAstudioGUI', respectively. Detailed methodological descriptions of tRNAstudio are available in Supplementary Methods. tRNAstudio can be run in standard computers but we recommend at least 8 cores, 16 Gb of RAM and 100 Gb of available ROM. Under these conditions, four samples (around 10 Gb of information per sample) can be analyzed in 2–3 h.

Processing of tRNA-Seq datasets is achieved in six simple steps. The first time tRNAstudio is implemented, the user will download the reference Human Genome hg38 (Fig. 1A). Next, samples (datasets: Fastq files) of interest are directly downloaded from the Gene Expression Omnibus public repository (https://www.ncbi.nlm.nih.gov/gds/) by providing the run accession ID (prefixes SRR . . ., ERR . . ., DRR . . .) (Fig. 1B). Users can also analyze manually downloaded datasets from other repositories, or their own datasets, upon incorporating the desired Fastq files into the 'Fastq_downloaded' folder of tRNAstudio. Samples that are to be aligned against the reference genomes are selected and labeled in a metadata file that is required for comparative analyses among samples (Fig. 1C). Labeling information includes the group to where the sample belongs to (e.g. 'Control' samples and 'Treated' samples) and whether the selected datasets have been generated by single- or paired-end sequencing. The alignment pipeline is then executed (Fig. 1D) and a pop-up dialogue will inform when the alignments are done. The user can choose which of the aligned samples will be used for data analysis (Fig. 1E). Previously aligned samples can also be selected. Last, the data analysis is performed (Fig. 1F), results are saved, and an html summary report is generated to help the user interpret the data. tRNAstudio and detailed instructions of use can be found at https://github.com/GeneTranslationLab-IRB/tRNAstudio.

3 Results

tRNAstudio performs serial alignments against the whole human genome and against custom genomes as depicted in Supplementary Figure S1. Custom genomes are built with all unique human tRNA sequences, either in the form of pre-tRNAs (genomic sequences with 5′- and 3′-flanking regions; only applicable for nuclear-encoded tRNAs) or mature tRNAs (tRNA sequences without flanking and intronic regions, with 3′-CCA trinucleotide addition, and, in the case of $tRNA^{His}$, with a 5′-G addition; applicable to nuclear- and mitochondrial-encoded tRNAs). To aid in the identification of tRNA genes, we provide a file that links each tRNA gene analyzed by tRNAstudio using hg38 to its corresponding gene in hg19, its tRNA 'license plate' (Pliatsika *et al.*, 2016), and a hyperlink to additional gene expression information from MINTbase (Pliatsika *et al.*, 2016; Supplementary Table S1).

Datasets are first aligned against the whole human genome. Reads mapping to non-tRNA genes are discarded and reads mapping to tRNA genes are classified as 'mitochondrial' (if mapped to mitochondrial-encoded tRNA genes) or 'cytosolic' (if mapped to nuclear-encoded tRNA genes). Given the polycistronic nature of mitochondrial transcripts (Ojala *et al.*, 1981), reads classified as derived from mitochondrial tRNAs (mt-tRNA) are then aligned against a custom mature mt-tRNA genome to remove unprocessed mitochondrial transcripts that may partially overlap with mt-tRNA genes. Remaining mapped reads are kept for further analyses. Reads classified as derived from cytosolic tRNAs, and unmapped reads resulting from the initial mapping against the whole genome undergo serial alignments against custom tRNA genomes as described in Supplementary Methods (Supplementary Fig. S1). As a virtue of these serial alignment strategy, tRNAstudio improves both the recovery of reads mapped to nuclear-encoded tRNA genes and the alignment quality of the reads when compared against alignment strategies that use single reference genomes (Supplementary Fig. S2A). Users of tRNAstudio obtain absolute read counts for every nuclear- and mitochondrial-encoded tRNA genes, and their corresponding mapping quality score (MAPQ; Supplementary Methods and Fig. S2B). Of note, tRNAstudio considers all mapped tRNA reads for analysis, regardless of their MAPQ or whether they are derived from tRNAs with or without natural post-transcriptional nucleotide additions (i.e. $tRNA^{His}$ 5′-G or partial or full 3′-CCA) (further details in Supplementary Methods).

Mapped reads are then used for differential tRNA gene expression analyses using two complementary methods: DESeq2 (Love *et al.*, 2014) and iso-tRNA-CP (Torres *et al.*, 2019). Iso-tRNA-CP evaluates the proportional contribution of each tRNA gene to its corresponding isodecoder tRNA set (i.e. individual analyses among all genes having the same tRNA anticodon sequence). Given that mt-tRNA genes are represented by a single isodecoder gene (Juhling *et al.*, 2009), iso-tRNA-CP is only applicable to nuclear-encoded (i.e. cytosolic) tRNAs. Results are accompanied by a principal component analysis and can be visualized through heatmaps and interactive graphs and tables (Supplementary Fig. S3).

tRNAstudio also classifies reads derived from cytosolic pre-tRNAs or processed tRNAs. We validated this function by analyzing datasets enriched in reads derived from pre-tRNAs (Torres *et al.*, 2015) or mature tRNAs (Zheng *et al.*, 2015; Supplementary Fig. S4). The custom tool for the trimming of soft-clipped bases implemented by tRNAstudio aids in the correct assignment of reads to each category, as it specifically detects reads bearing post-transcriptional 3′-CCA additions, or $tRNA^{His}$ 5′-G addition. These modifications are present on processed tRNAs but may otherwise be confused as nucleotides derived from pre-tRNA 3′-trailer or 5′-leader sequences, respectively (further details in Supplementary Methods). We benchmarked this function against a standard tool for trimming soft-clipped bases (Biostar84452 from JVarkit). We find that both 3′-CCA and 5′-G additions are removed from the reads when using JVarkit but are retained when using tRNAstudio's customized tool (Supplementary Fig. S5A and B). Furthermore, tRNAstudio classifies reads as 'likely derived from pre-tRNAs' or 'likely derived from mature tRNAs' (i.e. processed tRNAs), based on the genomic coordinates of the mapped tRNA reads and on the presence or absence of post-transcriptional nucleotide additions (see Supplementary Methods). Using datasets enriched in reads derived from mature tRNAs (Zheng *et al.*, 2015), we find that tRNAstudio assigns 99.3% of tRNA reads to the processed tRNA set when applying its custom classification strategy, while only 68.5% of tRNA reads are classified into this group when using a standard genomic coordinates-based classification method (Supplementary Fig. S5C).

tRNAstudio uses a base-calling function to evaluate tRNA modification levels. Reverse transcriptases generate mutations in the obtained cDNA (and hence, in their derived sequencing reads) when encountering modified tRNA bases. Analyses of datasets with tRNAstudio revealed sequence variations that coincide with tRNA positions known to undergo post-transcriptional modifications such as positions 9 (m^1G), 26 (m^2_2G), 32 (m^3C), 34 (I: inosine), 37 (m^1I) and 58 (m^1A: 1-methyladenosine; de Crecy-Lagard *et al.*, 2019;

Supplementary Fig. S6). Furthermore, analyses of datasets derived from human cell lines depleted of ADAT2, the catalytic subunit of the enzyme that catalyzes A-to-I conversion at positions 34 of tRNAs (Torres *et al.*, 2015), revealed a quantitative decrease in the modification ratio at these positions without changes in the modification ratio of unrelated positions such as 58 (m^1A) (Supplementary Fig. S7A). Likewise, we detected a specific decrease in the modification ratio at positions 58 (m^1A) when analyzing datasets derived from artificially demethylated RNAs (Zheng *et al.*, 2015), without alterations in the modification ratio at positions 34 (I) (Supplementary Fig. S7B). Interactive heatmaps aid in visualizing global base-calling results for each position in every tRNA gene, and evaluating changes in modification ratios at specific tRNA positions when samples are compared (see Supplementary Discussion and Fig. S8).

tRNAstudio also reports on tRNA gene sequence coverages, which can aid in the identification of *bona fide* tRFs. tRFs derived from the 3′-arm of $tRNA^{Arg}_{CCU}$ and $tRNA^{Arg}_{UCG}$, and from the 5′-arm of $tRNA^{Cys}_{GCA}$ were shown to be abundant in human brain (Torres *et al.*, 2019). Using tRNAstudio, we analyzed datasets from human brain and found that tRNA sequence coverages mapped to the abovementioned tRFs (Supplementary Fig. S9).

We show that tRNAstudio can extract biologically relevant information from tRNA-Seq datasets, while allowing the analyses to be performed in local mode and with a user-friendly GUI. This work brings bioinformatics closer to experimental laboratories and will be useful to accelerate the pace at which knowledge on canonical and non-canonical tRNA biology expands (see Supplementary Discussion).

Acknowledgements

We thank Oscar Reina from the Biostatistics and Bioinformatics Core Facility at IRB Barcelona for technical assistance and helpful discussions.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness [PID2019-108037RB-100 to L.R.d.P]. M.M.-R. is funded by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) [2021 FI_B 01053].

Conflict of Interest: none declared.

References

Behrens,A. *et al.* (2021) High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Mol. Cell*, **81**, 1802–1815.e1807.

Cozen,A.E. *et al.* (2015) ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat. Methods*, **12**, 879–884.

de Crecy-Lagard,V. *et al.* (2019) Matching tRNA modifications in humans to their known and predicted enzymes. *Nucleic Acids Res.*, **47**, 2143–2159.

Hoffmann,A. *et al.* (2018) Accurate mapping of tRNA reads. *Bioinformatics*, **34**, 1116–1124.

Hu,J.F. *et al.* (2021) Quantitative mapping of the cellular small RNA landscape with AQRNA-seq. *Nat. Biotechnol.*, **39**, 978–988.

Juhling,F. *et al.* (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Ojala,D. *et al.* (1981) tRNA punctuation model of RNA processing in human mitochondria. *Nature*, **290**, 470–474.

Pliatsika,V. *et al.* (2016) MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments. *Bioinformatics*, **32**, 2481–2489.

Shigematsu,M. *et al.* (2017) YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs. *Nucleic Acids Res.*, **45**, e70.

Su,Z. *et al.* (2020) Noncanonical roles of tRNAs: tRNA fragments and beyond. *Annu. Rev. Genet.*, **54**, 47–69.

Torres,A.G. *et al.* (2015) Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Res.*, **43**, 5145–5157.

Torres,A.G. *et al.* (2019) Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. *Proc. Natl. Acad. Sci. USA*, **116**, 8451–8456.

Zheng,G. *et al.* (2015) Efficient and quantitative high-throughput tRNA sequencing. *Nat. Methods*, **12**, 835–837.

Supplemental data

tRNAstudio: facilitating the study of human mature tRNAs from deep sequencing datasets.

Marina Murillo-Recio[1], Ignacio Miguel Martínez de Lejarza Samper[1], Cristina Tuñí i Domínguez[1], Lluís Ribas de Pouplana[1,2] and Adrian Gabriel Torres[1]

[1] Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Catalonia, 08028, Spain.

[2] Catalan Institution for Research and Advanced Studies, Barcelona, Catalonia, 08010, Spain.

- Supplementary Methods
- Supplementary Figures (S1-S9)
- Supplementary Table S1
- Supplementary Discussion

Supplementary Methods

Deep sequencing datasets

The following datasets have been used in this study: Standard Illumina-Seq of Human embryonic kidney 293T (HEK293T) control cells (shCV) and with ADAT2 KD (shADAT2) (PRJEB8019) (Torres, et al., 2015). Libraries prepared using cDNA circularization without demethylase treatment (CircRNA-Seq) and with demethylase-treated RNA (DM-tRNA-Seq) (GEO: GSE66550) (Zheng, et al., 2015). Datasets of human brain (GEO: GSE43335) (Ryvkin, et al., 2013).

Custom genomes

A total of 619 human hg38 nuclear tRNA gene predictions were downloaded from the Genomic tRNA database (GtRNAdb; release 18.1, August 2019) (Chan and Lowe, 2016). 98 sequences were discarded and 521 tRNA sequences were used to create custom genomes as previously done (Torres, et al., 2019). The 22 human mitochondrial tRNA (mt-tRNA) gene sequences were obtained from https://www.ncbi.nlm.nih.gov/nuccore/251831106 (NC_012920.1). All tRNA gene sequences were crosschecked against a list of human nuclear tRNA-lookalike genes (Telonis, et al., 2014) to ensure that these are not part of tRNAstudio's internal database. However, some nuclear tRNA-lookalike genes have the exact same sequence as that of *bona fide* mt-tRNA genes (Telonis, et al., 2014). Because reads bearing these sequences cannot be unambiguously assigned to *bona fide* mt-tRNA genes, tRNAstudio only evaluates mt-tRNA gene expression and report the data as "Mitochondrial tRNA genes or mitochondrial tRNA-lookalike nuclear sequences" to warn the user that results could be partially biased due to potential read missassignment.

The "pre-tRNA" custom genome for cytosolic tRNAs was generated with the nuclear genomic tRNA sequence and the addition of 50 nucleotides downstream and upstream of each corresponding tRNA gene (to mimic 5'-leader and 3'-trailer sequences), using BEDTools v2.30.0 (Quinlan and Hall, 2010). The "mature tRNA" custom genome was generated upon removal of intronic sequences, addition of a "CCA" at the 3'-end of all tRNA genes and incorporation of a "G" at the 5'-end of $tRNA^{His}$ sequences (G^{-1} modification) (Jackman and Phizicky, 2006), using Python v3.7.10 custom scripts. Mature tRNA and pre-tRNA sequences that had the same sequence (100% of identity) were grouped into families (Torres, et al., 2019). This resulted in 341 unique sequences of human cytosolic mature tRNAs and 490 unique sequences of human cytosolic pre-tRNAs. None of the 22 human mt-tRNA genes were required to be grouped into families.

tRNAstudio uses the tRNA gene name format reported by the GtRNAdb for the human hg38 genome (i.e. tRNA-AminoAcid-Anticodon-FamilyNumber-GeneNumber). However, this nomenclature can change in future updates of the GtRNAdb or when new assemblies of the human genome are reported. Therefore, **Supplementary Table S1** shows for each hg38 tRNA gene that tRNAstudio analyses, their correspondence with the tRNA gene set from the hg19 assembly (obtained using the UCSC LiftOver tool; https://genome.ucsc.edu/cgi-bin/hgLiftOver), and tRNA "license plates" (Pliatsika, et al., 2016). License plates are sequence-based and should remain unchanged over time. License plates were obtained by running the license plate code available via MINTbase (https://cm.jefferson.edu/MINTbase/)(Pliatsika, et al., 2016). Genes depicted in **Supplementary Table S1** are further hyperlinked to their corresponding entry in the MINTbase website (when available) so that the user of tRNAstudio can obtain additional information for their tRNA gene of interest.

Mapping and classification of tRNA reads.

The mapping strategy used by tRNAstudio is shown in **Supplementary Fig. S1**. When datasets generated by paired-end sequencing are used as input for the analyses, tRNAstudio will merge pair-end reads using PEAR v0.9.6 (Zhang, et al., 2014) and handle the generated merged dataset as one obtained by single-end sequencing. The first time tRNAstudio is implemented by the user, the whole human genome hg38 needs to be downloaded. This action also automatically builds the genome indexes for this file and for all other custom reference genomes using Bowtie2 (Bowtie2-build). Alignments against reference genomes are carried out with Bowtie2 v2.4.2 (Langmead and Salzberg, 2012) using the local alignment mode allowing 0 mismatches in the seed and using default options (bowtie2 --local -N 0). The last mapping step is done with Bowtie2 but allowing 1 mismatch in the seed and using default options (bowtie2 --local -N 1). This step identifies remaining reads derived from fully modified mature tRNAs that may contain a higher degree of mismatches, insertions or deletions due to post-transcriptional tRNA modifications that interfere with the accurate RT of tRNAs. Such reads may contain too many discrepancies against the reference genome that can prevent their alignment using the 0 mismatch in the seed alignment mode of bowtie2, but are more likely to map when using the 1 mismatch in the seed alignment mode of bowtie2.

Upon aligning datasets to the full human genome, reads mapping to tRNA genes are extracted using SAMtools v1.12 (Li, et al., 2009), while reads mapping to other portions of the genome are discarded. Because reads derived from mature tRNAs can contain a 3'-CCA (and they may also contain a 5'-G if derived from $tRNA^{His}$) that would not map when using the complete human genome, alignment tools can interpret these bases as bases that require soft clipping and removal. To overcome this issue, tRNAstudio performs a custom trimming of soft clipped bases (**Supplementary Fig. S1A**) by processing the alignment files using the Python utility Pysam v0.15.4 (github.com/pysam-developers/pysam). Using the genomic coordinates of tRNAs and the coordinates to which the read maps, soft clipped bases at the 3'-end that correspond to "CCA" for all tRNA genes and soft clipped bases that correspond to "G" at the 5'-end of $tRNA^{His}$ genes, are identified and retained in the sequencing read. We have benchmarked this function against the tool for trimming soft-clipped bases from JVarkit (Biostar84452) (Lindenbaum, 2015) (**Supplementary Fig. S5A-B**). The remaining soft clipped sequences are removed (adapters, low base quality, etc.). This approach allows the processing of adapter sequences without the need to specify the sequence of the adapter used for library preparations. We verified that none of the main commercially available library preparation adapters (NEB, Illumina, Qiagen and Lexogen) starts with "CCA". However, if the sequence of the adapters used for library preparation would start with "CCA", reads derived from some processed tRNAs (i.e. those derived from tRNA transcripts after 3'-trailer removal but before 3'-CCA addition) will be misclassified as reads derived from processed tRNAs after 3'-CCA addition. While this potential artifact could overestimate the presence of 3'-CCA on processed tRNAs, it will not affect the classification of pre-tRNA vs processed tRNA reads implemented by tRNAstudio (see below). After whole genome alignments, reads are classified as mt-tRNAs (if they map to mt-tRNA genes) or cytosolic tRNAs (if they map to nuclear-encoded tRNA genes). Reads mapping to mt-tRNA genes are then aligned against a custom mature mt-tRNA genome (see above) to remove reads derived from unprocessed mitochondrial polycistronic transcripts that only partially map to mt-tRNA genes. Reads mapping to cytosolic tRNAs are then classified as likely derived from pre-tRNAs or mature tRNAs.

Given that reads mapping to mature tRNA sequences can also frequently map to pre-tRNA sequences, we use the term "processed tRNAs" to classify reads likely derived from "mature tRNAs" (i.e. reads that do not possess 5'-leader, 3'-trailer or intronic regions), as previously proposed (Torres, et al., 2015). Classification of reads derived from pre-tRNAs or processed tRNAs is done using custom Python

scripts that implement Picard v2.25.5 (http://broadinstitute.github.io/picard/) and SAMtools. For some human tRNA genes, 3'-trailer regions (present in pre-tRNAs) can start with "C", "CC" or "CCA". Mapping of reads to these genes results in a misclassification of mature tRNA-derived reads (that contain a post-transcriptionally added 3'-CCA tail) into pre-tRNA-derived reads when using standard methods based on the genomic coordinates of mapped tRNA reads that do not take into consideration the presence of post-transcriptional tRNA nucleotide additions (**Supplementary Fig. S5C**). To overcome this issue, tRNAstudio performs a custom evaluation of the mapping coordinates of tRNA reads considering the potential presence of post-transcriptional tRNA nucleotide additions. This method assumes that 3'-trailer processing by RNase Z is exact (i.e. will not leave 1-3 nucleotides of the 3'-trailer sequence on processed tRNAs). Therefore, reads that map to the mature tRNA sequence and that only possess "C", "CC" or "CCA" 3'-extensions that can map to their 3'-trailer pre-tRNA sequence are considered as reads derived from mature tRNAs (with partial or full CCA addition) and thus classified as "processed tRNAs" (**Supplementary Fig. S5C**). Note however that although these instances are detected by tRNAstudio, tRNAstudio is not designed to quantitatively study the presence/absence of full or partial 3'-CCA tails on tRNAs. We also noticed that given the alignment algorithm of Bowtie2 that allows some degree of mismatches, the "CCA" extension could also be mapped to trailer sequences starting with "GCA", "TCA", or "ACA". The custom evaluation of the mapping coordinates of tRNA reads implemented by tRNAstudio detects these instances and classify these reads as derived from processed tRNAs. Of note, we did not observe cases where "CCA" extensions would be mapped to trailer sequences starting with "CGA", "CAA" or "CTA". An equivalent approach is implemented by tRNAstudio to consider 5'-G additions on reads derived from $tRNA^{His}$ genes.

Mapping quality (MAPQ) values for each tRNA gene are extracted from the Binary Alignment Map (BAM) file using SAMtools (samtools view –c –bSq 3). A MAPQ score above 2 is indicative of reads assigned with high confidence to a single gene, while a MAPQ score equal or below 2 suggests that the read is highly likely to map to multiple genes (Langmead and Salzberg, 2012). The proportion of reads for each MAPQ score is computed with R and plots are generated using ggplot2 v3.3.5 (Wickham, 2009). When reads can map equally well to different tRNA genes, Bowtie2 will randomly choose one of those genes to assign the reads in question, but this will be reflected in the MAPQ score for reads mapping to that given tRNA gene (i.e. lower MAPQ). In this sense, the mapping to sequential genomes used by tRNAstudio improves MAPQ scores by first assigning unambiguous reads using stringent mapping parameters and only in the last mapping step uses relaxed mapping parameters to assign remaining reads (see **Supplementary Discussion**, **Supplementary Figs. S1** and **S2**). Users of tRNAstudio receive as output the MAPQ scores of reads mapping to every tRNA gene to verify the robustness of the alignment to their tRNA gene of interest. For reads mapping to nuclear-encoded tRNA genes, tRNAstudio reports the MAPQ values obtained after their last mapping step (i.e. after alignment to the custom pre-tRNA genome for pre-tRNA reads and after alignment to the custom mature tRNA genome with relaxed parameters for processed tRNA reads). For reads mapping to mt-tRNA genes, tRNAstudio reports the MAPQ values obtained after whole genome alignment. This is done because i) the custom mature mt-tRNA genome contains only 22 sequences and thus MAPQ values will artifactually be above 2 for all reads, and ii) MAPQ values below 2 obtained after whole genome alignment can inform on reads that can potentially map to mt-tRNA lookalike nuclear sequences (see above).

Read counts of pre-tRNAs and processed tRNAs are obtained using the function featureCounts from the package Subread v2.0.1 (Liao, et al., 2014). Total number of reads for each tRNA gene is calculated by aggregating both pre-tRNA and processed tRNA counts. Report plots showing the total number of reads at isoacceptor, isodecoder, and gene level, together with plots showing the proportion between

processed and pre-tRNAs for each tRNA gene are generated using ggplot2. Expression results for mt-tRNA genes are represented in additional plots. When developing tRNAstudio, all steps were validated and crosschecked by visualizing results using the Integrative Genomes Viewer (IGV) (Robinson, et al., 2011). tRNAstudio scripts can be obtained from https://github.com/GeneTranslationLab-IRB/tRNAstudio.

Differential expression analysis

Differential expression of tRNA genes is performed using the R package DESeq2 v1.32.0 (Love, et al., 2014), based on a negative binomial GLM fit and Wald significance test. Principal component analysis (PCA) to verify the variability observed within and among samples, is also generated using DESeq2. Normalized gene expression values are used to generate heatmaps with the R package Pheatmap v1.0.12. Benjamini-Hochberg-adjusted Wald test p-values are used to control the false discovery rate (FDR). Differentially expressed genes that have an adjusted p-value below 0.05 are considered to be statistically significant. Benjamini-Hochberg-adjusted Wilcoxon signed rank test is applied to analyze differential tRNA gene expression by iso-tRNA-CP (Torres, et al., 2019). Differentially expressed genes that have an adjusted p-value lower than 0.05 are considered to be statistically significant. To maximize the ability to identify differentially expressed tRNA genes, tRNAstudio does not consider specific fold change thresholds for differential expression. The R packages Glimma v2.2.0 (Su, et al., 2017) and ReportingTools v2.32.0 (Huntley, et al., 2013) are used to generate interactive gene expression plots and display the obtained statistics. Given that the mitochondrial genome only encodes a single isodecoder for each mt-tRNA (Juhling, et al., 2009), tRNAstudio reports on differential expression of mt-tRNA genes only by means of DESeq2.

tRNA modification and sequence coverage analyses

The consensus nucleotide positions in each tRNA sequence was defined based on the tRNA secondary structure. Secondary structures of each tRNA were obtained using tRNAscan-SE v2.0.5 (Chan, et al., 2021). For each isodecoder set, the most common secondary structure was chosen to define the consensus tRNA residue positions from residues 1 to 76 (positions 74, 75 and 76 being the 3'-CCA sequence). Using such structure as reference, corrections were applied to other tRNAs within the same isodecoder set that presented different lengths. Nucleotides not mapping to the consensus 1-76 positions are denoted with an upper-case letter (e.g. residues in the variable loop can be indicated as positions 45, 45A, 45B, etc.). We further verified that the universally conserved positions U8, G10, A14, G18, G19, A21, U33, G53, U54, U55, C56, A58 and C61 were correctly assigned to each tRNA sequence. A few sequences presented minor exceptions to these conserved positions consistent with those reported in available databases (Lin, et al., 2019). Graphical representations of tRNA gene sequence coverage obtained by tRNAstudio show mature tRNA sequences without inclusion of 3'-CCA tails.

The base-calling function of tRNAstudio calculates the proportion of mismatches observed in sequencing reads with respect to the genomic mature tRNA sequence ('Modification ratio'). Base calling analysis to estimate tRNA modification levels is carried out using the Python utility pysamstats v1.1.2 (https://github.com/alimanfoo/pysamstats). All bases different from the reference sequence are used to compute the mismatches. Interactive heatmaps showing modification ratios along with the expected/observed nucleotide(s) are generated using the R package heatmaply v1.2.1 (Galili, et

al., 2018). ggplot2 is used to illustrate the modification ratios and sequence coverage pattern plots. For each tRNA sequence, tRNAstudio generates graphical outputs, numerical tables, and interactive heatmaps indicating modification ratios. Data from genes with a low number of mapped reads (< 50 reads) can be filtered out to increase the robustness of the analyses.

When analysing multiple samples, an interactive heatmap is also generated to compare the changes in modification ratios at validated tRNA modification sites (positions 9, 26, 32, 34, 37, 45F and 58) at tRNA gene level. Position 45F refers to a residue within the variable loop, note that this nomenclature may differ from other studies (de Crecy-Lagard, et al., 2019). To aid in heatmap visualization, changes in modification ratios are represented as a Relative Difference [(Group2 - Group1)/mean groups]. However, tRNAstudio also generates spreadsheets with log2FC values in modification ratios [log2(Group2) - log2(Group1)] and their associated statistical significance. Statistical significance is obtained with a Fisher exact test and Bonferroni adjusted p-values. Log2FC and their associated statistics can also be visualized in the interactive heatmap upon hoovering the mouse pointer on the desired tRNA position (see **Supplementary Fig. S8B**). To avoid misleading conclusions due to low number of mapped reads, the heatmap will show changes in modification ratio only for positions where the coverage is greater than 50 reads in both compared groups, and the proportion of modified reads is equal or greater than 10 % in at least one of the compared groups.

Supplementary Figures

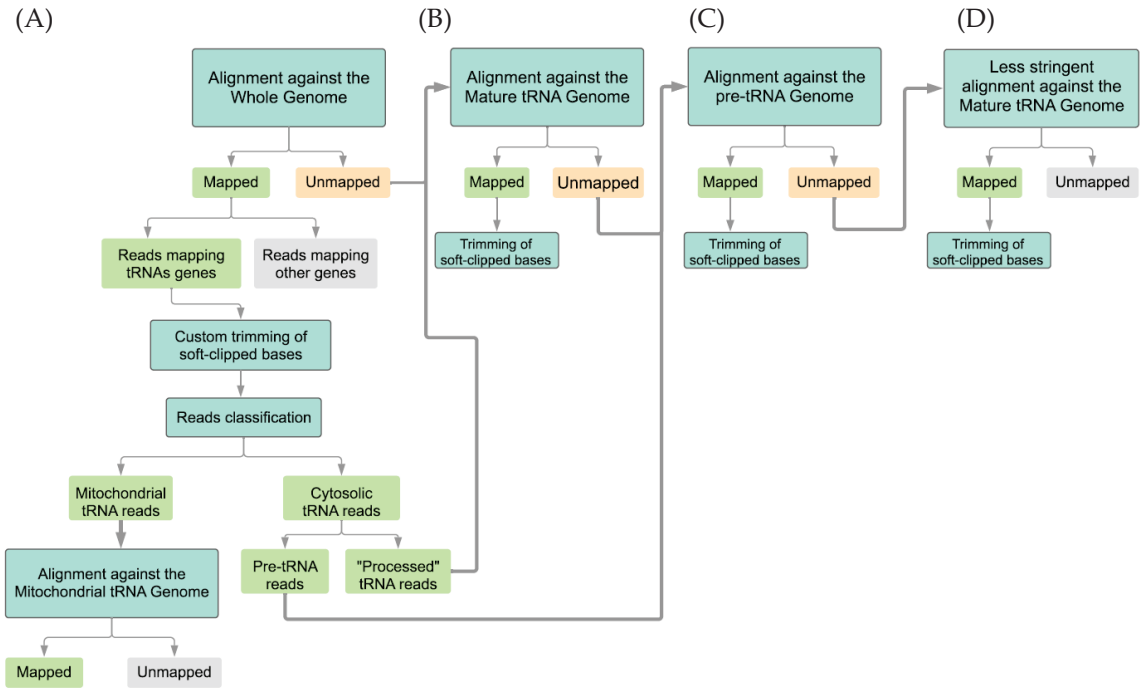


Figure S1. Schematic representation of the mapping workflow of tRNAstudio. **(A)** Mapping against the standard human hg38 genome. Reads mapping to tRNA genes are classified into “Mitochondrial” or “Cytosolic” tRNA reads. Cytosolic tRNA reads are further classified into “pre-tRNA reads” if their sequences map to 5’-leader, 3’-trailer or intronic regions of the tRNA; or into “processed tRNA reads” if their sequences map to the expected sequence of mature tRNAs. A custom function for the trimming of soft clipped bases aids in this classification step (see **Supplementary methods**). Mitochondrial tRNA reads are further aligned against a custom mitochondrial tRNA genome that contain the expected mature mitochondrial tRNA sequences. This step removes reads derived from unprocessed mitochondrial transcripts that have partially overlapped with mitochondrial tRNA genes. **(B)** “Processed tRNA reads” and reads left unmapped after whole genome alignment are then aligned against a custom genome built with mature tRNA sequences. **(C)** Reads that still remain unmapped, and reads previously classified as “pre-tRNA reads” are aligned against a custom genome built with pre-tRNA sequences. **(D)** Unmapped reads are aligned against the mature tRNA custom genome but reducing alignment stringency. Arrows indicate the direction of the data flow. Dark green boxes represent data processing actions. Light green, orange and grey boxes represent accepted reads, unmapped reads that undergo further mapping, and discarded reads, respectively. Additional information available in **Supplementary methods**.

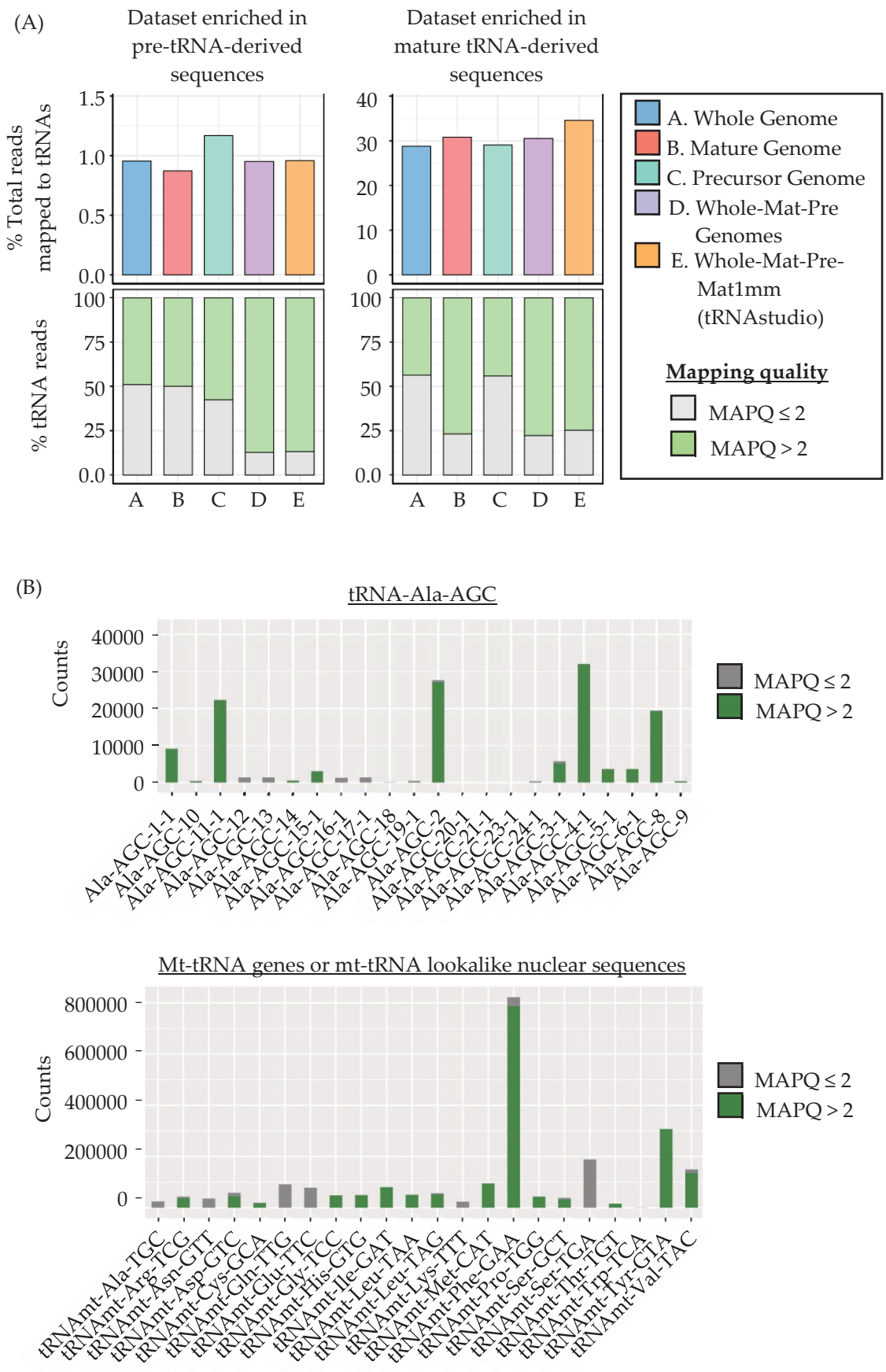
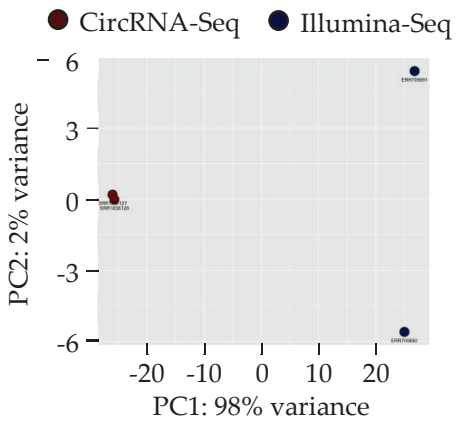

(A)
Dataset enriched in pre-tRNA-derived sequences
Dataset enriched in mature tRNA-derived sequences
% Total reads mapped to tRNAs
% tRNA reads
A. Whole Genome
B. Mature Genome
C. Precursor Genome
D. Whole-Mat-Pre Genomes
E. Whole-Mat-Pre-Mat1mm (tRNAstudio)
Mapping quality
MAPQ ≤ 2
MAPQ > 2
(B)
tRNA-Ala-AGC
Counts
Mt-tRNA genes or mt-tRNA lookalike nuclear sequences

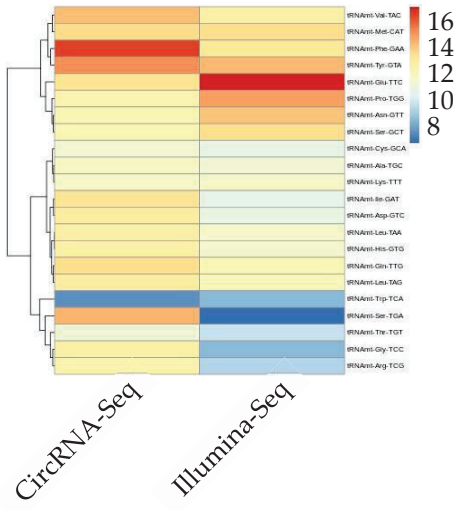
Figure S2. Percentage of total reads aligned to tRNA genes (upper panels) and, within those reads, the percentage of reads with high (MAPQ > 2) or low (MAPQ ≤ 2) mapping quality (lower panels). Datasets are aligned to single genomes (strategies A, B and C) or to multiple genomes following the tRNAstudio workflow (D and E). Results obtained for datasets enriched in reads derived from pre-tRNAs ("Illumina-Seq" datasets (Torres, et al., 2015); left panels) or from mature tRNAs ("CircRNA-Seq" datasets (Zheng, et al., 2015); right panels) are shown. Abbreviations: "Whole", whole human hg38 genome; "Mat", mature tRNA custom genome; "Pre", pre-tRNA custom genome; "Mat1mm", mature tRNA custom genome allowing 1 mismatch. **(B)** Examples of the graphical output for total counts of reads mapping to nuclear-encoded tRNA-Ala-AGC genes (i.e. cytosolic tRNAs) (upper panel) and mitochondrial tRNA (mt-tRNA) genes (lower panel) obtained for datasets enriched in reads derived from mature tRNAs (Zheng, et al., 2015). Note that some mt-tRNA genes cannot be distinguished from mt-tRNA lookalike nuclear sequences (see **Supplementary methods**). The proportion of reads with high (MAPQ > 2; green) or low (MAPQ ≤ 2; grey) mapping quality is indicated. Reads of low MAPQ are likely to map to other portions of the genome.

(A)



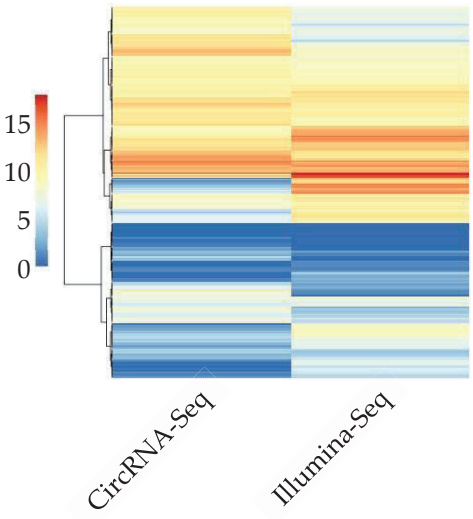
(B)

DESeq2 (mt-tRNA genes)

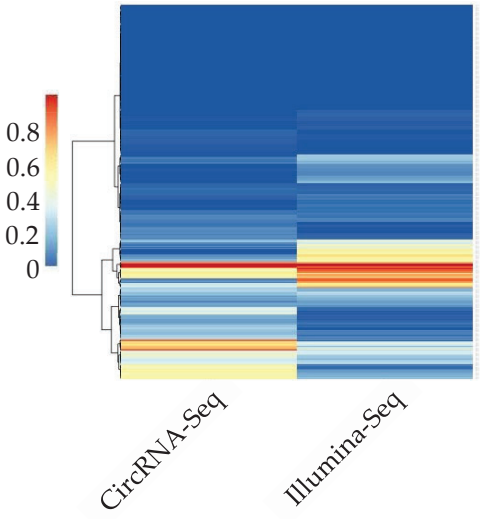


(C)

DESeq2



iso-tRNA-CP



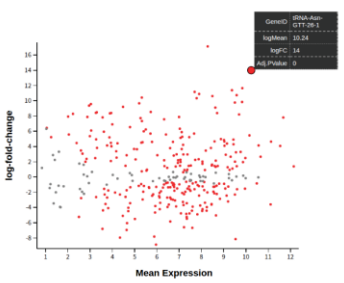
(D)

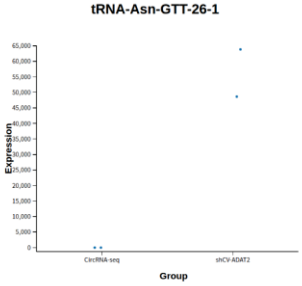
10 records per page Search all columns:

genes	Proportion_CircDM	Proportion_Illu	Fold_change	pvalue	Adjusted_pvalue
tRNA-Ala-AGC-1-1	0.06790000	0.0605000	0.89100	0.167	0.185
tRNA-Ala-AGC-10	0.00330000	0.0005470	0.16600	0.167	0.185
tRNA-Ala-AGC-11-1	0.16600000	0.0188000	0.11400	0.167	0.185
tRNA-Ala-AGC-12	0.01030000	0.0019900	0.19300	0.167	0.185
tRNA-Ala-AGC-13	0.01080000	0.0034700	0.32200	0.167	0.185
tRNA-Ala-AGC-14	0.00414000	0.0009160	0.22100	0.167	0.185
tRNA-Ala-AGC-15-1	0.02330000	0.0645000	2.77000	0.167	0.185
tRNA-Ala-AGC-16-1	0.01000000	0.0003690	0.03680	0.167	0.185
tRNA-Ala-AGC-17-1	0.01050000	0.0014500	0.13800	0.167	0.185
tRNA-Ala-AGC-18	0.00057900	0.0920000	159.00000	0.167	0.185

Showing 1 to 10 of 328 entries ← Previous 1 2 3 4 5 Next →

(E)





Search:

GeneID	logMean	logFC	Adj.PValue
tRNA-Asn-GTT-26-1	10.24	14	1.307e-221
tRNA-Val-CAC-11	9.373	11.38	4.743e-212
tRNA-Gly-GCC-3-1	9.505	9.764	1.192e-211
tRNA-Gly-CCC-6-1	9.585	10.73	5.166e-162

Figure S3. Examples of the data outputs generated for differential tRNA gene expression when comparing datasets generated by CircRNA-Seq (Zheng, et al., 2015) and Illumina-Seq (Torres, et al., 2015). **(A)** Principal component analysis (PCA) of the analyzed samples. **(B)** Heatmap displaying individual mitochondrial tRNA gene expression by means of DESeq2. **(C)** Heatmap displaying individual nuclear-encoded tRNA gene expression by means of DESeq2 and iso-tRNA-CP. Each line on the heatmap represents a tRNA gene. **(D)** Interactive table to browse results on differential tRNA gene expression by iso-tRNA-CP. **(E)** Interactive function to browse results on differential tRNA gene expression by DESeq2.

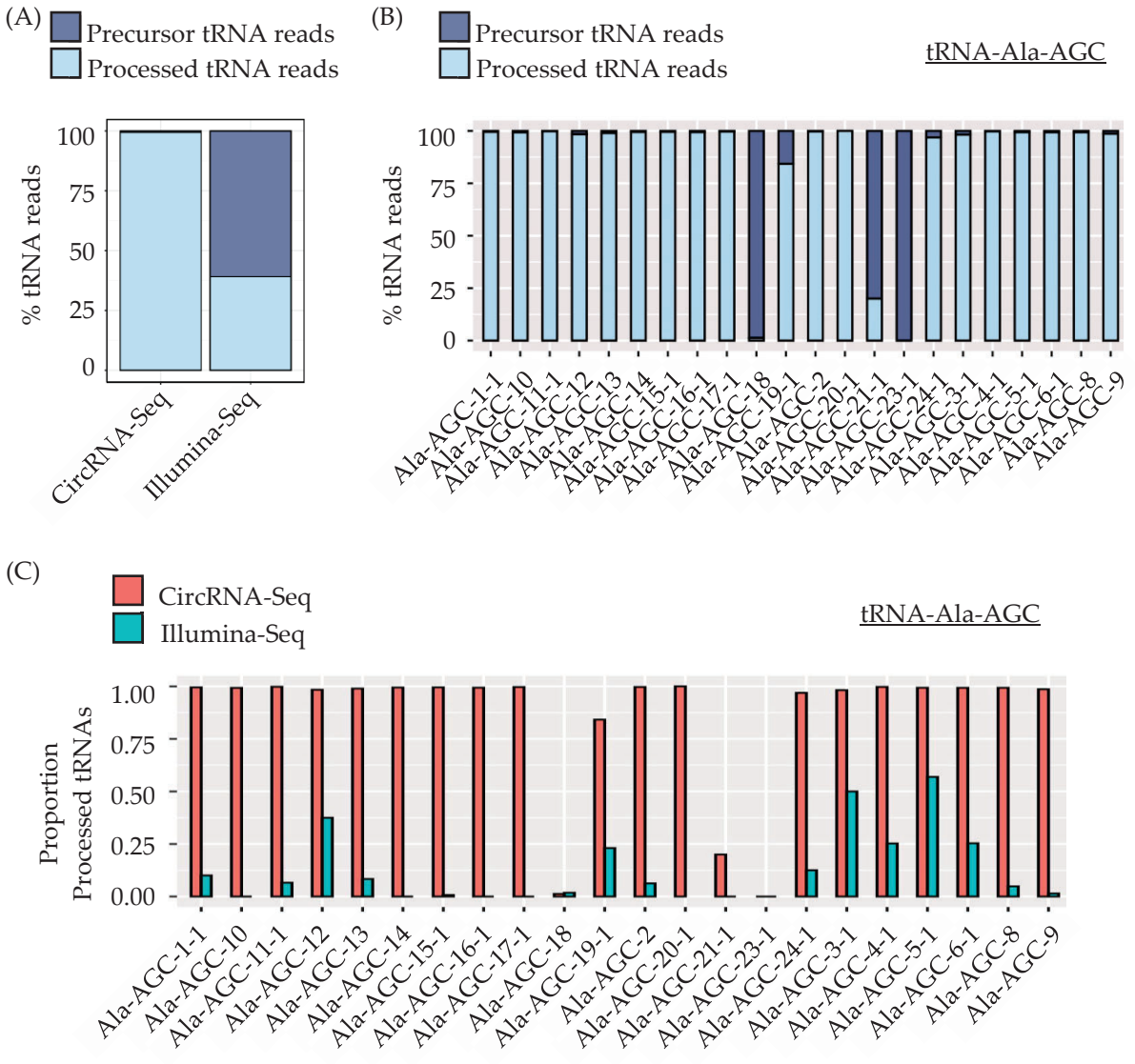
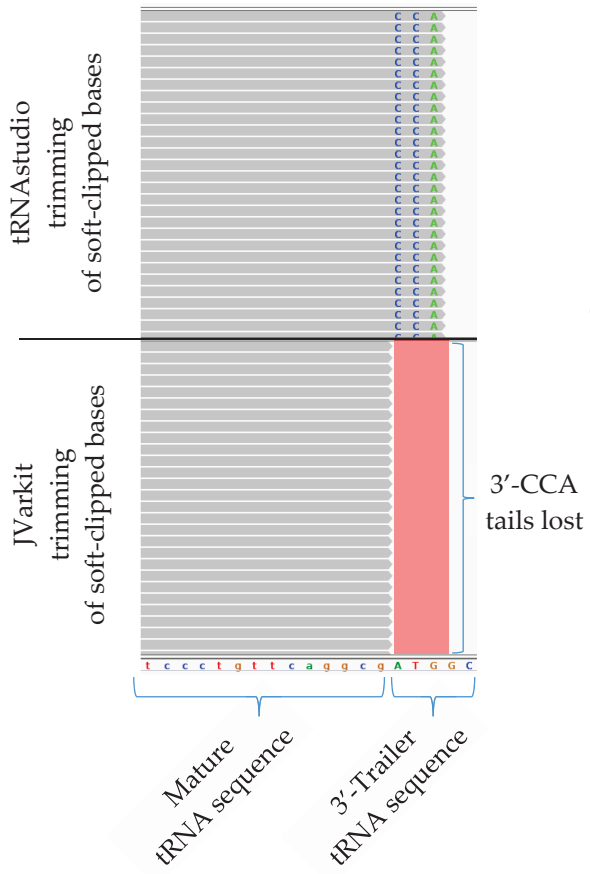


Figure S4. Analyses of the proportion of tRNA reads mapping to pre-tRNA or processed tRNA sequences. **(A)** Comparative analyses of detected tRNA reads from HEK293T cells when using two methods of sequencing library preparations CircRNA-Seq (Zheng, et al., 2015) and Illumina-Seq (Torres, et al., 2015). **(B)** An example of the graphical output generated with tRNAstudio upon analyzing genes encoding $tRNA^{Ala}_{AGC}$ using datasets enriched in detection of processed tRNA reads (Zheng, et al., 2015). **(C)** Comparative analysis of the proportion of processed tRNAs derived from each tRNA gene depicted in **(B)** in CircRNA-Seq (red) (Zheng, et al., 2015) and Illumina-Seq (green) (Torres, et al., 2015). tRNAstudio generates graphical outputs as shown in **(B)** and **(C)**, to evaluate pre-tRNA and processed tRNA proportions at isoacceptor, isodecoder and single gene levels; as well as accompanying spreadsheets with every read count in every group so that the user can perform their custom analyses.

(A)



(B)

	JVarkit Trimming	tRNAstudio Trimming
Reads with 3′-CCA	16.15 %	92.91 %
Reads with tRNAHis 5′-G	0.07 %	4.01 %

(C)

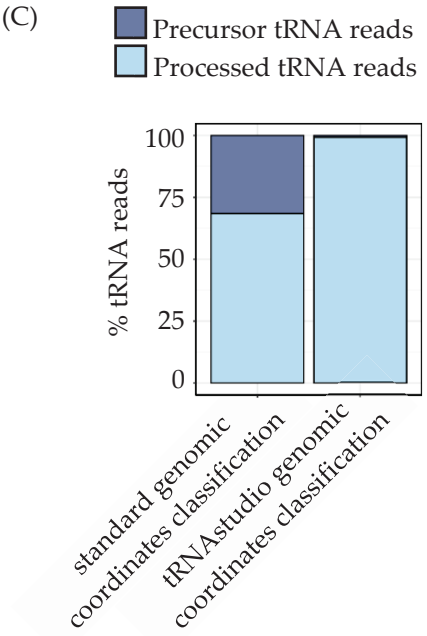


Figure S5. Benchmarking of tRNAstudio's custom functions for trimming soft-clipped bases and classifying tRNA reads into pre-tRNAs or processed tRNAs using datasets enriched in reads derived from mature tRNAs (Zheng, et al., 2015). **(A)** Representative section of an Integrated Genome Viewer (IGV) screenshot showing that post-transcriptional 3′-CCA additions are lost when using standard trimming tools (i.e. JVarkit) but are retained when using the custom trimming tool implemented by tRNAstudio. **(B)** Proportion of total tRNA reads containing 3′-CCA additions and proportion of $tRNA^{His}$ reads containing 5′-G additions when soft-clipped bases are trimmed with JVarkit or with tRNAstudio's custom trimming function. Note that reads retaining these nucleotide additions after JVarkit trimming are reads that have been unmapped against the whole human genome (first alignment step, see **Supplementary Fig. S1A**) but have been recovered upon subsequent alignments against the custom mature genome (that contain tRNA gene sequences with 3′-CCA addition on all tRNA genes and 5′-G addition on $tRNA^{His}$ genes) (**Supplementary Fig. S1B and S1D**). **(C)** Classification of cytosolic tRNA reads into likely derived from pre-tRNAs or from processed tRNAs using standard methods based on the genomic coordinates where the reads map, or using tRNAstudio's method based on the genomic coordinates where the reads map but that take into account the presence of 3′-CCA or $tRNA^{His}$ 5′-G nucleotide additions.

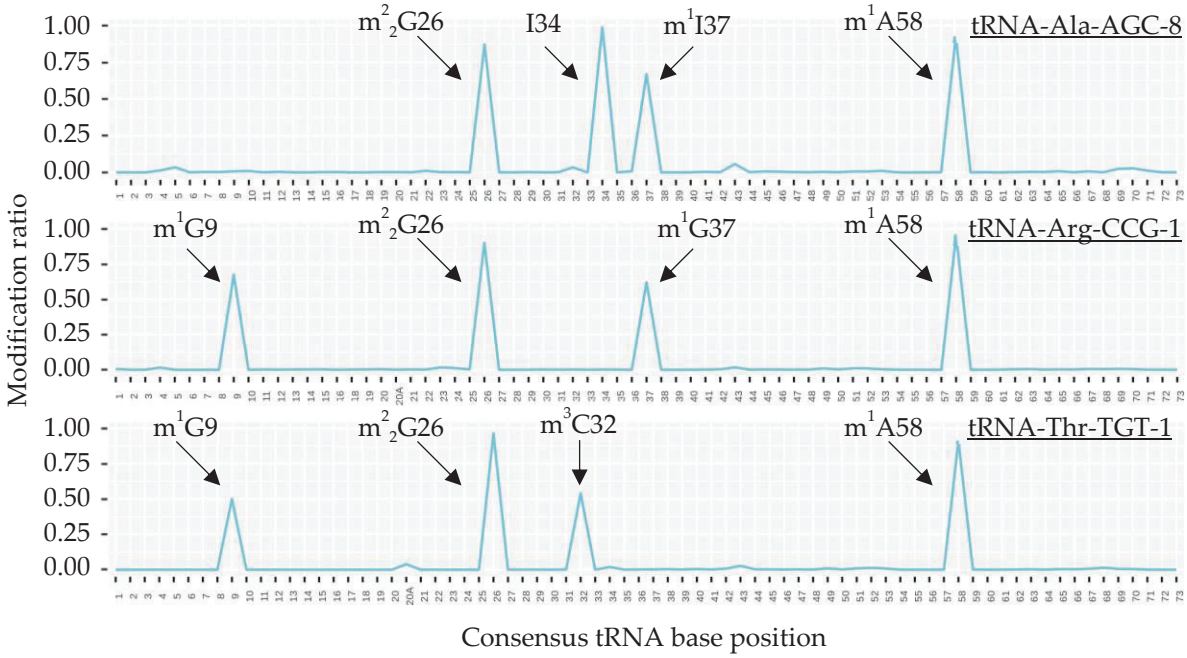


Figure S6. Base-calling method for detection of tRNA modifications by tRNAstudio. Examples of graphical outputs for three tRNA genes upon analysis of datasets from HEK293T cells (Zheng, et al., 2015) are shown. Residues known to be modified and their expected modification are indicated for reference. m^2_2G: N2,N2 dimethylguanosine; I: inosine; m^1I: 1-methylinosine; m^1A: 1-methyladenosine; m^1G: 1-methylguanosine; m^3C: 3-methylcytidine.

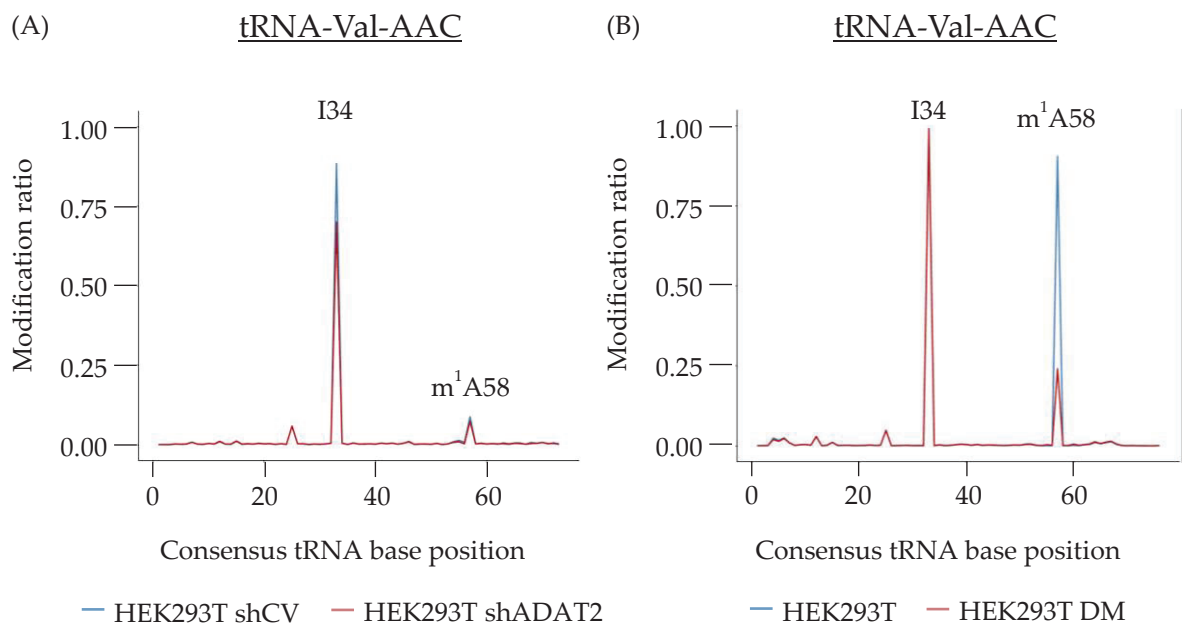


Figure S7. Base-calling analyses on datasets obtained from **(A)** RNA of HEK293T cells expressing a short-hairpin RNA against the enzyme that catalyzes I34 modifications (shADAT2; red line) or a control vector (shCV; blue line) (Torres, et al., 2015); and **(B)** RNA from HEK293T cells (HEK293T; blue line) and the same RNA treated with a demethylase to artificially remove methylations (HEK293T DM; red line) (Zheng, et al., 2015). Residues known to be modified and their expected modification are indicated. I: inosine; m^1A: 1-methyladenosine. Shown are custom analyses based on the numerical outputs obtained with the base-calling function of tRNAstudio for all tRNA-Val-AAC genes. Similar results were obtained for all other tRNAs known to present these modifications (not shown).

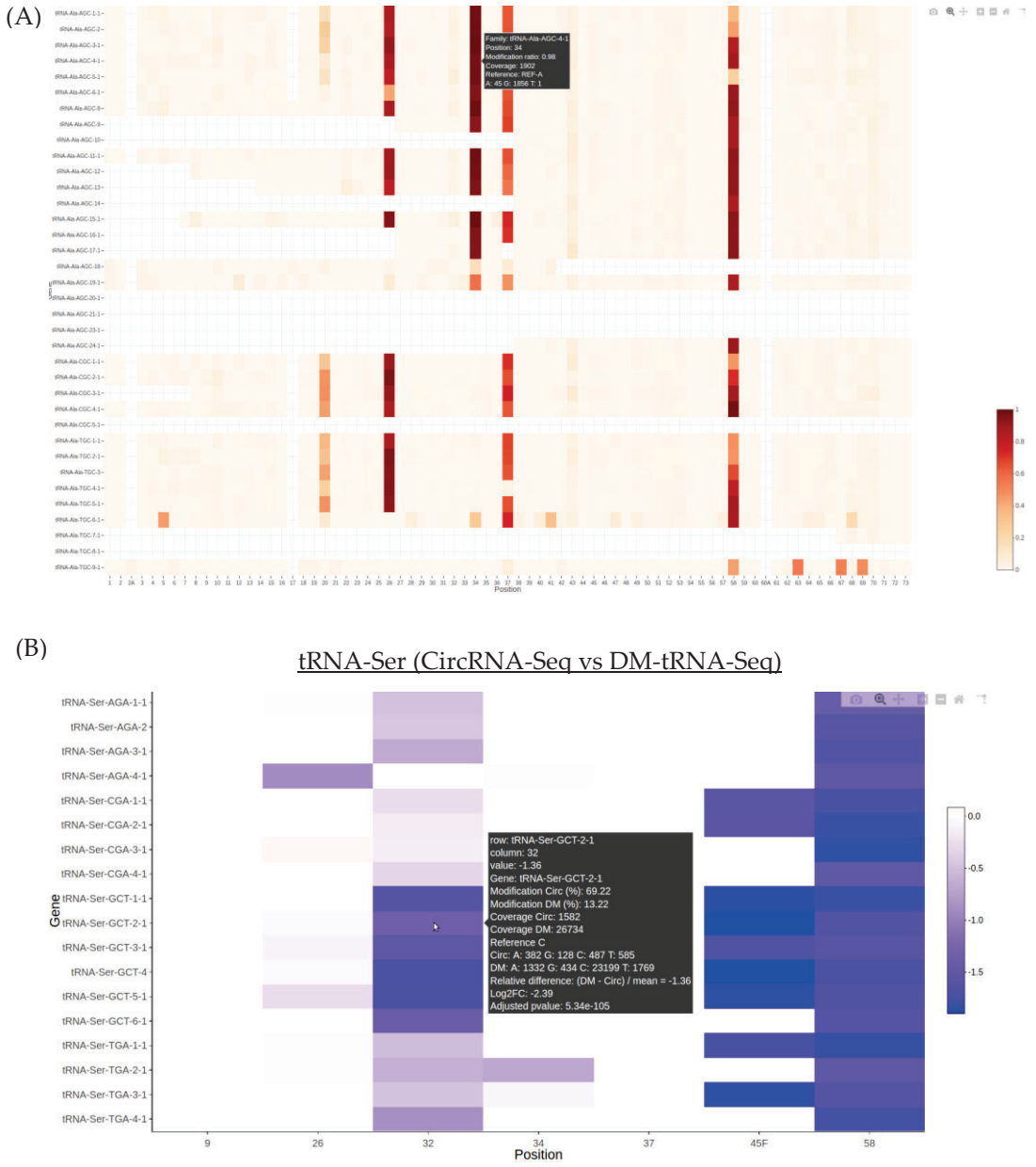


Figure S8. **(A)** Screenshot of the interactive heatmap obtained in the summary report generated by tRNAstudio. Mouse pointer is hovering position A34 on the tRNA-Ala-AGC-4-1 gene showing that 97 % of the reads are detected as a guanosine at this position, consistent with I34 detection (Torres, et al., 2015). **(B)** Interactive heatmap showing changes in modification ratios at positions 9, 26, 32, 34, 37, 45F and 58 of tRNA-Ser genes when comparing RNA extracted from HEK293T cells (CircRNA-Seq) and the same RNA upon demethylation (DM-tRNA-Seq) (Zheng, et al., 2015). The colouring scale adapts to the obtained values and goes from blue (down-regulated) to red (up-regulated). White colour represents undetected modification, unchanged modification levels between samples or lack of a sufficient number of mapped reads to obtain reliable results (see **Supplementary methods**). Note that, as expected, in this example there are no modifications up-regulated (red colouring) in DM-tRNA-Seq as compared to CircRNA-Seq. Mouse pointer hoovers on tRNA-Ser-GCT-2 and shows a statistically significant Log2FC = -2.39 at position 32. This represents a depletion of the modification ratio at this position upon RNA demethylation, and is consistent with a reduction in 3-methylcytidine (m^3C32) (de Crecy-Lagard, et al., 2019).

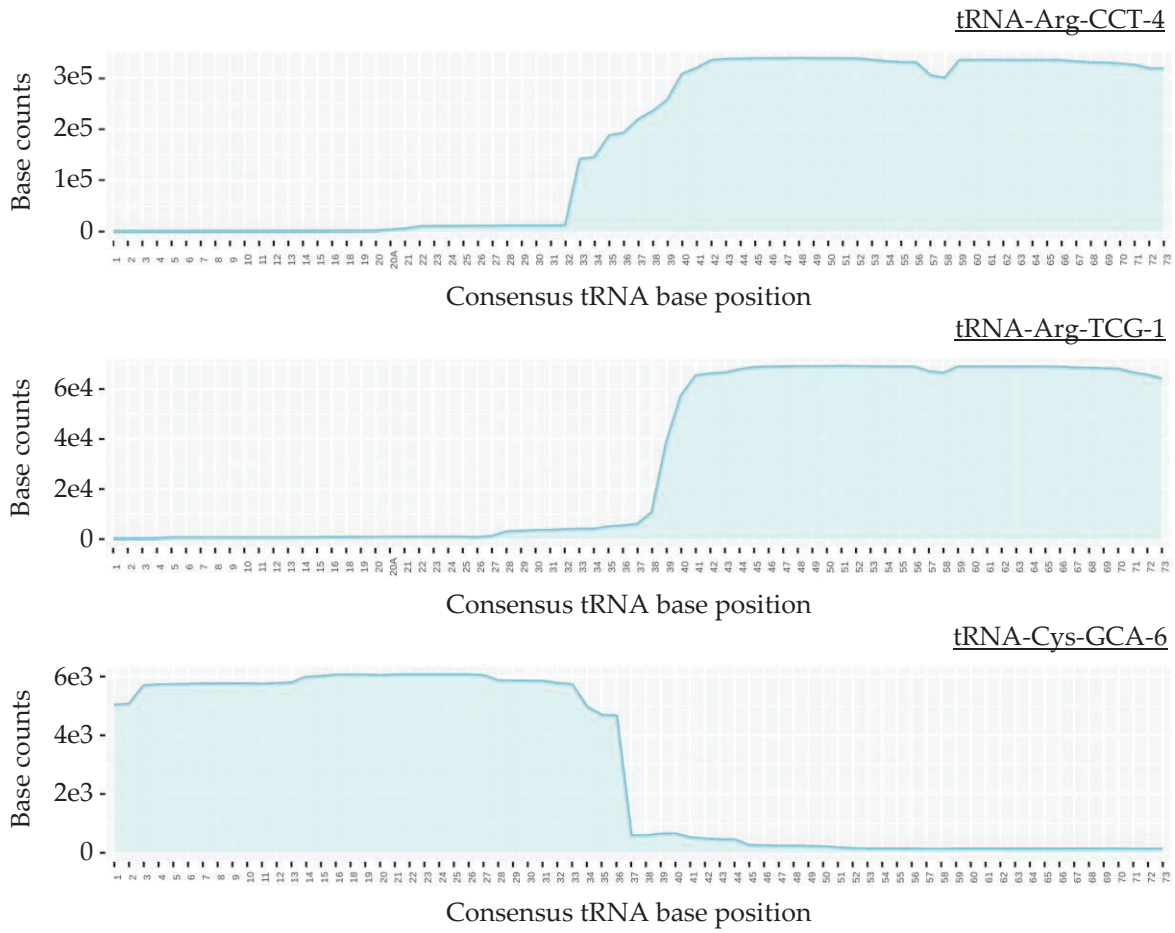


Figure S9. Examples of graphical outputs on tRNA gene coverage analyses obtained from human brain datasets (Ryvkin, et al., 2013). Shown tRNA genes are known to generate 3'-tRFs ($tRNA^{Arg}_{CCU}$ and $tRNA^{Arg}_{UCG}$) or 5'-tRFs ($tRNA^{Cys}_{GCA}$) in human brain (Torres, et al., 2019). We recommend employing the patterns of tRNA gene sequence coverage to detect *bona fide* tRFs only when analyzing tRNA-Seq libraries prepared with methods that do not sequence truncated transcripts due to tRNA modifications causing RT-blocks (i.e. standard small RNA-Seq library preparations) (Torres, et al., 2019).

Supplementary Discussion

tRNAstudio is a fully-automated pipeline that allow users to extract information on tRNA gene expression, tRNA processing and tRNA modification patterns by analyzing tRNA-Seq datasets through a user-friendly GUI. Because of the complexity of the process, tRNAstudio does not give users options to customize the analysis pipeline. We have thus made every effort to validate the performance of tRNAstudio through the analysis of several different datasets.

We show that the tRNAstudio pipeline (**Supplementary Fig. S1**) outperforms the mapping strategies that rely on mapping reads to single genomes (**Supplementary Fig. S2**). Importantly, this holds true regardless of the experimental method used to generate tRNA-Seq libraries, thus validating tRNAstudio for the general analysis of tRNA-Seq datasets. When applied to Illumina-Seq datasets, tRNAstudio did not increase the number of reads mapped to tRNA genes, but their mapping quality was improved. We observed 87 % of reads with a MAPQ > 2 (strategy E) as opposed to only ~50 % of genes having these MAPQ scores when a single-mapping step is done (strategies A, B and C) (**Supplementary Fig. S2A** left panels). Interestingly, the final mapping step performed with relaxed mapping parameters (strategy D) did not influence the outcome, consistent with most reads in this library originating from poorly modified tRNAs (i.e. pre-tRNAs and tRFs) (Torres, et al., 2015). Of note, aligning these datasets only to a custom pre-tRNA genome (strategy C) did increase the number of reads aligned to tRNA genes, but their poor mapping quality, and the fact that whole genome mapping leads to these reads being discarded (compare with strategy A) suggest that these reads are multimappers derived from non-*bona fide* tRNA genes (i.e. 'tRNA-lookalikes'), a known artifact (Telonis, et al., 2015; Telonis, et al., 2014; Telonis, et al., 2016). When analyzing the CircRNA-Seq datasets, tRNAstudio increased both the number of aligned reads and their mapping quality (**Supplementary Fig. S2A** right panels). Here, equivalent MAPQ scores were also obtained when aligning reads to a custom mature tRNA genome (compare strategies B, D and E), but relaxation of mapping parameters (strategy E) led to an increase in the number of aligned reads. This is consistent with reads in this library mostly deriving from post-transcriptionally modified mature tRNAs (Zheng, et al., 2015). There are several algorithms available that have proven useful for tRNA mapping such as Segemehl (Hoffmann, et al., 2018), tDRmapper (Selitsky and Sethupathy, 2015) or SHRiMP2 (Shigematsu, et al., 2017), among others. tRNAstudio uses bowtie2 (Langmead and Salzberg, 2012) as the alignment algorithm, which has its strengths and limitations (Behrens, et al., 2021). It remains to be seen whether tRNA-Seq analyses using alternative alignment algorithms will also benefit from a multi-genome mapping workflow of similar characteristics to that of tRNAstudio.

tRNAstudio can be used to study differential tRNA gene expression among samples (**Supplementary Fig. S3**). tRNAstudio implements two complementary approaches for differential tRNA gene expression: the widely used DESeq2 (Love, et al., 2014) and the tRNA-specific method iso-tRNA-CP (Torres, et al., 2019). 'Isodecoder-specific tRNA gene contribution profiling' (iso-tRNA-CP) compares the expression of tRNA genes within isodecoder sets, which are likely to have similar structure and modifications patterns and are thus subject to similar quantification biases. Importantly, given that iso-tRNA-CP compares the expression of tRNA genes relative to other tRNA genes within the same sample, this approach can also be applied to single samples. This could be relevant to identify tRNA genes that are silent or poorly expressed. We have previously proposed that tRNA genes that appear to contribute less than 1 % of their isodecoder pool are likely to be silent (Torres, 2019).

tRNAstudio can also effectively identify reads derived from pre-tRNAs or from processed tRNAs (**Supplementary Fig. S4**). This is primarily due to the custom trimming of soft-clipped bases implemented by tRNAstudio, which can accurately detect reads containing post-transcriptionally

added 3'-CCA tails (present in processed tRNAs) (**Supplementary Fig. S5A-B**), and that are further distinguished from reads containing 3'-trailer extensions (present in pre-tRNAs) by tRNAstudio's custom tRNA read classification function (**Supplementary Fig. S5C**). This classification may facilitate the study of tRNA biogenesis, including the order of tRNA processing events (Su, et al., 2013). We detected cases where pre-tRNA to processed tRNA ratios appear to be tRNA gene-specific (**Supplementary Fig. S4B**), suggesting that tRNA transcripts may be differentially processed. This observation is consistent with previous reports (Gogakos, et al., 2017; Torres, et al., 2015), but care must be taken when interpreting these results as a low number of reads mapping particular tRNA genes or the presence of modifications in individual mature tRNAs that may be hindering their detection, could be causing quantification artifacts leading to result misinterpretation.

The base-calling function of tRNAstudio can detect a subset of modified residues on tRNAs, and this detection can be used for relative quantification of post-transcriptional tRNA modifications in different samples (**Supplementary Figs. S6, S7** and **S8**). Given that some tRNA modifications are introduced as the tRNA matures (Gaston, et al., 2007; Jiang, et al., 1997; Li, et al., 2021; Torres, et al., 2015), combining this feature with the pre-tRNA vs processed tRNA classification provided by tRNAstudio may inform on tRNA modifications incorporated early during tRNA biogenesis and their potential structural requirements (e.g. removal of leader and/or trailer sequences, presence or absence of introns, post-transcriptional modifications required to incorporate other tRNA modifications, etc.). Of note, tRNAstudio does not apply a base quality filter when calling pysamstats for modification detection, which could lead to false positive errors. Thus, we do not recommend the use of tRNAstudio to evaluate potential modifications at novel tRNA positions. However, when different samples are compared to evaluate modifications at known tRNA positions (**Supplementary Fig. S8B**) tRNAstudio applies several thresholds to prevent potential false positive errors (see '*tRNA modification and sequence coverage analyses*' section on **Supplementary Methods**).

The profiles of tRNA gene coverage provided by tRNAstudio can be used as a tool to screen for potential tRFs (**Supplementary Fig. S9**). Importantly, tRNA gene coverages reported by tRNAstudio are limited to the expected mature tRNA sequence; thus, with its current set-up, tRNAstudio is not a suitable tool to examine pre-tRNA-derived tRFs. In fact, reads (potential tRFs) derived from pre-tRNAs may extend beyond the expected mature tRNA sequence shown in sequence coverage plots. This is important as the biological length of tRFs derived from pre-tRNAs may differ from the tRF length that can be inferred from the sequence coverage graphical outputs of tRNAstudio.

Although tRNAstudio is primarily oriented to the analysis of cytosolic tRNAs, it does also report on mt-tRNA gene expression. A number of mt-tRNA genes have identical counterparts encoded in the nuclear genome (i.e. mt-tRNA-lookalikes) (Telonis, et al., 2014). Thus, reads mapping to these genes cannot be unambiguously assigned to *bona fide* mt-tRNA genes. The MAPQ scores reported by tRNAstudio for mt-tRNA genes (**Supplementary Fig. S2B**, lower panel) can aid in the identification of mt-tRNA genes whose expression may be biased by the presence of mt-tRNA-lookalike nuclear sequences. In addition, mt-tRNA-lookalike reads can affect the analyses of mt-tRNA gene coverage or *bona fide* mt-tRNA modification patterns. Thus, tRNAstudio does not report on these parameters and we refer users of tRNAstudio to established databases such as MINTbase (Pliatsika, et al., 2016) (**Supplementary Table S1**) to extract additional information on mt-tRNA genes from their sequencing datasets.

We note that, currently, tRNAstudio can only analyse human datasets, although the pipeline could be adapted to other organisms through the generation of custom genomes for the species of interest. Adapting tRNAstudio to study other organisms does require programming skills, and we strongly recommend users to validate the pipeline thoroughly before attempting to use it in a fully-automated manner. Instructions for experienced bioinformaticians on how to generate custom genomes of reference and adapt tRNAstudio accordingly can be found in GitHub (https://github.com/GeneTranslationLab-IRB/tRNAstudio). Finally, tRNAstudio is computationally demanding, and it is particularly suited for low-throughput analyses of tRNA-seq data. However, with appropriate computational resources, tRNAstudio can be used to analyze datasets at larger scale.

References

Behrens, A., Rodschinka, G. and Nedialkova, D.D. (2021) High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq, *Mol. Cell*, **81**, 1802-1815 e1807.

Chan, P.P.*, et al.* (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes, *Nucleic Acids Res*.

Chan, P.P. and Lowe, T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes, *Nucleic Acids Res*, **44**, D184-D189.

de Crecy-Lagard, V.*, et al.* (2019) Matching tRNA modifications in humans to their known and predicted enzymes, *Nucleic Acids Res*, **47**, 2143-2159.

Galili, T.*, et al.* (2018) heatmaply: an R package for creating interactive cluster heatmaps for online publishing, *Bioinformatics*, **34**, 1600-1602.

Gaston, K.W.*, et al.* (2007) C to U editing at position 32 of the anticodon loop precedes tRNA 5' leader removal in trypanosomatids, *Nucleic Acids Res*, **35**, 6740-6749.

Gogakos, T.*, et al.* (2017) Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP, *Cell reports*, **20**, 1463-1475.

Hoffmann, A.*, et al.* (2018) Accurate mapping of tRNA reads, *Bioinformatics*, **34**, 1116-1124.

Huntley, M.A.*, et al.* (2013) ReportingTools: an automated result processing and presentation toolkit for high-throughput genomic analyses, *Bioinformatics*, **29**, 3220-3221.

Jackman, J.E. and Phizicky, E.M. (2006) tRNAHis guanylyltransferase adds G-1 to the 5' end of tRNAHis by recognition of the anticodon, one of several features unexpectedly shared with tRNA synthetases, *RNA*, **12**, 1007-1014.

Jiang, H.Q.*, et al.* (1997) Pleiotropic effects of intron removal on base modification pattern of yeast tRNAPhe: an in vitro study, *Nucleic Acids Res*, **25**, 2694-2701.

Juhling, F.*, et al.* (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes, *Nucleic Acids Res*, **37**, D159-162.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2, *Nature methods*, **9**, 357-359.

Li, H.*, et al.* (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078-2079.

Li, J.*, et al.* (2021) The occurrence order and cross-talk of different tRNA modifications, *Science China. Life sciences*.

Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics*, **30**, 923-930.

Lin, B.Y., Chan, P.P. and Lowe, T.M. (2019) tRNAviz: explore and visualize tRNA sequence features, *Nucleic Acids Res*, **47**, W542-W547.

Lindenbaum, P. (2015) JVarkit: java-based utilities for Bioinformatics, *figshare*, http://dx.doi.org/10.6084/m6089.figshare.1425030.
Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol*, **15**, 550.
Pliatsika, V.*, et al.* (2016) MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments, *Bioinformatics*, **32**, 2481-2489.
Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841-842.
Robinson, J.T.*, et al.* (2011) Integrative genomics viewer, *Nat. Biotechnol.*, **29**, 24-26.
Ryvkin, P.*, et al.* (2013) HAMR: high-throughput annotation of modified ribonucleotides, *RNA*, **19**, 1684-1692.
Selitsky, S.R. and Sethupathy, P. (2015) tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data, *BMC bioinformatics*, **16**, 354.
Shigematsu, M.*, et al.* (2017) YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs, *Nucleic Acids Res*, **45**, e70.
Su, A.A., Tripp, V. and Randau, L. (2013) RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri, *Nucleic Acids Res*, **41**, 6250-6258.
Su, S.*, et al.* (2017) Glimma: interactive graphics for gene expression analysis, *Bioinformatics*, **33**, 2050-2052.
Telonis, A.G.*, et al.* (2015) Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies, *Oncotarget*, **6**, 24797-24822.
Telonis, A.G.*, et al.* (2014) Nuclear and mitochondrial tRNA-lookalikes in the human genome, *Frontiers in genetics*, **5**, 344.
Telonis, A.G.*, et al.* (2016) Consequential considerations when mapping tRNA fragments, *BMC bioinformatics*, **17**, 123.
Torres, A.G. (2019) Enjoy the Silence: Nearly Half of Human tRNA Genes Are Silent, *Bioinformatics and biology insights*, **13**, 1177932219868454.
Torres, A.G.*, et al.* (2015) Inosine modifications in human tRNAs are incorporated at the precursor tRNA level, *Nucleic Acids Res*, **43**, 5145-5157.
Torres, A.G.*, et al.* (2019) Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments, *Proc. Natl. Acad. Sci. U. S. A.*, **116**, 8451-8456.
Wickham, H. (2009) *ggplot2. Elegant Graphics for Data Analysis*. Use R! Springer-Verlag New York.
Zhang, J.*, et al.* (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR, *Bioinformatics*, **30**, 614-620.
Zheng, G.*, et al.* (2015) Efficient and quantitative high-throughput tRNA sequencing, *Nature methods*, **12**, 835-837.

3.2

Human tRNAs with inosine 34 are essential to efficiently translate eukarya-specific low-complexity proteins

Published online 14 June 2021

Nucleic Acids Research, 2021, Vol. 49, No. 12 ***7011–7034***
https://doi.org/10.1093/nar/gkab461

Human tRNAs with inosine 34 are essential to efficiently translate eukarya-specific low-complexity proteins

Adrian Gabriel Torres[1], Marta Rodríguez-Escribà[1], Marina Marcet-Houben[1,2], Helaine Graziele Santos Vieira[3], Noelia Camacho[1], Helena Catena[1], Marina Murillo Recio[1], Àlbert Rafels-Ybern[1], Oscar Reina[1], Francisco Miguel Torres[1], Ana Pardo-Saganta[4], Toni Gabaldón[1,2,5], Eva Maria Novoa[3,6] and Lluís Ribas de Pouplana[1,5,*]

[1]Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Catalonia 08028, Spain, [2]Barcelona Supercomputing Centre (BSC-CNS), Barcelona, Catalonia 08034, Spain, [3]Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Catalonia 08003, Spain, [4]Centre for Applied Medical Research (CIMA Universidad de Navarra), Pamplona 31008, Spain, [5]Catalan Institution for Research and Advanced Studies, Barcelona, Catalonia 08010, Spain and [6]University Pompeu Fabra, Barcelona, Catalonia 08003, Spain

Received April 15, 2021; Revised May 07, 2021; Editorial Decision May 10, 2021; Accepted May 18, 2021

ABSTRACT

The modification of adenosine to inosine at the wobble position (I34) of tRNA anticodons is an abundant and essential feature of eukaryotic tRNAs. The expansion of inosine-containing tRNAs in eukaryotes followed the transformation of the homodimeric bacterial enzyme TadA, which generates I34 in $tRNA^{Arg}$ and $tRNA^{Leu}$, into the heterodimeric eukaryotic enzyme ADAT, which modifies up to eight different tRNAs. The emergence of ADAT and its larger set of substrates, strongly influenced the tRNA composition and codon usage of eukaryotic genomes. However, the selective advantages that drove the expansion of I34-tRNAs remain unknown. Here we investigate the functional relevance of I34-tRNAs in human cells and show that a full complement of these tRNAs is necessary for the translation of low-complexity protein domains enriched in amino acids cognate for I34-tRNAs. The coding sequences for these domains require codons translated by I34-tRNAs, in detriment of synonymous codons that use other tRNAs. I34-tRNA-dependent low-complexity proteins are enriched in functional categories related to cell adhesion, and depletion in I34-tRNAs leads to cellular phenotypes consistent with these roles. We show that the distribution of these low-complexity proteins mirrors the distribution of I34-tRNAs in the phylogenetic tree.

INTRODUCTION

Transfer RNAs (tRNAs) are essential components of the translation machinery that physically connect amino acids to their cognate nucleotide triplets (anticodons) according to the Genetic Code. Regulation of tRNA pools is a well-known adaptive mechanism that acts in combination with codon usage to implement translational responses to internal or external cues (1). Codon-anticodon interactions are optimized and modulated by post-transcriptional chemical RNA modifications such as inosine (2,3) that are species-specific, and vary greatly across the phylogenetic tree (4,5). Thus, although the genetic code is essentially universal, the mechanisms that decode it are not.

Inosine at position 34 of the tRNA (I34; first nucleotide of the tRNA anticodon) is produced in Bacteria and Eukarya through the deamination of adenosine (A34) (6,7) (Figure 1). Whereas tRNAs with A34 can only efficiently decode U-ended codons, tRNAs with I34 can decode U-, A- and C-ended codons (8) by wobble pairing (Figure 1A). In Bacteria I34 is produced by the homodimeric enzyme tRNA adenosine deaminase A (TadA), and, in Eukarya, by the heterodimeric adenosine deaminase acting on tRNA (ADAT). ADAT evolved from TadA early in eukaryotic evolution, through a duplication of the bacterial *tadA* gene that gave rise to the two genes coding for the two ADAT subunits (*ADAT2* and *ADAT3*) (6,9) (Figure 1B). In Bacteria, I34 can be found in two different tRNAs (almost universally on $tRNA^{Arg}_{ACG}$ and rarely on $tRNA^{Leu}_{AAG}$) (10,11). In contrast, eukaryotic I34 is found in eight tRNAs ($tRNA^{Thr}_{AGT}$, $tRNA^{Ala}_{AGC}$, $tRNA^{Pro}_{AGG}$,

*To whom correspondence should be addressed. Tel: +34 934034867; Fax: +34 934034870; Email: lluis.ribas@irbbarcelona.org
Present address: Àlbert Rafels-Ybern, National Centre for Genomic Analysis – Centre for Genomic Regulation (CNAG-CRG), Barcelona, Catalonia 08028. Spain.

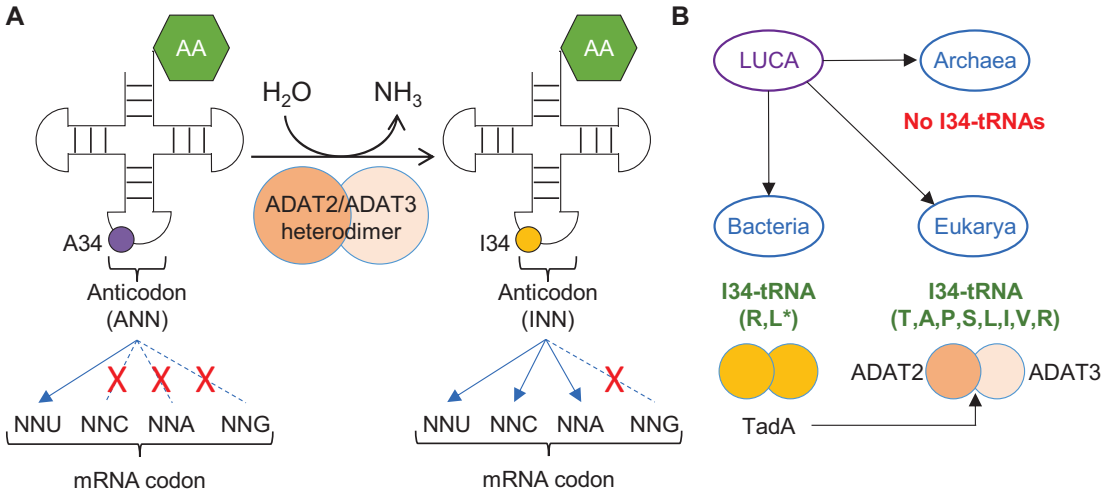


Figure 1. (**A**) Schematic representation of adenosine deamination to inosine at position 34 of eukaryotic tRNAs (A34-to-I34 editing), and the preferential codon:anticodon pairing for unmodified (left) and modified (right) tRNAs. 'AA': amino acid Thr, Ala, Pro, Ser, Leu, Ile, Val or Arg (TAPSLIVR). The anticodon region within the cloverleaf structure of tRNAs is indicated (see also Supplementary Figure S2B for a detailed example of anticodon structure). (**B**) I34 in Bacteria is catalysed by the homodimeric TadA enzyme, from which the eukaryotic ADAT2/ADAT3 heterodimeric enzyme evolved. I34-tRNAs are absent in Archaea, present in two substrates in Bacteria (*to date bacterial I34-tRNA^{Leu} has been reported only on *Oenococcus oeni* (10)), and present in eight substrates in Eukarya.

tRNA$^{Ser}_{AGA}$, tRNA$^{Leu}_{AAG}$, tRNA$^{Ile}_{AAT}$, tRNA$^{Val}_{AAC}$ and tRNA$^{Arg}_{ACG}$) (6,10,12–20) that are highly enriched in eukaryotic genomes (10,21) (Figure 1B). Interestingly, additional A34-tRNAs have been reported in Bacteria but they are not modified to I34-tRNAs, suggesting that the expansion of I34-tRNAs began with the emergence of unmodified A34-tRNAs (10).

In Bacteria, G34-tRNAs (with the exception of tRNA^{Arg}, see above) are globally used as the major isoacceptors to decode threonine, alanine, proline, serine, leucine, isoleucine, valine and arginine (TAPSLIVR). In Eukarya, however, the expansion of eukaryotic I34-containing tRNAs (I34-tRNAs) replaced G34-tRNAs as the preferred mechanism to decode these amino acids (4,5,10,22–25), and played a major role in defining the structure and codon composition of eukaryotic genomes (21). Structure-based hypotheses have been put forward to explain the eukaryotic expansion of I34-tRNAs that affected all three-, four-, and six-codon boxes of the Genetic Code (with the exception of glycine) (26,27). However, the selection forces that drove this phenomenon at the root of eukaryotic evolution remain unclear. Nevertheless, it stands to reason that selection of I34-tRNAs was linked to the mechanisms involved in the translation of codons for TAPSLIVR.

Protein regions of low amino acid diversity are commonly referred to as 'low-complexity' domains (28). Low-complexity domains may or may not be structured, depending on their amino acid composition (29–32), and are often important components of structural, extracellular matrix (ECM) and cell adhesion proteins (33–37). Translating low-complexity coding sequences is a challenge because their highly biased codon composition can slow down translation (38), induce frameshifts leading to mistranslation (39), or cause ribosome stalling and translational arrest (40). Both universal and species-specific adaptations exist to overcome these challenges and expand the protein repertoire (33,40–42), and it is possible that the selection and enrichment of I34-tRNAs in eukaryotic genomes is connected to the contribution of these tRNAs to the efficient translation of low-complexity TAPSLIVR-rich proteins.

Here, we investigate the functional relevance of I34-tRNAs in human cells. The complete elimination of I34 in tRNAs is lethal in all the species where this has been attempted (6,14,16,17,19,20). Thus, generating a cellular model completely devoid of I34-tRNAs is not possible. However, we find that partial I34-tRNA depletion is tolerated in human cells and does not affect translational efficiency or accuracy at global scale. Under these experimental conditions, pathways particularly sensitive to I34-tRNA levels may be identified. Indeed, we find that the impact upon gene translation of a partial reduction in I34-tRNA levels is codon-dependent, and mostly affects low-complexity TAPSLIVR-rich proteins that are prevalent in functional categories linked to cell-cell interactions and ECM-associated pathways. Chief among these proteins are polypeptides containing mucin-like domains. Consistently, I34-tRNA depletion results in abnormal cell morphology and impaired adhesion caused by the deficient translation of membrane proteins exposed to the extracellular environment.

Phylogenetic analyses reveal that TAPSLIVR-rich low-complexity proteins are essentially absent in Bacteria and Archaea, but are abundant in eukaryotes. Moreover, and consistent with their roles in cell adhesion, we find that these proteins are significantly enriched in multicellular species. Our results indicate that I34-tRNAs improve the translation efficiency of genes with highly biased codon compositions that would, otherwise, be inaccessible to the translation apparatus. We propose that the eukaryotic expansion of I34-tRNAs, and of related codons in eukaryotic genomes, was

driven by the increase in proteome diversity afforded by the modified tRNAs.

MATERIALS AND METHODS

Cell lines and cell culture

Human cell lines HEK293T (female; RRID:CVCL_0063), HeLa (female; RRID:CVCL_0030) and HT-29 M6 (female; RRID:CVCL_G077) were maintained in Dulbecco's modified Eagle's medium (DMEM) (41966029, Thermo Fisher), and NCI-H292 (female; RRID:CVCL_0030) cells were maintained in Roswell Park Memorial Institute (RPMI) 1640 Medium (ATCC modification) (A1049101, Thermo Fisher). All media were supplemented with 10% fetal bovine serum (FBS) (10270106, Thermo Fisher), 100 U/ml Penicillin–Streptomycin (15140122, Thermo Fisher) and 25 μg/ml plasmocin (ant-mpp, InvivoGen); herein 'Full media'. Cells were grown at 37°C in a humidified atmosphere with 5% CO_2 (37°C/5% CO_2), and were periodically checked for mycoplasma contamination by PCR. The cell line HT-29 M6 was a gift from Dr Eduard Batlle (IRB Barcelona), and the cell line NCI-H292 was provided by Dr Ana Pardo (CIMA, University of Navarra).

Generation of CRISPR-ADAT KD cell lines

Guide strands were designed using public resources (http://crispr.mit.edu) (see Supplementary Table S1 for detailed oligonucleotide sequences), and were cloned into px330 SV40-GFP vector (gift from Dr Eduard Batlle, IRB Barcelona) (43) as described in (44).

HEK293T cells growing in six-well plate format were transfected with 3 μg px330 SV40-GFP (CTRL), px330 SV40-GFP ADAT2 (ADAT2 KD) or px330 SV40-GFP ADAT3 (ADAT3 KD) constructs using lipofectamine 2000 (L2K) (11668027, Thermo Fisher) following the manufacturer's protocol (250 μl plasmid/lipid reaction in 2 ml DMEM Full Media). Forty-eight hours later, GFP-positive cells were sorted using a FacsAria I SORP sorter (Beckton Sickinson). Sorting on 96-well plates was done using an ACDU system, and one cell per well was sorted in wells containing 100 μl DMEM Full Media (see Supplementary methods). Out of 96 clones analysed per cell line, 55 and 74 were inviable when they were derived from px330 SV40-GFP ADAT2 or px330 SV40-GFP ADAT3 treated cells, respectively. Out of the viable clones, none presented a full KO of the targeted gene, suggesting that full ablation of ADAT2 or ADAT3 is lethal in this cell line. 100% of the single cell seeded clones derived from px330 SV40-GFP treated cells (CTRL) were viable. DNA edition was confirmed by sequencing.

Cell line generation by lentiviral infection

shCV and shADAT2 stable cell lines were generated as previously described (17). Plasmid for hADAT2 over expression was generated by Gateway cloning system following the manufacturer's protocol (hADAT2-pDONR221) using specific oligonucleotides (Supplementary Table S1; and see Supplementary methods). hADAT2 gene was amplified from HEK293T cDNA. hADAT2-pLenti construct was generated by performing an LR reaction using hADAT2-pDONR221 and pLenti vector (Adgene plasmid 19068: pLenti PGK Puro DEST W529-2) following the manufacturer's protocol. Plasmids for DOX-inducible shRNA expression (shNonTarget and shADAT2) were generated by cloning the respective sh sequences (see Supplementary methods) into pTRIPZ vector (Thermo Scientific Open Biosystems Expression Arrest TRIPZ Lentiviral shRNAmir) following the design guidelines reported previously (45).

All shCV, shADAT2, pLenti-hADAT2, and DOX-inducible shRNA cell lines were generated by lentiviral infection using the aforementioned plasmids as described in (17) (see also Supplementary methods). For the transduction of NCI-H292 cells, viral supernatants obtained from HEK293T cells were collected, cleared with a 0.45 μm filter and concentrated by ultracentrifugation through a 20% sucrose cushion at 26 000 g for 2 h at 4°C using a Beckman SW-28 rotor. Purified lentiviral particles were re-suspended in PBS, aliquoted and stored at −80°C. Lentiviral titer was determined using QuickTiter Lentivirus Quantitation Kit (Cell Biolabs, VPK-107). NCI-H292 cells were infected at a MOI of 6 for 24 h, and puromycin at 2 μg/ml was added to culture medium for selection of transduced cells two days later.

Protein extraction

Unless stated otherwise, all protein extractions were performed with 'RIPA buffer': 50 mM Tris pH 7.5, 150 mM NaCl, 1% NP-40, 0.1% SDS, 1× 'cOmplete' EDTA-free Protease Inhibitor Cocktail (PIC) (11873580001, Merck) (see Supplementary methods). Quantification of protein extracts was performed using Pierce BCA Protein Assay Kit (23227, Thermo Fisher) and measuring absorbance at 562 nm with a Synergy HTX Multi-Mode reader (BioTek). For differential extraction of RIPA-soluble and RIPA-insoluble proteins, a pellet of 16×10^6 cells was re-suspended in 250 μl of RIPA buffer, and RIPA-soluble fractions were obtained. The remaining pellet was washed once with 1 ml RIPA buffer and was then re-suspended in 250 μl of 'Solubilisation buffer': 50 mM Tris pH 7.4, 150 mM NaCl, 50 mM DTT, 2% SDS, 8 M urea, 1× PIC. The re-suspended pellet was incubated for 10 min at 95°C, centrifuged at maximum speed for 2 min at room temperature (RT°), and supernatant was recovered (RIPA-insoluble fraction). For Figure 3E, 250 μl of 'insoluble protein loading buffer 2×' (100 mM Tris pH 6.8, 0.1% Bromophenol blue, 20% glycerol) was added to each RIPA-soluble and RIPA insoluble fractions. Then, 20 μl of each sample was resolved by 10% PAGE and the gel was stained with BlueSafe (MB15201, NZYtech).

RNA extraction

Total RNA was isolated from cells with TRIzol (15596026, Thermo Fisher) and ethanol re-precipitated as described (17). For RNA extraction from high polysome fractions, samples were combined and concentrated using an Amicon Ultra-15 mL 100K Da (UFC910024, Merck) to a volume of approximately 200 μl and RNA was extracted using 500 μl

TRIzol LS (10296010, Thermo Fisher) following the manufacturer's protocol. Extracted RNA was quantified using a Nanodrop ND1000 spectrophotometer (Thermo Fisher). RNA integrity was evaluated with a 2100 Bioanalyzer Instrument (Agilent).

Western blots

Western blotting was performed by standard procedures as previously described (46). See Supplementary methods for details on antibodies used in this study. Blots were developed using an Odyssey Fc Imaging System (LI-COR) and analysed using Image Studio Lite v5.2. Raw Odyssey FC image files available upon request.

Real-Time quantitative PCR

RT-qPCR was performed as previously described (17,46) in a StepOnePlus Real-time PCR System (Applied Biosystems). Details on primers used are shown in Supplementary Table S1 (17,47–49).

Analyses of tRNA-Seq datasets

Inosine and 1-methylinosine quantifications by tRNA-Seq were performed as previously described (17), except that reads were aligned against the human reference genome hg38. Quantification of tRNA gene expression at tRNA isodecoder level was performed with DESeq2 v1.18 (50) as previously described (51). Datasets used in this study GSE114904 (51) and PRJEB8019 (17).

Pulse-chase analyses

Pulse-chase experiments were performed with cells at approximately 80% confluence. Growing media was removed, cells were washed twice with PBS, and incubated at 37°C/5% CO_2 for 30 min in Starvation media: DMEM No Cys, No Met, No Glu (21013024, Thermo Fisher), 10% FBS, 4 mM L-glutamine (25030024, Thermo Fisher). Media was then removed and cells were incubated for 30 min at 37°C/5% CO_2 with Pulse media: Starvation media containing 300 μCi/ml of ^{35}S-Met/^{35}S-Cys (EasyTag™ EXPRESS35S Protein Labeling Mix, NEG772007MC, Perkin Elmer) or ^{35}S-Met (NEG 009L005 MC, Perkin Elmer) and 0.2 mM L-Cys (non-radioactive) (C6852, Merck). Cells were washed twice with PBS and were incubated for 5 min at RT° with Chase media: DMEM Full media, 5 mM L-Cys (non-radioactive), 5 mM L-Met (non-radioactive) (M9625, Merck). Cells were then washed twice with PBS and harvested with PBS. When cycloheximide (CHX) treatments were required, Starvation, Pulse and Chase media contained 100 μg/ml CHX (C4859, Merck). Proteins were extracted with RIPA buffer and 10 μg of obtained proteins were resolved by 10% SDS-PAGE. The gel was stained with Coomassie (A1092, Panreac AppliChem), dried using a Slab Gel Dryer GD2000, and exposed to a Typhoon Screen for radioactivity detection.

Quantitative metabolic labeling was performed as previously described (52). Pulse-labeling medium contained 50 μCi/ml of ^{35}S-Met/^{35}S-Cys (EasyTag™ EXPRESS35S Protein Labeling Mix, NEG772007MC, Perkin Elmer). Cells were incubated with pulse-labeling medium for 15, 30 and 60 min, were washed and collected as described (52). Cell pellets were resuspended in 100 μl ice-cold PBS and 15 μl of cell suspension were spotted on 2.5 cm glass microfiber filter disks (Whatman GF/C; WHA1822025) (to measure total radioactivity) or to perform TCA precipitation (to measure TCA-precipitable label) as described (52). When CHX treatments were required, Starvation media, Pulse-labeling media and PBS contained 100 μg/ml CHX (C4859, Merck). Scintillation counting was measured in a Tri-Carb 2900 TR (Perkin Elmer) as described (52).

Analyses of cell growth

1×10^6 cells in 8 ml Full Media (time point Day 0) were seeded on a 10 cm Petri dish. Two days later, cells were washed once with PBS, and harvested with 2 ml Trypsin-EDTA (0.05%) (25300054, Thermo Fisher) that was later quenched with 2 ml Full Media. Total number of cells was counted (time point Day 2) using a Countess Automatic Cell Counter (Invitrogen). Then, 1×10^6 harvested cells were plated on a new 10 cm Petri dish and the process was repeated up until the last time point. Results represent the cumulative counting of cells from Day 0 to the last time point.

Cell cycle analyses

Cells were synchronised by a double thymidine block: 2 mM thymidine (T1895, Merck) in DMEM Full Media for 13 h, release (media without thymidine) for 8 h, and second block for 17 h. Determination of cell cycle stages were performed in an Epics Cyan ADP flow cytometer (Beckman Coulter) as previously described (46).

Cellular treatments with stress reagents

1×10^6 cells in DMEM Full Media were seeded in six-well plate format. Forty-eight hours later, media was replaced by 2 ml DMEM Full Media containing a either 700 μg/ml hygromycin B (HygroB) (10687010, Thermo Fisher), 700 nM emetine (E2375, Merck), 100 μg/ml blasticidin S (BlaS) (R21001, Thermo Fisher), 100 μg/ml cycloheximide (CHX), 50 mM $CaCl_2$ (1.02391, Merck), 0.45 M Sucrose (84097, Merck), or DMEM only (no FBS; starvation). Cells were visualised in an Eclipse Ts2-FL microscope (Nikon). Results depicted on Figure 5A were obtained at 2 h (Sucrose), 18 h (BlaS and CHX), 20 h (Emetine and HygroB) and 24 h ($CaCl_2$ and starvation) of treatment. Every condition had its own 'untreated control' per cell line; Figure 5A shows a representative untreated control. Activation of the UPR (Figure 3D) was performed by treating cells with 2 μM thapsigargin (T9033, Merck) for 3 h. Proteins were then extracted with RIPA buffer containing 1 mM Na_3VO_4, 5 mM $Na_4P_2O_7$ and 50 mM NaF to retain their phosphorylation status.

Cell viability assays

To prevent ADAT KD cells to detach upon treatments with stress reagents, 96-well culture plates were coated with 100

μg/ml rat-tail collagen type I (A1048301, Thermo Fisher). 1×10^4 cells were seeded per well and 48 h later were treated with stress reagents for 12 h. Cell viability was measured with reagent WST-1 (5015944001, Merck) following the manufacturer's protocol in a Synergy HTX Multi-Mode reader (BioTek).

Cellular adhesion to ECM components

Cells were harvested using Trypsin and re-plated in 175 cm^2 flasks. Twenty-four hours after plating, cells were washed once with pre-warmed PBS and harvested with pre-warmed PBS/2 mM EDTA (131026.1209, Panreac Quimica). 1.5×10^5 harvested cells in 100 μl Assay Buffer were plated for 2 h at 37°C/5% CO_2 on an ECM Array Plate (ECM cell adhesion array kit colorimetric, ECM540, Merck). Wells in ECM Array Plates are pre-coated with individual ECM components to test the binding preferences of seeded cells. The following steps of the assay were performed as described by the manufacturer's protocol. For control experiments depicted in Supplementary Figure S3E, cells were harvested with Trypsin instead of PBS/2 mM EDTA.

Polysome profiling

Cells growing in 175 cm^2 flasks were trypsinised and plated in three 10 cm Petri dishes. Each dish contained 1×10^7 cells (for experiments carried out 24 h after plating; Figure 6A) or 1×10^6 cells (for experiments carried out 72 h after plating; Supplementary Figure S4A). At 24 or 72 h after plating, cells were lysed following a protocol adapted from (53). All solutions were prepared fresh on the day to be used. Cells in each Petri dish were treated with 7 ml of 100 μg/ml CHX in DMEM Full Media at 37°C/5% CO_2 for 3 min. Cells were then washed with 4 ml ice-cold PBS containing 100 μg/ml CHX (PBS/CHX). PBS/CHX was removed, cells from all three Petri dishes were harvested by scrapping, combined to generate a single cell lysate, and kept on ice at all times. Cells were centrifuged at 1000 × g for 5 min at 4°C and the supernatant was discarded. Cell pellets were gently washed (one pipette 'up and down' stroke) with 1 mL PBS/CHX, and were centrifuged and the supernatant was removed as before. Cell pellets were re-suspended (five pipette strokes) in 500 μl 'Polysome extraction buffer' (PEB) (20 mM Tris pH 7.4, 100 mM KCl, 10 mM $MgCl_2$, 0.5% NP-40, 2 mM DTT, 100 μg/ml CHX, 100 U/ml RNasin (N2615, Promega), 1× PIC). Cells were incubated on ice for 10 min, vortexing briefly every 2 min. Cell lysate was then centrifuged at maximum speed for 10 min at 4°C and supernatant (approximately 600 μl) was recovered. 10% of this lysate was used for total RNA extraction, and the rest was used to obtain polysome profiles. Polysome profiling was carried out as described in (54) using a 10–50% linear sucrose gradient, with minor modifications (see Supplementary methods). P/M ratios were obtained from three independent replicates, after integrating the area under the curve of monosomes (peak corresponding to the 80S fraction) and polysomes (peaks corresponding to low- and high-polysome fractions).

RNA-Seq

Library preparations for RNA-Seq studies (total RNA and HP fractions, in biological triplicates, for both HEK293T CTRL and HEK293T ADAT2 KD cell lines) were prepared using the TruSeq mRNA library preparation kit (single indexes set A; 20020492, Illumina), following manufacturer's recommendations. Libraries were indexed, pooled, and then sequenced in a NextSeq Flow Cell machine as 2 × 150 bp paired-end reads. Datasets have been deposited at NCBI GEO, accession GSE150860.

Reads were aligned to the human genome (hg38) using STAR 2.3.0e with default options. Reads counts at gene level were generated with the featureCounts function from the Rsubread package version 1.28.1 using options annot.inbuilt = 'hg38', isPairedEnd = TRUE, requireBothEndsMapped = TRUE, checkFragLength = TRUE. Only protein coding genes (Ensembl biomart v97 July 2019) having >10 reads in at least half of the samples were considered for differential expression analyses. DESeq2 1.18 was used to detect differentially expressed genes with default options and using the following thresholds: Benjamini–Hochberg adjusted P-value < 0.1, |FC| > 1.5. The ROAST method (55) was used to perform Gene Set Enrichment Analyses using the MaxMean statistic (56). All gene set mapping was performed at Gene Symbol level (org.Hs.eg.db v3.0.0). Statistically significant categories were defined as those having an adjusted P-value < 0.05.

Statistical significance of the differences between empirical distributions of global TE was computed using the mded package version 0.1–2. (57) (Figure 6C). Downregulated genes for the interaction analysis (HP ADAT2 KD/HP CTRL)/(Total RNA ADAT2 KD/Total RNA CTRL) were detected using DESeq2 with FC HP versus Total < 1.5; P-value <0.05. Statistical significance of the enrichment of transcripts encoding low-complexity TAPSLIVR-rich proteins among those down-regulated in the interaction analysis was assessed via Fisher Exact Test. (Supplementary Table S4). Enrichment in proportion of transcripts encoding low-complexity TAPSLIVR-rich proteins with decreased TE in ADAT2 KD cells, among transcripts with TE CTRL >1.5 was assessed via permutation test ($n = 31$, $B = 10000$) (Figure 8B). P-values are computed as the proportion of permutations with more extreme statistics than the observed.

Construction of ADAT eGFP reporters

eGFP ADAT and eGFP nonADAT sequences flanked by EcoRI, XhoI and XbaI restriction sites (5-end) and PmeI, AgeI and EcoRI restriction sites (3′-end) were ordered from GenScript (GenScript HK Inc) and were cloned into a custom pLV-CMV-SV40-Puro plasmid. Correct sequence insertion for all constructs was verified by Sanger sequencing using the CMV-F universal primer (GATC Biotech). Details on eGFP ADAT sequence and eGFP nonADAT sequence are depicted in Supplementary methods. The eGFP open reading frame contains 239 codons, 88 of which encode for TAPSLIVR and are uniformly distributed across the gene. Thus codon differences between the reporters affected 36.8% of the eGFP sequence. Importantly, these dif-

ferences did not significantly affect the Codon Adaptation Index (CAI) of the genes (CAI-eGFP ADAT = 0.761; CAI-eGFP nonADAT = 0.759). In addition, since all TAPSLIVR codons were modified to either C-ended (eGFP ADAT) or G-ended (eGFP non-ADAT) triplets, the overall GC content of the genes remained unaltered (GC content for both eGFP sequences = 61.8%). Conservation of CAI and GC-content is important to rule out potential non-I34-tRNA dependent effects on eGFP expression.

Evaluation of ADAT eGFP production

Cells growing on 6-well plate format at approximately 80% confluence were transfected with 2.5 μg of eGFP ADAT or eGFP nonADAT plasmids using L2K (250 μl plasmid/lipid reaction in 2 ml DMEM Full Media), following the manufacturer's protocol. Negative control cells ('L2K only') received the same amount of lipid formulation without plasmids. Proteins were extracted with RIPA buffer 48 h after transfection and western blots were carried out as described above. For FACS analyses, 24 h after lipofection, cells were washed twice with PBS and the cell suspension was analysed in a Cytomics FC500 MPL flow cytometer (Beckman Coulter) (see also Supplementary methods). Data was analysed with Summit v4.3 or FlowJo v10.5.3.

In silico detection of low-complexity TAPSLIVR-rich genes

In this work we refer to 'low-complexity regions' as sections of a protein sequence bearing low amino acid diversity, a widely used definition (28). Based on the concept of a TAPSLIVR-rich region defined by Rafels-Ybern *et al*. (23), we consider that a low-complexity region is rich in TAPSLIVR if at least 80% of its amino acids (in any combination) belong to the TAPSLIVR category. Bioinformatics identification of low-complexity TAPSLIVR-rich regions were performed on the Human CCDS release 22 (14 June 2018), using a running window strategy as previously described (23), but the window size was modified to include regions of 30 or more amino acids to evaluate a larger number of proteins. Based on the reported *H. sapiens* codon usage (58), we applied a threshold of 65.743% to define a genetic sequence as significantly enriched in ADAT-dependent codons. Gene ontology analyses were performed with DAVID (Database for Annotation, Visualization and Integrated Discovery) v6.8 (59) using default options. Statistically significant categories were defined as those with a FDR <0.25 and Benjamini-Hochberg adjusted *P*-value <0.05. For the Functional Annotation Clustering sets, only those with an Enrichment Score >4 were included.

Construction of ADAT luciferase reporters

SDC3-RLuc and SDC3(G-end)-RLuc plasmids were generated using the backbone vector psiCHECK-2 (C8021, Promega). The psiCHECK-2 RLuc gene was PCR amplified from the vector and the desired portion of SDC3 was PCR amplified from HEK293T cDNA (or from a custom made SDC3(G-end) sequence, GeneArt, Life Technologies). A linker (reported in Promega's NanoLuc reporter plasmids) that serves as a spacer between the SDC3 section and the RLuc gene was present in the reverse (RVR) primer used to amplify SDC3/SDC3(G-end). The obtained SDC3/SDC3(G-end) and RLuc products were ligated and inserted into the psiCHECK-2 vector resulting in replacement of the original RLuc gene. Oligonucleotides used are depicted in Supplementary Table S1 (see Supplementary methods for further details).

Luciferase assays

6×10^4 cells in 100 μl DMEM Full Media per well were plated in a 96-well black plate with clear bottom (CLS 3603, Merck), and were lipofected with 100 ng plasmids on the next day following the manufacturer's protocol (50 μl of plasmid/lipid reaction in 100 μl DMEM Full Media per well). Cells were left at 37°C/5% CO_2 until luciferase measurements. Luciferase activity was monitored using the Dual-Glo Luciferase Assay System (E2920, Promega), following the manufacturer's protocol using a MicroLumat Plus LB96V luminometer (Berthold).

Purification of reporter proteins

Reporter proteins were purified following standard procedures. SDC3-RLuc and SDC3(G-end)-RLuc were purified using magnetic Dynabeads Protein A (10002D, Thermo Fisher) incubated with a Renilla luciferase antibody (PA5-32210, Thermo Fisher) and cross-linked with 5 mM BS^3 (21580, Thermo Fisher) in Conjugation Buffer (20 mM NaP, 150 mM NaCl). eGFP ADAT and eGFP nonADAT were purified using Protein G sepharose beads (17-0618-01, VWR) incubated with Green Fluorescent Protein antibody (DSHB-GFP-12A6, Developmental Studies Hybridoma Bank) using magnetic Dynabeads Protein A (10002D, Thermo Fisher), following the manufacturer's protocol. Purified proteins were visualized by SDS- PAGE, and confirmed by western blotting. See Supplementary methods for further details.

Mass Spectrometry analyses

Protein samples were reduced, alkylated and overnight tryptic digested (60). Digested peptide mixtures were desalted and clean-up using polyLC C18 and strong cation-exchange (SCX) filters. Samples were subject to nano-LC–MS/MS analysis. The nanochromatographic system used was either a Nanoacquity (Waters) or a Dionex Ultimate (Thermo Scientific). The Advion Triversa NanoMate (Advion Biosciences) was used as the nanosource and it was fitted on an LTQ-FT Ultra (Thermo Fisher) or an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher). The mass spectrometer was operated in a DDA mode, with survey scans acquired at 120 k and MS2 scans at 30 k in the orbitrap or IT resolution.

Data processing was performed with Proteome Discoverer software v2.1 or Bioworks v3.1.1 SP1 (Thermo Fisher) using Sequest HT search engine and SwissProt HUMAN, contaminants and the proteins of interest (SDC3-RLuc or eGFP) fasta databases. Search parameters included trypsin as enzyme, carbamidomethylation in cysteine as fixed modification and oxidation in methionine as variable modification. Peptide mass tolerance was 10 ppm and the

MS/MS tolerance was 0.6 Da (MS2 in the IT) or 0.02 Da (MS2 in the Orbitrap). Peptides with FDR <1% were considered as positive identifications with a high confidence level.

To identify possible mistranslation in SDC3-RLuc we performed *de novo*, database and homology searches using PEAKS v8.5 with search parameters as described above. *De novo* score (ALC %) threshold was set to 15 and peptide hit threshold (-10logP) was 30.0. *De novo* hits that did not match any database or homology searches and that have an ALC >90% were used in a BLAST (The Basic Local Alignment Search Tool) search against SDC3-RLuc protein in order to find regions of local similarity between sequences and highlight possible mutations.

Whole proteomics analyses was performed in biological triplicates. HEK293T CTRL and ADAT2 KD cells growing in 10 cm Petri dishes and at ∼80% confluence were washed once with PBS and harvested by cell scrapping with 0.5 ml proteomics extraction buffer (0.1 M Tris–HCl pH 7.5; 0.1 M DTT; 4% SDS). The lysate was further processed through a 20 G needle 20 times and then through a 15 G needle 15 times to shear DNA, and was quantified using the Bradford reagent (B6916, Merck). 100 μg of protein sample were then processed following the filter-aided sample preparation (FASP) method (61). Before trypsin digestion, urea buffer was removed and exchanged with triethylammonium bicarbonate (TEAB) buffer. Digested solutions were acidified to a final concentration of 0.1% formic acid. Samples were then dried in a speedvac and reconstituted in 46 μl TEAB 500 mM. 30 μl of sample was labeled with iTRAQ Reagent-8PLEX Multiplex Kit (4390812, Sciex) following the manufacturer's protocol. In addition, 11 μl of each sample were combined to generate 2 pools of all samples and were also labeled. The combined iTRAQ-labeled sample was cleaned up using polyLC C18 and SXC filters. Cleaned-up combined iTRAQ-labeled sample was fractionated using the Pierce High pH Reversed-Phased Peptide Fractionation kit (84868, Thermo Scientific) following the manufacturer's protocol. Fractions were dried with a speedvac and reconstituted in 58.8 μl 2% acetonitrile and 0.1% formic acid. LC–MS/MS analysis was done with the Advion Triversa Nanomate (Advion Bioscieneces) fitted on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher). Data processing was performed with Proteome Discoverer v2.1 as described above.

iTRAQ reporter ion intensities were used for protein quantifications. Contaminant sequences were removed. Unique and razor peptides with an average reporter ion signal to noise >1 were considered for further quantitative and statistical analysis. Within each iTRAQ experiment, peptide quantitation was normalized by summing the abundance values for each channel over all peptides identified within fractions. For each protein a linear model was fitted with or without random effects depending on available data. Condition was selected as fixed effect and peptide, fraction and replicate were set as random effects. Model fitting was accomplished with the lme4 R package version 1.1–23. Differentially expressed proteins were defined as those with |FC| > 1.5 and Benjamini & Hochberg adj. *P*-value <0.1.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (62) partner repository with the dataset identifier PXD025024.

Evaluation of MUC5AC production

For induction of mucous cell differentiation, NCI-H292 shCV and shADAT2 cells were seeded in six-well plates at 6×10^5 cells per well and incubated at 37°C/5% CO_2 in RPMI Full Media containing 2 μg/ml puromycin for 3 days or until confluent. Cells were then washed with PBS and incubated in puromycin-free fresh media containing either 50 nM amphiregulin (AREG) (A7080, Thermo Fisher) or PBS for 2 days. Treatment was then repeated for another 2 days. Immunofluorescence was performed on cells grown on glass coverslips. Cells were fixed with 4% paraformaldehyde for 15 min at RT°, washed twice with PBS, and permeabilised in Blocking buffer (0.3% Triton-X100, 1% BSA, 1X PBS) for 20 min at RT°. Cells were stained with MUC5AC (45M1) primary antibody (MA1-38223, Thermo Fisher) diluted 1:200 in blocking buffer overnight at 4°C, cells were then washed twice with PBS and incubated with 1:400 dilution of Anti-Mouse Alexa Fluor 555-conjugated secondary antibody (A-31570, Thermo Fisher) in the dark for 1 h at room temperature. Slides were then stained with DAPI (D9542, Merck), and were mounted with Vectashield (H-1000, Vector Laboratories). Images were acquired with a Zeiss LSM 780 confocal microscope and analysed using ImageJ software (63). All image adjustments were applied to all images equally for direct comparison. Quantification of MUC5AC signal was performed on at least 31 different images per condition taken with a Plan-Apochromat 10×/0.45 M27 objective, and using custom-made macros in ImageJ.

For FACS analyses, NCI-H292 cells were harvested using Ca^{+2}-free PBS/0.02% EDTA to avoid enzymatic degradation of extracellular proteins, washed in PBS, and re-suspended in Flow cytometry buffer (PBS containing 0.1% (w/v) saponin (S7900, Merck), 1% (w/v) sodium azide (S2002, Merck), and 10% FBS). Aliquots of 5×10^5 cells were then stained with MUC5AC (45M1) primary antibody (dilution 1:200 in flow cytometry buffer) at 4°C for 30 min, washed twice in flow cytometry buffer, and incubated with 1:250 dilution of Anti-Mouse Alexa Fluor 488-conjugated secondary antibody (A-21202, Thermo Fisher) in the dark at 4°C for 30 min. Cells were then washed twice in flow cytometry buffer, re-suspended in cold PBS and analysed on a FACSAria Fusion flow cytometer (BD Biosciences). Cells stained only with secondary antibody were used as negative control to set the gate. Representative plots showing the gating strategy are shown in Supplementary Figure S6D and E.

Homology search

The search for homologous sequences was performed using the OMA database (64) of 152 archaeal, 1674 bacterial, and 462 eukaryotic genomes. For consistency, we used the same version of the *Homo sapiens* genome (see 'In silico detection of low-complexity TAPSLIVR-rich genes'). A BlastP (v2.5.0) (65) search was performed with each human protein containing low-complexity TAPSLIVR-rich regions (2218 proteins: TAPSLIVR-set), and with the rest of the human proteins, against the OMA database. The number of accepted hits was 10 000. Blast results were filtered using an

e-value cut-off of 0.01 and an overlap threshold of 20%. Average number of hits per species was obtained by calculating, for each human protein, the number of hits in each group (prokaryotes and eukaryotes) divided by the number of genomes searched in each group. The average number of hits in prokaryotic species was then divided by the average number of hits in eukaryotic species to obtain 'average ratios'. Permutation tests were done by generating 5000 groups of 2218 proteins randomly selected from the whole human proteome. The average ratios for each group of proteins was calculated and their distribution compared to the average ratios obtained for the TAPSLIVR-set (Figure 10A), obtaining the z-score which was then used to calculate the *P*-value.

Protein sequences homologous to the TAPSLIVR-set were screened for the presence of low-complexity TAPSLIVR-rich regions as described above (see 'In silico detection of low-complexity of TAPSLIVR-rich genes'). Protein sequences containing TAPSLIVR-rich regions were used to generate the heatmap shown in Figure 10B. For data normalization, species were allocated to different taxonomic groups according to their phylum, and the number of species with at least one homolog sequence was divided by the number of species in the phylum. Species were also grouped based on whether they are unicellular or multicellular. R v3.5.3 was used to generate Boxplots (package ggplot2 v3.3.2) and calculate statistical significances the Mann-Whitney *U* test.

Analyses of A34-tRNA gene content

Total tRNA gene content for each species was obtained from the Genomic tRNA database v2.0 (58) and from (10), acquiring information for 168 of the eukaryotic species used for homology search (see above). The relationship between A34-tRNA gene content and abundance of found TAPSLIVR-rich homologs per species was evaluated by a Spearman's rank correlation coefficient using R v3.5.3 and the package ggpubr v0.2.5. For correlation tests reported in Figure 10C, seven species with an unusual number of A34-tRNA genes (>400 genes) were excluded from the analyses and are reported in Supplementary Table S6. Of note, the correlation strength reported in Figure 10C ($R = 0.53$; *P*-value = 3.7e–13) was maintained when the seven excluded eukaryotic species were included in the analyses ($R = 0.53$, *P*-value = 1e–13). For results depicted in Supplementary Figure S7, 1324 species from Bacteria and 114 species from Archaea were analysed (Supplementary Table S6).

Statistical analyses

Statistical analyses were performed with GraphPad Prism software v6.0 and R v3.5.3. Unless stated otherwise, data shows mean ± SD of at least three biological replicates. Statistical significance was obtained by a two-tailed *t* test (*P*-value < 0.05). For RNA-Seq and tRNA-Seq data, statistical significance was defined by Benjamini-Hochberg adjusted *P*-values (adj *P*-value < 0.1, and adj *P*-value < 0.05, respectively). For whole proteomics analyses (iTRAQ), statistical significance was defined by Benjamini-Hochberg adj *P*-value < 0.1). For the analyses of homolog proteins, statistical significance when comparing average ratios was obtained using the *z*-score. Correlation analyses were performed with a Spearman's rank correlation coefficient. Statistical significance for Figure 10E was obtained by Mann–Whitney *U* test. Statistical significance of the enrichment for candidate genes among gene lists was done via Fisher Exact Test and permutation analysis.

RESULTS

ADAT2 KD reduces I34 levels without affecting general protein synthesis

To study the biological relevance of I34-tRNAs in HEK293T cells we used CRISPR/Cas9 technology to disrupt the *ADAT2* or *ADAT3* genes. Both genes are essential in all model organisms studied to date, and eukaryotic I34-tRNAs are absolutely required to translate C-ended codons for TAPSLIVR in species that lack G34-tRNA isoacceptors (6,14,16,17,19,20) (see also Introduction). Thus, as expected, we were unable to obtain full ADAT2 or ADAT3 knockout clones (see Materials and methods), but we did obtain heterozygous clones carrying wild type (WT) and edited alleles ('HEK293T ADAT2 KD' or 'HEK293T ADAT3 KD') (Supplementary Figure S1). Editing of the ADAT2 allele resulted in the generation of a premature stop codon eight amino acids downstream of the edited site (Supplementary Figure S1A); while editing of the ADAT3 allele resulted in the elimination of seven residues mapping to the deaminase domain of the protein without changes in the translation reading frame (Supplementary Figure S1B).

Both KD cell lines presented reduced levels of the protein coded by the targeted gene (Figure 2A). ADAT2 KD did not affect the levels of ADAT3, but we observed a mild decrease in ADAT2 protein abundance upon ADAT3 KD (Supplementary Figure S2A). We did not detect changes in ADAT2 or ADAT3 mRNA levels in either cell line (Figure 2A). This is consistent with the effects of CRISPR/Cas9 targeting, and suggests that the artificially edited ADAT2 transcript with a premature stop codon can escape the nonsense-mediated decay pathway (66). HEK293T ADAT2 KD cells were stable in culture, but HEK293T ADAT3 KD cells rapidly reverted to the WT sequence (not shown).

Upon ADAT2 KD, we detected reduced levels of I34 on all its tRNA substrates, as seen by next generation sequencing of tRNAs (Figure 2B), without significant variations in tRNA transcript abundance (Figure 2C and Supplementary Table S2). As a control, we verified that the amount of the unrelated tRNA modification 1-methylinosine (m^1I) present at position 37 of $tRNA^{Ala}$, and catalysed by ADAT1 (67), was not affected (Supplementary Figure S2B). Similar results were observed upon shRNA-mediated KD of ADAT2 (17) (Supplementary Figure S2B-C). Because complete depletion of I34 is not possible, our cellular models allow us to identify cellular processes most sensitive to a reduction of I34-tRNAs.

Reduced levels of I34-tRNAs would be expected to impair cellular translation. However, pulse-chase analyses of

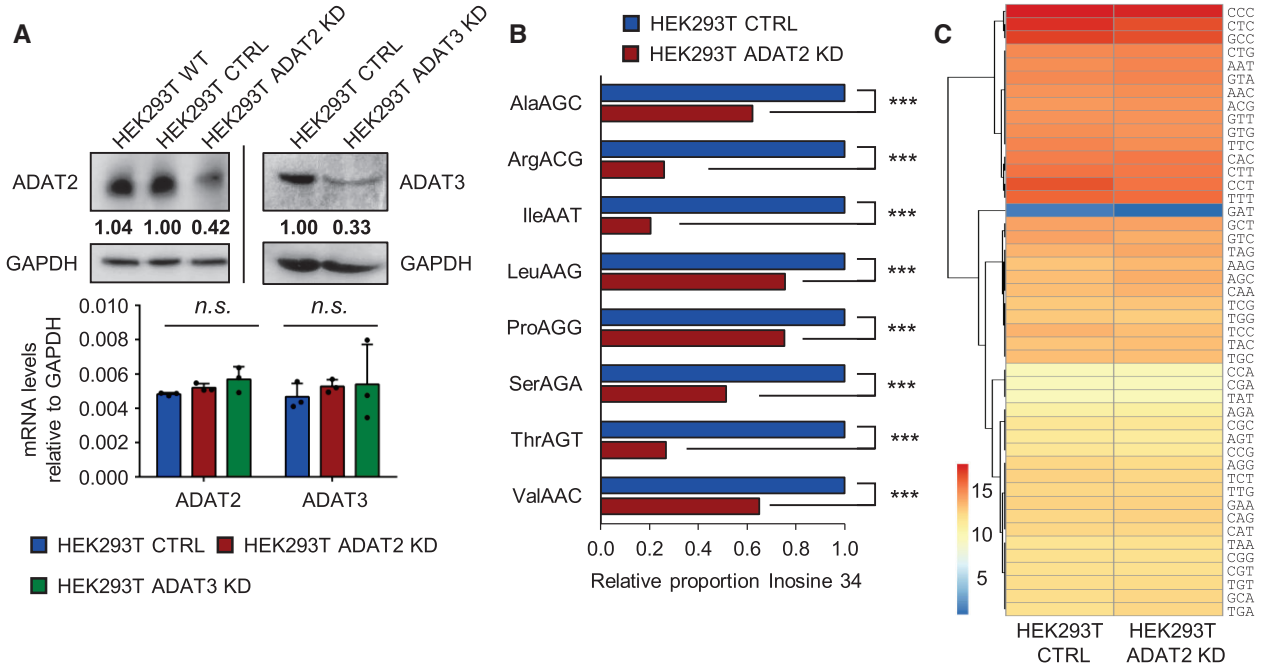


Figure 2. (**A**) Upper panel: ADAT2 and ADAT3 protein levels. Quantification of gel bands have been normalized to GAPDH and relative to HEK293T CTRL cells. Lower panel: ADAT2 and ADAT3 transcript levels in the indicated cell lines, relative to GAPDH. Shown are biological triplicates, their mean and standard deviations (SD). n.s.: not statistically significant (*t*-test). See also Supplementary Figure S2A. (**B**) I34 levels in HEK293T CTRL (blue) and ADAT2 KD (red) cells as evaluated by tRNA-Seq. Shown are I34 proportions relative to CTRL cells of two biological replicates calculated as in (17). ***: adj. *P*-value < 0.001 (Benjamini–Hochberg, Fisher Exact Test). See also Supplementary Figure S2B. (**C**) Heatmap visualization of tRNA gene expression at isodecoder level (tRNAs with the same anticodon) in HEK293T CTRL and ADAT2 KD cells as evaluated by tRNA-Seq. Colouring scale represents log2 DESeq2 normalized expression values based on two biological replicates calculated as in (51). No statistically significant differences were found (Benjamini–Hochberg, Fisher Exact Test, adj. *P*-value < 0.05). See also Supplementary Figure S2C and Supplementary Table S2.

general protein synthesis did not reveal defects in overall translation efficiency in ADAT2 KD and shADAT2 cells (Figure 3A and Supplementary Figure S2D). Quantitative metabolic labeling demonstrated that the amount of synthesized protein over time is similar in CTRL and ADAT2 KD cells (Figure 3B), and that the incorporation of free radiolabeled amino acid into proteins occurs at similar rates in both cell lines (Figure 3C). As a control, cycloheximide (CHX) treatment abolished the incorporation of radiolabeled amino acid into proteins in both cell lines (Supplementary Figure S2E).

As a proxy for studying mistranslation, we monitored activation of the unfolded protein response (UPR) (68), formation of RIPA-insoluble protein aggregates (69), and ubiquitination levels in whole protein extracts (70). Based on these parameters, we were unable to detect signs of mistranslation in ADAT2 KD cells (Figure 3D-F). We also performed mass spectrometry-based whole proteomics analyses (iTRAQ) and found only 7 differentially expressed proteins (|FC| > 1.5; adj. *P*-value < 0.1) among 2280 detected proteins (Figure 3G and Supplementary Table S3), indicating that 99.7% of detected proteins present unaltered levels under these conditions. Mass spectrometry data also showed that all detected peptides presented their expected mass for identification in all samples, indicating lack of mistranslation. Thus, depletion in I34-tRNAs caused by the inactivation of a single *ADAT2* allele (or by shRNA-mediated KD) does not cause appreciable defects in global translation efficiency or accuracy.

Reduced I34 levels affect cell growth, and cause morphology defects

We measured cell growth to assess the general physiological state of the cell after silencing ADAT2 or ADAT3, and found a reduced growth rate caused by a general deceleration of the cell cycle (Figure 4A, B and Supplementary Figure S3A). We observed similar phenotypes in different shADAT2 human cell lines (Supplementary Figure S3B and C), and we were able to fully recover growth rates in ADAT2 KD cells by introduction of a lentiviral ADAT2 expression system ('HEK293T ADAT2 KD pLenti-hADAT2′) (Figure 4C). Thus, the observed phenotypes are due to reduced levels of I34-tRNAs caused by ADAT depletion. We noticed that cells depleted of I34-tRNAs presented an abnormal morphology after being detached by trypsin treatment and re-plated in clean culture plates. This phenotype was transient (Figure 4D), and absent in cells detached using PBS-EDTA (Supplementary Figure S3D). This suggests that the silencing of ADAT2, and the resulting reduction in levels of I34-tRNAs, impair the ability of cells to recover from the proteolytic elimination of membrane proteins exposed to the extracellular milieu.

Depletion of I34-tRNAs impairs cell adhesion and sensitises cells to translation inhibitors

We then tested whether translation machinery inhibitors would have a synergistic effect with ADAT silencing. We

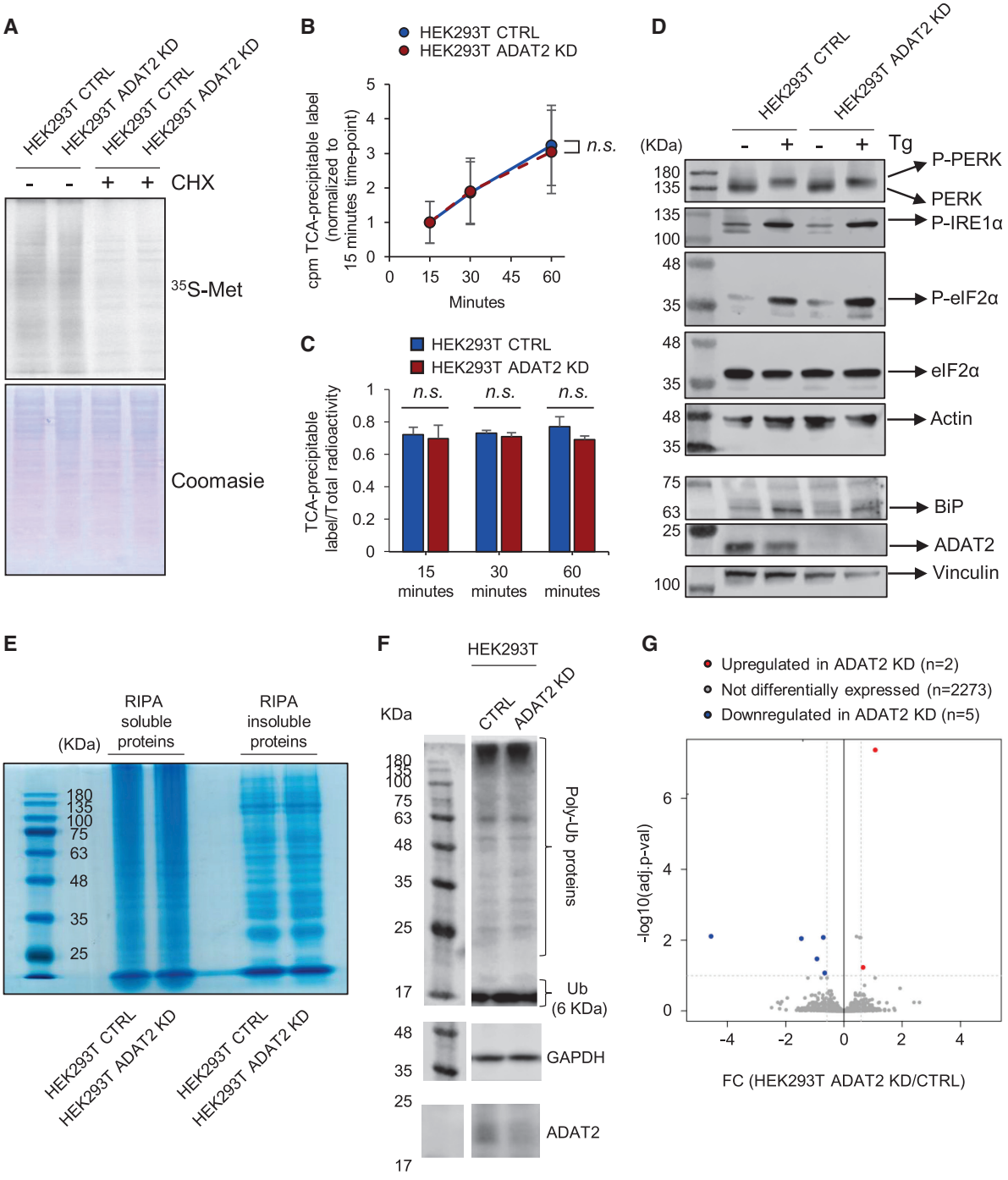


Figure 3. (**A**) Pulse-chase analyses on HEK293T CTRL and ADAT2 KD cells in the presence (+) or absence (−) of cycloheximide (CHX), that inhibits cytosolic translation. Coomassie staining is used as gel loading control. See also Supplementary Figure S2D. (**B**) Kinetics of protein synthesis in HEK293T CTRL (blue) and ADAT2 KD (red) cells upon monitoring the amount of TCA-precipitable label (^{35}S-Met/Cys) from each cell lysate at the indicated time-points. Shown are the mean and standard deviations from biological triplicates. n.s.: not statistically significant (*t*-test) (**C**) Quantitative evaluation of ^{35}S-Met/Cys incorporation into proteins upon monitoring the ratio between the total radioactivity (cpm) of the cell lysate and the amount of radioactivity (cpm) in TCA-precipitable fractions from samples in (B). Shown are the mean and standard deviations from biological triplicates. n.s.: not statistically significant (*t*-test). See also Supplementary Figure S2E. (**D**) Analyses of UPR markers on the indicated cell lines in the presence (+) or absence (−) of thapsigargin (Tg), that induces UPR. ADAT2 levels are shown for reference. Actin and Vinculin are used as gel loading controls. (**E**) Evaluation of RIPA-soluble and RIPA-insoluble protein abundance in HEK293T CTRL and ADAT2 KD cells by BlueSafe staining. (**F**) Evaluation of ubiquitination levels in total protein lysates from HEK293T CTRL and ADAT2 KD cell lines. Note that under these conditions free ubiquitin (6 KDa) runs with the front dye (∼ 17 KDa). GAPDH is used as gel loading control and ADAT2 levels are shown for reference. (**G**) Volcano plot showing differentially expressed proteins (|FC| > 1.5; adj. *P*-value < 0.1) in ADAT2 KD cells as compared to CTRL cells based on whole proteomics analysis by Mass Spec (iTRAQ). See also Supplementary Table S3.

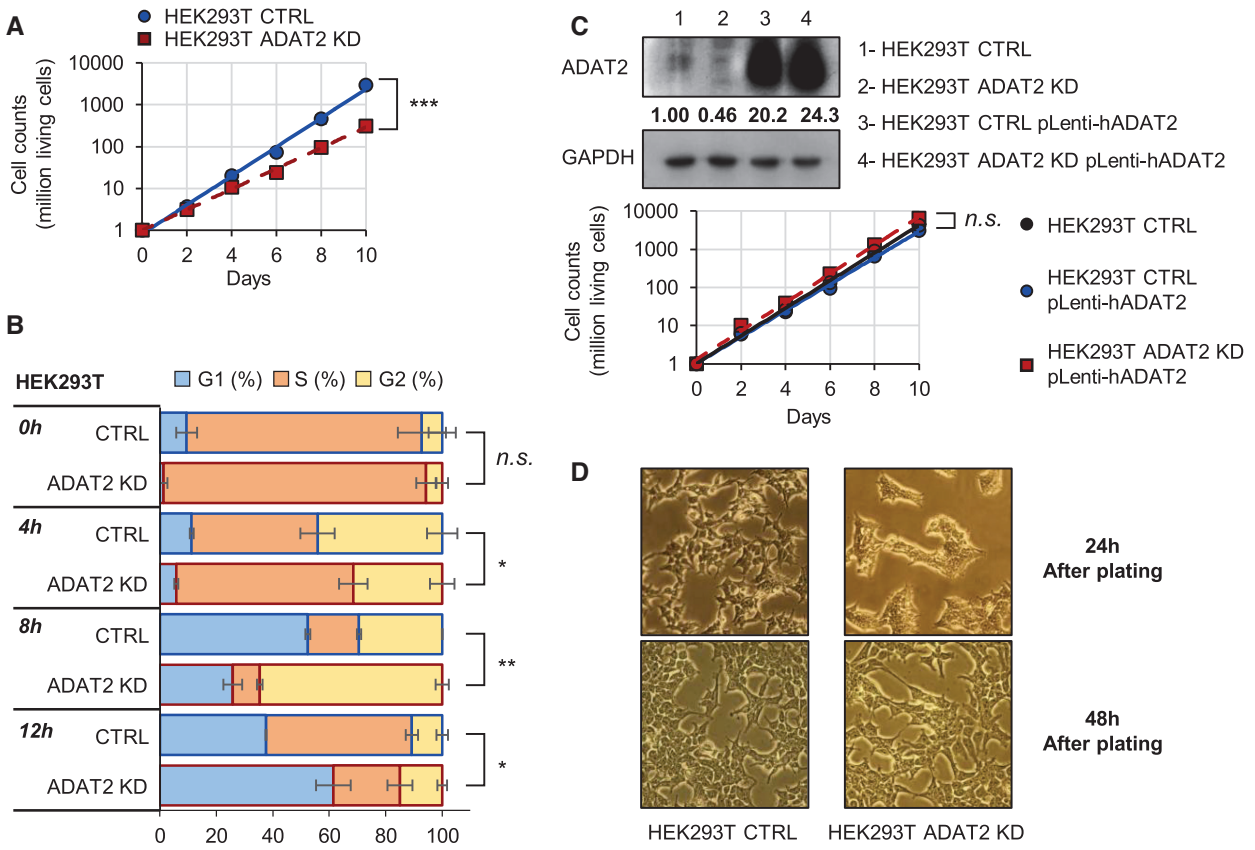


Figure 4. (**A**) Growth curves of HEK293T CTRL (blue) and ADAT2 KD (red) cells showing total living cells over time. Y-axis is set in logarithmic scale and an exponential trendline was fit to the data points. Data points correspond to the mean and SD of three biological replicates. ***: *P*-value < 0.001 (*t*-test). See also Supplementary Figure S3A–C. (**B**) Cell cycle analysis of HEK293T CTRL (blue-bordered bars) and ADAT2 KD (red-bordered bars) cells as evaluated by FACS over a 12-h period. Shown are the mean and SD of biological duplicates. Statistical significance is depicted only for percentage of cells in G1 phase at every time point. n.s.: not statistically significant. *: *P*-value < 0.05. **: *P*-value < 0.01 (*t*-test). (**C**) Upper panel: ADAT2 protein levels on the indicated cell lines. GAPDH is used as gel loading control. Quantification of ADAT2 bands have been normalized to GAPDH and relative to HEK293T CTRL cells. Lower panel: Growth curves as in (A) of HEK293T CTRL cells (black) and of HEK293T CTRL and ADAT2 KD cells upon stable re-expression of ADAT2 ('pLenti-hADAT2') (blue and red, respectively). n.s.: not statistically significant (*t*-test). (**D**) Representative light microscopy images of trypsin-treated HEK293T CTRL and ADAT2 KD cells at 24 (upper panels) and 48 (lower panels) h after plating. See also Supplementary Figure S3D.

found that both ADAT2 KD and ADAT3 KD cells, but not CTRL cells, spontaneously detached from culture plates upon treatment with antibiotics such as Hygromycin B (HygroB), Emetine, Blasticidin S (BlaS) and Cycloheximide (CHX) (Figure 5A). In contrast, all three cell lines remained adhered to culture plates when exposed to insults that do not directly affect translation, such as calcium chloride ($CaCl_2$), starvation, or incubation in hyperosmotic media (0.45 M Sucrose) (Figure 5A). We further found that detached cells treated with antibiotics were viable, grew normally if re-plated in clean culture plates (not shown), and were metabolically equal to CTRL cells (Figure 5B), indicating that their detachment was not due to a differential sensitivity to antibiotic toxicity. Thus, although our data shows that global translation is not affected in cells with reduced I34-tRNAs, we observe phenotypes consistent with impaired translation of specific functional protein families.

These results prompted us to investigate whether I34-tRNA depletion quantitatively impairs the adhesion capacity of cells. Furthermore, because cellular morphology and proliferation depends upon cellular adhesion (71,72), compromised cell adhesion can also explain the phenotypes observed in trypsin-treated ADAT2 KD cells (Figure 4). We reasoned that the observed phenotypes could be caused by impairment in the *de novo* synthesis of membrane proteins necessary for cell attachment. To evaluate the adhesion capacity of I34-depleted cells in a context where *de novo* translation is required for this function we: (i) treated ADAT2 KD and CTRL cells with trypsin to degrade plasma membrane proteins and stimulate their synthesis; (ii) plated the cells in standard culture plates for 24 h; (iii) harvested the cells with PBS-EDTA (preserving all newly synthesized membrane proteins) and (iv) placed the cells in plates previously coated with individual components of the extracellular cell matrix (ECM) to test the ability of the cells to bind to physiological substrates. We found that ADAT2 KD cells display impaired adhesion to collagens (Col II, Col IV) and vitronectin (VN), but not to fibronectin (FN), laminin (LN) or tenascin (TN) (Figure 5C and Supplementary Figure S3E). These results indicate that upon degradation of

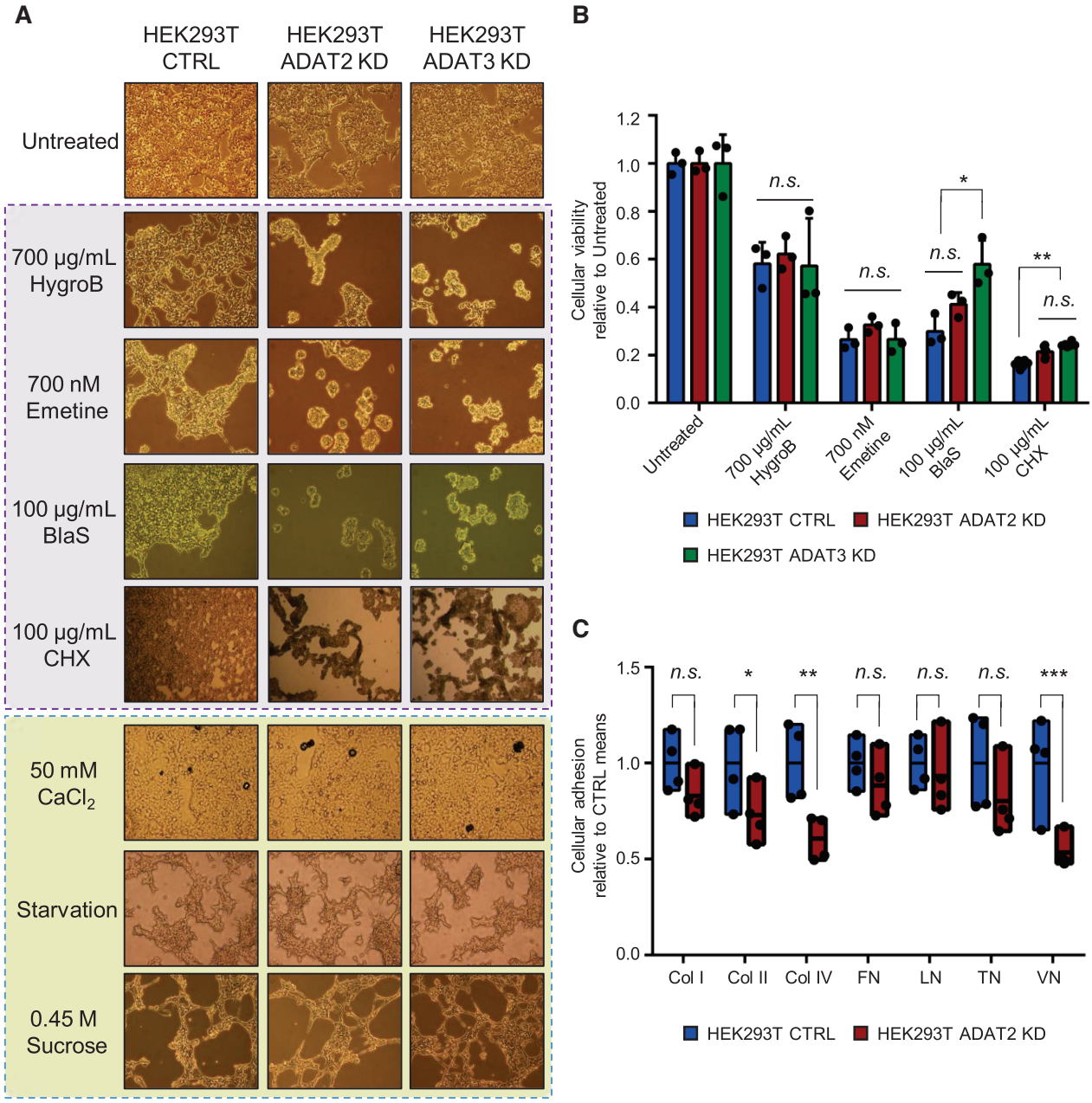


Figure 5. (**A**) Representative light microscopy images of HEK293T CTRL, ADAT2 KD and ADAT3 KD cells in the presence of the indicated stress agents. Those that directly affect the translation machinery are highlighted in purple. HygroB: hygromycin B. BlaS: blasticidin S. CHX: cycloheximide. (**B**) Cellular viability (metabolic activity) measured by WST-1 assays for the indicated cell lines in the presence of antibiotics depicted in (A). Shown are biological triplicates, their means and SD relative to untreated cells. n.s.: not statistically significant. *: *P*-value < 0.05. **: *P*-value < 0.01 (*t*-test). (**C**) Evaluation of cell adhesion to components of the extracellular matrix. Col I, II and IV: collagen I, II and IV respectively. FN: fibronectin. LN: laminin. TN: tenascin. VN: vitronectin. Box plots represent min-to-max cell adhesion relative to the means of HEK293T CTRL cells based on four biological replicates. n.s.: not statistically significant. *: *P*-value < 0.05. **: *P*-value < 0.01. ***: *P*-value < 0.001 (*t*-test). See also Supplementary Figure S3E.

membrane proteins cells depleted from I34-tRNAs fail to efficiently resynthesize proteins required for cellular attachment to specific components of the ECM.

Depletion of I34-tRNAs reduces ribosome occupancy on a subset of transcripts

To explore the impact of I34-tRNAs on the translatome we performed polysome profiling at 24 h after trypsin treatment and plating. We detected reduced levels of mRNAs in the high polysomal fractions and a consequent increase of mRNA abundance present in the low polysomal fractions in ADAT2 KD cells (Figure 6A), indicating reduced ribosome occupancy on transcripts. In addition, we found an increase in the 80S ribosomal fraction and a shift in the 40S-to-60S ribosomal fraction ratio (Figure 6A). At 72 h after trypsin treatment and plating, we observed these differences substantially reduced, consistent with proteome

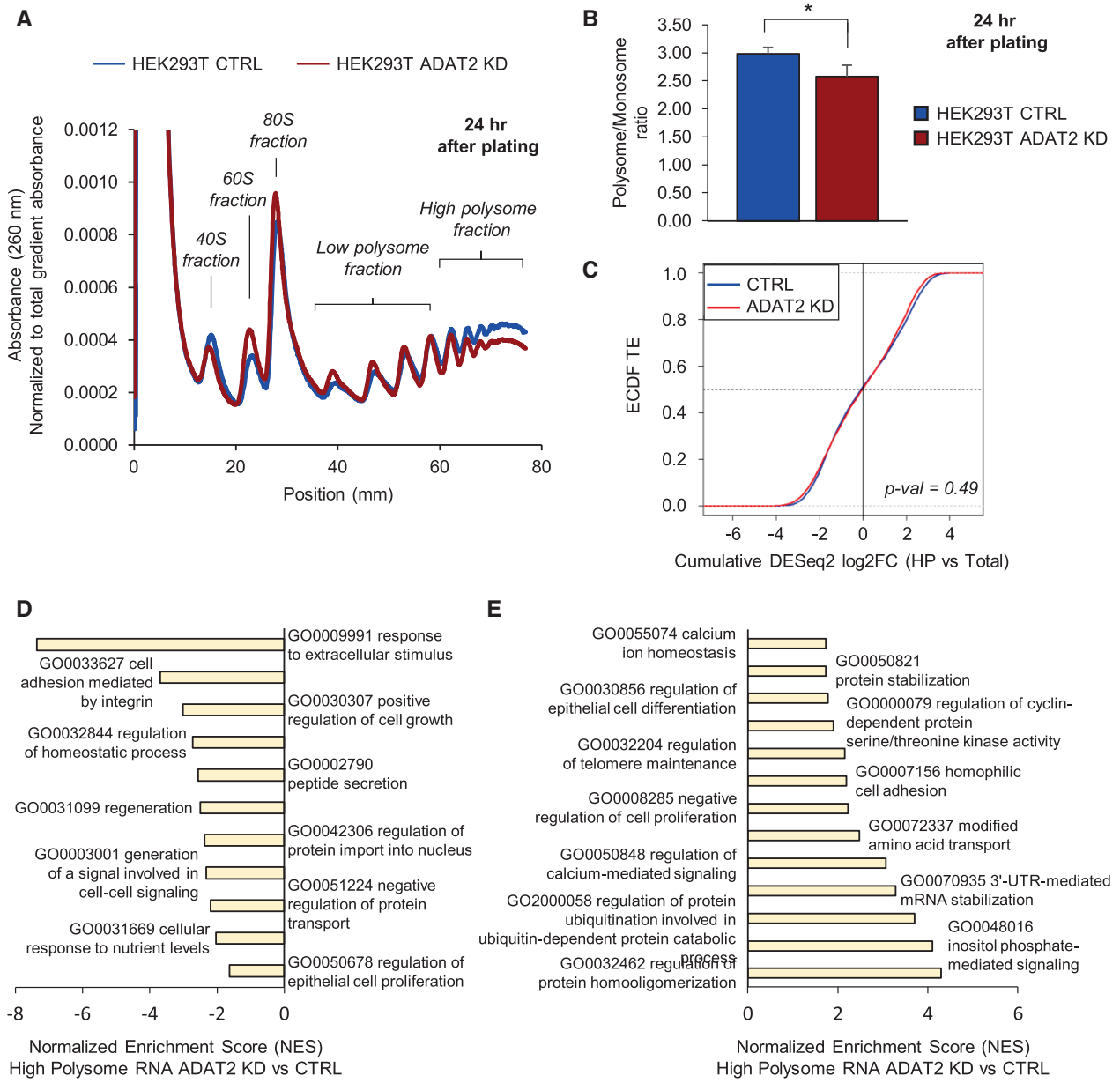


Figure 6. (**A**) Representative polysome profile of trypsin-treated HEK293T CTRL (blue) and ADAT2 KD (red) cells at 24 h after plating. The 40S, 60S, 80S, Low polysome, and High polysome (HP) fractions are indicated for reference. Experiments were done in biological triplicates. See also Supplementary Figure S4A. (**B**) Polysome to Monosome ratio (P/M) obtained from experiments as in (A). Shown are the mean and standard deviations of biological triplicates. *: *P*-value < 0.05 (*t*-test). (**C**) Evaluation of translation efficiency (TE) of transcripts based on their expression in total RNA (Total) and high polysome fraction (HP). Shown are the Empirical Cumulative Distribution Function (ecdf) of translation efficiencies for HEK293T CTRL (blue) and HEK293T ADAT2 KD (red) cells. Statistical test was performed using the mded function to compare empirical distributions (57). (**D** and **E**) ROAST Gene Set Enrichment Analysis using the Gene Ontology set Biological Process for transcripts depleted (D) or enriched (E) in the HP fraction of HEK293T ADAT2 KD cells. Selected categories are shown. Adj. *P*-value < 0.05. See also Supplementary Table S4.

normalization after protease treatment (Supplementary Figure S4A).

To quantitatively assess these differences, we calculated polysome to monosome ratios (P/M ratio). We found that ADAT2 KD cells presented a significant depletion in the P/M ratio at 24 h after trypsin treatment (Figure 6B), a difference that disappeared at 72 h after treatment (Supplementary Figure S4B). This data is consistent with the hypothesis that the modest effects on ribosome occupancy observed are due to translation impairment of a subset of genes, while global translation is generally not affected.

To characterize transcripts differentially translated in ADAT2 KD cells, we performed RNA-Seq at 24 h after trypsin treatment and plating, both from input RNA ('Total RNA' to assess transcriptomic changes) as well as from RNA obtained from the high polysome (HP) fractions. We

detected a significant depletion of ADAT2 transcripts (FC < 1.5; adj. *P*-value < 0.1) in the HP fraction of the ADAT2 KD cell line without significant changes in total RNA (Supplementary Figure S4C), indicating translational impairment. This is consistent with ribosomal drop-off caused by the premature stop codon introduced in this gene by CRISPR/Cas9 editing (Supplementary Figure S1A, see also Figure 2A). In agreement with previous observations, we did not observe alterations of ADAT3 transcript levels in total RNA or HP fractions in this cell line (Supplementary Figure S4C, see also Figure 2A and Supplementary Figure S2A).

Although we found 726 differentially enriched or depleted (|FC| > 1.5; adj. *P*-value < 0.1) protein-coding genes in the HP fractions of ADAT2 KD cells (Supplementary Table S4), a global analysis of translation efficiency (TE) (cumulative $\log_2$FC HP versus total) found no major differences in ADAT2 KD cells compared to CTRL cells (*P*-value = 0.49; Figure 6C). This is in agreement with our previous observations that general translation is not affected in ADAT2 KD cells, and indicates that most of the differential expression observed in HP fractions can be explained by changes in transcriptional rates.

Despite the fact that general translation efficiency is not affected by depletion of I34-tRNAs, gene ontology (GO) analyses revealed compositional differences in the HP fractions of ADAT2 KD cells. Indeed, HP fractions after ADAT2 KD are significantly depleted in transcripts associated to cellular proliferation, cell adhesion, cell-cell signalling, response to extracellular stimuli, protein transport and peptide secretion functions. On the other hand, HP fractions after ADAT2 KD are significantly enriched in transcripts linked to cellular differentiation, calcium signalling, protein and mRNA stabilization, telomere maintenance, and protein ubiquitination (Figure 6D, E and Supplementary Table S4). Thus, the depletion of I34-tRNAs does not affect general translation efficiency, but induces changes in the composition of transcript populations associated to ribosomes.

Impact of I34-tRNA depletion upon translation depends on codon composition and distribution

To gauge the relationship between codon composition and I34-tRNA dependence, we first assessed the impact of ADAT depletion upon translation of proteins with an even distribution of TAPSLIVR in their sequences. To that end we engineered two eGFP genes where codons for TAPSLIVR were either C-ended (ADAT-sensitive; 'eGFP ADAT'), or G-ended (ADAT-insensitive; 'eGFP nonADAT') (Figure 7A, see also Materials and methods). GFP reporters are frequently used for the analysis of codon-biased translation (73). Importantly, to prevent differences in translation rates caused simply by the changes in codon usage, we ensured that these two eGFP sequences would share a similar Codon Adaptation Index (CAI) (74).

We observed similar levels of total eGFP protein and fluorescence in ADAT2 KD and CTRL cells when transfected with either eGFP variant by western blotting and FACS analyses (Figure 7B, C, respectively). This indicates that both eGFP variants are translated at a similar rate and that they fold into their active form to produce fluorescence in both cell lines. Likewise, we detected equivalent eGFP production in HEK293T shCV and shADAT2 cells, and for both expression constructs (Supplementary Figure S5A and B). Thus, eGFP translation is not sensitive to partial I34-tRNA depletion, even if codon composition is maximally biased towards I34-tRNA use. In addition, we did not find signs of mistranslation based on peptide analysis by mass spectrometry (Supplementary Figure S5C). Therefore, a reduction in I34 levels had no effect upon the efficiency or the fidelity of translation of a soluble protein containing evenly distributed TAPSLIVR. This is consistent with the observation that translation of soluble proteins of average amino acid composition remains unaffected upon ADAT2 KD (Figure 3).

We have previously shown that the frequency of codons recognised by I34-tRNAs in eukaryotic genes positively correlates with the number of consecutive TAPSLIVR-encoding codons in the corresponding proteins (10,23,24). We therefore asked whether I34 levels are important for the synthesis of proteins with low-complexity TAPSLIVR-rich regions. First, we identified human transcripts encoding proteins with low-complexity TAPSLIVR-rich regions, and ranked them according to the size of these regions, and their relative enrichment in TAPSLIVR codons cognate for I34-tRNAs (Supplementary Table S5). Next, we performed an *in silico* functional characterization of the identified low-complexity TAPSLIVR-rich proteins. We found that this subset of the human proteome is associated to cellular structure, morphology, adhesion, cell signalling, and interaction with the extracellular space (Supplementary Table S5). We further found that these low-complexity regions are characteristic of mucin-like domains (MLDs) (75), and are abundant in Mucins (MUC) and other proteins involved in ECM regulation and adhesion (Supplementary Table S5).

We monitored endogenous levels of the MLD-containing protein Syndecan 3 (gene *SDC3*) (76) (Figure 7D), as a function of ADAT2 levels. We found that cells depleted of I34-tRNAs produce less SDC3 compared to CTRL cells, without significant changes in SDC3 transcript abundance (Figure 7E and F). To test if this effect was due to translation impairment of the low-complexity MLD region of SDC3 we generated a reporter gene where this section of the SDC3 transcript (Figure 7D) was cloned at the N-terminus of a Renilla luciferase (RLuc) gene (SDC3-RLuc). We also generated an equivalent construct (SDC3(G-end)-RLuc) where all ADAT-sensitive codons (U-, C- and A-ended codons) of the cloned region of SDC3 were replaced by G-ended codons, thus rendering them I34-tRNA-insensitive (decoded by C34-tRNAs) (see Materials and methods). Both constructs contain a Firefly luciferase (FLuc) that acts as an internal control for normalization of expression (Figure 7G).

We found a 20% reduction in SDC3-RLuc expression, but not of SDC3(G-end)-RLuc, in ADAT2 KD cells at 48 h after transfection (Figure 7H). A time-course analysis revealed a continued decrease of SDC3-RLuc in ADAT2 KD cells relative to CTRL cells (Figure 7I). We purified SDC3-RLuc and SDC3(G-end)-RLuc from all cell lines and found their protein sequences to be identical by mass spectrome-

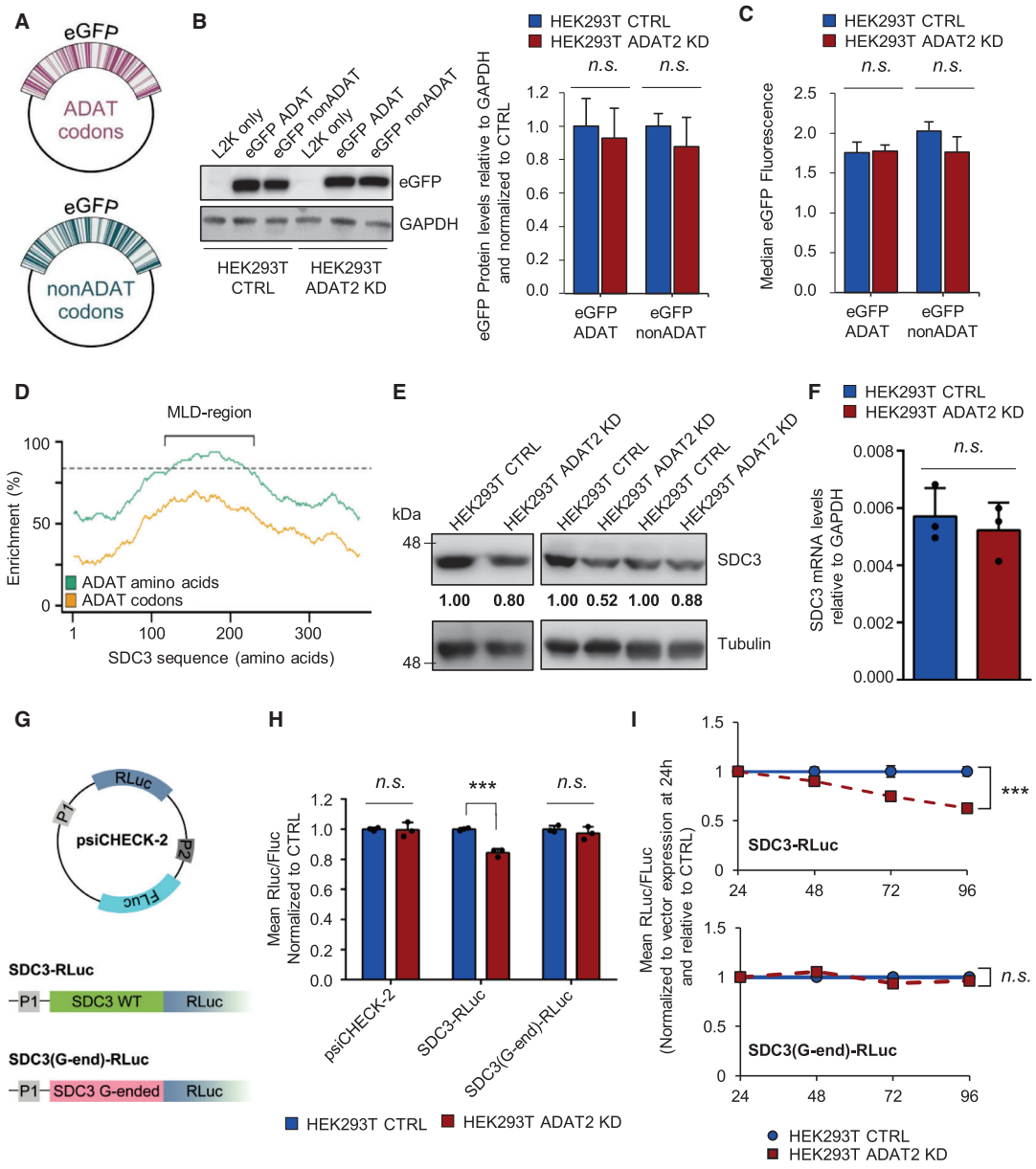


Figure 7. (**A**) Schematic representation of constructs encoding eGFP with TAPSLIVR codons that are C-ended (ADAT codons, upper panel in violet) or G-ended (nonADAT codons, lower panel in dark green). (**B**) eGFP protein levels in the indicated cell lines when transfected with eGFP constructs depicted in (A). GAPDH is used as gel loading control. 'L2K only': cells incubated with lipofectamine 2000 in the absence of eGFP constructs. Quantification of western blot bands for eGFP relative to GAPDH levels and normalized to CTRL cells was obtained from three biological replicates. Shown are the mean and SD. *n.s.*: not statistically significant (t-test). See also Supplementary Figure S5A–C. (**C**) Evaluation of eGFP fluorescence by FACS in the indicated cell lines when transfected with eGFP constructs depicted in (A). Shown is the median and SD of biological duplicates. n.s.: not statistically significant (*t*-test). (**D**) Distribution of TAPSLIVR (ADAT amino acids; green curve) and of codons decoded by I34-tRNAs (ADAT codons; orange curve) along the coding sequence of SDC3. Dotted line indicates the threshold of TAPSLIVR enrichment considered significant as defined in (23). The low-complexity TAPSLIVR-rich region containing the Mucin-like domain (MLD) is shown. (**E**) SDC3 protein levels in HEK293T CTRL and ADAT2 KD cells. Tubulin is used as gel loading control. Quantification of SDC3 bands relative to Tubulin and normalized to CTRL cells for each biological triplicate is shown. (**F**) SDC3 transcript levels relative to GAPDH in the indicated cell lines. Shown are biological triplicates, their mean and SD. n.s.: not statistically significant (*t*-test). (**G**) Schematic representation of psiCHECK-2-based SDC3 luciferase reporters bearing the wild-type low-complexity region of SDC3 ('SDC3 WT' in light green) or the same amino acid sequence but encoded with synonymous G-ended codons for all TAPSLIVR (codons not recognised by I34-tRNAs, 'SDC3 G-ended' in pink). The promoter 1 (P1) drives the expression of Renilla luciferase (RLuc). The second independent promoter (P2) drives the expression of Firefly luciferase (FLuc). (**H**) Evaluation of luciferase expression (RLuc/FLuc ratio normalized to CTRL cells) in HEK293T CTRL and ADAT2 KD cells transfected with the constructs depicted in (G). Shown are biological triplicates, their mean and SD. n.s.: not statistically significant. **: *P*-value < 0.01 (*t*-test). (**I**) Time-course analysis of luciferase expression (RLuc/FLuc ratio normalized to CTRL cells and to vector expression at the 24 h time-point) in the indicated cell lines transfected with SDC3-RLuc (upper panel) or SDC3(G-end)-RLuc (lower panel). Shown are the means (blue for CTRL and red for ADAT2 KD cells) and SD of biological triplicates. n.s.: not statistically significant. ***: *P*-value < 0.001 (*t*-test). See also Supplementary Figure S5D.

try (Supplementary Figure S5D). Thus, in contrast to transcripts with evenly distributed TAPSLIVR codons, a partial depletion of I34-tRNAs impairs the translation of low-complexity TAPSLIVR-rich transcripts.

In light of this evidence we revisited our polysome profiling data to evaluate the specific synthesis of low-complexity TAPSLIVR-rich proteins in ADAT2 KD cells. Using an interaction analysis (see Materials and methods) we found that 7 out of the 36 genes with impaired TE in ADAT2 KD cells (FC HP versus total < 1.5; P-value < 0.05) encoded proteins with TAPSLIVR-rich low-complexity regions (Figure 8A and Supplementary Table S4). Notably, we found that under these conditions, these transcripts are highly translated in CTRL cells (i.e. FC HP CTRL versus total CTRL > 1.5; P-value < 0.05) (Figure 8A). A permutation test revealed that the fraction of translationally impaired transcripts in ADAT2 KD cells that are highly translated in CTRL cells (31 genes) is enriched in low-complexity TAPSLIVR-rich coding sequences (10 000 sets of 31 random genes detected in the polysome profiling experiment, P-value $= 0.0121$; Figure 8B). These results suggest that ADAT2 KD causes impaired translation of transcripts that require I34-tRNAs and are under high translational demand.

To extend this analysis to a larger set of transcripts we relaxed the statistical constraints imposed on the above-mentioned interaction analysis, and evaluated the TE (i.e. upregulated (FC > 0) or downregulated (FC < 0)) upon ADAT2 KD without setting up a FC or P-value threshold. We found 989 transcripts encoding proteins with low-complexity TAPSLIVR-rich regions with downregulated TE, representing a statistically significant enrichment among all detected transcripts with downregulated TE (P-value $= 0.03036$; Fisher exact test). This significance is increased when the analysis is restricted to highly translated transcripts in CTRL cells (i.e. FC HP CTRL versus Total CTRL > 0) (550 transcripts encoding low-complexity TAPSLIVR regions; P-value $=$ 4.985e–14; Fisher exact test) (Supplementary Table S4). These analyses support the observation that transcripts encoding proteins with low-complexity TAPSLIVR-rich regions are enriched among those translationally impaired upon ADAT2 KD, particularly if such transcripts are under high translational demand.

Depletion of I34-tRNAs impairs translation of MLD-containing proteins in different cell lines

To rule out cell-specific effects we evaluated the endogenous levels of two additional low-complexity MLD-containing proteins in different cellular model systems. First, we examined the expression of Dystroglycan 1 (coded by the gene *DAG1*) (Figure 9A) in HT29-M6 shCV and shADAT2 cells (Supplementary Figure S3B, C and Supplementary Figure S6A). Dystroglycan 1 is translated from a single transcript as a propeptide that is post-translationally cleaved into two subunits: alpha-dystroglycan (α-DG) that has a low-complexity TAPSLIVR-rich MLD, and beta-dystroglycan (β-DG) (77) (Figure 9A). Therefore, defects in translation of α-DG should also impact translation of β-DG. We found that ADAT2 KD reduced the levels of both proteins (Figure 9B) without affecting DAG1 mRNA abundance (Figure 9C).

To investigate translation of mucins (MUC) in ADAT-silenced cells, we used a line of human pulmonary mucoepidermoid carcinoma cells (NCI-H292) where MUC production is induced by the epidermal growth factor-like protein amphiregulin (AREG) (78). Interestingly, AREG treatment induced ADAT2 expression in both NCI-H292 shCV and shADAT2 cells, although the latter continued to present reduced ADAT2 abundance compared to shCV cells (Figure 9D). This is consistent with the notion that ADAT activity is linked to the efficient synthesis of mucins. We evaluated a number of molecular markers of AREG-induced signalling and found that ADAT2 depletion did not generally affect the cellular response to AREG treatment (Supplementary Figure S6B-C).

We next examined the expression of mucin-5 Subtype AC (*MUC5AC*) (Figure 9E). As expected, AREG treatment sharply increased the levels of MUC5AC mRNAs (78). This activation was of $\sim$ 1000-fold and similar for shCV and shADAT2 cells (Figure 9F). However, we detected a strong reduction in MUC5AC protein levels in shADAT2 cells, both by FACS analyses ($\sim$ 30% reduction, Figure 9G and Supplementary Figure S6D-E) and immunohistochemistry ($\sim$70% reduction, Figure 9H), consistent with a severe translational defect.

Low-complexity TAPSLIVR-rich proteins are primarily Eukarya-specific and enriched in multicellular organisms

The enrichment of I34-tRNAs in Eukarya (4,10,21), and the fact that MLDs are found mostly in eukaryotes (75), prompted us to ask whether low-complexity TAPSLIVR-rich proteins are overrepresented in Eukaryotes. Using established methods (79) we searched for homologous sequences to these human TAPSLIVR-rich proteins in all three domains of life. Evaluating homology on the basis of low-complexity regions is subject to numerous biases (80), thus we first identified homologs using the full sequence of human proteins containing TAPSLIVR-rich regions. In this way we were able to identify all proteins evolutionary related to the human query set, independently of their low-complexity TAPSLIVR-rich region.

We found that the average per-species abundance of homologous sequences to low-complexity human TAPSLIVR-rich proteins is 66-fold higher in eukaryotes than in prokaryotes (homologs in prokaryotes/homologs in eukaryotes $= 0.015$; see Materials and methods) (Figure 10A). To evaluate the significance of this result, we performed a permutation test with 5000 sets of randomly chosen human sequences. This confirmed that low-complexity TAPSLIVR-rich protein homologs are exceedingly rare in prokaryotic organisms (P-value $<$ 1e–10) (Figure 10A). We then examined the presence of low-complexity TAPSLIVR-rich regions within these proteins to find that they are almost absent in prokaryotes (Figure 10B).

Interestingly, we found an uneven distribution of homologs of these sequences within eukaryotes (Figure 10B). We asked whether this could correlate with the number of tRNA genes coding for precursors of I34-tRNAs (i.e. A34-

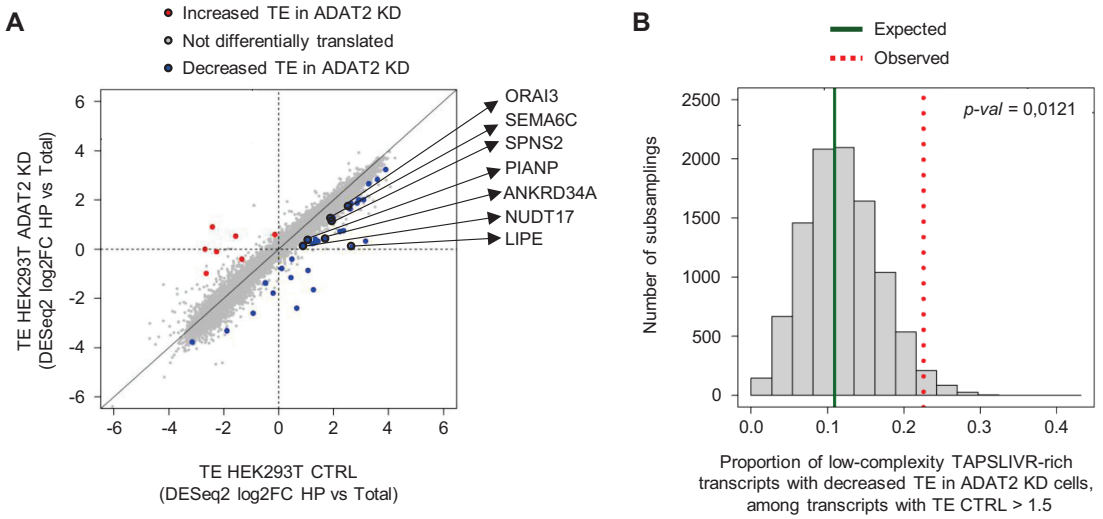


Figure 8. (**A**) Scatter plot showing the evaluation of translation efficiency (TE) for individual transcripts based on an interaction analysis between transcript expression in total RNA (Total) and high polysome (HP) fractions obtained for experiments as in Figure 6. Blue, Red and Grey dots indicate genes with decreased TE, increased TE and unchanged TE in HEK293T ADAT2 KD cells as compared to HEK293T CTRL cells, respectively. Note that the x- and y-axis show Log2FC values (log2FC 1.5 = 0.58). Translationally impaired transcripts that encode proteins with low-complexity TAPSLIVR-rich regions are indicated. See also Supplementary Table S4. (**B**) Histogram of a permutation test using 10 000 sets of 31 random transcripts detected in experiments shown in (A). The expected (dark solid green line) and observed (red dotted line) proportion of transcripts encoding low-complexity TAPSLIVR-rich proteins with decreased TE in ADAT2 KD cells among transcripts highly translated in HEK293T CTRL cells (FC HP CTRL versus Total CTRL > 1.5) is indicated. *P*-values are computed as the proportion of permutations with more extreme statistics than the observed. See also Supplementary Table S4.

tRNA genes) present in these species. We found a positive trend ($R = 0.53$, P-value = 3.7e–13, Spearman) between the abundance of low-complexity TASPLIVR-rich proteins and A34-tRNA gene content (Figure 10C and Supplementary Table S6). Furthermore, eukaryotes that lack A34-tRNA genes for any of the TAPSLIVR (i.e. reduced A34-tRNA gene diversity) are also depleted in low-complexity TAPSLIVR-rich homologs (Figure 10C). Of note and as expected, proteins with low-complexity TAPSLIVR-rich regions are rare in bacterial and archaeal genomes (Figure 10B), which are depleted of I34-tRNAs (Supplementary Figure S7).

The capacity to synthesize cell adhesion molecules was instrumental for the origin of multicellularity (81–84), and low-complexity TAPSLIVR-rich proteins are involved in cell adhesion and extracellular matrix interactions (Supplementary Table S5). We calculated the average per-species abundance of homologs to human low-complexity TAPSLIVR-rich proteins in unicellular and multicellular eukaryotes and found that they are severely depleted in unicellular species (homologs in unicellular eukaryotes/homologs in multicellular eukaryotes = 0.096; P-value < 1e–34 compared to 5000 sets of randomly chosen human sequences) (Figure 10D). We further evaluated the presence of low-complexity TAPSLIVR-rich regions within these proteins and detected a 7.3-fold enrichment in multicellular species (Figure 10E and Supplementary Table S7).

These results show that the scarcity or abundance of I34-tRNAs in eukaryotes correlate with the capacity of these species to synthesize proteins with low-complexity TAPSLIVR-rich regions involved in cell adhesion, and with their unicellular or multicellular condition. Interestingly, we found four unicellular eukaryotic species with an unusually high number of proteins with low-complexity TASPLIVR-rich stretches (i.e. *Capsaspora owczarzaki, Salpingoeca rosetta, Monosiga brevicollis* and *Spizellomyces punctatus*) (Figure 10E). All these species are considered model organisms to study the transition towards metazoan multicellularity (25,83–86).

DISCUSSION

The selective pressures that drove the evolution of the translation apparatus, and their impact upon the functional and structural diversity of proteomes are unknown. More specifically, the replacement of G34-tRNAs for I34-tRNAs in eukaryotic genomes is a major event during early eukaryotic evolution that remains unexplained (4,5,21). Extant eukaryotic I34-tRNAs are essential to translate C-ended codons due to the lack of genes coding for isoacceptor G34-tRNAs (6,14,16,19,20,58). However, although this highlights an essential function of these tRNAs, it does not inform on the selective advantage that drove their expansion early in eukaryote evolution.

The expansion of I34-tRNAs during eukaryotic emergence needs to be considered in the context of the physical constraints surrounding codon-anticodon interactions. G34-tRNAs generate high-energy codon-anticodon pairings (87) which, in bacteria, require an internal base pairing between bases 32 and 38 of the anticodon loop to reduce the codon anticodon affinity through structural strains upon the loop structure (26). In eukaryotic translation systems, G34-tRNAs induce miscoding and are toxic, presumably because of non-cognate pairing of G34-tRNA anticodons with C-ended codons (26). It is conceivable that bacterial G34-tRNAs would cause a fitness conflict when used by an

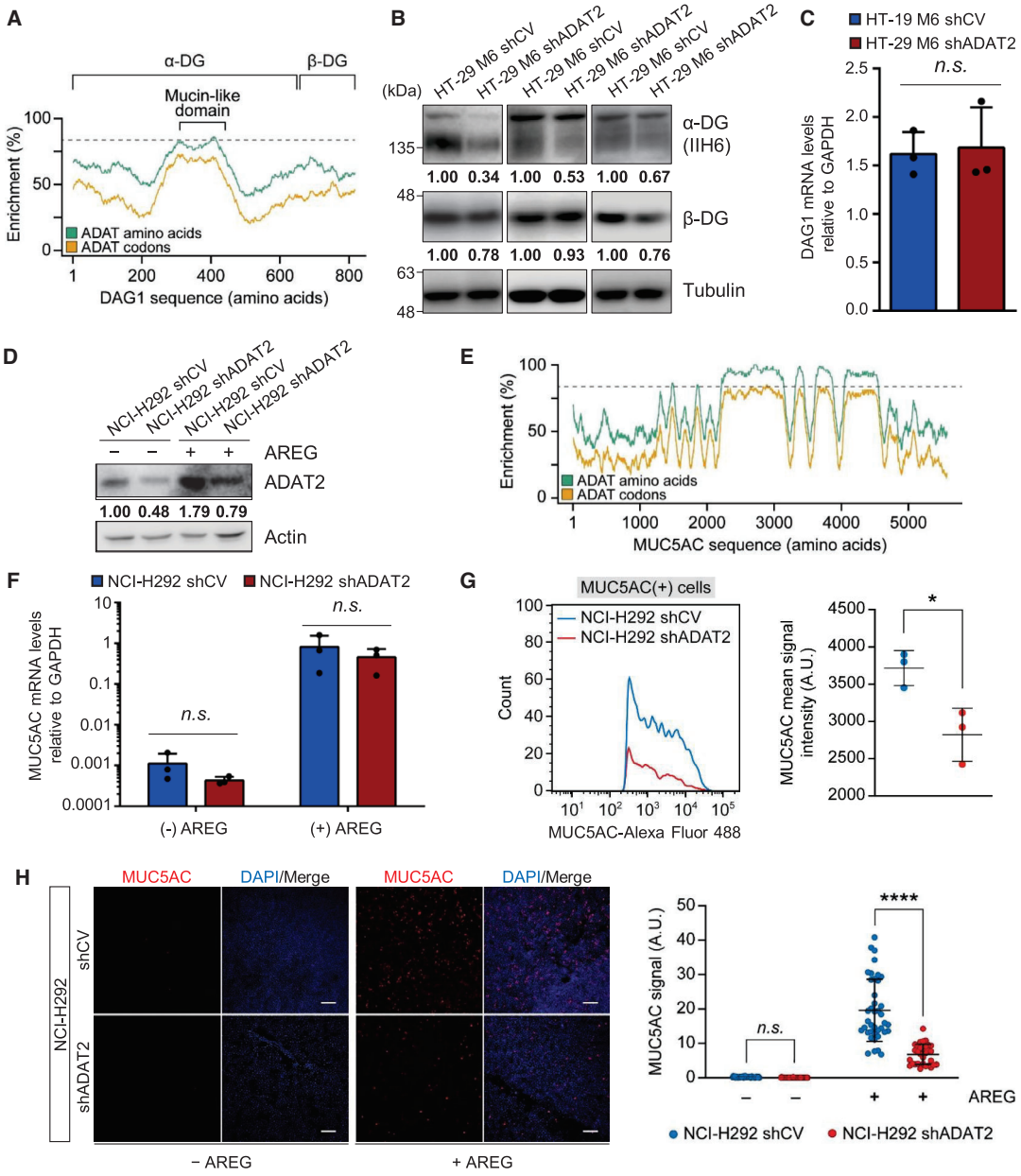


Figure 9. (**A**) Distribution of TAPSLIVR (ADAT amino acids; green curve) and of codons decoded by I34-tRNAs (ADAT codons; orange curve) along the coding sequence of DAG1. Dotted line indicates the threshold of TAPSLIVR enrichment considered significant as defined in (23). The low-complexity TAPSLIVR-rich region containing the Mucin-like domain, and the α and β dystroglycan (DG) subunits are shown. (**B**) α-DG and β-DG protein levels in HT-29 M6 shCV and shADAT2 cells. Tubulin is used as gel loading control. Quantification of DG bands relative to Tubulin and normalized to shCV cells for each biological triplicate is shown. Note that for each replicate α-DG and β-DG were measured from the same protein extract. ADAT2 levels in each protein extract is shown in Supplementary Figure S6A. (**C**) DAG1 transcript levels relative to GAPDH in the indicated cell lines. Shown are biological triplicates, their mean and SD. n.s.: not statistically significant (*t*-test). (**D**) ADAT2 protein levels in NCI-H292 shCV and shADAT2 cells treated (+) or not (−) with amphiregulin (AREG). Actin is used as gel loading control. Quantification of ADAT2 bands relative to Actin and normalized to shCV cells (−) AREG is shown. See also Supplementary Figure S6B-C. (**E**) Distribution of TAPSLIVR (ADAT amino acids; green curve) and of codons decoded by I34-tRNAs (ADAT codons; orange curve) along the coding sequence of MUC5AC. Dotted line indicates the threshold of TAPSLIVR enrichment considered significant as defined in (23). (**F**) MUC5AC transcript levels relative to GAPDH in the indicated cell lines treated (+) or not (−) with AREG. Shown are biological triplicates, their mean and SD. n.s.: not statistically significant (*t*-test). (**G**) MUC5AC expression levels upon AREG treatment in NCI-H292 shCV (blue) or shADAT2 (red) cells, quantified by FACS. Only MUC5AC(+) cells are shown. Left panel: representative assay showing number of MUC5AC(+) cells (y-axis) and MUC5AC signal intensity (x-axis). Right panel: MUC5AC mean signal intensity obtained for three biological replicates. The mean of the means and SD is also shown. *: *P*-value < 0.05 (*t*-test). See also Supplementary Figure S6D and E. (**H**) MUC5AC expression in NCI-H292 shCV and shADAT2 cells treated (+) or not (−) with AREG, quantified by confocal microscopy. Left panel: representative confocal microscopy images. MUC5AC (red) and nuclei staining (DAPI, blue) are shown. Scale bar corresponds to 10 μm. Right panel: quantification of microscopy images ($n > 30$) as those shown in Left panel for the indicated cell lines with (+) or without (−) AREG treatment with their mean and SD. n.s.: not statistically significant. ****: *P*-value < 0.0001 (*t*-test).

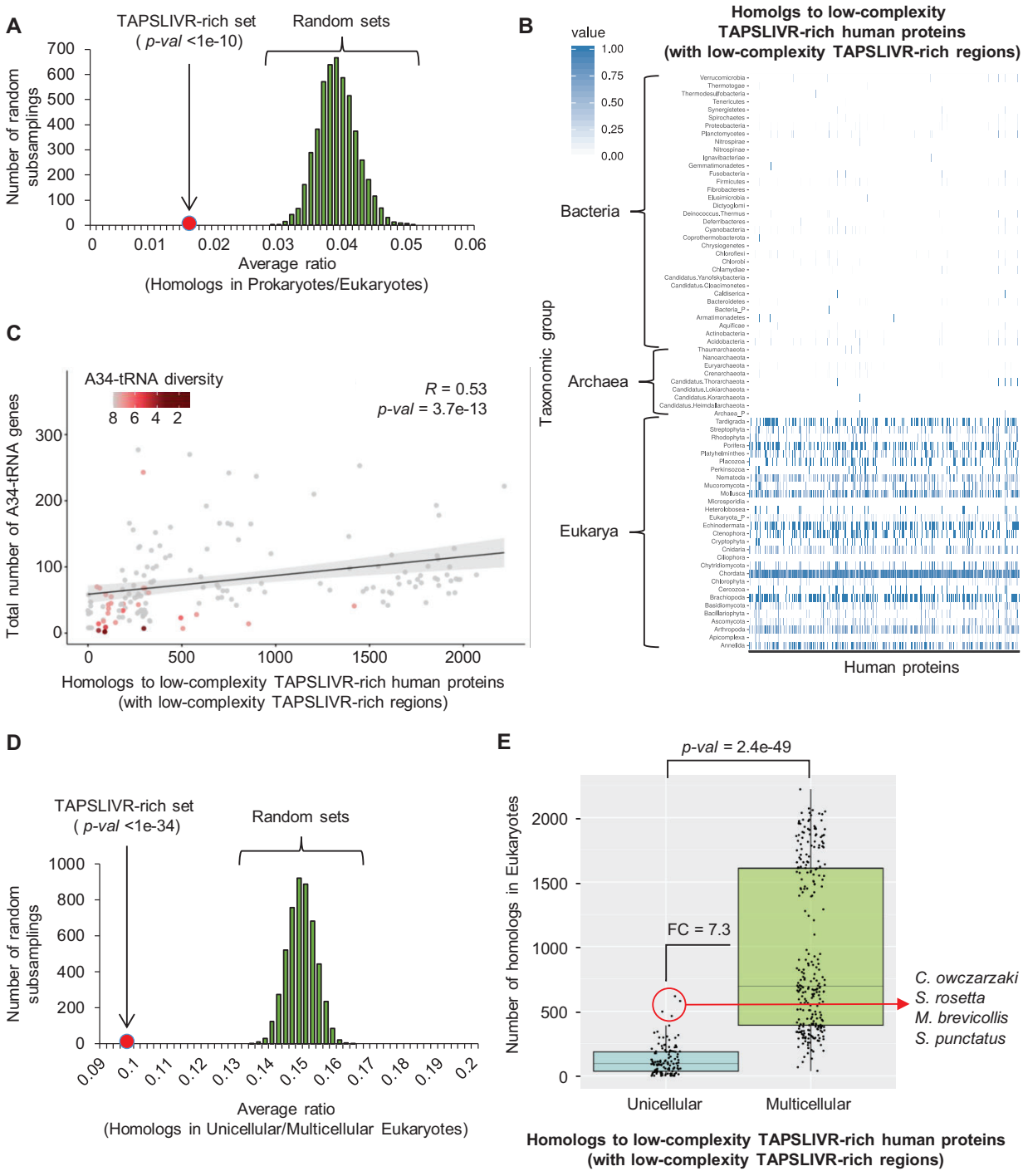


Figure 10. (**A**) Histogram of permutation tests using 5000 sets of random human proteins, showing the distribution of homologs in prokaryotes relative to those in eukaryotes (average ratio) for each random set analysed (subsamplings). The average ratio obtained for the human TAPSLIVR-rich set is shown (red dot). *P*-value < 1e-10 (based on z-score). (**B**) Heatmap visualization of the abundance of homologs to low-complexity TAPSLIVR-rich human proteins, and that contain low-complexity TAPSLIVR-rich regions, in taxonomic groups belonging to Archaea, Bacteria and Eukarya. Colouring scale represents hits normalized to the number of species analysed in each taxonomic group. X-axis: human proteins used as input for these analyses. (**C**) Spearman's rank correlation coefficient between the total number of A34-tRNA genes and the number of low-complexity human TAPSLIVR-rich proteins with homologs in eukaryotes. Color-coding indicates A34-tRNA diversity being 8 the maximum (i.e. at least one A34-tRNA gene to decode each TAPSLIVR). See also Supplementary Table S6 and Supplementary Figure S7. (**D**) Histogram of permutation tests using 5000 sets of random human proteins, showing the distribution of homologs in unicellular eukaryotes relative to multicellular in eukaryotes (average ratio) for each random set analysed (subsamplings). The average ratio obtained for the human TAPSLIVR-rich set is shown (red dot). *P*-value < 1e–34 (based on *z*-score). See also Supplementary Table S7. (**E**) Boxplot representation of the abundance of homologs to low-complexity TASPLIVR-rich human proteins, and that contain low-complexity TAPSLIVR-rich regions, in unicellular and multicellular eukaryotes. The identity of outlier unicellular species is shown. FC: Fold-change of the medians. Statistical significance was obtained by Mann–Whitney *U* test. See also Supplementary Table S7.

archaeal-type translation machinery, leading to the substitution of G34-tRNAs by an alternative tRNA. However, this scenario does not explain why I34-tRNAs would be the preferred solution to this conflict. The toxicity of G34-tRNAs in human cells could be alleviated by single base changes at positions 32 or 38 (26), moreover, I34-tRNAs may impose other constraints upon tRNA sequences. For example, eukaryotic $tRNA^{Ala}_{AGC}$ presents peculiar tertiary structures unique to this kingdom (88). Thus, additional selective forces may have contributed to the dramatic expansion of I34-tRNAs in nucleated cells.

In this work we use cellular models that are partially depleted from I34-tRNAs to levels equivalent or lower than those previously reported in human cell lines or other species upon ADAT downregulation (13,15–17,20,49,89). This depletion is achieved without causing additional alterations to the tRNA pool, and the resulting cells are still viable (Figures 2 and 4, Supplementary Figure S2B-C and Supplementary Table S2). Under these conditions, we might expect to identify cellular processes particularly sensitive to I34 depletion. We find that depletion of I34-tRNAs impairs cell adhesion (Figure 5C), and ADAT KD cells tend to detach when exposed to protein synthesis inhibitors, but not to other cellular insults (Figure 5A and B). Moreover, we observe that reduced I34 levels cause an abnormal cellular morphology upon trypsin treatments (Figure 4D and Supplementary Figure S3D), indicating that I34-tRNAs are required for the *de novo* synthesis of membrane proteins involved in interactions with the extracellular environment. We also observe reduced proliferation and a slower cell cycle (Figure 4A–C), two phenotypes commonly caused by defects in cellular adhesion (71,72) and previously reported in other cellular systems upon ADAT depletion (6,15,16,19,20). Silencing of ADAT2 also causes a notable decrease in transcripts with high ribosomal occupancy after trypsin treatment (Figure 6A), and within this group, we find an overrepresentation of genes linked to cell adhesion, response to extracellular stimuli and cell-cell signalling (Figure 6D and Supplementary Table S4), indicating that extracellular polypeptides predominate among those affected by a partial depletion of I34-tRNAs.

These observations are consistent with the hypothesis that partial I34-tRNA depletion leads to translational impairment of a specific subset of transcripts. Indeed, we do not find translation to be generally compromised when I34-tRNA levels are reduced, as shown by monitoring protein synthesis rates through metabolic labelling, analysing soluble and insoluble protein fractions, and evaluating UPR markers and levels of protein ubiquitination (Figure 3A–F and Supplementary Figure S2D). Furthermore, whole proteome mass spectrometry analyses revealed only 7 proteins out of 2280 to be differentially expressed (2 proteins upregulated and 5 downregulated) (Figure 3G and Supplementary Table S3). We favour the hypothesis that these changes are due to modulation of transcriptional rates. Likewise, no global alterations in translation efficiency were observed in ADAT2 KD cells by RNA-Seq in polysome profiling experiments (Figure 6C), and only 36 genes out of 12 447 were found to be translationally impaired (Figure 8A, and see below). This indicates that the majority of the differential abundance of transcripts associated to ribosomes found in ADAT2 KD cells could be explained by changes in transcriptional rates (Figure 6 and Supplementary Table S4).

On the other hand, we do find impaired translation of low-complexity TAPSLIVR-rich proteins that are encoded by transcripts enriched in codons cognate for I34-tRNAs. An *in silico* analysis of low-complexity TAPSLIVR-rich proteins functionally links this subset of the human proteome to cellular integrity, adhesion, and generation of, and interaction with the ECM, among others (Supplementary Table S5). We analysed endogenous expression of human transcripts encoding low-complexity TAPSLIVR-rich proteins in three different cellular systems. We detected translational defects in membrane-associated proteins such as LIPE, SPNS2, ORAI3, PIANP and SEMA6C (Figure 8A and Supplementary Table S4) (90), and in proteins containing MLDs (75–77) such as SDC3, Dystroglycan and MUC5AC (Figures 7D–F and Figure 9). Notably, not all low-complexity TAPSLIVR-rich proteins are transmembrane or secretory proteins (Supplementary Table S5), thus I34-tRNA depletion may affect translation of proteins both in the cytosol and the endoplasmic reticulum (91). The most striking translational phenotype caused by ADAT2 silencing was the reduction in the *de novo* synthesis of MUC5AC protein in NCI-H292 cells stimulated with AREG, despite a ∼1000-fold transcriptional activation of the *MUC5AC* gene (Figure 9D–H and Supplementary Figure S6B–E).

We also find that transcripts encoding low-complexity TAPSLIVR-rich proteins under high translational demand are more sensitive to I34-tRNA depletion. Our polysome profiling data detects translational impairment on transcripts encoding low-complexity TAPSLIVR-rich proteins that are highly translated in CTRL cells (Figure 8 and Supplementary Table S4). Likewise, we observe severe translational defects for MUC5AC in a cellular context where MUC5AC is required to be highly translated (i.e. upon AREG stimulation) (Figure 9D–H). On the other hand, translational impairment of SDC3 or Dystroglycan under standard growth conditions is milder (Figures 7D–F and 9A–C).

We determined that depletion of I34-tRNAs primarily affects translational efficiency, but not accuracy, of low-complexity TAPSLIVR-rich proteins, as seen by the time-dependent recovery of phenotypes (Figures 4D and 6A-B and Supplementary Figure S4A-B), time-course analysis of translation (Figure 7I), and mass spectrometry analyses (Supplementary Figure S5D). Furthermore, we showed that translation impairment is codon-dependent, as defects are not detected when TAPSLIVR codons are mutated to triplets not recognized by I34-tRNAs (Figure 7G–I). These results do not question a general role for I34-tRNAs in efficient translation (13,15), particularly because I34-tRNAs are required to decode all C-ended codons for TAPSLIVR (see Discussion above), a fact that explains why a full depletion of ADAT is lethal in all eukaryotic models (also this work, Figure 2A, B and see Materials and methods). Rather, our data support the hypothesis that a full complement of I34-tRNAs is essential for the efficient translation of low-complexity TAPSLIVR-rich coding sequences.

The fact that a partial downregulation of ADAT2 affects the translation of a specific subset of proteins without affecting overall protein fidelity or abundance could be

used to develop new therapies designed to treat conditions caused by the accumulation of low-complexity TAPSLIVR-rich proteins, such as asthma or chronic obstructive pulmonary disease (92); or to control infection by viruses that may use MLDs for immunoevasion (93,94). Mutations in human ADAT have been associated to a complex syndrome that includes intellectual disability, microcephaly, and strabismus (95–100). The depletion of I34-tRNAs in our cellular models (Figure 2B and (17)) is similar to the reported levels of I34-tRNAs in patients carrying mutations in ADAT (97). Our results suggest that a defective synthesis of low-complexity TAPSLIVR-rich proteins might contribute to these phenotypes.

We characterized the phylogenetic distribution of homologs to human proteins with low-complexity TAPSLIVR-rich regions that depend on I34-tRNAs for their synthesis, and found that they are almost limited to eukaryotic species that abundantly utilize I34-tRNAs (Figure 10A, B and Supplementary Figure S7). It has previously been shown that tRNA genes encoding for I34-tRNA precursors (i.e. A34-tRNAs) are more abundant in eukaryotes than prokaryotes, a fact accompanied by a concomitant enrichment in eukaryotic codon usage bias towards codons cognate for I34-tRNAs (4,5,10,21,24,25). In addition, the abundance of A34-tRNA genes also correlates with the presence of TadA/ADAT required for A34-to-I34 editing (10,101–103). Here we show that, although I34-tRNAs exist in deeply-rooted eukaryotic groups (10,24,25), their abundance correlates with that of proteins with low-complexity TAPSLIVR-rich regions (Figure 10C, Supplementary Table S6 and Supplementary Figure S7). Strikingly, we find such proteins to be scarce in unicellular eukaryotes, with the sole exception of holozoan protists (the closest known relatives of metazoans (25)), where the abundance of low-complexity TAPSLIVR-rich proteins is comparable to that of multicellular species (Figure 10D-E and Supplementary Table S7)

Our data supports the hypothesis that I34-tRNAs contributed to expand eukaryotic proteome diversity, facilitating the synthesis of a specific set of low-complexity proteins involved in cellular interactions with the extracellular environment. There is evidence for adaptations of the translation machinery required for the synthesis of proteins of highly biased amino acid content. For example, the bacterial EF-P (eukaryotic eIF5A) is an elongation factor that allows the synthesis of poly-proline stretches (40). Likewise, in the salivary glands of certain arthropods, modulation of the tRNA pool is essential for the production of silk fibres that are alanine, glycine and serine rich (33). Other tRNA modifications have been reported important for decoding short stretches of consecutive codons (104,105). However, I34 is the first example of a translation machinery adaptation linked to the emergence of a new set of functionally-related proteins. We posit that the enrichment in I34-tRNAs provided organisms with the opportunity to translate low-complexity TAPSLIVR-rich proteins, which were then selected and expanded because of the functional advantages they provide in extracellular functions such as cellular adhesion. This is consistent with the proposal that unicellular ancestors to extant metazoans already possessed genetic features required for multicellularity (81–84). It is tempting to speculate that I34-tRNAs contributed to the burst of low-complexity TAPSLIVR-rich proteins in holozoan protists, which may have facilitated the advent of metazoan multicellularity.

DATA AVAILABILITY

The datasets generated during this study are available at NCBI GEO (accession GSE150860), and at the ProteomeXchange Consortium via the PRIDE (dataset identifier PXD025024).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the Biostatistics and Bioinformatics Unit (IRB Barcelona), the Mass Spectrometry and Proteomics Service (IRB Barcelona), and the Cytometry Facility (U. of Barcelona) for technical assistance and advice. We thank Dr E. Batlle (IRB Barcelona) for providing the cell line HT-29 M6, the parental px330-GFP plasmid and technical advice on CRISPR/Cas9 cloning and targeting.

Author contributions: Conceptualization, A.G.T. and L.R.dP.; Methodology, A.G.T., M.R.-E., A.P.-S. and F.M.T.; Formal Analysis, A.G.T., M.R.-E., M.M.-H, H.C., M.M., A.R.-Y. and O.R.; Investigation, A.G.T., M.R.-E., M.M.-H., H.G.S.V., N.C.; Writing – Original Draft, A.G.T. and L.R.dP.; Writing – Review & Edit, A.G.T., M.R.-E., M.M.-H., A.R.-Y., O.R., F.M.T., A.P.-S., E.M.N., T.G. and L.R.dP.; Visualization, A.G.T., M.R.-E., M.M.-H.; Resources, A.P-S., E.M.N., T.G. and L.R.dP.; Supervision, A.G.T., E.M.N., T.G. and L.R.dP.; Project Administration, L.R.dP.; Funding Acquisition, E.M.N., T.G. and L.R.dP.

FUNDING

Spanish Ministry of Economy and Competitiveness [PID2019-108037RB-100 to L.R.d.P., PGC2018-099921 to T.G., PGC2018-098152-A-100 to E.M.N]; Australian Research Council [DP180103571 to E.M.N]; European Union's Horizon 2020 research and innovation programme [ERC-2016–724173 to T.G.]. Funding for open access charge: Spanish Ministry of Economy and Competitiveness.

Conflict of interest statement. None declared.

REFERENCES

1. Goodarzi,H., Nguyen,H.C.B., Zhang,S., Dill,B.D., Molina,H. and Tavazoie,S.F. (2016) Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell*, **165**, 1416–1427.
2. Chan,C.T., Pang,Y.L., Deng,W., Babu,I.R., Dyavaiah,M., Begley,T.J. and Dedon,P.C. (2012) Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat. Commun.*, **3**, 937.
3. Licht,K., Hartl,M., Amman,F., Anrather,D., Janisiw,M.P. and Jantsch,M.F. (2019) Inosine induces context-dependent recoding and translational stalling. *Nucleic Acids Res.*, **47**, 3–14.
4. Grosjean,H., de Crecy-Lagard,V. and Marck,C. (2010) Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett.*, **584**, 252–264.

5. Maraia,R.J. and Arimbasseri,A.G. (2017) Factors that shape eukaryotic tRNAomes: processing, modification and anticodon-codon use. *Biomolecules*, **7**, 26.
6. Gerber,A.P. and Keller,W. (1999) An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science*, **286**, 1146–1149.
7. Srinivasan,S., Torres,A.G. and Ribas de Pouplana,L. (2021) Inosine in biology and disease. *Genes*, **12**, 600.
8. Crick,F.H. (1966) Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**, 548–555.
9. Torres,A.G., Pineyro,D., Filonava,L., Stracker,T.H., Batlle,E. and Ribas de Pouplana,L. (2014) A-to-I editing on tRNAs: Biochemical, biological and evolutionary implications. *FEBS Lett.*, **588**, 4279–4286.
10. Rafels-Ybern,A., Torres,A.G., Camacho,N., Herencia-Ropero,A., Roura Frigole,H., Wulff,T.F., Raboteg,M., Bordons,A., Grau-Bove,X., Ruiz-Trillo,I. *et al.* (2019) The expansion of inosine at the wobble position of tRNAs, and its role in the evolution of proteomes. *Mol. Biol. Evol.*, **36**, 650–662.
11. Wolf,J., Gerber,A.P. and Keller,W. (2002) tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. *EMBO J.*, **21**, 3841–3851.
12. Arimbasseri,A.G., Blewett,N.H., Iben,J.R., Lamichhane,T.N., Cherkasova,V., Hafner,M. and Maraia,R.J. (2015) RNA polymerase III output is functionally linked to tRNA dimethyl-G26 modification. *PLoS Genet.*, **11**, e1005671.
13. Bornelov,S., Selmi,T., Flad,S., Dietmann,S. and Frye,M. (2019) Codon usage optimization in pluripotent embryonic stem cells. *Genome Biol.*, **20**, 119.
14. Liu,H., Wang,Q., He,Y., Chen,L., Hao,C., Jiang,C., Li,Y., Dai,Y., Kang,Z. and Xu,J.R. (2016) Genome-wide A-to-I RNA editing in fungi independent of ADAR enzymes. *Genome Res.*, **26**, 499–509.
15. Lyu,X., Yang,Q., Li,L., Dang,Y., Zhou,Z., Chen,S. and Liu,Y. (2020) Adaptation of codon usage to tRNA I34 modification controls translation kinetics and proteome landscape. *PLoS Genet.*, **16**, e1008836.
16. Rubio,M.A., Pastar,I., Gaston,K.W., Ragone,F.L., Janzen,C.J., Cross,G.A., Papavasiliou,F.N. and Alfonzo,J.D. (2007) An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. *Proc. Natl Acad. Sci. U.S.A.*, **104**, 7821–7826.
17. Torres,A.G., Pineyro,D., Rodriguez-Escriba,M., Camacho,N., Reina,O., Saint-Leger,A., Filonava,L., Batlle,E. and Ribas de Pouplana,L. (2015) Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Res.*, **43**, 5145–5157.
18. Torres,A.G., Wulff,T.F., Rodriguez-Escriba,M., Camacho,N. and Ribas de Pouplana,L. (2018) Detection of inosine on Transfer RNAs without a reverse transcription reaction. *Biochemistry (Mosc.)*, **57**, 5641–5647.
19. Tsutsumi,S., Sugiura,R., Ma,Y., Tokuoka,H., Ohta,K., Ohte,R., Noma,A., Suzuki,T. and Kuno,T. (2007) Wobble inosine tRNA modification is essential to cell cycle progression in G(1)/S and G(2)/M transitions in fission yeast. *J. Biol. Chem.*, **282**, 33459–33465.
20. Zhou,W., Karcher,D. and Bock,R. (2014) Identification of enzymes for adenosine-to-inosine editing and discovery of cytidine-to-uridine editing in nucleus-encoded transfer RNAs of Arabidopsis. *Plant Physiol.*, **166**, 1985–1997.
21. Novoa,E.M., Pavon-Eternod,M., Pan,T. and Ribas de Pouplana,L. (2012) A role for tRNA modifications in genome structure and codon usage. *Cell*, **149**, 202–213.
22. Novoa,E.M. and Ribas de Pouplana,L. (2012) Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet.*, **28**, 574–581.
23. Rafels-Ybern,A., Attolini,C.S. and Ribas de Pouplana,L. (2015) Distribution of ADAT-dependent codons in the human transcriptome. *Int. J. Mol. Sci.*, **16**, 17303–17314.
24. Rafels-Ybern,A., Torres,A.G., Grau-Bove,X., Ruiz-Trillo,I. and Ribas de Pouplana,L. (2018) Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla. *RNA biology*, **15**, 500–507.
25. Southworth,J., Armitage,P., Fallon,B., Dawson,H., Bryk,J. and Carr,M. (2018) Patterns of ancestral animal codon usage bias revealed through holozoan protists. *Mol. Biol. Evol.*, **35**, 2499–2511.
26. Pernod,K., Schaeffer,L., Chicher,J., Hok,E., Rick,C., Geslain,R., Eriani,G., Westhof,E., Ryckelynck,M. and Martin,F. (2020) The nature of the purine at position 34 in tRNAs of 4-codon boxes is correlated with nucleotides at positions 32 and 38 to maintain decoding fidelity. *Nucleic Acids Res.*, **48**, 6170–6183.
27. Saint-Leger,A., Bello,C., Dans,P.D., Torres,A.G., Novoa,E.M., Camacho,N., Orozco,M., Kondrashov,F.A. and Ribas de Pouplana,L. (2016) Saturation of recognition elements blocks evolution of new tRNA identities. *Sci. Adv.*, **2**, e1501860.
28. Mier,P., Paladin,L., Tamana,S., Petrosian,S., Hajdu-Soltesz,B., Urbanek,A., Gruca,A., Plewczynski,D., Grynberg,M., Bernado,P. *et al.* (2020) Disentangling the complexity of low complexity proteins. *Brief. Bioinform.*, **21**, 458–472.
29. Cascarina,S.M., Elder,M.R. and Ross,E.D. (2020) Atypical structural tendencies among low-complexity domains in the Protein Data Bank proteome. *PLoS Comput. Biol.*, **16**, e1007487.
30. Kumari,B., Kumar,R. and Kumar,M. (2015) Low complexity and disordered regions of proteins have different structural and amino acid preferences. *Mol. Biosyst.*, **11**, 585–594.
31. Saqi,M. (1995) An analysis of structural instances of low complexity sequence segments. *Protein. Eng.*, **8**, 1069–1073.
32. Suveges,D., Gaspari,Z., Toth,G. and Nyitray,L. (2009) Charged single alpha-helix: a versatile protein structural motif. *Proteins*, **74**, 905–916.
33. Chevallier,A. and Garel,J.P. (1982) Differential synthesis rates of tRNA species in the silk gland of Bombyx mori are required to promote tRNA adaptation to silk messages. *Eur. J. Biochem.*, **124**, 477–482.
34. Luo,H. and Nijveen,H. (2014) Understanding and identifying amino acid repeats. *Brief. Bioinform.*, **15**, 582–591.
35. So,C.R., Fears,K.P., Leary,D.H., Scancella,J.M., Wang,Z., Liu,J.L., Orihuela,B., Rittschof,D., Spillmann,C.M. and Wahl,K.J. (2016) Sequence basis of barnacle cement nanostructure is defined by proteins with silk homology. *Sci. Rep.*, **6**, 36219.
36. Harrison,P.M. (2006) Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. *BMC Bioinformatics*, **7**, 441.
37. Schaper,E., Gascuel,O. and Anisimova,M. (2014) Deep conservation of human protein tandem repeats within the eukaryotes. *Mol. Biol. Evol.*, **31**, 1132–1148.
38. Frugier,M., Bour,T., Ayach,M., Santos,M.A., Rudinger-Thirion,J., Theobald-Dietrich,A. and Pizzi,E. (2010) Low complexity regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. *FEBS Lett.*, **584**, 448–454.
39. Davies,J.E. and Rubinsztein,D.C. (2006) Polyalanine and polyserine frameshift products in Huntington's disease. *J. Med. Genet.*, **43**, 893–896.
40. Lassak,J., Wilson,D.N. and Jung,K. (2016) Stall no more at polyproline stretches with the translation elongation factors EF-P and IF-5A. *Mol. Microbiol.*, **99**, 219–235.
41. Ribas de Pouplana,L., Torres,A.G. and Rafels-Ybern,A. (2017) What froze the genetic code? *Life (Basel)*, **7**, 14.
42. Schuller,A.P. and Green,R. (2018) Roadblocks and resolutions in eukaryotic translation. *Nat. Rev. Mol. Cell Biol.*, **19**, 526–541.
43. Cortina,C., Turon,G., Stork,D., Hernando-Momblona,X., Sevillano,M., Aguilera,M., Tosi,S., Merlos-Suarez,A., Stephan-Otto Attolini,C., Sancho,E. *et al.* (2017) A genome editing approach to study cancer stem cells in human tumors. *EMBO Mol. Med.*, **9**, 869–879.
44. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
45. Paddison,P.J., Cleary,M., Silva,J.M., Chang,K., Sheth,N., Sachidanandam,R. and Hannon,G.J. (2004) Cloning of short hairpin RNAs for gene knockdown in mammalian cells. *Nat. Methods*, **1**, 163–167.
46. Picchioni,D., Antolin-Fontes,A., Camacho,N., Schmitz,C., Pons-Pons,A., Rodriguez-Escriba,M., Machallekidou,A., Guler,M.N., Siatra,P., Carretero-Junquera,M. *et al.* (2019) Mitochondrial protein synthesis and mtDNA levels coordinated

through an aminoacyl-tRNA synthetase subunit. *Cell Rep.*, **27**, 40–47.
47. Mitrovic,S., Nogueira,C., Cantero-Recasens,G., Kiefer,K., Fernandez-Fernandez,J.M., Popoff,J.F., Casano,L., Bard,F.A., Gomez,R., Valverde,M.A. *et al.* (2013) TRPM5-mediated calcium uptake regulates mucin secretion from human colon goblet cells. *eLife*, **2**, e00658.
48. Sahu,B., Laakso,M., Ovaska,K., Mirtti,T., Lundin,J., Rannikko,A., Sankila,A., Turunen,J.P., Lundin,M., Konsti,J. *et al.* (2011) Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *EMBO J.*, **30**, 3962–3976.
49. Wulff,T.F., Arguello,R.J., Molina Jordan,M., Roura Frigole,H., Hauquier,G., Filonava,L., Camacho,N., Gatti,E., Pierre,P., Ribas de Pouplana,L. *et al.* (2017) Detection of a subset of posttranscriptional transfer RNA modifications in vivo with a restriction fragment length polymorphism-based method. *Biochemistry (Mosc).*, **56**, 4029–4038.
50. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
51. Torres,A.G., Reina,O., Stephan-Otto Attolini,C. and Ribas de Pouplana,L. (2019) Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 8451–8456.
52. Bonifacino,J.S. (2001) Metabolic labeling with amino acids. *Curr. Protoc. Cell Biol.*, https://doi.org/10.1002/0471140864.ps0307s17.
53. Pringle,E.S., McCormick,C. and Cheng,Z. (2019) Polysome profiling analysis of mRNA and associated proteins engaged in translation. *Curr Protoc Mol Biol*, **125**, e79.
54. Martinez-Nunez,R.T. and Sanford,J.R. (2016) Studying isoform-specific mRNA recruitment to polyribosomes with Frac-seq. *Methods Mol. Biol.*, **1358**, 99–108.
55. Wu,D., Lim,E., Vaillant,F., Asselin-Labat,M.L., Visvader,J.E. and Smyth,G.K. (2010) ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, **26**, 2176–2182.
56. Efron,B. and Tibshirani,R. (2001) On testing the significance of sets of genes. *Annals of Applied Statistics*, **1**, 107–129.
57. Poe,G.L., Giraud,K.L. and Loomis,J.B. (2005) Computational methods for measuring the difference of empirical distributions *Amer. J. Agr. Econ.*, **87**, 353–365.
58. Chan,P.P. and Lowe,T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.
59. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
60. Kinter,M. and Sherman,N.E. (2000) In: *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. John Wiley, NY.
61. Wisniewski,J.R., Zougman,A., Nagaraj,N. and Mann,M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods*, **6**, 359–362.
62. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
63. Schindelin,J., Arganda-Carreras,I., Frise,E., Kaynig,V., Longair,M., Pietzsch,T., Preibisch,S., Rueden,C., Saalfeld,S., Schmid,B. *et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**, 676–682.
64. Altenhoff,A.M., Glover,N.M., Train,C.M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., de Farias,T.M., Zile,K., Stevenson,C., Long,J. *et al.* (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.
65. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
66. Supek,F., Lehner,B. and Lindeboom,R.G.H. (2020) To NMD or Not To NMD: nonsense-mediated mRNA decay in cancer and other genetic diseases. *Trends Genet.*, https://doi.org/10.1016/j.tig.2020.11.002.
67. Keegan,L.P., Gerber,A.P., Brindle,J., Leemans,R., Gallo,A., Keller,W. and O'Connell,M.A. (2000) The properties of a tRNA-specific adenosine deaminase from Drosophila melanogaster support an evolutionary link between pre-mRNA editing and tRNA modification. *Mol. Cell. Biol.*, **20**, 825–833.
68. Geslain,R., Cubells,L., Bori-Sanz,T., Alvarez-Medina,R., Rossell,D., Marti,E. and Ribas de Pouplana,L. (2010) Chimeric tRNAs as tools to induce proteome damage and identify components of stress responses. *Nucleic Acids Res.*, **38**, e30.
69. Sekiya,M., Maruko-Otake,A., Hearn,S., Sakakibara,Y., Fujisaki,N., Suzuki,E., Ando,K. and Iijima,K.M. (2017) EDEM function in ERAD protects against chronic ER proteinopathy and age-related physiological decline in Drosophila. *Dev. Cell*, **41**, 652–664.
70. Shcherbakov,D., Teo,Y., Boukari,H., Cortes-Sanchon,A., Mantovani,M., Osinnii,I., Moore,J., Juskeviciene,R., Brilkova,M., Duscha,S. *et al.* (2019) Ribosomal mistranslation leads to silencing of the unfolded protein response and increased mitochondrial biogenesis. *Commun. Biol.*, **2**, 381.
71. Dix,C.L., Matthews,H.K., Uroz,M., McLaren,S., Wolf,L., Heatley,N., Win,Z., Almada,P., Henriques,R., Boutros,M. *et al.* (2018) The role of mitotic cell-substrate adhesion re-modeling in animal cell division. *Dev. Cell*, **45**, 132–145.
72. Gloerich,M., Bianchini,J.M., Siemers,K.A., Cohen,D.J. and Nelson,W.J. (2017) Cell division orientation is coupled to cell-cell adhesion by the E-cadherin/LGN complex. *Nat. Commun.*, **8**, 13996.
73. Torrent,M., Chalancon,G., de Groot,N.S., Wuster,A. and Madan Babu,M. (2018) Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Sci. Signal*, **11**, eaat6409.
74. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
75. Pinzon Martin,S., Seeberger,P.H. and Varon Silva,D. (2019) Mucins and pathogenic mucin-like molecules are immunomodulators during infection and targets for diagnostics and vaccines. *Front. Chem.*, **7**, 710.
76. Erdman,R., Stahl,R.C., Rothblum,K., Chernousov,M.A. and Carey,D.J. (2002) Schwann cell adhesion to a novel heparan sulfate binding site in the N-terminal domain of alpha 4 type V collagen is mediated by syndecan-3. *J. Biol. Chem.*, **277**, 7619–7625.
77. Barresi,R. and Campbell,K.P. (2006) Dystroglycan: from biosynthesis to pathogenesis of human disease. *J. Cell Sci.*, **119**, 199–207.
78. Huang,L., Pu,J., He,F., Liao,B., Hao,B., Hong,W., Ye,X., Chen,J., Zhao,J., Liu,S. *et al.* (2017) Positive feedback of the amphiregulin-EGFR-ERK pathway mediates PM2.5 from wood smoke-induced MUC5AC expression in epithelial cells. *Sci. Rep.*, **7**, 11084.
79. Pearson,W.R. (2013) An introduction to sequence similarity ("homology") searching. *Curr Protoc. Bioinformatics*, https://doi.org/10.1002/0471250953.bi0301s42.
80. Forslund,K. and Sonnhammer,E.L. (2009) Benchmarking homology detection procedures with low complexity filters. *Bioinformatics*, **25**, 2500–2505.
81. Abedin,M. and King,N. (2010) Diverse evolutionary paths to cell adhesion. *Trends Cell Biol.*, **20**, 734–742.
82. Cavalier-Smith,T. (2017) Origin of animal multicellularity: precursors, causes, consequences-the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **372**, 20150476.
83. King,N., Hittinger,C.T. and Carroll,S.B. (2003) Evolution of key cell signaling and adhesion protein families predates animal origins. *Science*, **301**, 361–363.
84. King,N., Westbrook,M.J., Young,S.L., Kuo,A., Abedin,M., Chapman,J., Fairclough,S., Hellsten,U., Isogai,Y., Letunic,I. *et al.* (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature*, **451**, 783–788.
85. Ruiz-Trillo,I., Burger,G., Holland,P.W., King,N., Lang,B.F., Roger,A.J. and Gray,M.W. (2007) The origins of multicellularity: a multi-taxon genome initiative. *Trends Genet.*, **23**, 113–118.
86. Sebe-Pedros,A., Ballare,C., Parra-Acero,H., Chiva,C., Tena,J.J., Sabido,E., Gomez-Skarmeta,J.L., Di Croce,L. and Ruiz-Trillo,I. (2016) The dynamic regulatory genome of capsaspora and the origin of animal multicellularity. *Cell*, **165**, 1224–1237.

87. Grosjean,H. and Westhof,E. (2016) An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res.*, **44**, 8020–8040.
88. Westhof,E., Liang,S., Tong,X., Ding,X., Zheng,L. and Dai,F. (2020) Unusual tertiary pairs in eukaryotic tRNA(Ala). *RNA*, **26**, 1519–1529.
89. Berthel,E., Vincent,A., Eberst,L., Torres,A.G., Dacheux,E., Rey,C., Marcel,V., Paraqindes,H., Lachuer,J., Catez,F. *et al.* (2020) Uncovering the translational regulatory activity of the tumor suppressor BRCA1. *Cells*, **9**, 941.
90. UniProt. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
91. Reid,D.W. and Nicchitta,C.V. (2015) Diversity and selectivity in mRNA translation on the endoplasmic reticulum. *Nat. Rev. Mol. Cell Biol.*, **16**, 221–231.
92. Samsuzzaman,M., Uddin,M.S., Shah,M.A. and Mathew,B. (2019) Natural inhibitors on airway mucin: Molecular insight into the therapeutic potential targeting MUC5AC expression and production. *Life Sci.*, **231**, 116485.
93. Andersen,K.G., Rambaut,A., Lipkin,W.I., Holmes,E.C. and Garry,R.F. (2020) The proximal origin of SARS-CoV-2. *Nat. Med.*, **26**, 450–452.
94. Bagdonaite,I. and Wandall,H.H. (2018) Global aspects of viral glycosylation. *Glycobiology*, **28**, 443–467.
95. Alazami,A.M., Hijazi,H., Al-Dosari,M.S., Shaheen,R., Hashem,A., Aldahmesh,M.A., Mohamed,J.Y., Kentab,A., Salih,M.A., Awaji,A. *et al.* (2013) Mutation in ADAT3, encoding adenosine deaminase acting on transfer RNA, causes intellectual disability and strabismus. *J. Med. Genet.*, **50**, 425–430.
96. El-Hattab,A.W., Saleh,M.A., Hashem,A., Al-Owain,M., Asmari,A.A., Rabei,H., Abdelraouf,H., Hashem,M., Alazami,A.M., Patel,N. *et al.* (2016) ADAT3-related intellectual disability: Further delineation of the phenotype. *Am. J. Med. Genet. A*, **170A**, 1142–1147.
97. Ramos,J., Han,L., Li,Y., Hagelskamp,F., Kellner,S.M., Alkuraya,F.S., Phizicky,E.M. and Fu,D. (2019) Formation of tRNA wobble inosine in humans is disrupted by a millennia-old mutation causing intellectual disability. *Mol. Cell. Biol.*, **39**, e00203-19.
98. Sharkia,R., Zalan,A., Jabareen-Masri,A., Zahalka,H. and Mahajnah,M. (2019) A new case confirming and expanding the phenotype spectrum of ADAT3-related intellectual disability syndrome. *Eur J Med Genet*, **62**, 103549.
99. Thomas,E., Lewis,A.M., Yang,Y., Chanprasert,S., Potocki,L. and Scott,D.A. (2019) Novel Missense Variants in ADAT3 as a Cause of Syndromic Intellectual Disability. *J Pediatr Genet*, **8**, 244–251.
100. Torres,A.G., Batlle,E. and Ribas de Pouplana,L. (2014) Role of tRNA modifications in human diseases. *Trends Mol. Med.*, **20**, 306–314.
101. de Crecy-Lagard,V., Marck,C., Brochier-Armanet,C. and Grosjean,H. (2007) Comparative RNomics and modomics in Mollicutes: prediction of gene function and evolutionary implications. *IUBMB Life*, **59**, 634–658.
102. Diwan,G.D. and Agashe,D. (2018) Wobbling forth and drifting back: The evolutionary history and impact of bacterial tRNA modifications. *Mol. Biol. Evol.* **35**, 2046–2059.
103. Yokobori,S., Kitamura,A., Grosjean,H. and Bessho,Y. (2013) Life without tRNAArg-adenosine deaminase TadA: evolutionary consequences of decoding the four CGN codons as arginine in Mycoplasmas and other Mollicutes. *Nucleic Acids Res.*, **41**, 6531–6543.
104. El Yacoubi,B., Hatin,I., Deutsch,C., Kahveci,T., Rousset,J.P., Iwata-Reuyl,D., Murzin,A.G. and de Crecy-Lagard,V. (2011) A role for the universal Kae1/Qri7/YgjD (COG0533) family in tRNA modification. *EMBO J.*, **30**, 882–893.
105. Muller,M., Legrand,C., Tuorto,F., Kelly,V.P., Atlasi,Y., Lyko,F. and Ehrenhofer-Murray,A.E. (2019) Queuine links translational control in eukaryotes to a micronutrient from bacteria. *Nucleic Acids Res.*, **47**, 3711–3727.

Torres et al. **Human tRNAs with inosine 34 are essential to efficiently translate eukarya-specific low-complexity proteins. Supplementary Methods, Supplementary Figures and Legends for Supplementary Tables.**

Supplementary Methods

Additional information on the generation of CRISPR-ADAT KD cell lines

Sorting of GFP-positive cells was done by excitation of the sample using a 488 nm laser for forward scatter (FSC) and GFP (530nm); a 561 nm laser was used for the excitation of side scatter (SSC) signal. Single GFP expressing cells were selected according to the scatter and fluorescence parameters; doublets were excluded using the FSC-W signal.

To validate CRISPR cell lines (**Figure S1**).DNA was extracted using the DNeasy Blood and Tissue kit (69504, Qiagen) following the manufacturer's protocol. Obtained DNA was quantified using a Nanodrop ND1000 and was used to PCR amplify ADAT genes. PCR amplifications were carried out using the Phusion High-Fidelity DNA Polymerase (F530S, Thermo Fisher) using primers depicted in **Supplementary Table S1**, and an annealing temperature of 60ºC for both set of primers. The obtained PCR products were gel-purified and sent for sequencing using as sequencing primers for the ADAT2 gene the same primers used for PCR amplification, and for ADAT3 another set of internal primers (see **Supplementary Table S1**).

Additional information on the cell line generation by lentiviral infection

For hADAT2-pDONR221 plasmid generation, hADAT2 gene was amplified using oligonucleotides depicted in **Supplementary Table S1** (hADAT2 with attB sites) following a standard PCR reaction with pfu ultra polymerase (600384, Agilent). All PCR reactions described in this work were done in a MyCycler Thermal Cycler System (Bio-Rad). hADAT2 attB PCR product was purified and cloned into pDONR 221 vector (12536017, Invitrogen) following the manufacturer's protocol (BP reaction).

For DOX-inducible shRNA expressing vectors the shRNA sequences were as follows:

shNonTarget sequence: 5'-TGCTGTTGACAGTGAGCGACCTCCACCCTCACTCTGCCATTAGTGAAGCCACAGATGTAATGGCAGAGTGAGGGTGGAGGGTGCCTACTGCCTCGGA-3'.

shADAT2 sequence: 5'-TGCTGTTGACAGTGAGCGCACACACTGTGTTGTATGTCACTAGTGAAGCCACAGATGTAGTGACATACAACACAGTGTGTTTGCCTACTGCCTCGGA-3'.

Lentivirus were produced in HEK293T cells growing in T75 flask format. Cells at approximately 80 % confluence were transfected with 75 µL of polyethylenimine (PEI) (23966-1, Polysciences), 1.5 µg envelope vector (VSVG), 1.5 µg Rev-expressing vector (RTR2), 4.5 µg packaging vector (PKG-PIR), 7.5 µg transfer vector and 1.5 mL 150 mM NaCl that was added to the cell culture media. 16 h later, cells were washed with PBS (14190094, Thermo Fisher), fresh Full Media was added and cells were left growing at 33 ºC/5 % CO_2. 48 hours post-transfection, viral supernatants were collected, cleared using a 0.45 µm filter, supplemented with 10 % FBS and 8 µg/mL polybrene (hexadimethrine bromide) (H9268, Sigma-Aldrich), and added to the cell line to be infected. Infected cells were left growing at 37 ºC/5 % CO_2. Fresh Full Media was added to the virus-producing HEK293T cells and were left growing at 33 ºC/5 % CO_2 for an additional day. The next day (72 hours post-transfection) viral supernatants were collected as before and were used to infect a second time the targeted cell line; at this point virus-producing cells were disposed. 24 hour after the second viral infection, cells were washed once with PBS and fresh Full Media was added. On the next day, 2 µg/mL puromycin (ant-pr-1, Invivogen) was added to culture medium for selection of transduced cells.

Additional information on protein extraction

Cells were harvested and washed once with 1 mL cold PBS. The cell pellet was re-suspended with an appropriate volume of RIPA buffer (50-250 µL). Cells were vortexed and incubated on ice for 30 minutes, followed by centrifugation at maximum speed at 4 ºC for 30 minutes. Supernatant (protein extract) was recovered and the remaining pellet was disposed.

Antibodies used in this study

Antibody used for western blotting and performed dilutions: PERK (C33E10) (3192, Cell Signaling Technology; RRID:AB_2095847) 1:500; IRE1 (phospho S724) (ab48187, Abcam; RRID:AB_873899) 1:1000; Phospho-eIF2α (Ser51) (9721, Cell Signaling Technology; RRID:AB_330951) 1:1000; eiF2α (9722, Cell Signaling Technology; RRID:AB_2230924) 1:1000; Actin (JLA20, Developmental Studies Hybridoma Bank; RRID:AB_528068) 1:1000; BiP (C50B12) (3177, Cell Signaling Technology; RRID:AB_2119845) 1:1000; ADAT2 (C-13) (sc-107385, Santa Cruz Biotechnology; RRID:AB_2273604) 1:1000; ADAT3 (N-term) (AP17369a, Abgent/Abcepta; RRID:AB_11136249) 1:200; Vinculin (SPM227) (ab18058, Abcam; RRID:AB_444215) 1:5000; GAPDH (G3PDH) (2275-PC-100, Trevigen; RRID:AB_2107456) 1:10000; Syndecan-3 (G-2) (sc-398194, Santa Cruz Biotechnology; RRID:AB_2732022) 1:500; Green Fluorescent Protein (06-896, Merck; RRID:AB_11214044) 1:1000; α-Dystroglycan (IIH6C4) (05-593, Merck; RRID:AB_309828) 1:250; β-Dystroglycan (MANDAG2(7D11)), Developmental Studies Hybridoma Bank; RRID:AB_2618140) 1:500; β-Tubulin (E7, Developmental Studies Hybridoma Bank; RRID:AB_528499) 1:1000; Phospho-EGF Receptor (Tyr1045) (2237, Cell Signaling Technology; RRID:AB_331710) 1:100; EGFR Antibody Cocktail (AHR5062, Thermo Fisher; RRID:AB_2536360) 1:250; Renilla Luciferase (PA5-32210, Thermo Fisher;

RRID:AB_2549683) 1:1000; Ubiquitin (Ub (P4D1)) (sc-8017, Santa Cruz Biotechnology; RRID:AB_628423) 1:1600. Secondary antibodies: Donkey Anti-Goat IgG H&L (HRP) (ab6885, Abcam; RRID:AB_955423) 1:10000; Sheep Anti-Mouse IgG (HRP) (NA931, VWR; RRID:AB_772210) 1:10000; Donkey Anti-Rabbit IgG (HRP) (NA934, VWR; RRID:AB_772206) 1:10000; Donkey Anti-Chicken IgY (HRP) (AP194P, Merck; RRID:AB_92682) 1:10000.

Antibodies used for protein purifications: Renilla luciferase antibody (PA5-32210, Thermo Fisher; RRID:AB_2549683); Green Fluorescent Protein antibody (DSHB-GFP-12A6, Developmental Studies Hybridoma Bank; RRID:AB_2617417).

Antibodies used for MUC5AC detection: MUC5AC (45M1) primary antibody (MA1-38223, Thermo Fisher; RRID:AB_2266697) diluted 1:200 and Anti-Mouse Alexa Fluor 555-conjugated secondary antibody (A-31570, Thermo Fisher; RRID:AB_2536180) for immunohistochemistry or Anti-Mouse Alexa Fluor 488-conjugated secondary antibody (A-21202, Thermo Fisher; RRID:AB_141607) for FACS analyses.

Additional information on cell cycle analyses

Cells growing on 10 cm Petri dish format at approximately 60 % confluence were incubated with 7 mL DMEM Full Media containing 2 mM Thymidine (T1895, Merck) for 13 hours at 37 ºC/5 % CO_2 (first thymidine cell cycle block). Media was then removed, cells were washed twice with 4 mL PBS, and were left growing in DMEM Full Media for 8 hours at 37 ºC/5 % CO_2. Media was then replaced by 7 mL DMEM Full Media containing 2 mM Thymidine and cells were left at 37 ºC/5 % CO_2 for 17hours (second thymidine cell cycle block). Cells were then washed twice with 4 mL PBS and were left growing on 6 mL of DMEM Full Media at 37 ºC/5 % CO_2 until harvesting time point.

Cells were harvested at time 0, 4, 8, and 12 hours after removal of the second thymidine block. Cells were detached using 5 mL PBS. Cell suspension was centrifuged for 3 minutes at 800 xg, the supernatant was removed and cells were re-suspended in 0.5 mL PBS. Then 4.5 mL ethanol 70 % was added to fix the cells and they were left at 4 ºC overnight.

The next day, the fixing solution was removed, cells were washed once with 5 mL PBS, and were re-suspended in 550 µL propidium iodide (PI) staining solution: 0.1 % Triton-X100, 2 mg/mL RNase A (Qiagen), 40 µg/mL PI (P4864), 1x PBS. Cells were incubated with staining solution for 3 hours at 37 ºC and were then analysed by FACS.

Additional information on polysome profiling

Sucrose gradients were generated in Open-top ultra clean tubes (344059, Beckman Coulter) with a Biocomp Gradient Station. 300 µL of cell lysate was loaded on top of the gradient and was centrifuged

using an SW41 rotor (Beckman Coulter) for 2 hours 30 minutes at 4 ºC at 35,000 RPM. After centrifugation the gradient was placed in the Gradient Station, scanned at 260 nm, and 20 fractions were collected per gradient. Gradient plots (**Figure 6A** and **Supplementary Figure S4A**) were normalized to the total 260 nm signal of the gradient to compensate for minor differences derived from unequal loading of cell lysates onto the gradient.

Additional information on construction of ADAT eGFP reporters

eGFP ADAT sequence:

GAATTCCTCGAGTCTAGAATGGTCAGCAAGGGCGAGGAGCTCTTCACCGGGGTCGTCCCCATCC
TCGTCGAGCTCGACGGCGACGTCAACGGCCACAAGTTCAGCGTCTCCGGCGAGGGCGAGGGC
GATGCCACCTACGGCAAGCTCACCCTCAAGTTCATCTGCACCACCGGCAAGCTCCCCGTCCCCT
GGCCCACCCTCGTCACCACCCTCACCTACGGCGTCCAGTGCTTCAGCCGCTACCCCGACCACAT
GAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTC
TTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTCAAGTTCGAGGGCGACACCCTCGTC
AACCGCATCGAGCTCAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTCGGGCACAAGCTC
GAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGG
TCAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTCCAGCTCGCCGACCACTACCAGCA
GAACACCCCCATCGGCGACGGCCCCGTCCTCCTCCCCGACAACCACTACCTCAGCACCCAGTC
CGCCCTCAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTCCTCGAGTTCGTCACCGC
CGCCGGGATCACCCTCGGCATGGACGAGCTCTACAAGTAAGTTTAAACACCGGTGAATTC

eGFP nonADAT sequence:

GAATTCCTCGAGTCTAGAATGGTGAGCAAGGGCGAGGAGCTGTTCACGGGGGTGGTGCCGATC
CTGGTGGAGCTGGACGGCGACGTGAACGGCCACAAGTTCAGCGTGTCGGGCGAGGGCGAGGG
CGATGCGACGTACGGCAAGCTGACGCTGAAGTTCATCTGCACGACGGGCAAGCTGCCGGTGCC
GTGGCCGACGCTGGTGACGACGCTGACGTACGGCGTGCAGTGCTTCAGCCGGTACCCGGACCA
CATGAAGCAGCACGACTTCTTCAAGTCGGCGATGCCGGAAGGCTACGTGCAGGAGCGGACGAT
CTTCTTCAAGGACGACGGCAACTACAAGACGCGGGCGGAGGTGAAGTTCGAGGGCGACACGCT
GGTGAACCGGATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAA
GCTGGAGTACAACTACAACAGCCACAACGTGTATATCATGGCGGACAAGCAGAAGAACGGCATC
AAGGTGAACTTCAAGATCCGGCACAACATCGAGGACGGCAGCGTGCAGCTGGCGGACCACTAC
CAGCAGAACACGCCGATCGGCGACGGCCCGGTGCTGCTGCCGGACAACCACTACCTGAGCACG
CAGTCGGCGCTGAGCAAAGACCCGAACGAGAAGCGGGATCACATGGTGCTGCTGGAGTTCGTG
ACGGCGGCGGGGATCACGCTGGGCATGGACGAGCTGTACAAGTAAGTTTAAACACCGGTGAAT
TC

Additional information on evaluation of ADAT eGFP production

For FACS analyses, excitation of the sample was done using a 488 nm air-cooled argon-ion laser at 15 mW power. The instrument was set up with the standard configuration: Forward scatter (FS), side scatter (SS), and green (525 nm) fluorescence for GFP. Fluorescence was collected in logarithmic scale. Optical alignment was checked using 10 nm fluorescent beads (Flow-Check fluorospheres, Catalog number 6605359; Beckman Coulter). Cell population was selected gating in a FS vs. SS dot plot, excluding aggregates and cell debris. To determine the percentage of GFP cells, a non-transfected control was used as negative reference.

Experiments depicted on **Supplementary Figure S5A-C** were performed on HEK293T DOX-inducible sh cell lines given that eGFP ADAT and eGFP nonADAT plasmids presented variable transfection efficiencies on HEK293T shCV and HEK293T shADAT2 cell lines. Cells were grown in DMEM Full Media with or without 1 µg/mL doxycycline (DOX) (D9891, Merck). At the time of lipofection, media from all cells was replaced and lipofection of eGFP ADAT and eGFP nonADAT plasmids was done without DOX. 4 hours after lipofection, cells were washed once with PBS and media was replaced by DMEM Full Media with or without DOX accordingly. 48 hours later cells were visualised in an Eclipse Ts2-FL microscope (Nikon).

Additional information on construction of ADAT luciferase reporters

First, the desired portion of SDC3/SDC3(G-end) was PCR amplified using a forward (FWD) primer containing a NheI restriction site and a start codon; and a reverse (RVR) primer containing a BamHI restriction site and the NanoLuc linker. Next, the psiCHECK-2 RLuc gene was PCR amplified from the vector using a FWD primer containing a BamHI restriction site and a deletion of the ATG-start codon of the RLuc gene; and a RVR primer having a XhoI restriction site. All PCR amplifications were carried out using high-fidelity polymerases (e.g. pfu ultra polymerase). Both amplicons were gel-purified, digested with BamHI and ligated following standard procedures. The ligated product (insert) was further PCR amplified to obtain more material using primers SDC3 (or SDC3(G-end)) FWD and RLuc RVR, and the amplicon was gel-purified. The amplified insert and the parental psiCHECK-2 plasmid were digested with NheI and XhoI, and gel-purified (note that this digestion removes the RLuc gene from the psiCHECK-2 plasmid). Digested vector and insert where then ligated using standard procedures to incorporate the SDC3-RLuc or SDC3(G-end)-RLuc insert into the vector.

SDC3 wild type low-complexity TAPSLIVR-rich region nucleotide sequence:

GAAGAGCTCCCCTCTGAGCGCCCCACCCTGGAGCCAGCCACCAGCCCCCTGGTGGTGACAGAA
GTCCCGGAAGAGCCCAGCCAGAGAGCCACCACCGTCTCCACTACCATGGCTACCACTGCTGCC
ACAAGCACAGGGGACCCGACTGTGGCCACAGTGCCTGCCACAGTGGCCACCGCCACCCCCAGC
ACCCCTGCAGCACCCCCTTTTACGGCCACCACTGCTGTTATAAGGACCACTGGCGTACGGAGGC
TTCTGCCTCTCCCACTGACCACAGTGGCTACGGCACGGGCCACTACCCCCGAGGCGCCCTCCC

CGCCCACCACGGCGGCTGTCTTGGACACCGAGGCCCCAACACCCAGGCTGGTCAGCACAGCTA
CCTCCCGGCCAAGAGCCCTTCCCAGGCCGGCCACCACCCAGGAGCCTGACATCCCTGAGAGGA
GCACCCTGCCCCTGGGGACCACTGCCCCTGGACCCACAGAGGTGGCTCAGACCCCAACTCCAG
AGACCTTCCTGACCACA

SDC3(G-end) low-complexity TAPSLIVR-rich region nucleotide sequence:

GAAGAGCTGCCGTCGGAGCGGCCGACGCTGGAGCCGGCGACGAGCCCGCTGGTGGTGACGGA
AGTGCCGGAAGAGCCGAGCCAGAGAGCGACGACGGTGTCGACGACGATGGCGACGACGGCGG
CGACGAGCACGGGGGACCCGACGGTGGCGACGGTGCCGGCGACGGTGGCGACGGCGACGCC
GAGCACGCCGGCGGCGCCGCCGTTTACGGCGACGACGGCGGTGATAAGGACGACGGGCGTGC
GGAGGCTGCTGCCGCTGCCGCTGACGACGGTGGCGACGGCGCGGGCGACGACGCCGGAGGC
GCCGTCGCCGCCGACGACGGCGGCGGTGTTGGACACGGAGGCGCCGACGCCGAGGCTGGTG
AGCACGGCGACGTCGCGGCCGAGAGCGCTGCCGAGGCCGGCGACGACGCAGGAGCCGGACA
TACCGGAGAGGAGCACGCTGCCGCTGGGGACGACGGCGCCGGGACCGACGGAGGTGGCGCA
GACGCCGACGCCGGAGACGTTCCTGACGACG

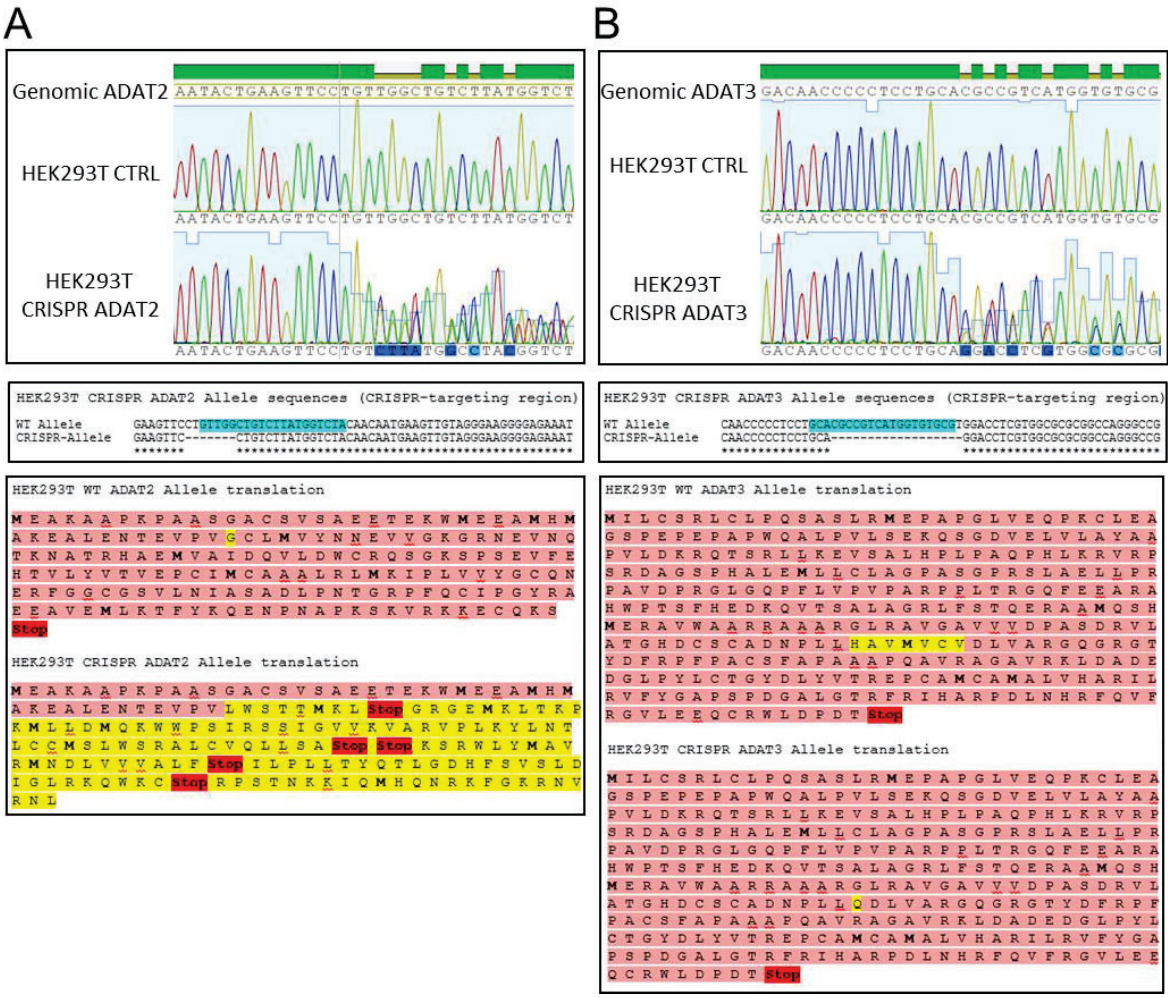
Additional information on purification of reporter proteins

HEK293T cells growing in 150 mm Petri dish format at around 90 % confluence were transfected with 10 µg SDC3-RLuc or SDC3(G-end)-RLuc constructs using L2K following the manufacturer's protocol (500 µL plasmid/lipid reaction in 21 mL DMEM Full Media). 48 hours after lipofection, cells were washed once with 5 mL cold PBS and were then collected into a pre-chilled 10 mL Falcon tube, by cell-scrapping using 3 mL cold PBS. Cells were then centrifuged at 500 xg for 3 minutes at 4 ºC, supernatant was removed and cell pellets were washed once with 3 mL cold PBS. Cells were centrifuged again, supernatant was removed and cell pellets were re-suspended in 500 µL Lysis buffer (20 mM Tris pH 8, 2 mM EDTA, 1 % NP-40, 150 mM NaCl, 0.1 mM NaVaO, 1X PIC), and incubated in a rotating wheel for 30 minutes at 4 ºC. Cells were then centrifuged at maximum speed at 4 ºC for 30 minutes and the supernatant was recovered and transferred to a new pre-chilled tube ("cell lysate"). Cell lysate was incubated with 20 µL magnetic Dynabeads Protein A (10002D, Thermo Fisher) in a rotating wheel for 2 hours at 4 ºC. Beads were separated using a magnet and the supernatant was recovered and transferred to a new pre-chilled tube ("pre-cleared cell lysate"). 50 µL dynabeads A were incubated for 10 minutes in a rotating wheel at RTº with 5 µg Renilla luciferase antibody (PA5-32210, Thermo Fisher) in 200 µL Wash Buffer (100 mM Tris pH 7.5, 1 mM EDTA, 1 mM EGTA, 1 % NP-40, 350 mM NaCl, 0.2 mM NaVaO, 1x PIC) ("Ig-Dynabeads"). Supernatant was removed and Ig-Dynabeads were washed twice with 200 µL Wash Buffer. Ig-Dynabeads were then washed twice with 200 µL Conjugation Buffer and were then cross-linked by re-suspending them in freshly prepared 250 µL 5 mM BS^3 (21580, Thermo Fisher) in Conjugation Buffer, and incubating them for 30 minutes at RTº in a rotating wheel. Then, 12.5 µL of 1 M Tris pH 7.5 was added to the reaction, which was further incubated for 15 minutes at RTº in a rotating wheel to quench it. Cross-linked Ig-Dynabeads were then washed 3 times with 200

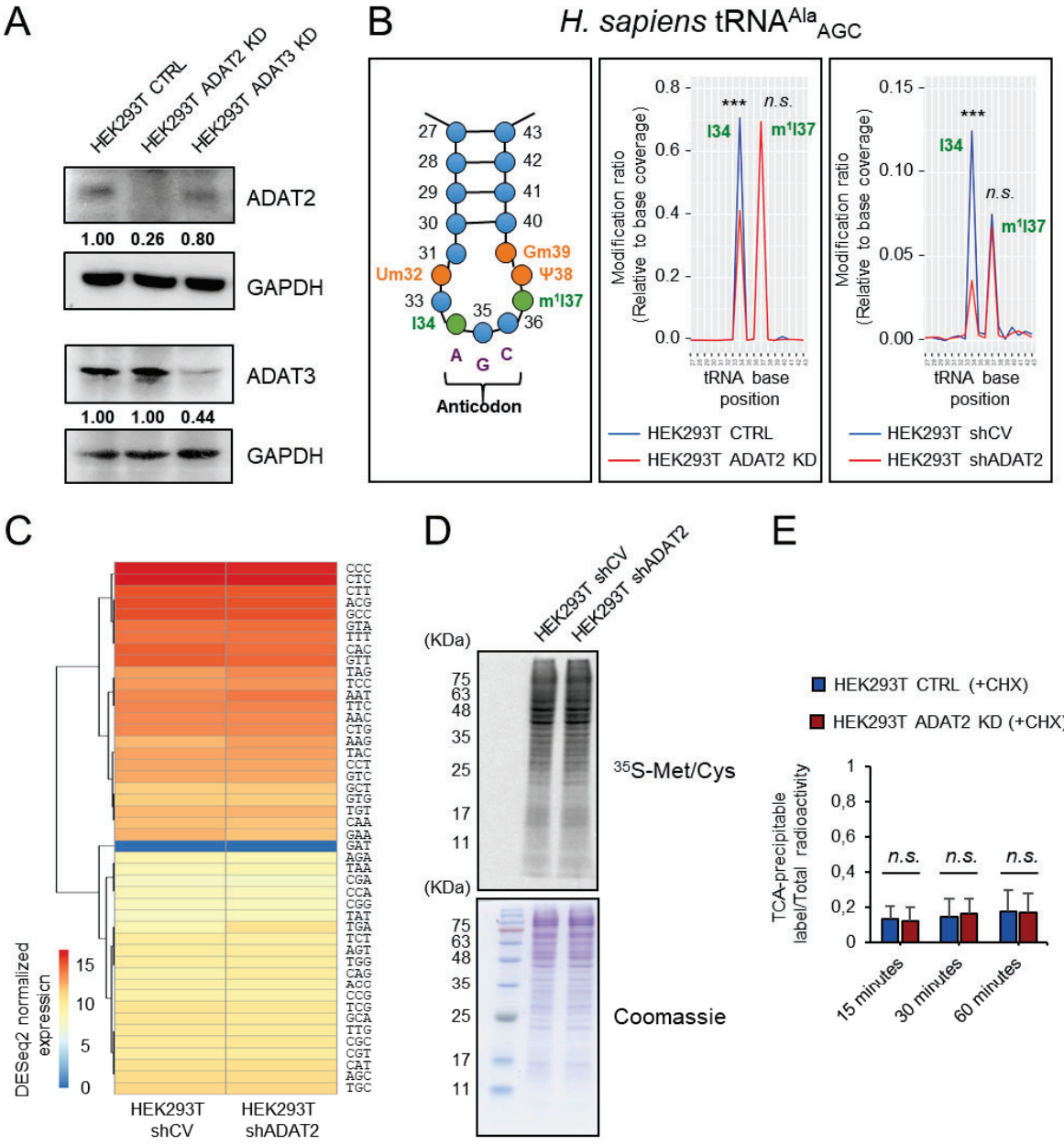
µL Wash Buffer and were left overnight incubating with pre-cleared cell lysate at 4 ºC in a rotating wheel. The next day, beads were recovered and washed twice with 500 µL Wash Buffer. Elution was performed by incubating beads with 40 µL 0.25 % trifluoroacetic acid (TFA) for 5 minutes at RTº. The solution was then neutralized with Tris pH 8. Eluted samples were mixed with Protein Loading Buffer (100 mM Tris pH 6.8, 4 % SDS, 0.1 % Bromophenol Blue, 20 % Glycerol, 100 mM DTT) and resolved in a 10 % PAGE that was further stained with BlueSafe. Protein bands corresponding to the SDC3-RLuc and SDC(G-end)-RLuc were gel purified and submitted to mass spectrometry analyses.

Purification of eGFP reporter proteins was performed essentially as described above but using Protein G sepharose beads (17-0618-01, VWR). Cells in lysis buffer were kept in ice and were disrupted by passing them through a 27G syringe five times. Cells were then centrifuged for 20 minutes at maximum speed at 4 ºC and the supernatant was recovered ("cell lysate"). 100 µL Protein G sepharose beads were equilibrated with 2 volumes of lysis buffer, centrifuged for 2 minutes at 700 xg at 4 ºC and re-suspended in lysis buffer for a final 50 % vol/vol. Cell lysate was incubated with 30 µL of equilibrated Protein G sepharose beads (eqProtG) in a rotating wheel for 2 hours at 4 ºC, and was then centrifuged for 20 minutes at 700 xg at 4 ºC. Supernatant was recovered ("pre-cleared cell lysate"). 50 µL eqProtG were incubated for 1 hour in a rotating wheel at RTº with 5 µg Green Fluorescent Protein antibody (DSHB-GFP-12A6, Developmental Studies Hybridoma Bank) in 200 µL lysis buffer ("Ig-eqProtG"). Ig-eqProtG were then incubated with pre-cleared cell lysate overnight at 4 ºC in a rotating wheel. The next day, samples were centrifuged for 10 minutes at 400 xg at 4 ºC, and beads were washed twice with 1 mL ice-cold lysis buffer. Elution was carried out by re-suspending the beads in 40 µL SDS sample buffer 1x (50 mM Tris pH 6.8, 2 % SDS, 0.1 % Bromophenol blue, 10 % glycerol) and incubating them for 15 minutes at 50 ºC. Samples were then centrifuged for 1 minute at 1000 xg at RTº and 10 % DTT was added to the eluted samples. Proteins were resolved in a 10 % PAGE, gel was stained with BlueSafe and bands corresponding to eGFP were gel purified and submitted to mass spectrometry analyses.

Supplementary Figures

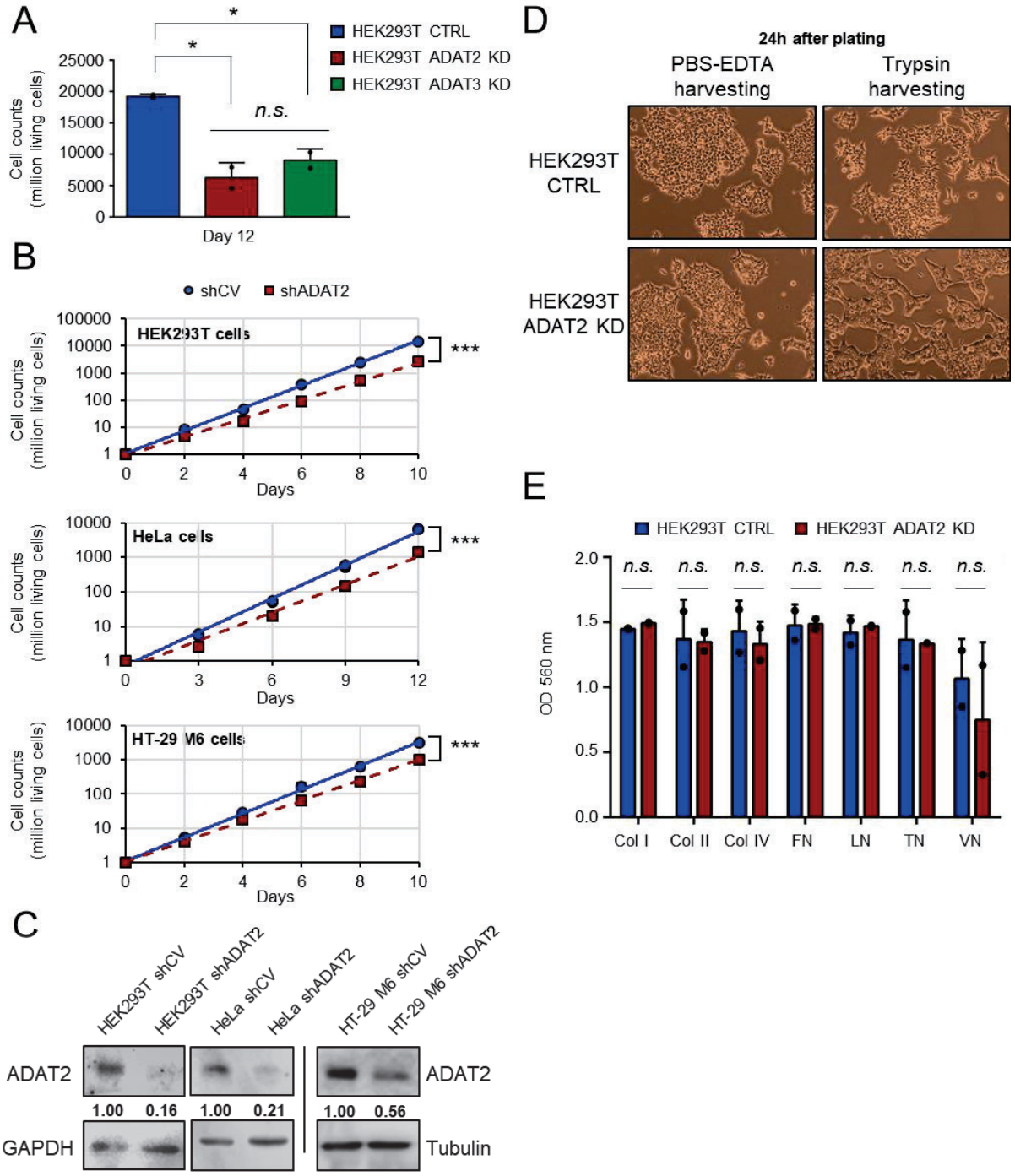


Supplementary Figure S1. (A) and (B) Upper panels: Sanger-sequencing spectra for amplified genes in HEK293T CTRL, and HEK293T ADAT2 KD (A) or HEK293T ADAT3 KD (B) cells. Centre panels: Sequence alignments of wild type and CRISPR/Cas9-edited alleles. Shaded sequence correspond to the section targeted by the guide strands. Lower panels: predicted translation of wild type and CRISPR/Cas9-edited alleles. Affected regions are highlighted in yellow. The deletion in the ADAT2 gene causes a frame-shift leading to a premature STOP codon. The deletion in the ADAT3 gene removes part of the deaminase domain.



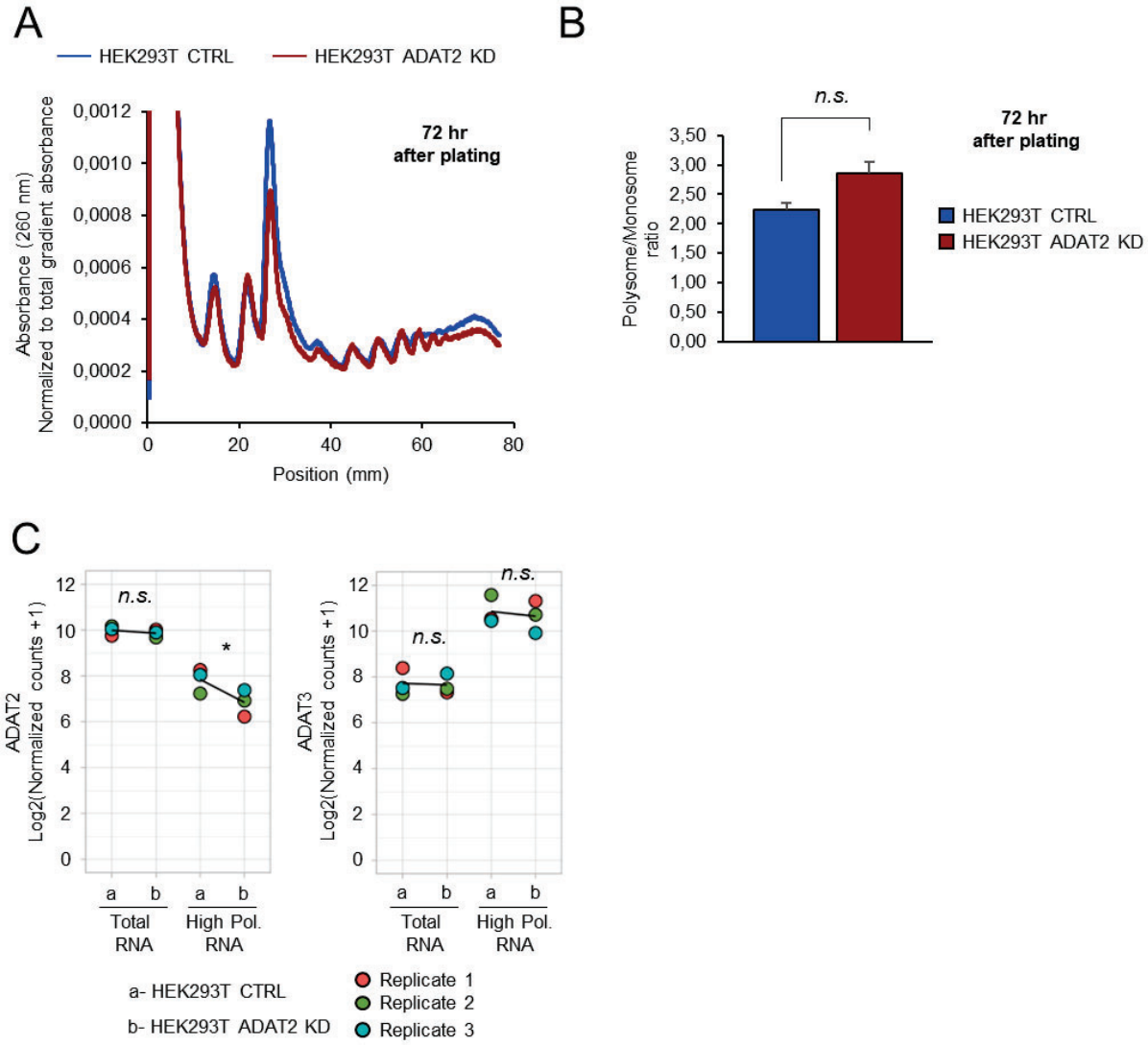
Supplementary Figure S2. (A) ADAT2 and ADAT3 protein levels evaluated by western blotting in the indicated cell lines. GAPDH is used as gel loading control. Quantification of gel bands relative to GAPDH and normalized to HEK293T CTRL cells are shown. (B) Left panel: schematic representation of the anticodon stem-loop structure of *Homo sapiens* $tRNA^{Ala}_{AGC}$ depicting known modified residues. Modifications detected (green) and not detected (orange) by tRNA-Seq are shown. Numbers indicate the tRNA base position. Anticodon position and sequence (AGC) is shown for reference (purple). The first anticodon position (A34) is deaminated to I34 (see also **Figure 1A**). Um (2′-O-methyluridine); I (Inosine); m^1I (1-methylinosine); Ψ (Pseudouridine); Gm (2′-O-methylguanosine). Centre and Right panels: modification ratio (proportion of sequencing mismatches relative to base sequencing coverage) observed in sequencing reads mapping to human $tRNA^{Ala}_{AGC}$ genes as seen by tRNA-Seq in the indicated cell lines. The expected tRNA modification at the indicated base position (I34 and m^1I37) is shown in green for reference (see also left panel). Centre panel: modification ratio in HEK293T CTRL

(blue) and HEK293T ADAT2 KD (red) cells. Right panel: modification ratio in HEK293T shCV (blue) and HEK293T shADAT2 (red) cells (see also (17)). *n.s.*: not statistically significant. ***: adj. p-val < 0.001 (Benjamini-Hochberg, Fisher Exact test). (C) Heatmap visualization of tRNA gene expression at isodecoder level (tRNAs with the same anticodon) in HEK293T shCV and shADAT2 cells as evaluated by tRNA-Seq. Colouring scale represents log2 DESeq2 normalized expression values based on two biological replicates calculated as in (51). No statistically significant differences were found (Benjamini-Hochberg). (D) Pulse-chase analyses on HEK293T shCV and shADAT2 cells. Coomassie staining of the same gel is used as loading control. (E) Quantitative evaluation of ^{35}S-Met/Cys incorporation into proteins in the presence of cycloheximide (CHX) upon monitoring the ratio between the total radioactivity (cpm) of the cell lysate and the amount of radioactivity (cpm) in TCA-precipitable fractions. Shown are the mean and standard deviations from biological triplicates. *n.s.*: not statistically significant (t-test). See also **Figure 3C**.

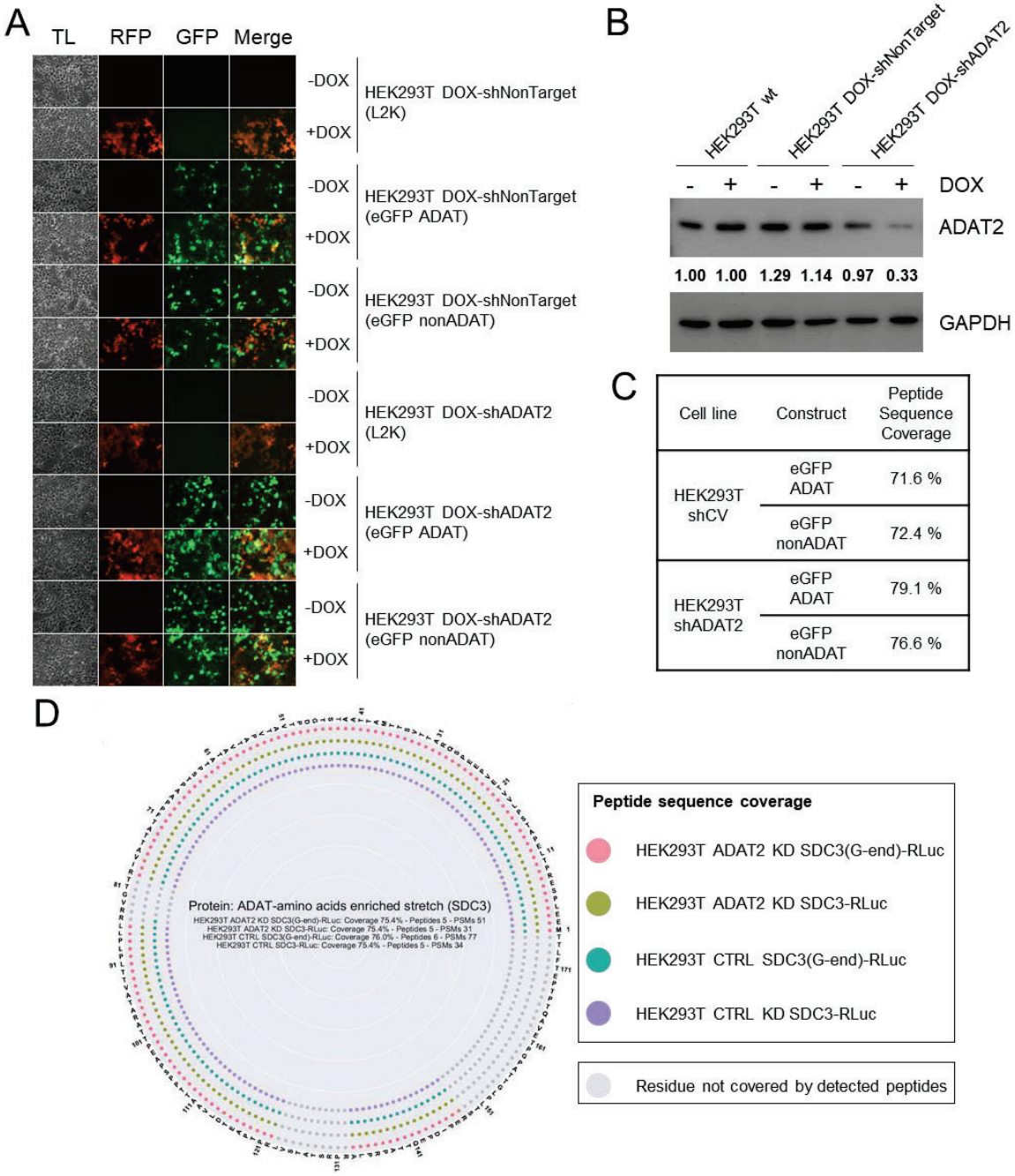


Supplementary Figure S3. (A) Comparative analysis of cell growth at Day 12, represented as total cell counts, for the indicated cell lines. Shown are biological duplicates, their mean and SD. *n.s.*: not statistically significant. *: p-val < 0.05 (t-test). (B) Growth curves of HEK293T, HeLa and HT-29 M6 shCV (blue) and shADAT2 (red) cells represented as total counts of living cells over time. Y-axis is set in logarithmic scale and an exponential trendline was fit to the data points. Data points correspond to the mean and SD of three biological replicates. ***: p-val < 0.001 (t-test). (C) ADAT2 protein levels evaluated by western blotting in the indicated cell lines. GAPDH or Tubulin is used as gel loading control. Quantification of ADAT2 bands relative to GAPDH or Tubulin and normalized to shCV cells is shown (D) Representative light microscopy images of HEK293T CTRL and ADAT2 KD cells harvested

with PBS-EDTA (left panels) or Trypsin (right panels) at 24 hours after plating them in clean culture plates. (E) Evaluation of cell adhesion capacity to components of the extracellular matrix when cells are collected for this analysis with trypsin instead of PBS-EDTA (cell adhesion profiles are expected to be similar for both cell lines because newly synthesised adhesion proteins are removed using trypsin). Col I, II and IV: collagen I, II and IV respectively. FN: fibronectin. LN: laminin. TN: tenascin. VN: vitronectin. Data shows the mean and SD of the obtained absorbance at 560 nm (OD 560 nm) for HEK293T CTRL (blue) and ADAT2 KD (red) cells. Experiments were done in biological duplicates. *n.s.*: not statistically significant (t-test).

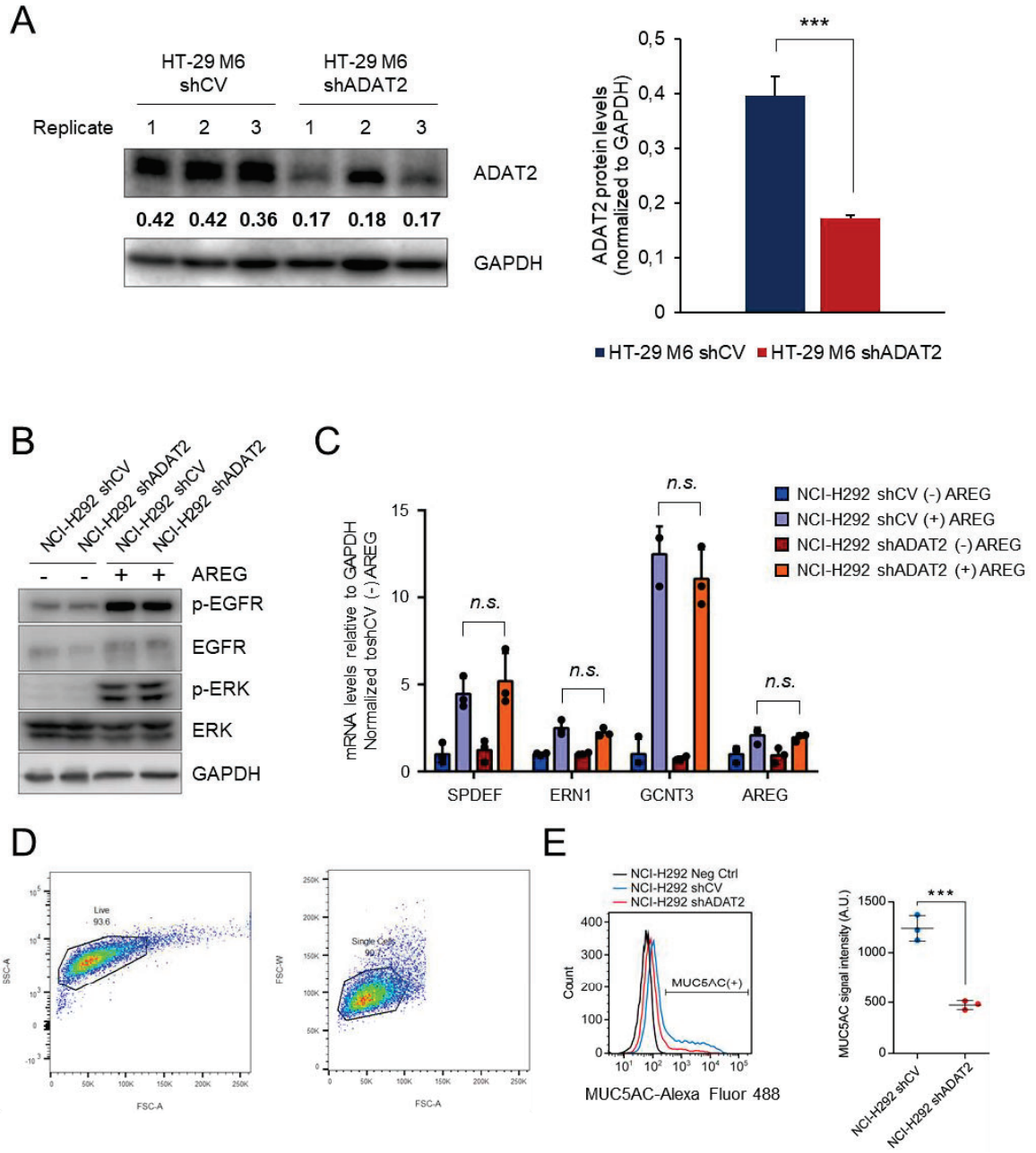


Supplementary Figure S4. (A) Representative polysome profile of trypsin-treated HEK293T CTRL (blue) and ADAT2 KD (red) cells at 72 hours after plating. Experiments were done in biological triplicates. (B) Polysome to Monosome ratio (P/M) obtained from experiments as in (A). Shown are the mean and standard deviations of biological triplicates. *n.s.*: not statistically significant (t-test). (C) Quantification of ADAT2 (left panel) and ADAT3 (right panel) transcript abundance (sequencing reads: log2(normalized counts + 1)) in Total RNA and the HP fractions on the indicated cell lines. Dots represent the obtained quantification for each biological triplicate. The line connects the means of the compared groups. *n.s.*: not statistically significant. *: adj. p-val < 0.1 (DESeq2, Benjamini-Hochberg).



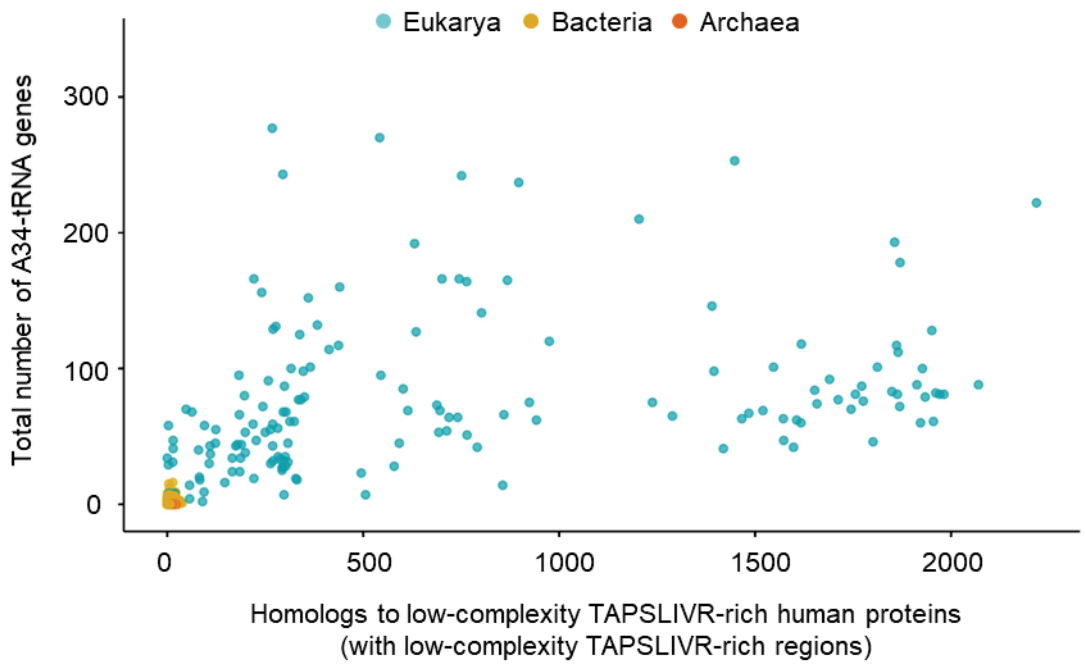
Supplementary Figure S5. (A) Transmittance light (TL) and epifluorescence (Red Fluorescent Protein, RFP, and Green Fluorescent Protein, GFP) microscopy images of doxycycline (DOX)-inducible ADAT KD cell lines transfected with eGFP ADAT or eGFP-nonADAT constructs. Addition of DOX (+) activates the generation of the inducible shRNA coupled to production of RFP. (B) ADAT2 protein levels evaluated by western blotting in the indicated cell lines in the presence (+) or absence (-) of doxycycline (DOX). GAPDH is used as gel loading control. Quantification of ADAT2 bands relative to GAPDH and normalized to WT cells with DOX (for (+) DOX treatments) or without DOX (for (-) DOX treatments) is shown. (C) Percentage of eGFP peptide sequence coverage as detected by mass spectrometry of purified eGFPs obtained upon expression of eGFP ADAT or eGFP nonADAT constructs in the indicated cell lines. (D) Evaluation of mistranslation of the low-complexity TAPSLIVR-rich region of SDC3 by *de*

novo protein sequencing of purified SDC3-RLuc or SDC3(G-end)-RLuc expressed in HEK293T CTRL or ADAT2 KD cells. The wheel represents the full sequence of the TAPSLIVR-rich region of SDC3. Amino acids covered by the obtained peptides by mass spectrometry in each sample are color-coded. Grey circles are undetected residues. Percentages of sequence coverage for each sample are shown.



Supplementary Figure S6. (A) ADAT2 protein levels in extracts depicted in **Figure 9B**. Quantification of ADAT2 bands relative to GAPDH is shown. Bar plot represents the mean and SD obtained from quantifying the shown blots (biological triplicates). ***: p-val < 0.001 (t-test) (B) Evaluation by western blotting of the expression of markers of AREG stimulation in the indicated cell lines treated (+) or not (-) with AREG. GAPDH is used as gel loading control. (C) Real time qPCR of markers of AREG stimulation relative to GAPDH in the indicated cell lines treated (+) or not (-) with AREG. Data is normalized to untreated NCI-H292 shCV cells for each target gene. Shown are biological triplicates, their mean and SD. Statistical significance is shown only for the comparisons between NCI-H292 shCV and shADAT2 in the presence of AREG (both cell lines show similar transcriptional activation of the evaluated targets upon AREG treatments). *n.s.*: not statistically significant (t-test). (D) Flow cytometry analysis of the experiments depicted in **Figure 9G** showing representative pseudocolor plots exemplifying the gating strategy used, and including both MUC5AC(+) and MUC5AC(-) cells. (E) Left

panel: NCI-H292 negative control cells (black line; cells incubated only with the secondary antibody) was used to set up the MUC5AC(+) gating (shown in figure). NCI-H292 shCV cells shown in blue and NCI-H292 shADAT2 cells shown in red. Right panel: quantification of MUC5AC signal intensity as shown in **Figure 9G** but including the signal obtained for the whole live/single cell suspension (i.e. without MUC5AC(+) gating).



Supplementary Figure S7. Distribution of species from Eukarya (turquoise), Bacteria (yellow) and Archaea (dark orange) based on the total number of A34-tRNA genes and the number of homologs to low-complexity TAPSLIVR-rich human proteins present in their genomes. Note that the data on eukaryotic species is the same one depicted on **Figure 10C,** and is used here for reference.

Legends for Supplementary Tables

Supplementary Table S1. Oligonucleotides used in this study.

Supplementary Table S2. Differential tRNA gene expression analyses in HEK293T CTRL, HEK293T ADAT2 KD, HEK293T shCV, and HEK293T shADAT2 cell lines.

Supplementary Table S3. Proteomic analyses on HEK293T CTRL and HEK293T ADAT2 KD (iTRAQ). Table with fold changes, p-values, mean estimates and confidence intervals.

Supplementary Table S4. Differential transcript expression in Total RNA and RNA from High Polysome fractions between HEK293T CTRL and HEK293T ADAT2 KD cells; GSEA (GO Biological Process); Interaction analysis and Fisher Exact Test for proportion of transcripts encoding low-complexity TAPSLIVR-rich regions with impaired translation upon ADAT2 KD.

Supplementary Table S5. Identified human proteins containing low-complexity TAPSLIVR-rich regions, and GO analyses obtained by DAVID (59).

Supplementary Table S6. Abundance of A34-tRNA genes and number of low-complexity TAPSLIVR-rich human proteins with homologs in eukaryotic and prokaryotic species. Highlighted eukaryotic species contain an unusually large (> 400) number of A34-tRNA genes and were removed from the analyses (See also *Materials and methods*).

Supplementary Table S7. Abundance of homologous sequences to human low-complexity TAPSLIVR-rich proteins in unicellular and multicellular organisms.

3.3

Genomic organization, transcription, and somatic mutagenesis of tRNA genes: Implications for proteome integrity

Genomic organization, transcription, and somatic mutagenesis of tRNA genes: Implications for proteome integrity

Marina Murillo-Recio[1], Marina Salvadores[1,2], Aina Vaquer-Picó[1], Lina Tsapanou[1], Adrián Gabriel Torres[1] , Fran Supek[1,2,3*], Lluís Ribas de Pouplana[1,3*]

1. Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Catalonia, 08028, Spain.

2. Biotech Research & Innovation Centre (BRIC), University of Copenhagen, 2200 Copenhagen N, Denmark.

3. Catalan Institution for Research and Advanced Studies, Barcelona, Catalonia, 08010, Spain.

* To whom correspondence may be addressed.

Abstract:

Transfer RNA genes (tRNAs) are essential non-coding RNAs that safeguard translational fidelity. tRNA genes (tDNAs) form a unique class of hyper-abundant essential loci in most organisms. In this study, we mapped and characterized tDNA distribution in the gapless human reference genome (T2T-CHM13) and analyzed their individual transcriptional activities. Our analysis revealed that tDNA clustering may positively influence tDNA transcriptional activity. Then we have characterized the impact of somatic mutations in human tDNAs and its relationship to the transcriptional status of each gene. We confirm that tDNAs are hotspots for somatic mutagenesis and show that they display mutational loads that are directly proportional to their transcription rates. We demonstrated that tDNAs are hotspots for somatic mutagenesis in both tumor and healthy tissues. Mutational loads at tDNAs are tumor-specific and increase with patient age. Analysis of mutational signatures identified APOBEC activity as the main contributor to tDNA somatic mutagenesis. Mutations at structurally conserved tRNA positions appear to be under negative selection. Strikingly, other hypermutated positions could lead to alterations in tRNA biogenesis and impair tRNA functions, which could alter translation by systematically introducing amino acid substitutions across the proteome. Our results reveal a previously unrecognized source of somatic heterogeneity in human cancer and aging tissues that may directly impact translation efficiency and fidelity and cause cell-specific proteostasis degeneration.

Introduction

The translational machinery of cells is finely tuned to maintain proteostasis, a key factor in cellular survival and proliferation (Gingold et al., 2014; Goodarzi et al., 2016; Torrent et al., 2018). Transfer RNAs (tRNAs) are central to translation, pairing their anticodons to codons in messenger RNAs and delivering the corresponding amino acids during translation. The human genome encodes over 600 tDNAs (Chan & Lowe, 2016), which are transcribed by RNA polymerase III (Pol III) via the recognition of two internal promoter regions that are highly conserved among all tDNAs **(Fig. 1)** (R. Giegé et al., 1998). It is now well established that, despite containing suitable promoter structures, not all tDNAs are equally expressed, and even many tDNAs are constitutively silent (Torres et al., 2019). Moreover, identified cis-elements and trans-factors regulating Pol III-dependent tDNA transcription are insufficient to fully explain the cell type- or tissue-specific expression of individual tDNAs (Hummel et al., 2019; Ishimura et al., 2014; Torres et al., 2019). Indeed, there is evidence suggesting a link between tDNA expression regulation and their genomic distribution (Van Bortle et al., 2017), supporting the idea that tDNA proximity and clustering of active tDNAs may enhance Pol III-transcription (Gao et al., 2024).

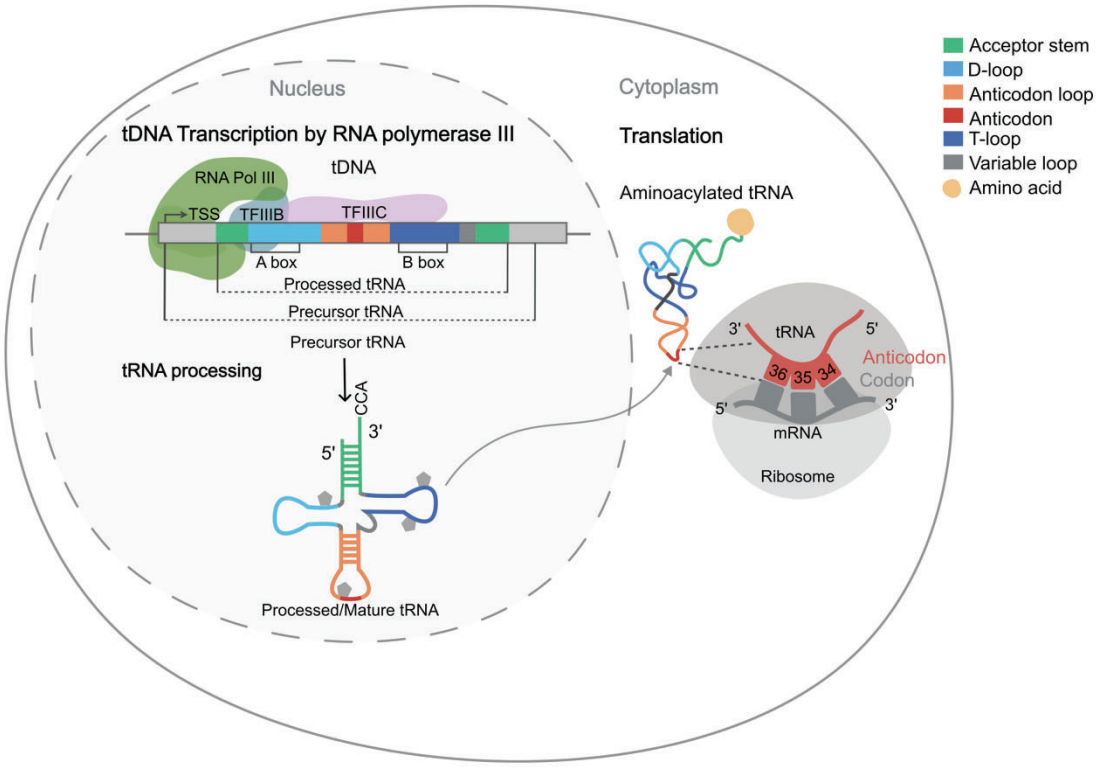


Figure 1. Schematic representation of tDNA biology. Including tDNA transcription by Pol III, tRNA processing, tRNA structure and tRNA canonical function in translation. The diagram illustrates the conserved internal promoter regions (A box and B box), which are essential for RNA Polymerase III-mediated transcription. These regions facilitate the recruitment of transcription factors TFIIIB and TFIIIC. The resulting precursor tRNA undergoes several processing steps, including the addition of a 3′ CCA tail required for amino acid attachment by aminoacyl-tRNA synthetases, along with post-transcriptional chemical modifications. In the cytoplasm, mature tRNA participates in translation by base-pairing the

anticodon (positions 34 to 36) with complementary mRNA codons and delivering the correct amino acid to the ribosome for protein synthesis.

Alterations in tRNA populations are linked to numerous human diseases (Orellana et al., 2022). For example, quantitative and qualitative alterations in tRNA populations, including tRNA-derived small RNAs (tsRNAs), are widespread in human cancers (Cabrelle et al., 2024; Huang et al., 2018), which adapt their tRNA expression to optimize the translation of specific genetic programs (Goodarzi et al., 2016; Gupta et al., 2022; Pinzaru & Tavazoie, 2023; Zhang et al., 2018). Specific tRNA isoacceptors and post-transcriptional base modifications regulate cancer cell survival and influence metastatic potential by controlling the translation of proliferation-related genes (Earnest-Noble et al., 2022; García-Vílchez et al., 2023). Stress conditions can disrupt components of the translation machinery, including tRNAs, resulting in errors during protein synthesis (mistranslation) (Mohler & Ibba, 2017; Ribas de Pouplana et al., 2014; Schuntermann et al., 2024). For example, metabolic and therapeutic stresses, such as chemotherapy, induce codon-biased aberrant protein production through altered tRNA function and decoding. Thus, variations in tRNA populations reshape the cancer proteome and support tumor adaptation and therapy resistance (Kochavi et al., 2024; Wernaart et al., 2024; Yang et al., 2024).

Somatic mutagenesis causes changes in the DNA sequence of somatic cells, drives the emergence of cancers, and is an important contributor to the aging process (Vijg & Dong, 2020). The nature and dynamics of human somatic mutagenesis has been studied mostly in protein-coding genes, and its impact upon non-coding RNA genes is less understood. The activity of members of the APOBEC3 subfamily of cytidine deaminase enzymes is an important contributor to somatic mutagenesis (Langenbucher et al., 2021; Mas-Ponte & Supek, 2020; Sanchez et al., 2024; Taylor et al., 2013). APOBEC3 enzymes function, primarily, as an innate immunity mechanism to defend against viruses and mobile genetic elements but, in some cancers, APOBEC3A and APOBEC3B paralogs can generate a high mutational burden (Burns et al., 2013; Petljak et al., 2022; Roberts et al., 2013; Supek & Lehner, 2017). APOBECs can access only single-stranded nucleic acids and are known to act upon ssDNA intermediates during DNA repair, as well as ssDNA segments within structured, stem-loop DNA (Buisson et al., 2019; Mas-Ponte & Supek, 2020; Roberts et al., 2013).

Studies in bacteria and yeast have reported that non-coding genes can accumulate mutations also caused by the activity of APOBEC3 enzymes. In an *E. coli* model, mutational impact of heterologous expressed APOBEC3A is highest in genes transcribed by Pol III, such as tDNAs (Sakhtemani et al., 2019). Mutational studies in a yeast model also suggested that human APOBEC3B overexpression causes severe DNA damage in Pol III-transcribed genes (Saini et al., 2017), suggesting a transcription-related mechanism. In human germline cells, mutations in nuclear tDNAs are subject to strong purifying selection (Thornlow et al., 2018), but analysis of tRNA sequences obtained from the plasma of healthy human volunteers revealed significant levels of human heterogeneity (Berg et al., 2019; Parisien et al., 2013). Human tDNAs are known to accumulate somatic mutations at high rates, and APOBEC has been identified as a major contributor to this phenomenon (Sakhtemani et al., 2019). However, a detailed

analysis of tDNA somatic mutations, including their distribution at single-base resolution and their potential physiological impact is lacking.

Therefore, to elucidate the multifaceted roles of tRNAs in the pathogenesis of human diseases, further research into tDNA genomics and transcriptomic dynamics is needed. This includes characterizing genomic localization and organization, transcriptional regulation, and vulnerability to somatic mutagenesis. Here we first mapped all tDNAs in the latest human genome assembly, T2T-CHM13 (Nurk et al., 2022), and used this information to study the relationship between tDNA distribution and tDNA expression. As previously reported, tDNAs can be found as isolated genes or as part of large gene clusters (Bermudez-Santana et al., 2010; Van Bortle et al., 2017). Combining our data with POLR3D ChIP-seq and 'biotin-capture of nascent RNA' data, we find that both isolated and clustered tRNAs can be actively transcribed, but tDNAs in clusters exhibit higher activity levels than isolated genes. We additionally found that tDNA cluster composition can result from tandem repetitions of specific sets of tDNAs, pointing to a potential tDNA cluster that may be subjected to tDNA copy number variation (tgCNV) between individuals.

We then characterized in detail the somatic mutations affecting tDNAs in human tissues, which are prevalent both in tumors and in healthy tissues. Indeed, tDNAs constitute hotspots of somatic mutagenesis whose intensity is directly linked to the transcriptional activity of each gene. Mutational load at tDNAs accumulate at rates up to nine-fold higher than in protein coding genes and are higher than in other genes transcribed by Pol III. Mutational loads at tDNAs vary greatly between tumor types, display mutational signatures that are different in cancer or in healthy tissues, and readily accumulate with age. Nucleotides important for tRNA structure accumulate mutations at rates lower than expected, but all three anticodon positions have average mutational frequencies. Somatic mutations at anticodon bases at highly transcribed tDNAs may result in mutant chimeric tRNAs capable of introducing specific and ubiquitous amino acid substitutions throughout the proteome.

Results

Genome-wide identification of tDNAs in the human genome

Genome assemblies previously used to identify tDNAs, such as hg19 and hg38, were constructed using short-read sequencing technologies, which fail to resolve repetitive regions (Hoyt et al., 2022). However, the telomere-to-telomere (T2T) assembly (Nurk et al., 2022), was generated using long-read sequencing, provides a complete and gapless representation of the human genome and reveals genomic regions that were incomplete. To obtain an accurate picture of the set of human tDNAs, we first mapped the positions of these genes in the most recent version of the human genome assembly generated with long-read sequencing (T2T-CHM13 v2.0/hs1) (Hoyt et al., 2022) (**Fig. 2a**). Our analysis sets the total number of identified tDNAs by tRNAscan-SE to 733, of which 521 were confidently predicted as functional tRNAs (~112 more high-confidence genes than previously reported with older assemblies) (**Fig. 2b**) (Chan et al., 2021). We successfully localized all known tDNAs and removed previous uncertainties regarding the localization of some of these genes (**Fig. 2b**).

Human tDNAs frequently group in specific genomic locations (Bermudez-Santana et al., 2010; Mungall et al., 2003; Van Bortle et al., 2017) (**Fig. 2a**). In the T2T-CHM13 assembly, two regions with the highest concentration of tDNAs were located on chromosome 1 (Chr 1) and Chr 6, which together account for more than 60% of all human tDNAs: 34.1% (250/733 tDNAs) and 25.38% (186/733 tDNAs), respectively (**Fig. 2a and Supp. Fig. 1a**). Notably, Chr 1 harbors the highest number of tDNAs in the T2T-CHM13 assembly, whereas in older assemblies Chr 6 had the highest tDNA count (Mungall et al., 2003), which is consistent with our analysis of the hg38 and hg19 assemblies (e.g., in hg38, Chr 6 has 187 tDNAs compared to 147 on Chr 1) **(Supp. Fig. 1a)**. These discrepancies suggest that previous assemblies likely underestimated the number of tDNAs on Chr 1. We also observed a decrease in tDNA counts on Chr X and Chr 6 in the T2T-CHM13 assembly compared to older assemblies. For other chromosomes, the number of tDNAs remained consistent across assemblies**.** Interestingly, some chromosomes are still nearly devoid of tDNAs, such as Chr 4, Chr 18, Chr 20, Chr 21, Chr 22 and Chr Y **(Fig. 2a and Supp. Fig. 1a).**

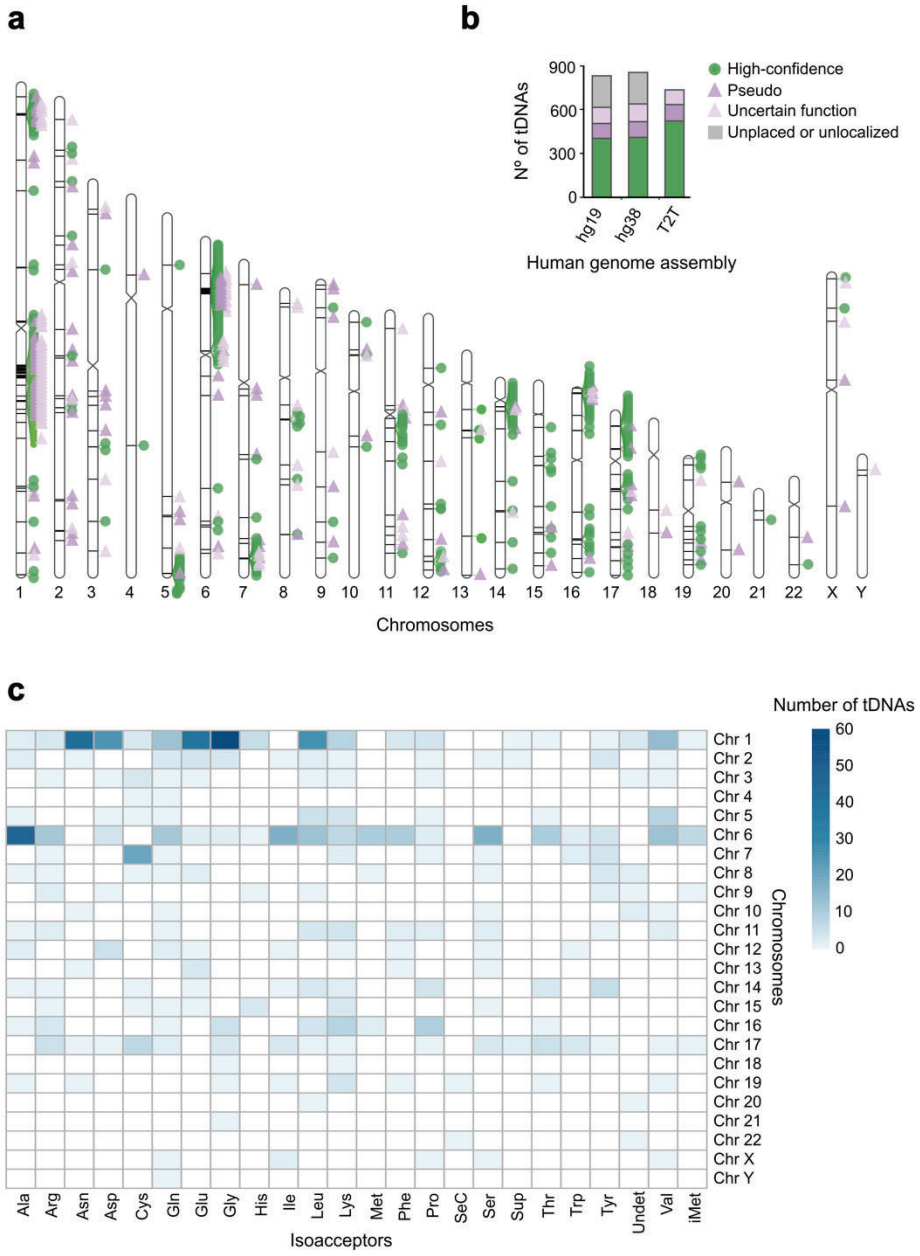


Figure 2. Localization of tDNAs in the genome. (a) Ideogram showing the chromosomal distribution and localization of human tDNAs in the T2T-CHM13 genome assembly, including tRNA-scan-SE 2.0 prediction classification (high-confidence, pseudogenes, and tDNAs with uncertain function). Black spots within the chromosomes indicate tDNA localization. **(b)** Quantification and classification of tDNAs across different human genome assemblies (hg19, hg38, T2T) based on analyses conducted using tRNAscan-SE 2.0. tDNAs identified within the unplaced or unlocalized sequences of the reference genome are shown in grey. In the most recent human genome assembly (Jan. 2022, T2T-CHM13 v2.0/hs1), a total of 733 tDNAs were identified, including 521 high-confidence tDNAs, 112 pseudo-tDNAs, and 100 tDNAs with uncertain functions. The Dec. 2013 assembly (GRCh38/hg38) reported 637 localized tDNAs, comprising 409 high-confidence, 107 pseudo, and 121 uncertain function tDNAs, along with 215 unlocalized tDNAs. The Feb. 2009 assembly (GRCh37/hg19) included 614 localized tDNAs, with 402 high-confidence, 102 pseudo, 110 uncertain function, and 215 unlocalized tDNAs. **(c)** Distribution and number of tDNAs across different chromosomes in the T2T-CHM13 genome for each isoacceptor.

Next, we examined the tDNA distribution by isoacceptor groups (set of tRNAs that carry the same amino acid). When comparing different genome assemblies, we found that the number of detected tDNAs for Gly, Leu, Glu, and Asp was higher in the T2T-CHM13 assembly than in hg38 and hg19 **(Supp. Fig. 1b)**. Focusing specifically on the T2T-CHM13 genome, we detected that for some isoacceptor families, tDNAs are often enriched in one specific chromosome **(Fig. 2c and Supp. Fig. 2)**. For example, the Gly isoacceptor family is the largest one with 75 tDNAs, and ~80 % of them localize at Chr 1, then we have Leu with 59 tDNAs, with 44 % in Chr 1 and 20% in Chr 6. Notably, Cys isoacceptors were predominantly concentrated on Chr 7q (the long arm of Chr 7). Other isoacceptor families are as well strongly represented at the high tDNA density regions 6p (the short arm of Chr 6) and 1q (the long arm of Chr 1), but several of their tDNA members further group at additional chromosomal sites. For example, Pro and Lys tDNAs are enriched at 16p, Leu and Val tDNAs populate the 5p region, and Tyr and Asp tDNAs are the strongest represented at 14p and 12p, respectively (**Supp. Fig. 2**). As expected, we do find a positive correlation between the amount of tDNAs within an isoacceptor family and the presence of tDNA members of the family in different chromosomes (Pearson correlation coefficient = 0.61; *P*-value = 0.0026, **Supp. Fig. 3).**

tDNAs genomic organization

tDNAs can either accumulate in specific chromosomal regions, organized in clusters, or be more dispersed across the genome (Bermudez-Santana et al., 2010; Van Bortle et al., 2017). To characterize the clusters distribution in the T2T-CHM13 genome, we first evaluated the genomic distance between consecutive tDNAs and plotted their cumulative distribution **(Supp. Fig. 4a)**. Inflection points on the curve were observed when tDNAs are separated by a maximum of 1 kilobase (kb) and 20 kb. A previous study compared the genomic organization of tDNA pairs for several species and concluded that it is not expected to have tDNA pairs when the distance between tDNAs is less than 1 kb. The authors thus defined two adjacent tDNAs to be clustered if their distance was within this limit (Bermudez-Santana et al., 2010). Another study defined adjacent tDNAs located within a 20 kb distance from each other (Van Bortle et al., 2017). These two tDNA cluster definitions are in agreement with our cumulative distribution profiles **(Supp. Fig. 4a).** Therefore, we decided to perform identical analyses in parallel, defining a tDNA cluster when adjacent tDNAs are within a 1 kb ("1 kb cluster") or a 20 kb ("20 kb cluster") distance **(Fig. 3a).** Note that this terminology refers to the maximum distance between two adjacent tDNAs and not the total size in kb of the cluster (e.g., three tDNAs separated each by 1 kb will form one "1 kb cluster" with a total size of ~3 kb). For each analysis, tDNAs that did not fall within a cluster were identified as isolated tDNAs.

Notably, the number of tDNAs within each cluster varied considerably: 1 kb clusters typically contain 2-5 tDNA genes, while most of the 20 kb clusters also fell within the 2-5 tDNA range, they contained larger clusters composed of 23, 29, or even 112 tDNAs **(Fig. 3b)**. Notably, the cluster of 112 tDNAs was previously reported in older assemblies but with a smaller number of genes **(Fig. 4b)**.

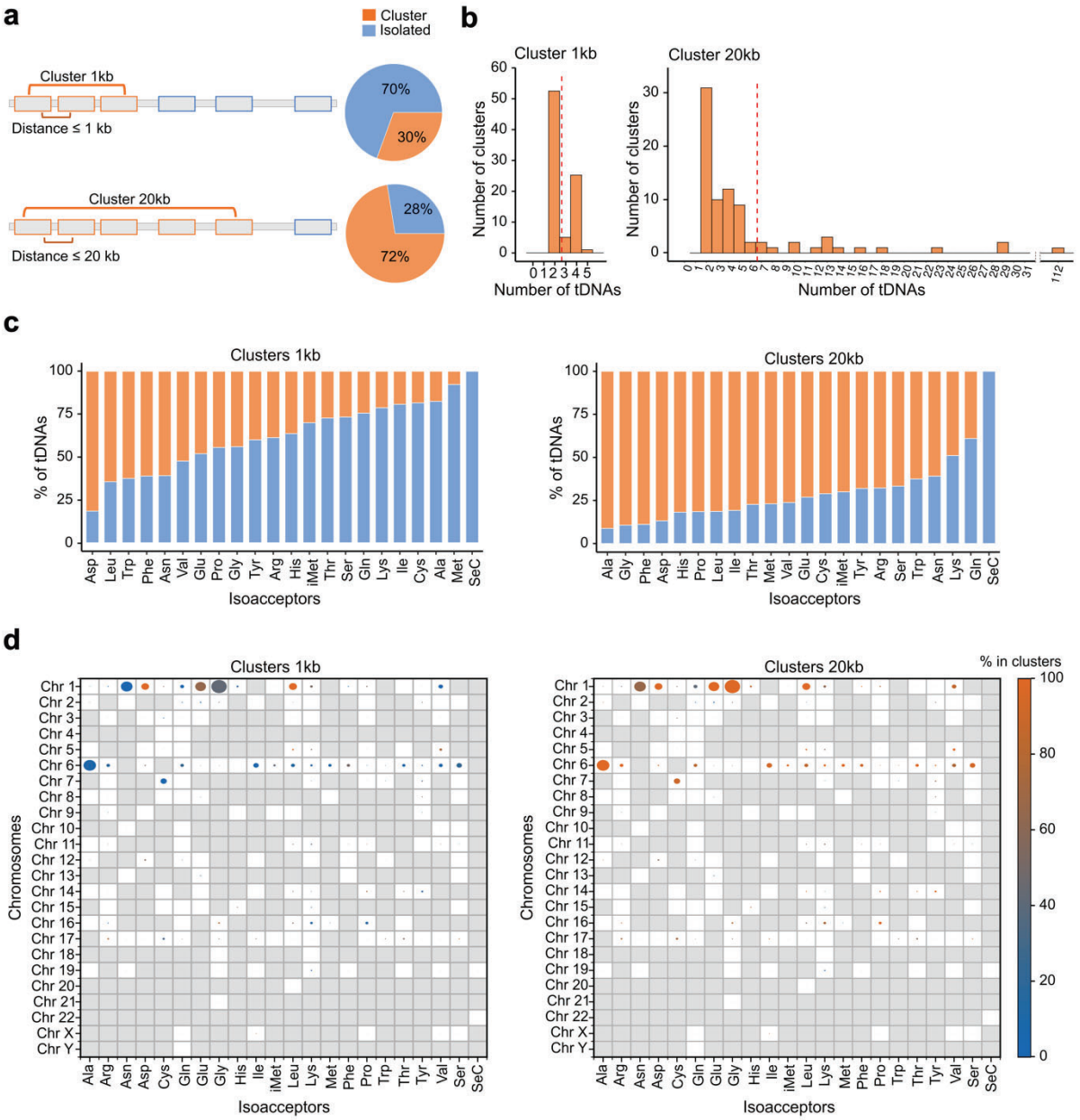


Figure 3. Clusters definition and characterization. (a) Definition of tDNA clusters based on two distance thresholds: 1 and 20 kb. tDNAs that did not fall within these thresholds were classified as isolated. The pie charts for each cluster definition illustrate the percentage of tDNAs classified within clusters and the percentage of isolated tDNAs **(b)** Distribution of tDNA clusters categorized by the number of tDNAs in each cluster. The red line indicates the average number of tDNAs in each cluster size. **(c)** Percentage of tDNA in clusters (orange) and tDNAs isolated (blue) by isoacceptor. **(d)** Chromosomal distribution of tDNAs classified by isoacceptor. For each isoacceptor, the percentage of tDNAs found in clusters (not necessarily the same cluster) is indicated.

When examining the length of these clusters, we observed, as expected, a correlation between cluster length and the number of tDNAs (Spearman correlation coefficient = 0.84, P-value < 2.2×10−16 for both 1 kb and 20 kb distances) **(Supp. Fig. 4b)**. However, in some cases, we observed variations in tDNA density, indicating that certain regions of the genome have a higher concentration of tDNAs packed into

a smaller space. For example, the cluster containing 112 tDNAs is shorter in length than the clusters with 23 or 29 tDNAs, indicating a much higher tDNA density within a more compact genomic region **(Supp. Fig. 4b)**. Interestingly, there are some isoacceptors that are found mostly in clusters **(Fig. 3c)**, while some others are more isolated, in the case of SeC, which is always found to be isolated.

Some of these clusters can be heterogeneous (composed by one isoacceptor type) or homogeneous (composed by more than one isoacceptor type). The proportion of homogeneous clusters is higher among 20 kb clusters (21%) than among 1 kb clusters (16%) **(Supp. Fig. 4c)**. By randomization the chance to have tDNA clusters homogeneous is lower than what we observe **(Supp. Fig. 4d)**. This provides insights into tDNA evolution and the generation of tDNA copies since the presence of homogeneous tDNA clusters could be tandem duplications, where copies or duplications of the same gene are placed in close proximity. Over time, evolutionary changes in one of the gene copies may result in slight modifications, leading to clustered regions with either identical or similar tDNAs.

Upon closer examination of the cluster in the T2T-CHM13 genome, we detected some clusters that seemed to be composed of a repetition of a set of genes. These events often involve duplication, inversion, and reinsertion *in cis* (occurring nearby on the same chromosome) **(Fig. 4a)**.

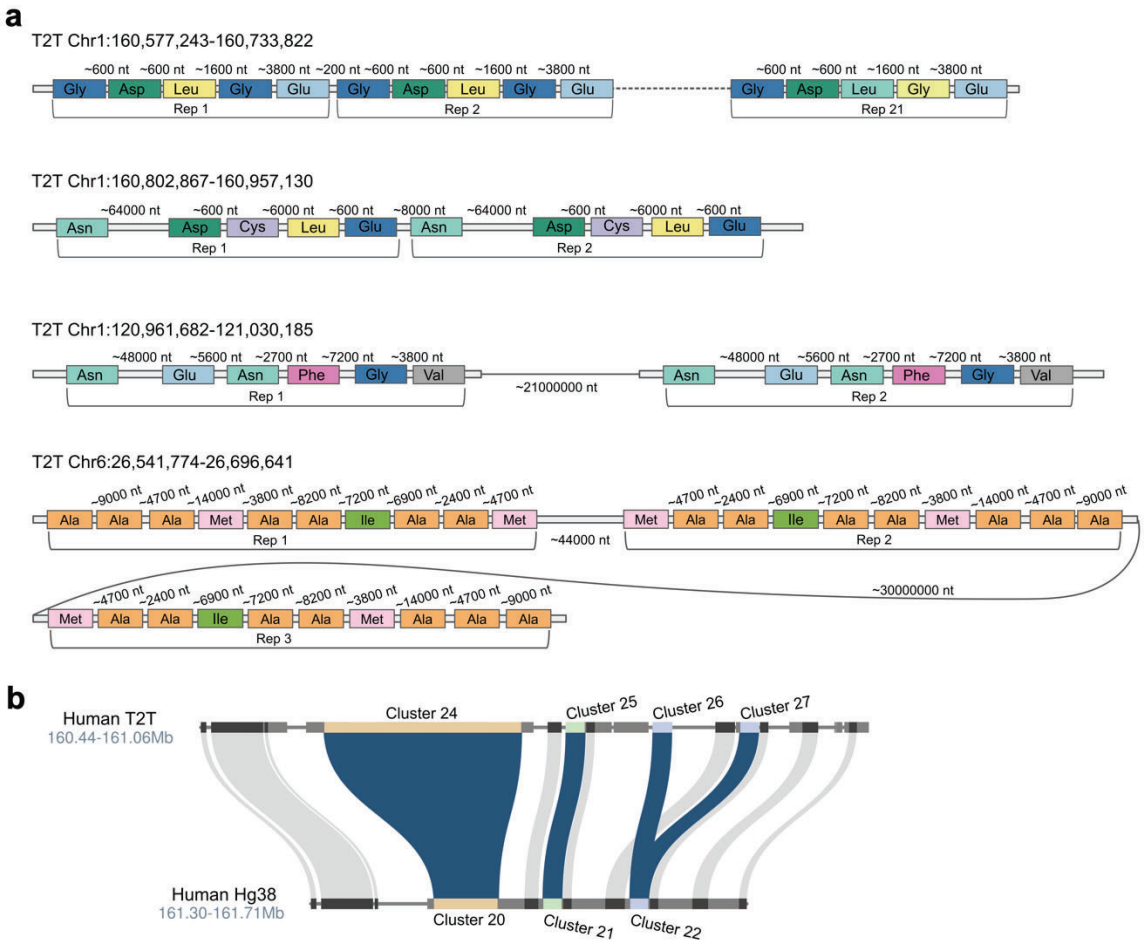


Figure 4. Tandem repetitions in tDNA clusters. (a) Representation of the tandem repetitions found on Chr 1 and 6. The first tandem repetition represented in the image, Chr1:160,577,243-160,733,822, is found in the large cluster of 112 tDNAs described previously. Consists of 21 tandem repetitions, each

containing the tDNA set Gly-TCC, Asp-GTC, Leu-CAG, Gly-GCC, and Glu-CTC. Each repeat unit is separated by only ~200 nucleotides, and the genomic distances between the tDNAs are conserved across repeats, indicating that the entire gene array was duplicated as a unit. We also found other examples of tandem duplications in Chr 1, such as in Chr1:160,802,867-160,957,130, where we detected two tandem repeats comprising the gene set Asn-GTT, Asp-GTC, Cys-GCA, Leu-CAA, and Glu-TCC. These repeats were separated by approximately 8 kb, representing tandem repeats with short intergenic spacing. Additionally, on Chr 1, the tDNA set Asn-GTT, Glu-TTC, Asn-GTT, Phe-GAA, Gly-CCC, and Val-CAC is located at Chr1:120,961,682-121,030,185 and again at Chr1:142,808,043-142,876,098, with the two occurrences separated by approximately 21,000 kb. Furthermore, another tandem repeat composed of ten tDNAs, Ala-AGC, Ala-AGC, Ala-AGC, Met-CAT, Ala-AGC, Ala-AGC, Ile-AAT, Ala-AGC, Ala-AGC, and Met-CAT is present on Chr6:26,541,774-26,696,641. This set of genes is also found in an inverted orientation approximately 44 kb away at Chr6:26,639,436–26,696,641, and another copy is located approximately 30,000 kb away at Chr6:57,647,398–57,709,136. **(b)** Microsynteny analysis between the human T2T-CHM13 and hg38 reference genomes. The region on Chr 1, specifically the large cluster 24 of 112 tDNAs (Chr1:160,577,243-160,733,822) in the T2T-CHM13 genome, matched cluster 20 in the hg38 genome.

The distribution and content of tDNA clusters are strongly conserved among certain primates (Bermudez-Santana et al., 2010), as we found in human, chimpanzee, and lemur **(Supp. Fig. 5)**. Despite this conservation, previous research has found significant variation in the abundance across Eukarya (Bermudez-Santana et al., 2010).

The genomic organization of tDNAs and their transcriptional activity

Evidence suggests that chromosomal localization of tDNAs can influence their transcriptional activity (Gao et al., 2024; Mungall et al., 2003; Van Bortle et al., 2017). We further examined the relationship between tDNA localization and transcription using our initial cluster characterization in the T2T-CHM13 genome assembly, to test the hypothesis that clustered tDNAs in close proximity to other tDNAs may exhibit higher expression levels than isolated tDNAs.

tDNA transcriptional activity can be estimated by combining RNA Pol III occupancy (POLR3D) ChIP-seq data with biotin-capture of nascent RNA (Van Bortle et al., 2017). Building upon this approach (see Methods), we found that while both isolated and clustered tDNAs can be actively transcribed, the tDNAs located within clusters show higher expression levels than those located within isolated genes **(Fig. 5a)**. For most isoacceptors, higher expression levels corresponded to tDNAs located in clusters **(Supp. Fig. 6c and Supp. Fig. 6d)**. Taken together, our results suggest that tDNA proximity and, consequently, tDNA clustering can favor tDNA expression (Gao et al., 2024; Van Bortle et al., 2017).

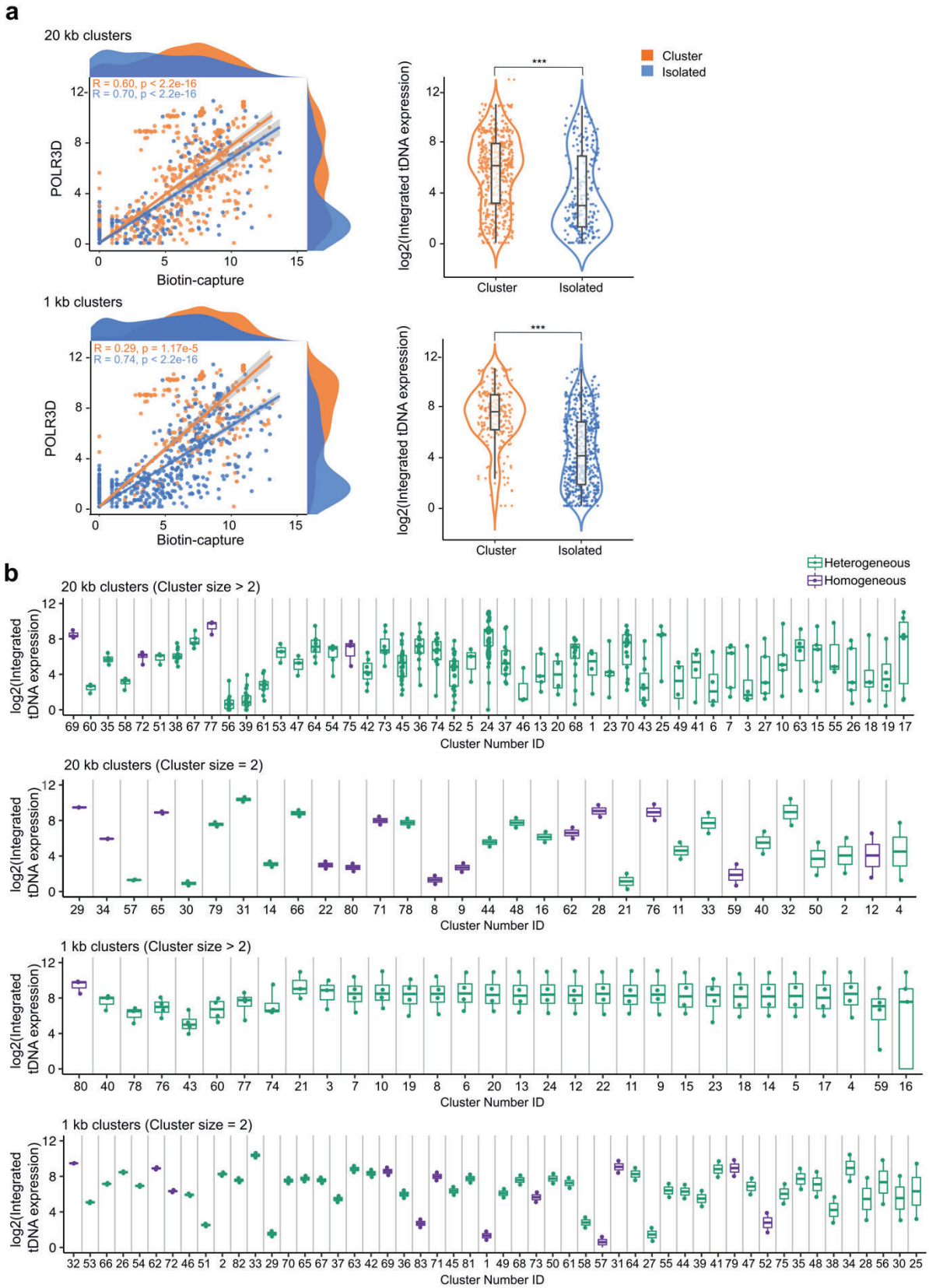


Figure 5. tDNAs localization and transcription. (a) Left scatter plots showing the correlation between POLR3D and biotin-capture expression values for tDNAs classified as either clustered (orange) or isolated (blue). The plots include Spearman's correlation (R) and the associated *P*-values (p). On the right, the violin plots show the log2 integrated tDNA expression for clustered and isolated tDNAs, with one-sided Wilcoxon rank-sum test (one-sided, alternative greater) significance levels indicated as ***, P

≤ 0.001. For the 20 kb clusters, *P*-value < 2.2×10−16, and for the 1 kb clusters, *P*-value = 1.714×10^{-10}. Top row: Analysis based on clusters within 20 kb. Bottom row: Analysis based on clusters within 1 kb window. **(b)** Plots showing the variability of integrated tDNA expression values within each tDNA cluster. Each dot represents the expression value of one tDNA. Clusters were classified as either homogeneous (composed of the same isoacceptor) or heterogeneous (composed of different isoacceptors). The two plots at the top show the clusters within a 1 kb window, whereas the bottom plots show those within a 20 kb window. For both types of clusters, the data were divided into clusters containing two tDNAs (i.e., tDNA pairs) and clusters containing more than two genes.

To investigate whether tDNAs within these clusters exhibit coordinated expression, that is, whether genes within a cluster exhibit similar expression levels, we analyzed the variability and dispersion of expression values within each cluster **(Fig. 5b)**, with lower variability suggesting more uniform expression among tDNAs within a cluster. We classified the clusters by size, distinguishing between those with two genes (i.e., tDNA pairs) and those with more than two. Furthermore, we categorized the clusters into two types: homogeneous clusters containing tDNAs encoding the same isoacceptor and heterogeneous clusters comprising tDNAs encoding different isoacceptors, to account for the impact of sequence similarity on the results. Our analysis revealed that for tDNA pairs, both 20 kb and 1 kb clusters tended to exhibit similar expression levels or expression levels within the same range **(Fig. 5b)**, a phenomenon observed in both heterogeneous and homogeneous clusters. We observed a similar trend for clusters containing more than two genes. This pattern was notably more pronounced in the 1 kb clusters, whereas the 20 kb clusters tended towards greater variability **(Fig. 5b)**.
Altogether, this result indicates that tDNA proximity can help coordinate tDNA expression, but distances greater than 1 kb might be insufficient to consistently promote a uniform expression level between tDNAs. It is crucial to note that the tDNA sequence itself, its genomic and epigenetic context, and other factors also play a key role in defining tDNA expression.

tDNAs are located in early-replicating regions

Another important factor influenced by genomic localization that provides insight into gene activity is the order of DNA replication, which occurs specifically during the S phase (synthesis phase) of the cell cycle (Maric & Prioleau, 2010; Müller & Nieduszynski, 2017; Rhind & Gilbert, 2013). According to the transcriptional activity, genomic regions replicate either early or late during S phase. Regions containing actively transcribed genes typically exhibit open chromatin (euchromatin), which is less condensed and more accessible to both the transcription and replication machinery, and thus are replicated early. In contrast, late-replicating regions are often associated with repressed gene expression and are characterized by closed chromatin (heterochromatin) that delays access to the replication machinery. This relationship suggests that replication timing can serve as a proxy for gene activity (Supek & Lehner, 2015).

To determine the replication timing of specific genomic regions, sequencing-based methods such as Repli-Seq (Hansen et al., 2010) are employed. This approach involves labeling and sequencing newly

synthesized DNA during S phase. Replication timing values are then calculated based on the relative abundance of nascent DNA in early versus late fractions, typically ranging from 0 (late-replicating) to 100 (early-replicating), providing a quantitative measure of replication timing (Supek & Lehner, 2015). By analyzing Repli-Seq data (Hansen et al., 2010), we found that most tDNAs are located in regions with Repli-Seq values characteristic of mid and early-replicating regions **(Fig. 6)** (96% of the tRNAs with values higher than 40).

Replication timing was calculated by dividing the genome in 200 kb (see Methods). Therefore, tDNAs that are in the same cluster fall within the same domain and share the same replication timing. Moreover, individual tDNA clusters display different replication times **(Supp. Fig. 7a)**. Consistent with the established link between replication regions and transcriptional activity, our analysis of tDNAs revealed that their transcription levels indeed align with their replication timing **(Supp. Fig. 7b)**.

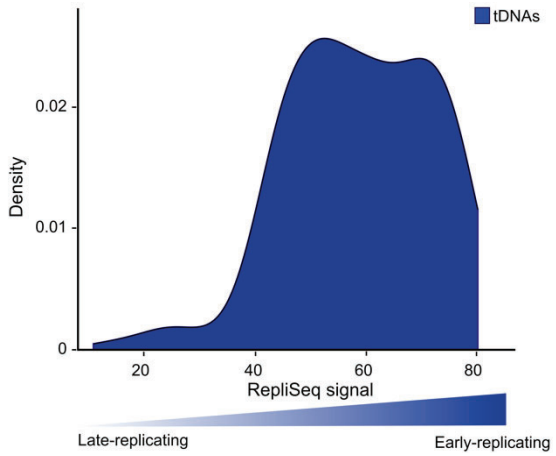


Figure 6. Replication timing of tDNAs. Distribution of Repli-Seq signal values for all tDNAs, with a higher signal indicating an earlier replication timing.

One of the characteristics of early-replicating regions is their tendency to exhibit a lower mutational rate compared to late-replicating regions (Gonzalez-Perez et al., 2019). This difference in mutational burden is partly explained by the fact that early-replicating regions tend to be more accessible and transcriptionally active, which facilitates more efficient DNA repair mechanisms, as described for protein-coding genes (Supek & Lehner, 2015). This pattern does not align with observations in tDNAs, as they have been previously reported to present high levels of mutagenesis (Saini et al., 2017; Sakhtemani et al., 2019; Seplyarskiy et al., 2023; Thornlow et al., 2018), and are located in early-replicating regions. However, most studies have focused on the analysis of germline cells and little is known about somatic mutagenesis in tDNAs. These observations prompted us to further investigate the somatic mutational patterns in tDNAs and the mechanisms underlying their mutagenesis.

tDNAs are hotspots of somatic mutagenesis

To quantify the extent of somatic mutagenesis in tDNAs, we used whole-genome sequencing (WGS) data from 9596 samples of cancerous tissues. Variant calling data was previously obtained using the hg19 genome. To confirm tDNA variation, we used LiftOver to convert T2T-CHM13 tDNAs coordinates to hg19, and we were able to localize 537 tDNAs. We analyzed the mutation density in tDNAs and their flanking regions by dividing the sequence into windows, where the tDNA itself is designated as window 0, and the flanking regions are divided into 100-nts windows.

We confirm that tDNAs in somatic cells are mutational hotspots in tumor samples (**Fig. 7a**) (Sakhtemani et al., 2019). tDNAs are not located in broadly hypermutated regions of the genome, the mutations are specifically concentrated within the tDNAs. However, we observed that mutations accumulated in tDNAs and, to a lesser extent, in their immediately adjacent flanking regions, corresponding to the 100-nts windows -1 and 1 **(Fig. 7b)**. We examined whether tDNAs categorized by their integrated tDNA expression levels exhibited differential accumulation of somatic mutations (see Methods). Our analysis revealed that the internal mutational burden of tDNAs correlated directly with their transcriptional activity (Spearman correlation coefficient = 0.58, P-value < 2.2×10−16) (**Fig. 8a, Fig. 8c and Supp. Fig. 8**). This suggests that tDNAs experience transcription-associated mutagenesis (TAM) (Saini et al., 2017; Seplyarskiy et al., 2023; Thornlow et al., 2018).

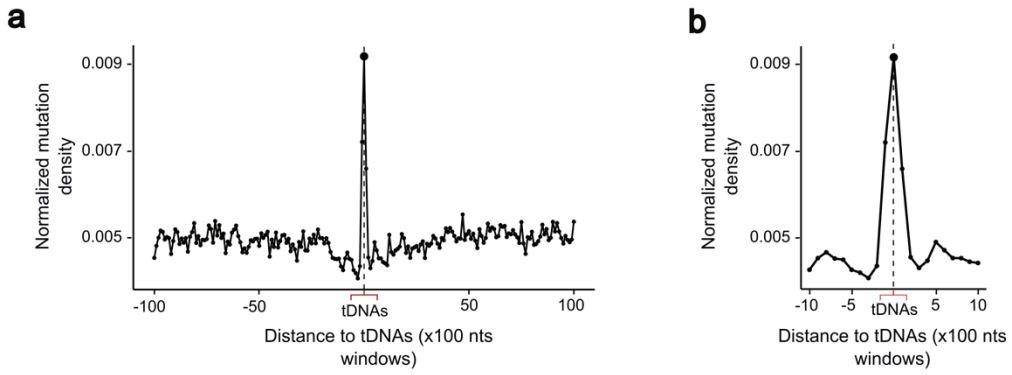


Figure 7. Somatic mutational profile in tDNAs. (a) Normalized mutation density in tDNAs and their flanking regions (10 kb downstream and upstream divided into windows of 100-nts) of cancer samples and **(b)** Close-up of normalized mutation density in tDNAs and their flanking regions (1 kb upstream and downstream, divided into 100-nts windows). The adjacent flanking regions to the tDNA correspond to windows -1 (downstrem) and 1 (upstream).

Protein-coding genes exhibit the opposite pattern to the one observed in tDNAs, with highly expressed genes showing the lowest level of mutagenesis **(Fig. 8b and Fig. 8d)**. Notably, highly transcribed tDNAs accumulate internal mutations at rates up to nine times higher than those observed in highly transcribed protein-coding genes. Moreover, tDNAs always exhibit higher mutational rates compared to protein-coding genes in all the different expression scenarios **(Fig. 8e)**. This difference suggests a direct link between Pol III activity and the mutation accumulation in tDNA genes. While tDNAs transcription is driven by Pol III, protein-coding genes are transcribed by RNA polymerase II (Pol II), which recruits DNA repair mechanisms such as transcription-coupled repair (TCR) (Hanawalt & Spivak, 2008). In

comparison, Pol III transcription lacks known coupled-repair, lacking mechanisms to deal with the TAM phenomena (Dammann & Pfeifer, 1997; Seplyarskiy et al., 2023).

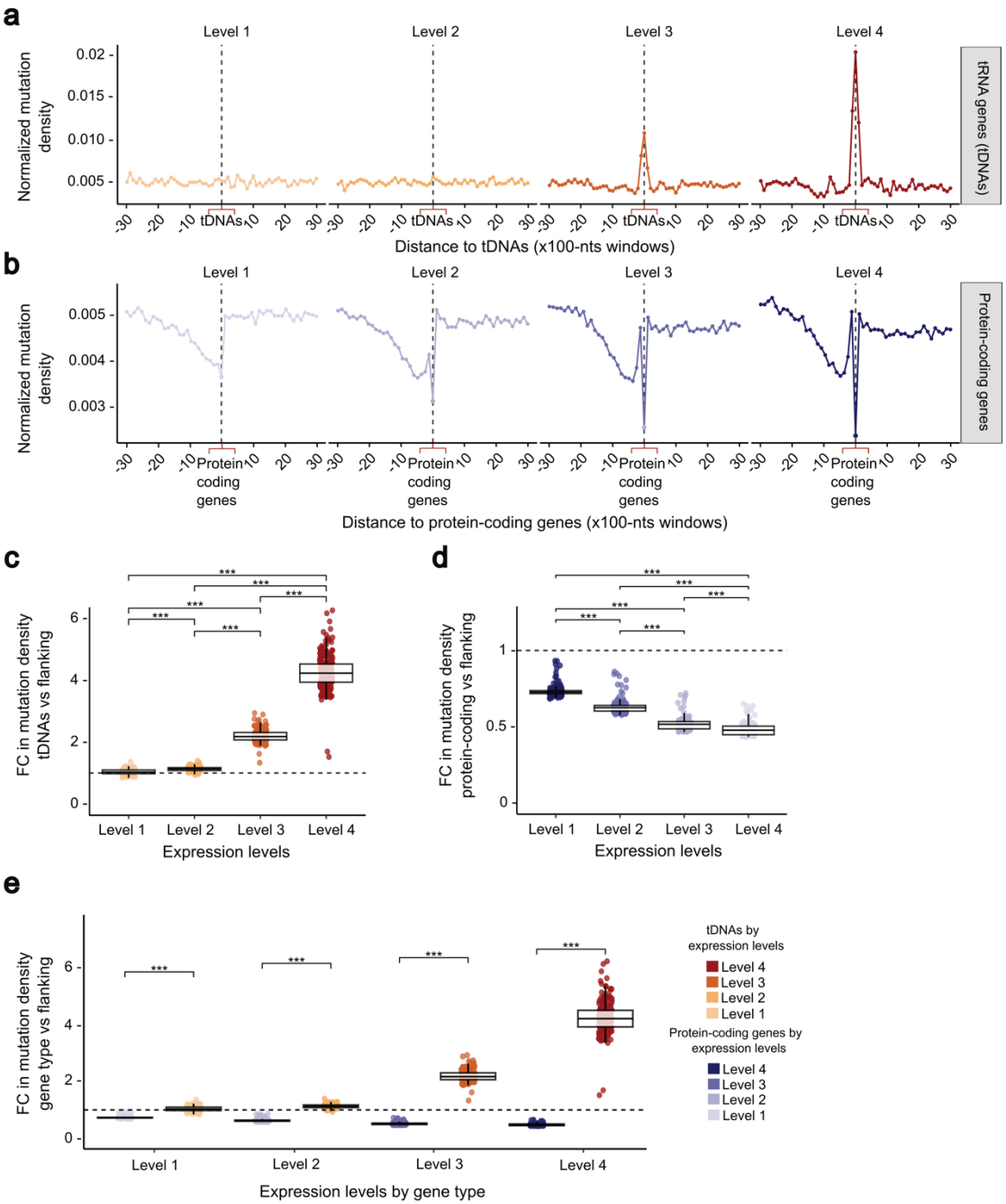


Figure 8. tDNAs and protein-coding genes somatic mutational profile by transcription activity **(a)** Normalized mutation density in tDNAs classified by transcription activity with expression levels from highest (Level 4) to lowest (Level 1), including flanking regions (3 kb upstream and downstream divided into 100 nt windows). **(b)** Normalized mutation density in protein-coding genes classified by expression levels, including flanking regions (3 kb upstream and downstream divided into 100 nt windows). **(c)** Fold changes (FC) between mutation density in tDNAs and each 100-nts window in the flanking regions of 10 kb. The data is divided by gene expression level. ***(d)*** FC between mutation density in protein-coding genes and each 100-nts window in the flanking regions of 10 kb. The data is divided by gene expression level. **(e)** Comparison of mutational densities in tDNAs and protein-coding genes. Each dot represents the fold change (FC) between the mutation density in each gene type (tDNAs or protein-coding genes) and their corresponding 100-nts window in the flanking regions. Statistical significance in all plots was

obtained using the Wilcoxon rank-sum test (one-sided with alternative hypothesis: greater) with Benjamini–Hochberg adjusted P-values (***: P-value ≤ 0.001).

We asked whether the observed somatic mutation load at tDNAs was also detectable in other Pol III-transcribed genes, as was reported for human germline mutations (Seplyarskiy et al., 2023). We found that other Pol III-transcribed genes, including small nuclear RNA genes (RNUs), 5S ribosomal RNA gene (rDNA) and unclassified/miscellaneous RNA genes (miscRNA genes), also accumulated mutations **(Fig. 9a)**, but at significantly lower rates than tDNAs **(Fig. 9b)**. Moreover, these Pol III-transcribed genes exhibit distinct mutational profiles compared to tDNAs, as the adjacent flanking regions (windows -1 and 1) show a higher mutational rate than the gene itself.

Among these biotypes, RNUs showed a higher mutation density than their flanking regions, with most flanking-to-gene mutation density ratios exceeding 1 **(Fig. 9b)**. Notably, the accumulation of mutations in RNUs genes has also been described in germline cells (Seplyarskiy et al., 2023). However, among Pol III-transcribed genes, tDNAs appear to experience a higher somatic mutation rate, suggesting that they may possess specific characteristics that make them particularly prone to mutagenesis during transcription **(Fig. 9b)**.

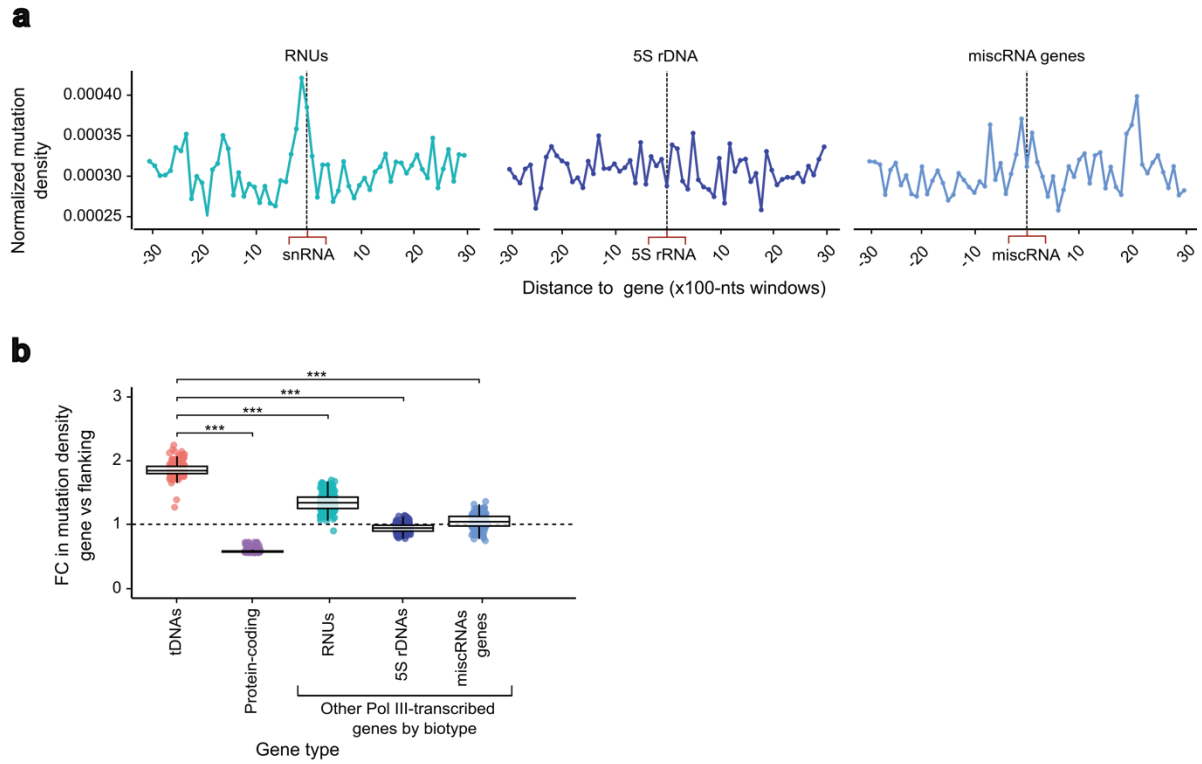


Figure 9. Somatic mutational profile in other Pol-III transcribed genes. (a) Normalized mutation density in other Pol III-transcribed genes across cancer samples. The flanking regions are 10 kb upstream and downstream, divided into 100-nts windows. Genes previously identified as Pol III-transcribed genes are classified by biotype, including 5S ribosomal RNA genes (5S rDNA), small nuclear RNA genes (RNUs), and unclassified or miscellaneous RNA genes (miscRNA genes). **(b)** Normalized mutation density in Pol III-transcribed genes in cancer samples. Each dot represent the fold change (FC) between the mutation density in each gene type (and their corresponding 100-nts window in the flanking

regions (10 kb nts upstream and downstream). Statistical significance was obtained using the Wilcoxon rank-sum test (one-sided with alternative hypothesis: greater) with Benjamini–Hochberg adjusted *P*-values (***: *P*-value ≤ 0.001).

Tumor type- and age-dependence of tDNA somatic mutagenic rates

The overexpression of tDNAs in human cancers has been repeatedly reported (Goodarzi et al., 2016; Gupta et al., 2022; Orellana et al., 2022; Pinzaru & Tavazoie, 2023; Zhang et al., 2018), and linked to adaptive mechanisms of codon-anticodon optimization that prioritize the translation of genetic programs important for tumor growth (Goodarzi et al., 2016). Moreover, tDNA expression can be cell type- or tissue-specific, often resulting in differential expression of not only mature tRNAs but also of tRNA fragments (Dittmar et al., 2006; Torres et al., 2019). To test whether somatic tDNA mutagenesis is a distinguishing feature of different human tissues, we compared tDNA somatic mutagenesis by cancer type, analyzing a total of 24 tissues **(Supp. Fig. 9)**. First, to determine whether tDNAs are mutagenesis hotspots, we normalized the mutation density in tDNAs of each cancer type relative to their flanking regions **(Fig. 10a)**. We then calculated the percentage of samples that had at least one mutated tDNA per cancer type to determine the prevalence of tDNA mutations across different tumor contexts **(Supp. Fig. 10)**.

These analyses revealed that tDNA mutation rates varied by cancer type and tissue **(Fig. 10a and Supp. Fig. 10)**. Bladder cancer (BLCA) exhibited the highest rates of tDNA mutagenesis, with more than 50% of the samples having at least one mutated tDNA **(Fig. 10a and Supp. Fig. 10)**. In agreement with our previous finding that high levels of transcription correlate with mutational levels at tDNAs, cancer types characterized by upregulated levels of tDNA expression (including cancer types from the bladder, uterus, breast, lung, esophagus, head and neck, and prostate) (Zhang et al., 2018), also exhibit the highest mutational burdens in tDNAs **(Fig. 10a)**. Among all the tissues analyzed, over 5% of samples in most cases exhibited at least one mutated tDNA. Interestingly, brain and lymphoid samples showed the lowest percentage of mutated tDNAs **(Supp. Fig. 10)**.

As expected, uterus and colorectal samples from patients with previously described genomic instability patterns, such as DNA polymerase epsilon, catalytic subunit (POLE) proofreading deficiency (hypermutated samples defined as HYPER), or DNA mismatch repair (MMR) deficiency leading to microsatellite instability (MSI) (Haradhvala et al., 2018; Kim et al., 2013), did not show a higher accumulation of mutations in tDNAs than in their flanking regions **(Fig. 10a)**. This is likely because the causative mutagenic agents, in this case POLE-deficient and MMR-deficient, affect both the gene itself and its flanking region in the same way.

Next, we investigated whether mutational loads in tDNAs accumulate with age by comparing samples from cancer patients across different age groups. While it is well established that somatic mutations accumulate with age throughout the genome (Cagan et al., 2022), our analysis revealed that tDNA mutation rates not only follow this age-related trend but also remain consistently higher than those in the flanking regions, as observed in multiple tissue types, including the breast, colorectum, esophagus,

kidney, liver, lung, pancreas, prostate, and lymphoid tissues **(Fig. 10b)**. These results suggest that the accumulation of somatic mutations in tDNAs is a characteristic of aging human tissues and that tDNAs may exhibit accelerated mutational aging compared to other genomic regions.

To investigate whether somatic mutagenesis in tDNAs also occurs in healthy tissues, we quantified the mutational burden in tDNAs in 1192 samples of clonal healthy tissues (including brain, colon, liver, and lung) using WGS data, and again found a significant accumulation of somatic mutations in tDNAs **(Supp. Fig. 11)**. Thus, somatic tDNA mutations are also prevalent in healthy tissues, where their time-wise accumulation may contribute to the proteostasis defects characteristic of aging individuals. The available data from healthy human tissues are insufficient to determine the age-related distribution of somatic mutations in tDNAs. However, the behavior observed in tumors and the fact that the datasets from healthy tissues reveal somatic mutations in tDNAs suggest that these mutations could accumulate in healthy cells as humans age.

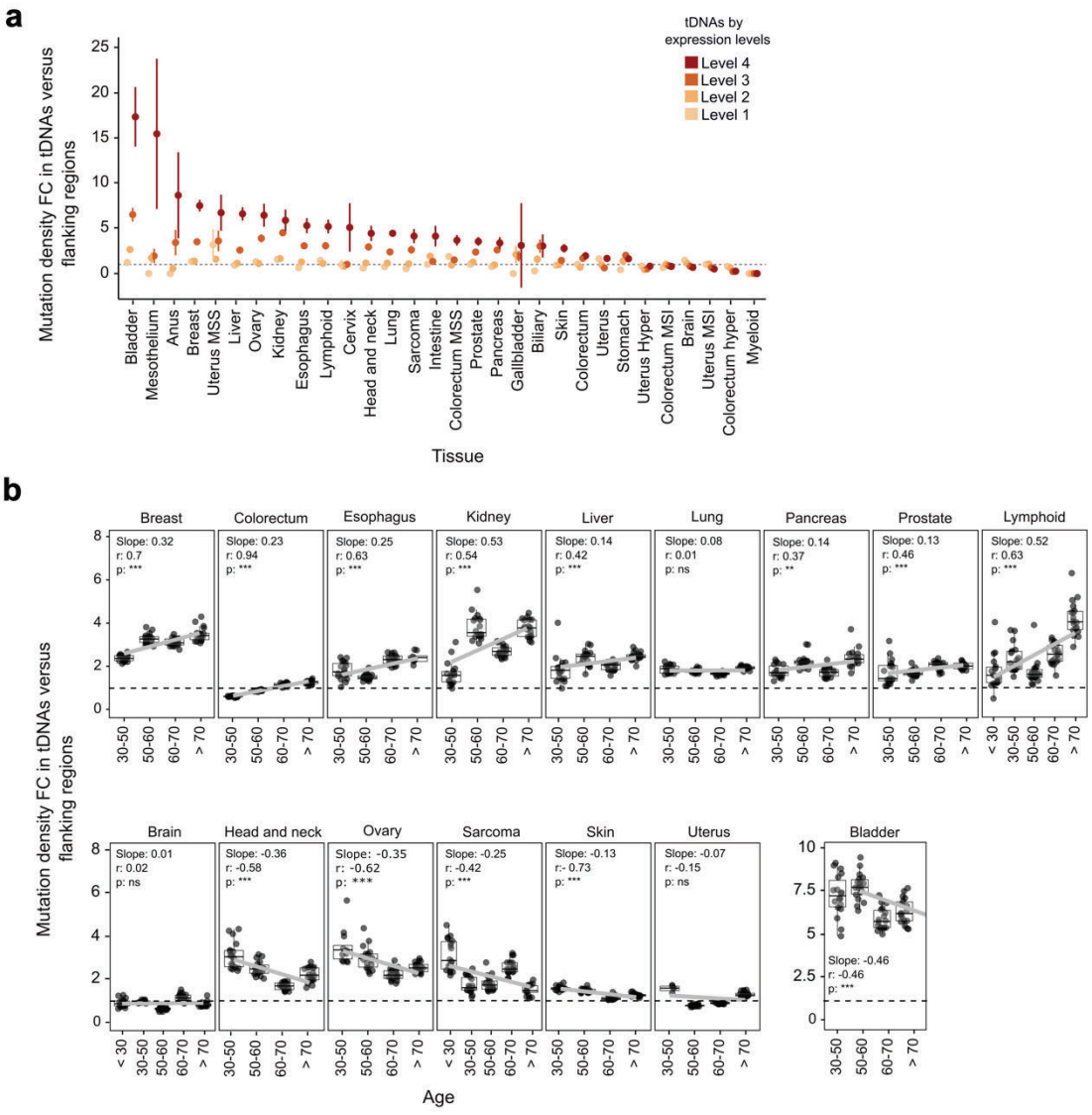


Figure 10. Age-related and tissue-specific somatic mutational burdens in tDNAs. (a) Mutation profiles in tDNAs across tissue types (including tissues with n > 20) relative to the flaking regions. Each dot represents the mean FC between the mutation density values for tDNAs and the flanking regions

(see Methods). tDNAs were categorized based on their expression levels. For colorectal and uterine samples, we categorized them into subgroups based on previously described genomic instability: hypermutated samples (HYPER) due to POLE deficiency, microsatellite instability (MSI) samples, or microsatellite stable (MSS) samples. For the remaining samples, the original tissue names, uterus and colorectum, were retained. **(b)** Correlation between the mutational density in tDNAs and age. The data was grouped by tissue and age range (including tissues with n > 200). Each dot represents the FC between the mutational densities of tDNAs and a flanking 10-window set (see Methods). The plots include values for the slope, Spearman correlation coefficient (R), and statistical significance of the correlation (***: *P*-value ≤ 0.001; **: *P*-value ≤ 0.01; *: *P*-value ≤ 0.05; ns: *P*-value > 0.05).

Mutational signatures and role of APOBEC3 in tDNA Mutagenesis

The nature and frequency of somatic mutations generate a signature that can be used to identify the agents responsible for the observed mutations, based on the frequencies of the trinucleotide contexts of the mutations. Over hundred mutational signatures exist in human samples, and mutational agents have been identified for approximately half of them (Alexandrov et al., 2020; Degasperi et al., 2022). All mutational signatures that have been identified and associated with a mutagenic agent are annotated in the curated COSMIC mutational signature catalogue reference (Alexandrov et al., 2020; Sondka et al., 2024). Because tDNA somatic mutagenesis is underexplored, we performed a de novo Non-negative Matrix Factorization (NMF) analysis to identify novel signatures that were not previously reported and annotated in the COSMIC reference and were thus unique or characteristic of tDNAs. NMF is a dimensionality reduction technique that decomposes a matrix of observed trinucleotide mutation counts into two reduced matrices: one representing a set of distinct mutational signatures and another quantifying the sample exposure or contribution of each signature to the mutational burden of individual samples (Alexandrov et al., 2020) (see Methods).

As a result, we obtained 12 NMF-extracted signatures, which we denoted by letters A to L (e.g., Signature A). Subsequently, we determined the resemblance between the NMF-extracted signatures and COSMIC reference signatures (**Fig. 11a**). If an NMF-extracted signature exhibits a high cosine similarity (e.g., typically ≥ 0.85) to a COSMIC reference signature, it indicates that the underlying mutational process or agent associated with that COSMIC signature is also represented by the NMF-extracted signature. All 12 signatures matched at least one known COSMIC signature, indicating that our analysis did not reveal any novel signatures. Nevertheless, this comparison allowed us to validate the results and subsequently characterize these 12 signatures to identify the mutational mechanisms underlying tDNAs mutagenesis **(Fig. 11a)**.

From this analysis, we identified that signatures L and C correspond to COSMIC signatures associated with sequencing artifacts **(Fig. 11a)**. Signatures D, H, I, and J are very sparse, nearly single-peak signatures that do not match the spectrum profile (distribution of the 96 possible trinucleotide mutation types) of their corresponding COSMIC signatures **(Supp. Fig. 15)**. In addition, signature K had a low cosine similarity of 0.50 **(Fig. 11a)**. Consequently, signatures L, C, D, H, I, J, and K were classified as non-relevant signatures, and we further explored signatures E, B, F, G, and A.

We identified signatures that were specifically enriched in tDNAs using sample exposures **(Fig. 11b)**. In protein-coding genes, we observed that nearly all mutational signatures affect both the coding sequences and their flanking regions similarly **(Fig. 11b),** and in some cases, the exposure in protein-coding genes was slightly lower than that in the flanking regions. As a control, this observation acknowledges that coding regions and their flanking sequences commonly face the same endogenous and exogenous mutagenic pressures. However, it is important to note that protein-coding genes can experience lower mutation levels owing to protective mechanisms, such as TCR (Seplyarskiy et al., 2023). In contrast, some signatures were more prevalent in tDNAs **(Fig. 11b)**. Strikingly, the exposure levels of signatures E and B increased with tDNA expression in cancer samples, aligning with the previous association of the mutagenic process with TAM **(Fig. 11b)**.

Furthermore, we identified the mutational signatures for each cancer/tissue type **(Fig. 11c and Supp. Fig. 13)**, demonstrating that the underlying mechanism driving mutations in tDNAs can vary between different cancer types and can affect healthy tissues.

NMF-extracted signature E corresponds to COSMIC signatures SBS13 and SBS2 (SBS13 + SBS2, cosine similarity = 0.94) **(Fig. 11a)**, which are associated with the activity of AID/APOBEC family enzymes (Sondka et al., 2024)**.** Upon analyzing the sample exposures for signature E, we observed that exposure levels increased with tDNA expression in cancer samples **(Fig. 11b)**, aligning with our previous findings on the association between tDNAs mutagenesis and TAM (the exposure levels were 3.28 times higher in highly transcribed tDNAs). APOBEC mutations generate a clear signature of C>T and C>G changes acting on TCN motifs (where N is any nucleotide) with a preference for TCW contexts (where W is T or A) (Roberts et al., 2013), as observed in the spectrum for signature E **(Supp. Fig. 14)**. Under normal conditions, the APOBEC enzymes function in antiviral defense. However, members of the APOBEC3 subfamily, particularly APOBEC3A (A3A) and APOBEC3B (A3B), have been implicated in cancer mutagenesis. Their activity can become off-target toward the host genome, particularly under conditions of replication stress or when DNA is highly exposed during transcription (Langenbucher et al., 2021)**.** We also observed this signature in tDNAs in both tumors and healthy non-cancerous tissues **(Fig. 11b)**.

Therefore, high transcription rates and/or other factors contributing to ssDNA occurrence must act at tDNA loci to facilitate APOBEC3 mutagenesis therein. Other genes transcribed by Pol III display different mutational loads, indicating that APOBEC3-driven mutagenesis at tDNAs is not driven only by high levels of transcription, but likely requires additional factors specific to tDNAs. We suspected secondary structure prone DNA sequences (Buisson et al., 2019; McCann et al., 2023; Roberts et al., 2013; Sui et al., 2020), and we used hairpin-localization algorithms to determine the presence of APOBEC3-preferred mutational sites in tDNAs (Buisson et al., 2019). This analysis revealed a significant enrichment of DNA hairpin formations in tDNAs compared to the rest of the genome (Chi-squared test $p = 0.01578$) or to other Pol III-transcribed genes (Chi-squared test $p = 0.03653$).

When examining the exposures by tissue type from cancer samples (**Fig. 11c and Supp. Fig. 13**), we identified this APOBEC3 mutational signature in tissues such as bladder, breast, uterus, head and neck,

and lung (**Fig. 11c and Supp. Fig. 13**), which aligns with previous findings where enriched APOBEC3 activity was reported for these cancer types (Alexandrov et al., 2020; Burns et al., 2013). Our data indicate that although APOBEC3 activity has been reported in these tissues, its activity appears to be increased in tDNAs **(Supp. Fig. 13)**. The mutational signature was also present in other tissues, such as the esophagus, intestine, liver, pancreas, sarcoma, ovary, and uterus **(Fig. 11c)**; however, the association with expression levels was less pronounced **(Supp. Fig. 13)**.

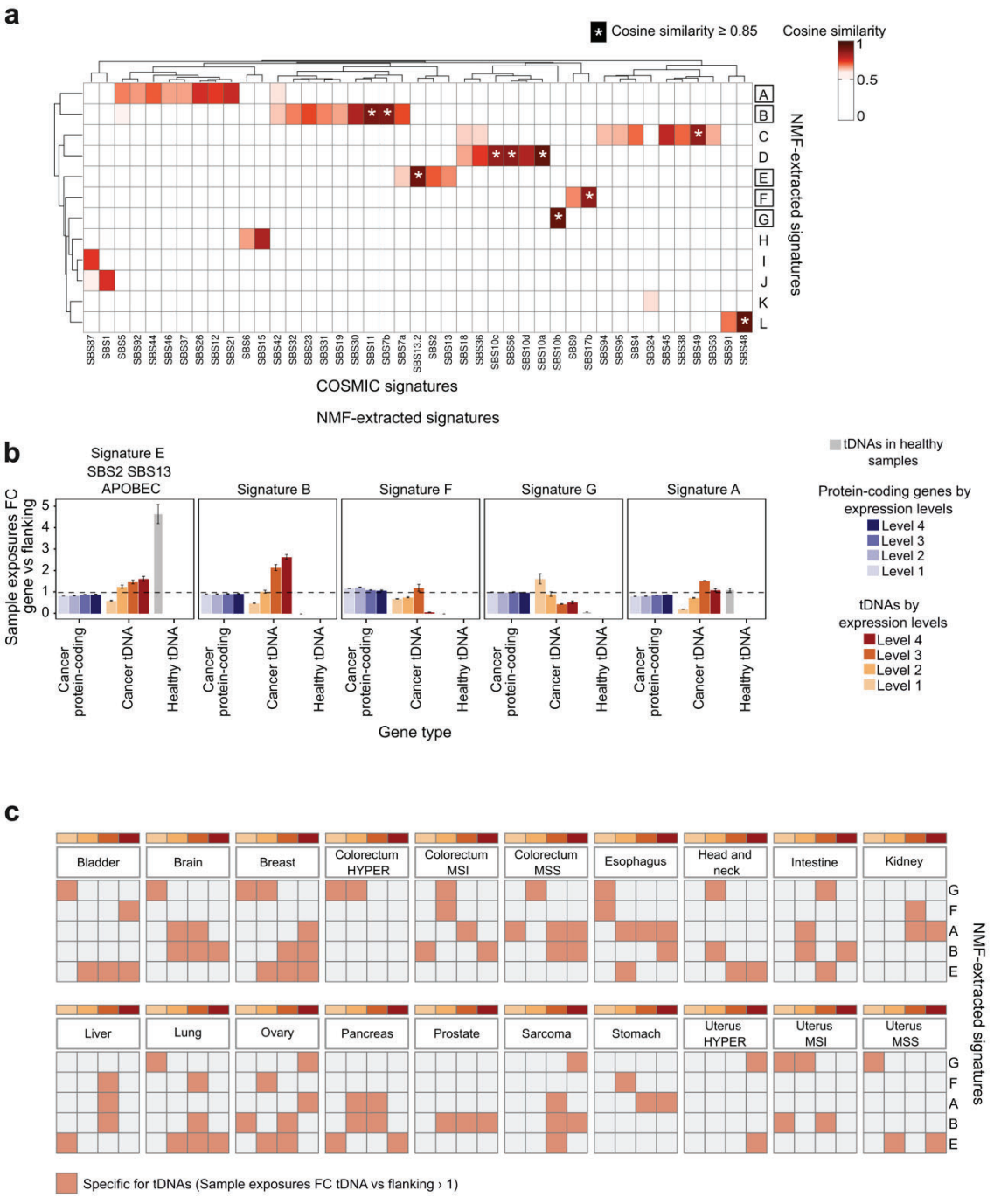


Figure 11. Mutational signatures in tDNAs. (a) Comparison of NMF-extracted signatures with COSMIC signatures. Heatmap showing cosine similarity values between the trinucleotide spectra of our de novo NMF-extracted signatures (rows) and the reference COSMIC signatures (columns). Only

COSMIC signatures with a cosine similarity ≥ 0.5 are shown. A COSMIC signature was assigned when the cosine similarity was ≥ 0.85, indicated by an asterisk (*). **(b)** Sample exposure FC between genes (stratified by type into tDNAs or protein-coding genes) and their flanking regions for each NMF-extracted signature. Exposures were analyzed for tDNAs and protein-coding genes in cancer samples and for tDNAs in healthy samples. Data for both gene types in cancer samples were stratified by expression levels, from low (level 1) to high (level 4). **(c)** Tissue-specific mutational processes found in tDNAs. Red indicates the activity of each NMF-extracted signature in the tDNAs. This activity was quantified by calculating the fold change (FC) between sample exposures in tDNAs and their flanking regions in different cancer types. A detailed analysis of sample exposure by cancer type is presented in Supplementary Figure 13.

Among the NMF-extracted signatures, Signature E, corresponding to APOBEC3, exhibited the most robust profile. Although other signatures were ambiguous in some cases **(Fig. 11a)**, we analyzed them to extract all possible information.

Signature F: Matches SBS17b with a cosine similarity of 0.88 (we note a similarity with SBS9, below the 0.85 threshold) **(Fig. 11a)**. Previous studies have associated SBS17b with prior treatment using alkylating agents such as 5-Fluorouracil (5-FU) chemotherapy and damage caused by reactive oxygen species (Sondka et al., 2024). The chemotherapy drug 5-FU has been used to treat various solid cancers, particularly colorectal and breast cancers (Christensen et al., 2019). However, we did not observe this signature in tDNAs neither in colorectal nor breast cancers. Therefore, as SBS17b has been associated with both 5-FU and oxidative stress damage, it is more likely that this signature is the result of oxidative damage. In this regard, we observed that this signature was more prevalent in the esophagus **(Supp. Fig. 13)**, a tissue type associated with high oxidative damage caused by gastroesophageal reflux, increasing the risk of esophageal adenocarcinoma (ESAD) (Killcoyne & Fitzgerald, 2021).

Interestingly, this signature shows decreased activity in highly-transcribed tDNA genes **(Fig. 11b)**. This suggests a protective effect of active, open chromatin at highly-transcribed tDNAs, possibly through recruiting DNA repair (Polak et al., 2014). Even though tDNA mutations cannot be repaired by TCR mechanisms, based on previously reported experiments, base excision DNA repair (BER) may be enriched in this situation (Saini et al., 2017).

Signature G: This signature matches SBS10b (cosine similarity = 0.95) **(Fig. 11a)**, which is associated with DNA Polymerase epsilon (POLE) exonuclease domain mutations that lead to an increased rate of base substitution errors during DNA replication (Sondka et al., 2024). The SBS10b signature is common in colon and uterus tissues exhibiting hypermutation caused by POLE (Alexandrov et al., 2020). We also detected these signatures in samples previously reported to have POLE deficiency (Colorectum HYPER and Uterus HYPER) **(Fig. 11c and Supp. Fig. 13)**. However, this signature is also present in other tissues, such as the esophagus, brain, bladder, and head and neck **(Fig. 11b and Supp. Fig. 13)**. This mutational signature was identified in human tumors but was absent in healthy cells, corresponding to cancer-related processes, and was negatively correlated with transcriptional intensity **(Fig. 11b)**. Similar

to Signature F, this also could suggests the activity of other repair mechanisms, such as BER, in highly transcribed genes (Saini et al., 2017).

Signature A: This signature matched multiple COSMIC signatures, with a cosine similarity of all of them below 0.85 **(Fig. 11a)**. Interestingly, three of these signatures (SBS26, SBS21, and SBS44) are associated with defective DNA mismatch repair (MMR). This signature also matches SBS92 (associated with tobacco smoking), SBS46 (possibly a sequencing artifact, commonly found in earlier TCGA colorectal cancer data released before 2013), and SBS37, SBS12, and SBS5 (of unknown etiology) (Sondka et al., 2024). This signature was also active in healthy samples **(Fig. 11b)**. Given that MMR is a known feature of cancers and is expected to be largely absent in healthy cells, together with its ambiguous source, our data suggest that this signature originates from another, currently unclear cause.

Signature B: This signature resembled SBS11 (cosine similarity = 0.92), SBS7b (cosine similarity = 0.91) and correlated with other signatures but with a cosine similarity lower than 0.85 **(Fig. 11a)**. SBS11 is associated with previous chemotherapy treatments using temozolomide, whereas SBS7b is linked to ultraviolet light exposure (Sondka et al., 2024). Temozolomide is commonly used to treat brain cancer, particularly high-grade gliomas (Ortiz et al., 2021). Notably, this signature was clearly observed in the brain **(Fig. 11c and Supp. Fig. 13)**, consistent with the origin of the treatment. However, this signature was observed not only in brain cancers but also in several other cancer types, such as intestinal, sarcoma, colorectal, and liver cancers **(Fig. 11c and Supp. Fig. 13)**.

Distribution of somatic mutations within tDNAs and evidence for negative selection

While negative selection acting on mutations in somatic cells is overall quite subtle, some essential genes and oncogenes do display lower mutation rates than expected (Besedina & Supek, 2024; Martincorena et al., 2018). We reasoned that the increased rate of somatic mutagenesis at tDNA loci, coupled with the fact that hundreds of tDNA genes can be easily aligned to compare variation at individual nucleotides, might provide good statistical power to search for evidence of negative selection acting against somatic mutations at critical positions of the tRNA molecule.

Thus, we quantified tDNA mutational rates at single-base resolution for human tDNAs to search for negative selection acting upon mutations at specific sites. This analysis revealed that most universally conserved positions in tRNA sequences (Beuning & Musier-Forsyth, 1999; Biela et al., 2023; Richard Giegé et al., 2012), display mutation rates significantly lower than the rest of nucleotides in tRNAs **(Fig. 12a, Fig. 12b ad Supp. Fig. 16a)**, indicating that these positions experience negative selection that eliminates mutations at these sites. Interestingly, this analysis revealed other positions that, while not universally conserved, also display lower than expected mutagenic rates, suggesting that these nucleotides play important, but as yet unidentified, functional roles in tRNAs **(Fig. 12a and Supp. Fig. 16a)**.

Strikingly, positions critical for translation elongation such as the anticodon triplet, where mutations can severely disrupt the efficiency and fidelity of protein synthesis (Geslain et al., 2010; Pinzaru & Tavazoie, 2023; Reverendo et al., 2014), showed no evidence of decreased mutation rates **(Fig. 12a and Fig. 12b)**, and we identified several instances of mutations at anticodon positions (see below).

In order to analyze the regional mutation load in structural domains of the tRNA, or at the two internal promoters of tDNAs, we used a rolling windows method to calculate the relative mutation load in ten-nucleotide sections of these molecules. This revealed that tDNA genes experience higher mutational loads at their 3' halves **(Supp. Fig. 16b)**. Interestingly, the second promotor region (B-box) is where most mutations accumulate **(Fig. 12a and Fig. 12b)**. This mutational distribution might reflect a locally protective effect of the Pol III complex of the transcription factor TFIIIB over a section of tDNA positive strands during their transcription. Alternatively, the presence of sequences or motifs required for the activity of APOBEC3 in the 3' half of tDNAs could be responsible for the relative accumulation of mutations in these regions of tDNAs **(Fig. 12a)**. Consistent with this second possibility, we find that position 56, within the region of the B box, is the highest mutated position in all tDNAs and is occupied by a highly conserved cytosine in all human tRNAs.

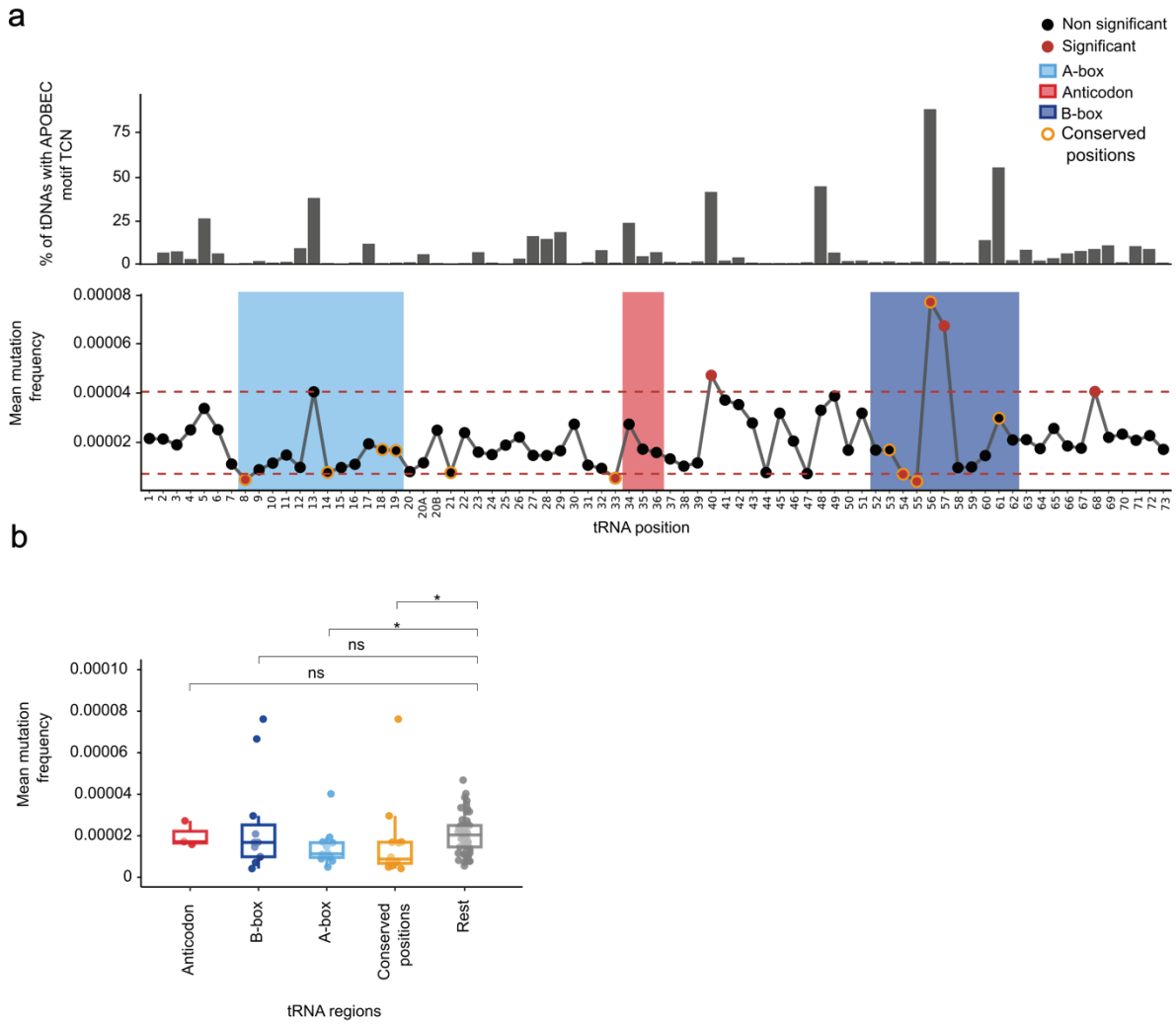


Figure 12. Somatic tDNA mutation frequencies at single-base resolution and distribution of APOBEC3 motif presence across tDNAs genomic positions (a) The lower panel shows mutation

rates at single-base resolution across tRNA sequences, highlighting universally conserved positions (U8, A14, G18, G19, A21, U33, G53, T54, U55, C56, A58, C61, C74, C75), internal promoter regions such as the A-box (positions 8-19), the B-box (positions 52-62), and the anticodon (positions 34-36). Statistically significant positions are indicated by red dots. Specific tDNA positions were considered significant when adj. $p \leq 0.01$ and if the mean mutation density exceeded the 0.95 quantile or fell below the 0.05 quantile of the overall distribution (quantile thresholds are indicated by red dashed lines). The upper panel shows the percentage of tDNAs with APOBEC3-associated TCN motifs (where N = T, C, G, or A) across the tDNA gene sequences. **(b)** Comparison of mutation densities in functionally and structurally important tRNA positions versus other positions. Wilcoxon significance levels are indicated in the plot: A-box vs. rest (adj. *P*-value = 0.025 *),* conserved positions vs. rest *(adj. P-value = 0.025*), anticodon vs rest (adj. *P*-value = 0.971, ns), and B-box vs. the rest (adj. *P*-value = 0.541, ns).

Interestingly, conserved positions that show little to no mutagenesis across most tRNAs tend to lack APOBEC3 motifs. From an evolutionary perspective, it is possible that these regions were selected against containing APOBEC3 motifs, as mutations in these critical areas would have a major negative impact on tDNA function. This suggests that there may have been evolutionary pressure to avoid APOBEC3 target motifs in key tDNA regions.

Somatic mutations at tDNAs generate chimeric tRNAs

The attachment of amino acids to their cognate tRNA molecules by aminoacyl-tRNA synthetases (ARS) is a crucial step for accurate protein translation. For the majority of ARS, tRNA recognition requires interactions with both the acceptor stem and anticodon loop motifs (R. Giegé et al., 1998; Richard Giegé & Eriani, 2023; Hou & Schimmel, 1988; Park & Schimmel, 1988; Rubio Gomez & Ibba, 2020). However, ARS cognate for amino acids alanine (Ala), leucine (Leu), serine (Ser), and tyrosine (Tyr) recognize only the acceptor stem of their cognate tRNA substrates (Richard Giegé & Eriani, 2023). In these cases, somatic mutations at tRNA anticodons can lead to ‘chimeric tRNAs’ whose anticodon triplet will recognize codons that do not correspond to the amino acid carried by the chimeric tRNA. Such molecules lead to translation errors throughout the proteome, caused by the misincorporation of Ala, Ser, Leu, or Tyr at non-cognate codons within the ribosome. Strikingly, in the datasets used in this study we could identify numerous examples of somatic mutations leading to the generation of chimeric tRNAs previously shown to induce proteome-wide amino acid substitutions **(Fig. 13d)** (Geslain et al., 2010; Pinzaru & Tavazoie, 2023; Santos et al., 2018). For example, we detected several examples of mutations at position 35 of tRNA-Ser-CGA that result in Leu to Ser substitutions through the proteome **(Fig.13d)**. The detection of mutations that compromise the fidelity of the Genetic Code at intensely transcribed tDNAs suggests that tumor human cells could produce chimeric tRNAs capable of introducing widespread amino acid changes throughout the proteome.

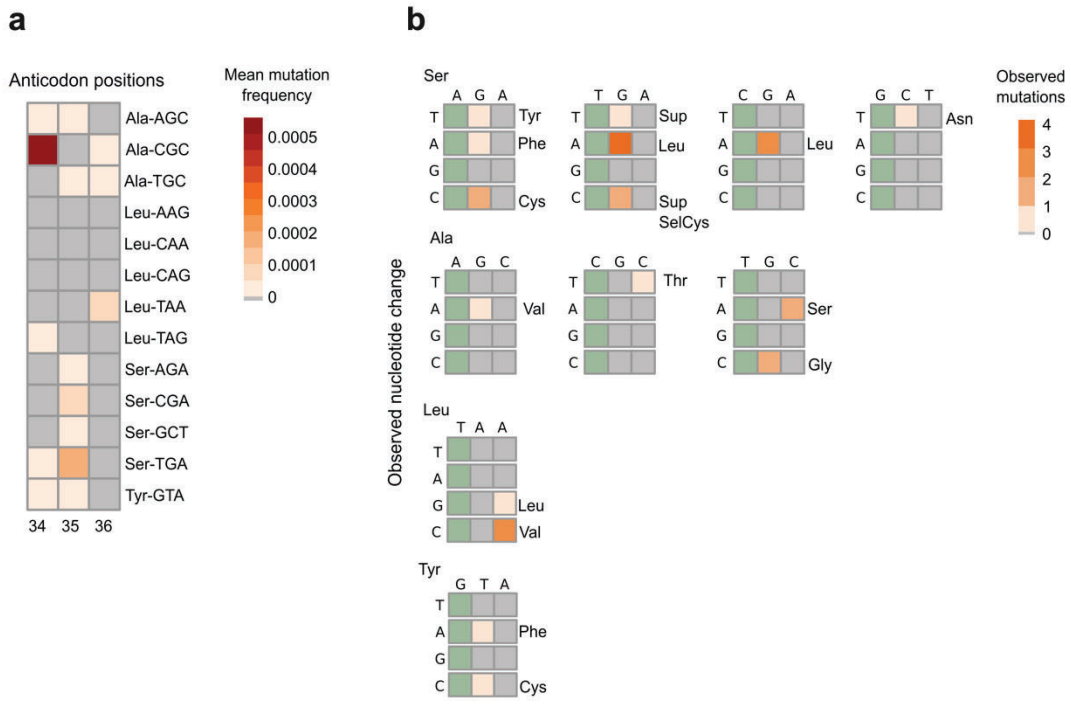


Figure 13. Characterization of mutational profiles in the anticodon region of potential chimeric tRNAs. (a) The heatmap shows the mean mutation density at each anticodon position, classified by isodecoders for alanine (Ala), serine (Ser), leucine (Leu), and tyrosine (Tyr). **(b)** The heatmap on the right illustrates the number of times specific mutational changes were observed in the anticodon region for specific isodecoders of Ala, Ser, Leu, and Tyr. Position 34 is highlighted in green because mutations at this position are always synonymous. Next to each mutation (observed nucleotide change), the codon recognized by the resulting chimeric tRNA is indicated in the table. For example, tRNA Ala-AGC mutates to Ala-AAC, resulting in a tRNA that still transports alanine but now binds the codon for valine.

Discussion

Proteome variability can have both negative and positive impacts upon cellular homeostasis. In microbes, for example, several adaptive mechanisms based on the induction of widespread proteome variation have been described, several of which are based on the use of mistranslating tRNAs (Ribas de Pouplana et al., 2014). In animals, adaptive proteome variations aided by changes in cellular tRNA populations are well known, and always linked to adaptive genetic switches that are important for both healthy tissues and tumor development (Earnest-Noble et al., 2022; Gingold et al., 2014; Pavon-Eternod et al., 2013). Adaptive mistranslation (the deliberate induction of translation errors during protein synthesis) has not been reported in animals, and it is assumed that such errors will always be deleterious to human health.

Frailty, sarcopenia, and many forms of neurodegeneration common during the aging process are linked to losses in proteome quality (Anisimova et al., 2018). Similarly, human neoplasies that develop resistance to existing chemotherapies often do so through point mutations that thwart drug-recognition sites, but the possibility that such modifications arise through mistranslation events has not been explored (Roszkowska, 2024).

We have studied in detail the dynamics of somatic mutagenesis in human tDNAs, a phenomenon previously observed in several model organisms and in human datasets (Saini et al., 2017; Sakhtemani et al., 2019; Seplyarskiy et al., 2016; Sui et al., 2020). Our data shows that human tRNA genes, but not other genes transcribed by Pol III, are hotspots of somatic mutagenesis that accumulate genetic changes directly as a function of their transcriptional rate. The fact that other Pol III-transcribed genes behave differently in terms of somatic mutagenesis indicates that some tDNA-specific factors drive these mutational events.

In agreement with previous studies, our mutational signature analysis indicates that members of the APOBEC3 family are the main contributors to this mutational load. Our topological analyses, together with published interactome data (Jang et al., 2024), indicate that the recognition of sequence motifs within tDNAs, coupled to interactions between APOBEC3 enzymes and tRNA modification enzymes, may be driving the specific mutational process at tDNAs. Given that tRNA transcription and processing occur simultaneously, the specificity of somatic mutations at tDNA loci may be the result of specific interactions of APOBEC3 enzymes with components of the tRNA processing machinery in the vicinity of positive strands of tDNAs during transcription.

Analysis of the mutational load within tDNAs at single-nucleotide resolution shows that universally conserved positions in tRNAs display significantly reduced heterogeneity, indicating that these sites may be under negative (purifying) selection in somatic cells. Unexpectedly, we find anticodon positions to experience mutational loads no different from the average values in the whole sequence of tDNAs. This fact leads to the remarkable conclusion that somatic mutagenesis at actively transcribed tDNAs generate chimeric tRNAs with anticodon mutations previously shown to cause generalized substitutions

in the human proteome. It stands to reason that age-related accumulation of mutations at tDNAs will result in an increase of such chimeric tRNAs in aging cells and tissues.

Our analysis of tDNA somatic mutations in different human tumors shows tumor-specific differences in the levels of mutation accumulation. Bladder cancer (BLCA), which displays the highest accumulation of such mutations, is characterized by highly augmented levels of APOBEC3A activity, and a several APOBEC3-associated driver hotspot mutations have been identified in this malignancy (Lindskrog et al., 2021; Shi et al., 2020). This suggests that the high levels of tDNA mutations in BLCA are the consequence of the high levels of APOBEC3 activity in these tumors and opens the possibility of a functional relationship between mutated tRNAs and BLCA development. Interestingly, brain cancer lays at the opposite spectrum in terms of tDNA somatic mutations, which are almost absent in the datasets from these tumors.

Whether tDNA mutations play a general role in carcinogenesis is an important question. This could happen, first, at the level of proteome variations that might compromise the activity of tumor suppressors or generate proteins that favor tumor growth, such as oncoproteins produced by translational variations in existing transcription factors. Secondly, mutant tRNAs with defective sequences or aberrant post-transcriptional modification patterns could alter the population of tRNA fragments (tRFs) in cells. It has been described that a dysregulation in the levels of specific tRFs is related with the aggressiveness of BLCA tumors (Papadimitriou et al., 2020). Finally, proteome variability induced by mutagenic tRNAs may provide tumoral cells with fitness advantages such as avoidance of immune recognition. Although such possibility has not been explored in human cells, in *Candida albicans* proteome variability induced by mistranslating tRNAs induces morphology changes that shield this organism against the immune response of its human host (Bezerra et al., 2013).

The extent to which tDNA somatic mutations accumulate in healthy tissues, their mutational signatures, and their physiological impact remain to be determined. However, the accumulation of defective tRNAs can rationally be linked to the many functional connections between tRNA dynamics and cancer (Gupta et al., 2022; Santos et al., 2019). The age-dependent accumulation of mutations at tDNAs is apparent from the analysis of cancer datasets stratified by patient age, a factor that may directly impact upon the proteostasis of aging tissues (Anisimova et al., 2018; López-Otín et al., 2023; Schmidt & Schimmel, 1993). Inactive tRNAs in aging cells may induce an attenuation of protein synthesis efficiency and fidelity, either through their interference with components of the translation machinery. On the other hand, chimeric tRNAs would lead to mistranslation, possibly causing more acute problems such as protein aggregation, stress responses and inflammation, or the generation of aberrant antigenic peptides.

* Note that this chapter's discussion is limited to the findings from the original manuscript prepared for publication on tDNA mutagenesis. A broader discussion of all the topics is presented in the General Discussion.

Methods

tDNAs prediction

Reference human genomes were obtained from the UCSC Genome Browser. The assemblies analyzed included Jan. 2022 (T2T-CHM13 v2.0/hs1), Dec. 2013 (GRCh38/hg38), and Feb. 2009 (GRCh37/hg19). The tDNAs for each assembly were predicted and annotated using tRNAscan-SE 2.0 (v2.0.9, July 2021) (Chan et al., 2021), with parameters set to search for eukaryotic tRNAs and to display both primary and secondary structure components in the covariance model bit scores to distinguish functional tRNAs from potential pseudogenes. For the hg38 and hg19 assemblies, tDNAs that mapped to unplaced sequences (DNA fragments associated with a specific chromosome but whose order or orientation could not be determined) or unassigned sequences (DNA fragments not assigned to any chromosome) were identified and designated as unlocalized tDNAs. Subsequently, for the remaining tDNAs, the EukHighConfidenceFilter script from tRNAscan-SE 2.0 was used to assess the fidelity of tDNA predictions. This step classified the predicted tDNAs into several categories: high confidence (predictions that exhibit strong sequence and structural alignment with known active tRNA profiles), pseudogenes (which encode inactive products due to sequence alterations, such as deletions), and uncertain function (genes whose function or identity is not clearly established, possibly due to non-canonical structures or sequence deviations from known functional tRNAs). Note that tDNAs refer to DNA sequences that encode mature tRNA.

tDNAs genomic distribution and clusters definition

tDNAs annotations from the most updated human genome (T2T-CHM13 v2.0/hs1) were used to analyze tDNA distribution. To visualize the genomic coordinates of tDNAs, we used the RIdeogram v0.2.2 package from R. The genomic distances between consecutive tDNAs were computed and used to plot their cumulative distribution using the empirical cumulative distribution function (ecdf) **(Supp. Fig. 4a)**. Considering the profile obtained from the ecdf and previous definitions of tDNA clusters (Alexandrov et al., 2020; Bermudez-Santana et al., 2010), tDNA clusters were defined as groups of consecutive tDNAs separated by a specific genomic distance of either 1 kb (“1 kb cluster”) or 20 kb (“20 kb cluster”).

Synteny analysis

Reference genomes for human, chimpanzee, were obtained from the UCSC Genome Browser. For the human genome (*Homo sapiens*) T2T-CHM13 v2.0/hs1 Jan. 2022 assembly and for the chimpanzee (*Pan troglodytes*) genomeClint_PTRv2/panTro6 Jan. 2018. The reference genome for lemur (*Microcebus murinus*) was downloaded from NCBI, the assebly February 2017 RefSeq (GCF_000165445.2/Mmur_3.0). To perform synteny analysis, we obtained Gene transfer format

(GTF) files and DNA sequences for each species from the Ensembl database (https://www.ensembl.org/Homo_sapiens/Info/Index). These files were modified by removing existing tDNA annotations and incorporating data for 20 kb tDNA clusters, including the start and end coordinates of each cluster and the full nucleotide sequences for each species. The Python version of MCScanX (Wang et al., 2012) (https://github.com/tanghaibao/jcvi/wiki/Mcscan-(python-version)) was used to perform synteny analyses. Macro-and microsynteny analyses were used to compare the tDNAs clusters distibution. The results were visualized using the karyotype format provided by MCScanX software.

tDNAs expression data

Sequencing data for Pol III occupancy (POLR3D ChIP-seq) and nascent transcription profiling (biotin-capture) assays were obtained from the Gene Expression Omnibus (GEO) database and are accessible through the GEO Series accession number GSE96800 (Van Bortle et al., 2017). Specifically, the subset of samples used in this study was obtained from the monocyte dataset and corresponded to the following GEO Sample accession numbers: POLR3D (GSM2544232, GSM2544233) and biotin capture (GSM2544240, GSM2544241). Following the methodology outlined by Van Bortle et al. (2017), preprocessing of the sequencing data involved trimming low-quality bases and removing adapter sequences using CutAdapt v4.1 (Martin, 2011). The processed reads were aligned to the complete human reference genome T2T-CHM13 using Bowtie2 v2.4.2 in paired-end and local mode with allowance for one mismatch in the seed (-N 1) (Langmead & Salzberg, 2012). Given the repetitive nature of tDNAs, sequencing reads frequently map to multiple genomic sites, complicating the accurate quantification of tDNA transcription levels in the resulting data. To overcome this, multi-mapping reads were allocated to a single “best” alignment position. This method helps reduce the risk of overestimating tDNA transcription levels that could result from counting multiple alignments while preventing potential underestimation by excluding all multi-mapping reads. Gene counts were obtained using featureCounts from Subread v2.0.1 (Liao et al., 2014). In our analysis, gene annotation files in GTF format, derived from tDNA coordinates predicted by tRNAscan-SE 2.0, were used to report tDNA counts. To simulate the precursor tDNA sequences, tDNA coordinates were extended by adding 50 bases both upstream and downstream. Subsequently, gene count normalization was performed using the estimateSizeFactors function from the DESeq2 package. The log 2 mean between the biological replicates was obtained for each experimental data POLR3D and biotin-capture.

Finally, integrated tDNA expression values were obtained by computing the average of the POLR3D and biotin-capture results **(Supp. Fig. 6a)**. tDNAs were categorized into four expression levels defined by quartiles: genes with integrated tDNA expression values up to the 25th percentile were designated as Level 1 (lowest expression); those between the 25th and 50th percentiles as Level 2; those between the 50th and 75th percentiles as Level 3; and those with values above the 75th percentile as Level 4 (highest expression).

To ensure the reliability of the integrated tDNA expression data, which originated from a single study on macrophages, we verified tDNA activity. The tRNA Activity Predictor (tRAP) (Thornlow et al., 2020) pipeline was employed to assess general tDNA activity. This tool evaluates tDNA sequences and their genomic context to classify them as active or inactive. The tDNA predictions from tRNAscan-SE, applied to the T2T-CHM13 genome, were used as input for the tRAP analysis. The analysis revealed that tDNAs predicted to be active had significantly higher integrated tDNA expression values than those predicted to be inactive (one-sided Wilcoxon rank-sum test, alternative greater, *P*-value < 2.2×10^{-16}) (**Supp. Fig. 6b**).

Replication timing data

The process for obtaining replication timing data followed the approach described in a previous study (Supek & Lehner, 2015). Briefly, Repli-Seq measurements (Hansen et al., 2010) were obtained for ENCODE cell lines from the UCSC Genome Browser. These data are publicly available in GEO under the accession number GSE34399. The Repli-Seq signal was calculated by dividing the genome into 200 kb genomic domains based on the average across 11 cell lines. This domain size was chosen because replication signals typically vary within regions ranging from 200–400 kb. To identify tDNA genes within these domains, we mapped their coordinates from the T2T-CHM13 assembly to hg19 using LiftOver to determine the specific 200 kb domain in which each tDNA gene resided. The final Repli-Seq signal values for each domain ranged from 0 to 100, with higher values indicating earlier replication timing. Experimentally, this phenomenon occurs because the Repli-Seq technique involves labeling nascent DNA with a nucleotide analog, such as bromodeoxyuridine (BrdU), followed by sequencing of the labeled DNA. Consequently, regions that replicate earlier in the S-phase have a higher abundance of labeled fragments. Notably, not all tDNAs fell into regions with proper signal quantification, resulting in a final dataset of 391 tDNAs with associated Repli-Seq values.

WGS datasets for somatic mutation analysis

For the cancer samples, somatic single nucleotide variants (SNVs) were identified using WGS data collected from six different datasets representing over 20 distinct cancer types. The datasets included the Pan-cancer Analysis of Whole Genomes (PCAWG) study (n = 1950), Hartwig Medical Foundation (HMF) project (n = 4823), Personal Oncogenomics (POG) project (n = 570), Cancer Genome Atlas (TCGA) (n = 724), Clinical Proteomic Tumor Analysis Consortium (CPTAC) (n = 781), and the MMRF COMMPASS project (n = 758). Detailed data information, somatic variant calling methodology, and data processing are described in the Methods section ("WGS Mutation Data Collection and Processing," Salvadores and Supek, 2024). Somatic variants were identified using the hg19 reference genome. For the analysis of healthy tissues, SNVs were obtained from single-cell-derived colonies from four different datasets (including the human brain, colon, liver, and lung), with a total of 1192 samples (Blokzijl et al., 2016; Brunner et al., 2019; Lodato et al., 2015; Yoshida et al., 2020).

Genomic data by gene type for somatic mutation analysis

The genomic coordinates of each tDNA were obtained using the UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert the previously obtained tRNA-scanSE 2.0 annotations from the T2T-CHM13 assembly to hg19. This approach utilizes the completeness of the T2T-CHM13 assembly, which is gapless and corrects previous uncertainty and scaffold mistakes, ensuring that the analysis benefits from the most accurate and comprehensive data available. A total of 537 tDNAs were identified in hg19.

The data for other Pol III-transcribed genes (excluding tDNAs) was extracted from Pol3Base (Cai et al., 2022). Some genes in the database appeared in multiple experimental datasets, for that reason they were merged to maintain a set of unique genes. The Alu genes were also discarded. To convert the coordinates from hg38 to hg19 and to classify the genes by biotypes BioMart (https://www.ensembl.org/info/data/biomart/index.html) was used. Finally, pol III-transcribed genes used for the analysis were classified into specific biotypes: small nuclear RNA genes (RNUs), 5S ribosomal RNA genes (rDNAs) and unclassified/miscellaneous RNA genes (miscRNA genes), with a total of 106, 125, 80 genes, respectively.

Protein-coding gene coordinates for GRCh37/hg19 were retrieved from Ensembl. Gene expression data, provided as adjusted Transcripts Per Million (TPM) values, were downloaded from Hartwig Medical Foundation (Priestley et al., 2019). To categorize protein-coding genes according to their expression, we divided the genes into four groups based on their adjusted TPM values. These divisions were defined by quartiles: genes with TPM values up to the 25th percentile were designated as Level 1 (lowest expression); those between the 25th and 50th percentiles as Level 2; genes between the 50th and 75th percentiles as Level 3; and genes with TPM values above the 75th percentile were classified as Level 4 (highest expression).

Mutation density

Analysis was performed using the hg19 reference genome. This procedure was used to analyze the mutational density in cancerous samples for tDNAs, protein-coding genes, and other Pol III-transcribed genes (rDNAs, RNUs, and miscRNA genes), as well as in non-cancerous samples for tDNAs. First, a set of quality filters was applied to the genomic data to ensure data quality. To minimize errors due to misalignment of short reads and select uniquely mappable regions, reducing potential mapping ambiguities, all regions in the genome defined in the 'CRG Alignability 75' track (Derrien et al., 2012) with an alignability < 1.0 were masked out. In addition, regions that are unstable when converting between GRCh37 and GRCh38 (Ormond et al., 2021) and the ENCODE blacklist of problematic regions of the genome (Amemiya et al., 2019) were removed. Somatic SNVs were analyzed for different gene types: tDNAs, protein-coding genes, and other Pol III-transcribed genes, along with their flanking regions (e.g., 10 kb upstream and downstream, divided into 100-nt windows, with the gene itself designated as 'window 0'). To normalize mutational densities by the tri-nucleotide context, we adjusted the mutation counts based on the local sequence composition within each window, ensuring that differences in

mutation density reflect the underlying mutational processes rather than biases in sequence composition. This involved calculating the frequency of each trinucleotide context and determining the mutation count for that context within each sample. By dividing the mutation count by the context's frequency, we obtained a context-specific mutational density.

Finally, we obtained the total mutation density for each window by adding the mutation densities across all trinucleotide contexts. The mutational density within each group (e.g., by transcription levels or gene type) was normalized to account for variations in the total mutation count across groups by dividing the observed mutation density in each window by the total mutation density of that group.

Analysis for mutational density in tDNAs by tissue and age

To assess the relative mutation density in each tissue type, we calculated the fold change (FC) by dividing the mutation density of the tDNAs (window 0) by the average mutation density in sets of 10 consecutive windows in the flanking regions (e.g., windows 1–10, 11–20, ..., 91–100), both upstream and downstream. Only tissue types with a sample size greater than 20 (n > 20) were included in the analysis **(Supp. Fig. 10)**. Additionally, samples from the colorectum and uterus were stratified into subgroups based on previously described genomic instability patterns. These subgroups included microsatellite instability samples (MSI), microsatellite stable (MSS) samples, and hypermutated samples (HYPER) due to POLE deficiency. For age-related analysis, we stratified the data by tissue; only tissues with a sample size greater than 200 (n > 200) were considered **(Supp. Fig. 10)**. We only included the age group above 30 (<30) years for sarcoma, brain, and lymphoid cancers, as these were the only cancers with a sample size greater than 20. To assess FC by age, the same approach was used to compare the mutational density between tDNAs and flanking. Spearman's correlation was used to assess the relationship between mutation levels and age by tissue. Thyroid samples were removed from the analysis since the data was noisy.

Mutation frequencies in tDNAs at single base resolution

To analyze mutation patterns along tRNA sequences and identify mutation hotspots (specific nucleotides within tRNA sequences that accumulate mutations more frequently) the mutation frequency at each position in each tDNA was calculated. This was done by dividing the number of samples exhibiting mutations at a given position by the total number of samples analyzed. The set of tDNA positions that did not pass the quality filters, including alignability < 1.0, regions that are unstable when converting between GRCh37 and GRCh38, and ENCODE blacklist problematic regions, were removed from the analysis. For each tDNA the genomic coordinates were adjusted to align with the consensus tRNA reference positions (Amemiya et al., 2019; Derrien et al., 2012; Ormond et al., 2021), ensuring accurate nomenclature and allowing for the effective grouping of results across different tRNAs. To determine positions with significantly higher or lower mutation frequencies, *P*-values were calculated using the Wilcoxon test, comparing mutational frequencies for each position from all the tRNAs against all other positions, with adjustments for multiple comparisons using the Benjamini-Hochberg method.

To assess the mutagenic load within the structural domains or regions within the tRNAs, a sliding window approach was applied. An 11-nucleotide window was moved along the tRNA sequence one nucleotide at a time, and the mutation frequencies within each window were combined to have a unique value by windows. The same statistical procedure used for the comparison by position was applied to calculate the differences by windows. In both analyses to consider a position or windows to be significant a threshold of adj. *P*-value < 0.01 and a mean higher or lower than the quantiles of 0.95 and 0.05 was considered. The mean mutational frequency at each position/windows across tDNAs was obtained and used to represent the results. To identify chimeric tRNAs, we first calculated the mutational density at the anticodon position (34, 35, 36) for tRNA-Ala, tRNA-Leu, tRNA-Ser, and tRNA-Tyr isodecoders. To detect specific mutations at the anticodon, absolute mutation counts and the annotation of base changes (e.g., C<G) were used to identify mutations occurring at the anticodon of tDNAs.

Mutational signature

To deconvolute the possible processes that cause mutations specifically at tRNAs, we applied the standard mutational signatures methodology (Alexandrov et al., 2020; Supek & Lehner, 2015), with several modifications. First, as input instead of calculating the mutation frequencies across the 96 trinucleotides using WGS, we calculated the tri-nucleotide mutation frequencies in subsets of specific sites of the genome, specifically in 4 groups of tRNA genes (divided by levels of gene expression) and their flanking regions (10 kb upstream and downstream, split in 4 chunks of 2.5 kb). As well as, 4 groups of protein coding genes (divided by levels of gene expression) and their flanking regions (10 kb upstream and downstream, split in 4 chunks of 2.5 kb). To account for confounders in the flanking regions, we removed 1 kb upstream of the tRNA (TSS) and all genes overlapping the flanking regions (i.e. protein coding exons and tRNAs).

Second, since we are studying a very small subset of the genome (e.g. tRNA genes), to gain statistical power we merged all mutations across different samples from the same cancer type, except samples that are MSI or POLE mutant that we kept as a different group. Additionally, we included as a separate group all somatic mutations from healthy samples. Third, because the genomic sites are different for each tri-nucleotide feature depending on the sample, for each sample and trinucleotide feature combination, we normalized the counts by the tri- nucleotide composition of each specific genomic subset.

In summary, the samples we considered for this analysis are a combination of tissue (cancer type or healthy), gene type (protein coding or tRNA), gene expression (levels 1 to 4) and position (tRNA, flanking upstream or downstream). To the matrix of normalized mutation frequencies across the 96 tri-nucleotides for these samples we applied the standard mutational signatures methodology (Alexandrov et al., 2020; Supek & Lehner, 2015).

Specifically, we first applied bootstrap resampling (R function UPmultinomial from package sampling) to the normalized mutation frequencies. Next, we applied the non-negative matrix factorization (NMF) algorithm (R function nmf from package NMF) to the bootstrapped matrices, testing different values of

the rank parameter (2 to 30), herein referred to as nFact. We repeated the bootstrapping and NMF 100 times for each nFact. We pooled all the results by nFact and performed a k-medoids clustering (R function pam from package cluster), with different number-of-clusters k values (2 to 30). We calculated the silhouette index value, a clustering quality score (which here measures, effectively, how reproducible are the NMF solutions across runs), for each clustering to select the best nFact and k values.

Based on the silhouette index (Second worst SI score) we selected the results from nFact=13 and k=13 **(Supp. Fig. 12)**. From the selected option signatures with SI < 0.4 were removed, and 12 signatures remained. These results include two matrices: H and W. The W matrix (our mutational signatures) describes the weight of each tri-nucleotide for every extracted signature. By comparing our signatures with the COSMIC database signatures (COSMIC_v3.3.1_SBS_GRCh37), we can match our signatures with the corresponding signature from COSMIC. Of note, we merged the two APOBEC signatures from COSMIC (SBS2+SBS13), since in our analyses they cannot be separated, likely because of our merging of samples. The H matrix describes the exposures or activities of each specific signature in every sample. Additionally, we stratified the exposures by tissue type and expression level to assess tissue-dependent and transcription-associated rate differences. To select relevant NMF-extracted signatures, different factors were examined. First, a cosine similarity threshold was applied to quantify the resemblance between NMF-extracted signatures and COSMIC signatures: cosine similarity values above 0.5 indicate moderate similarity, while values exceeding 0.85 denote strong similarity to COSMIC signatures. Additionally, we examined the signatures spectrum to confirm their consistency with the spectra reported by COSMIC Furthermore, signature sample exposures and their relevance to specific cancer types were evaluated to ensure that the identified signatures are biologically meaningful. To identify signatures enriched in tDNAs, we calculated the fold change (FC) in the signatures exposure/activity levels between each gene type (tDNAs or protein-coding genes) and their respective flanking regions.

Skin cancer samples were excluded from the analysis due to their distinct mutation profiles. Specifically, skin cancers exhibit a high prevalence of mutations caused by ultraviolet (UV) light exposure. These UV-induced mutations create distinctive signatures that can overshadow or confound the detection of other mutational processes when analyzing diverse cancer types.

Detection of hairpin structures in tDNAs

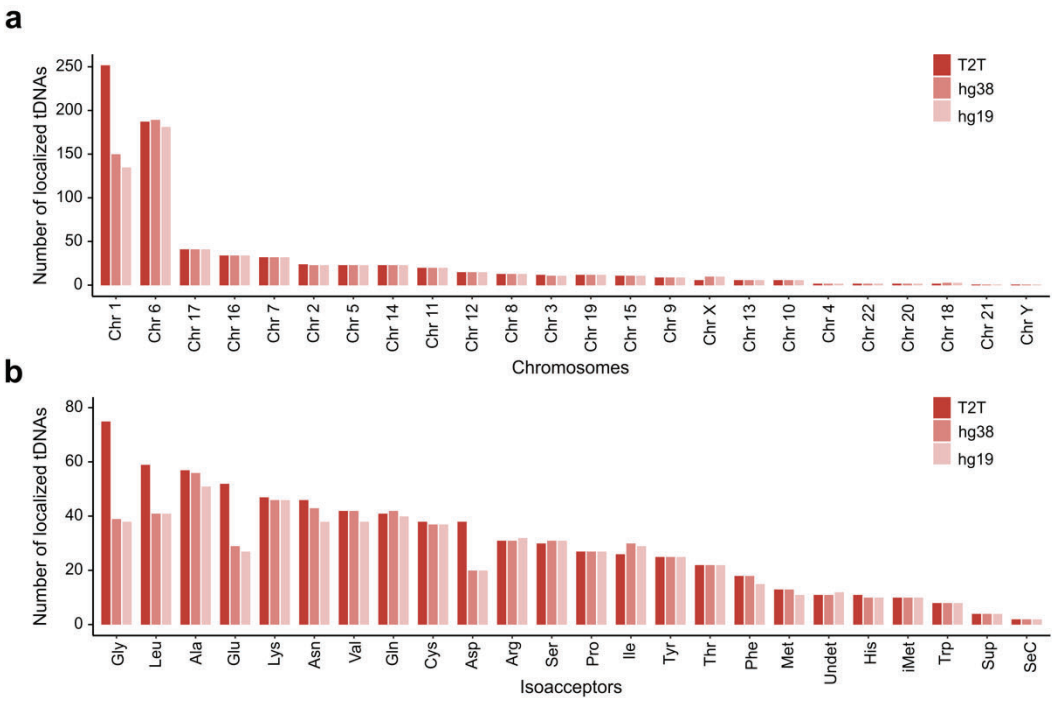
To assess the abundance of hairpin loop structures in tDNA genes relative to the rest of the genome and other Pol III-transcribed genes, we utilized a genome-wide dataset of predicted 3-5 nucleotide hairpin sites with APOBEC TCW motifs (Buisson et al., 2019). Using the GenomicRanges package in R, we identified overlaps between hairpin regions and genomic coordinates of annotated tDNAs genes in the hg19 version, as well as with a separate dataset of other Pol III-transcribed genes. From this, we quantified the number of nucleotides within tDNA genes, other Pol III genes, and the remainder of the genome that were part of the predicted hairpin structure. To calculate the frequencies of DNA hairpin structures in tDNAs, the number of nucleotides not involved in hairpin structures was obtained by subtracting hairpin-associated nucleotides from the total nucleotide content of each corresponding

genomic region (tDNA genes, other Pol III genes, and non-tDNA regions of the genome). To determine whether hairpin structures were statistically enriched or depleted in tDNA genes, we conducted two chi-square tests using the chisq.test() function in R: one comparing tDNA genes to the rest of the genome and another comparing tDNA genes to other Pol III-transcribed genes.

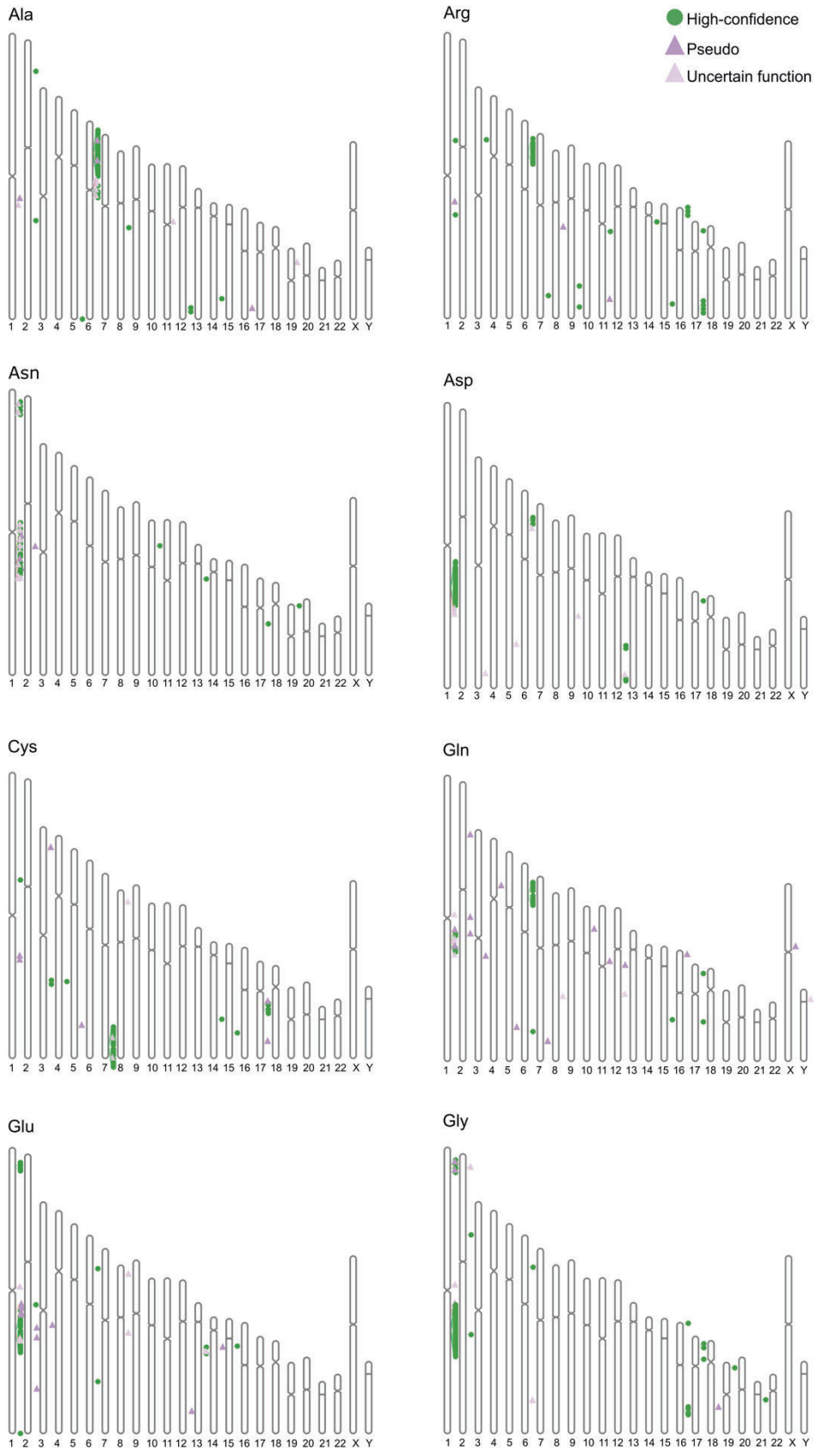
Software and packages

The mutational analyses were performed using R v.4.1.2. Relevant R packages used were liftOver v.1.18.0, GenomicRanges v.1.46.1, Biostrings v.2.62.0, sampling v.2.10, matrixStats v.1.4.1, NMF v.0.26, ComplexHeatmap v.2.10.0, cluster v.2.1.2, dplyr v.1.1.4, tidyr v.1.3.0, ggplot2 v.3.5.1.

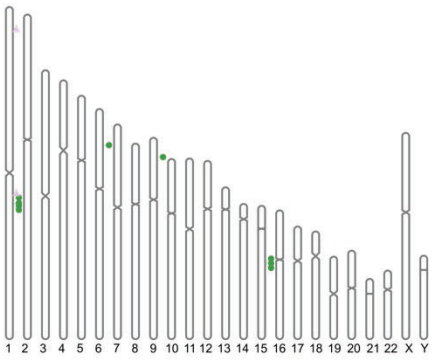
Supplementary



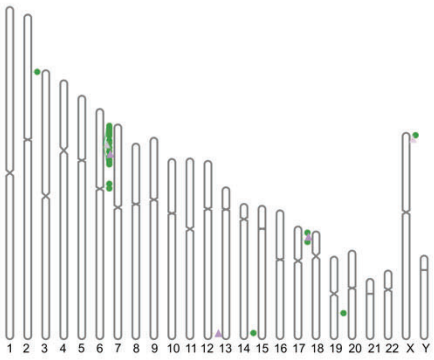
Supplementary Figure 1. Chromosomal and isoacceptor distributions of tDNAs detected by tRNAscan-SE. (a) Number of tDNAs per chromosome across the T2T-CHM13, hg38, and hg19 reference genomes. **(b)** Number of tDNAs by isoacceptor across the T2T-CHM13, hg38, and hg19 reference genomes.


Ala
Arg
High-confidence
Pseudo
Uncertain function
Asn
Asp
Cys
Gln
Glu
Gly
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

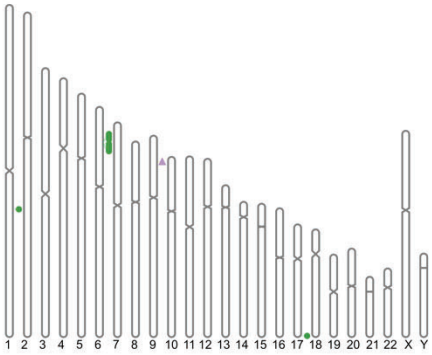
His



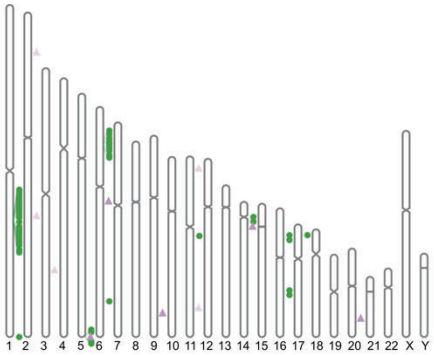
Ile



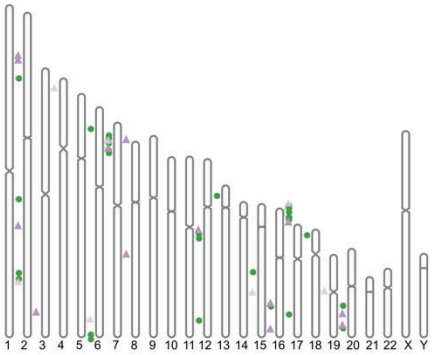
iMet



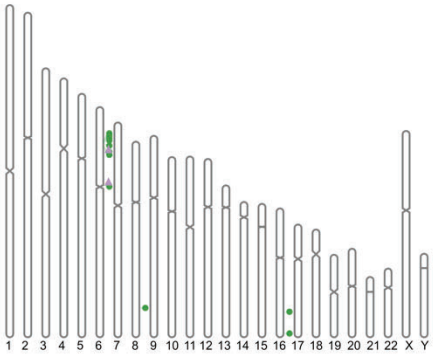
Leu



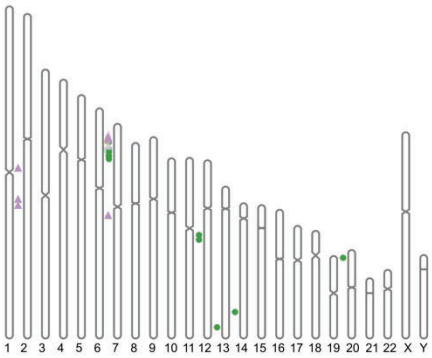
Lys



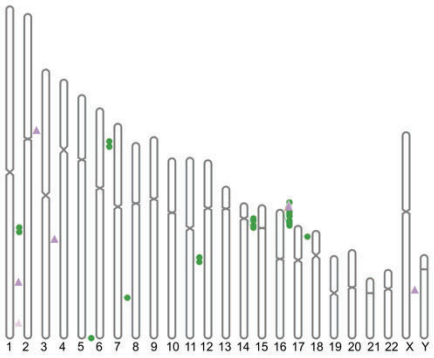
Met

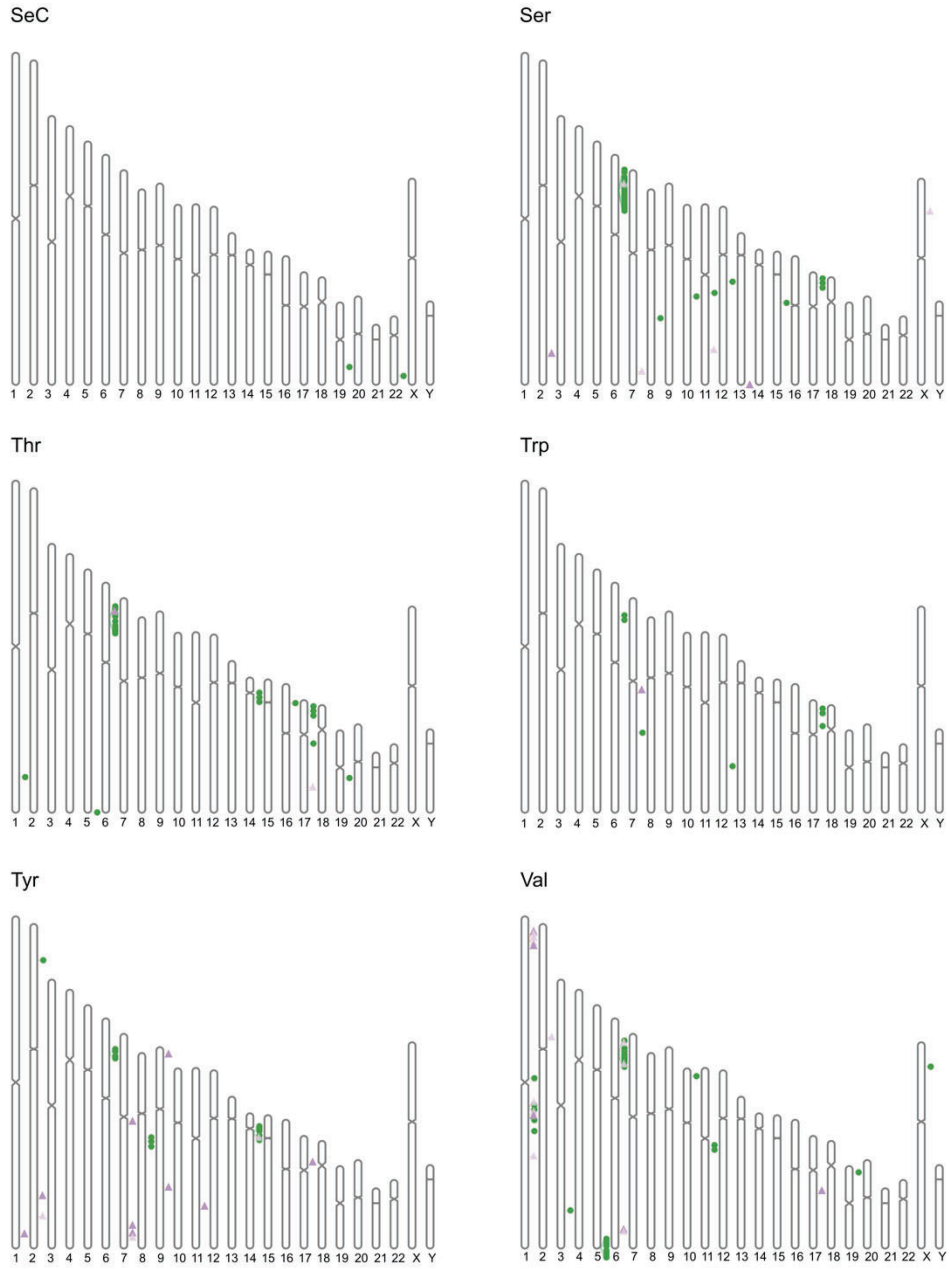


Phe

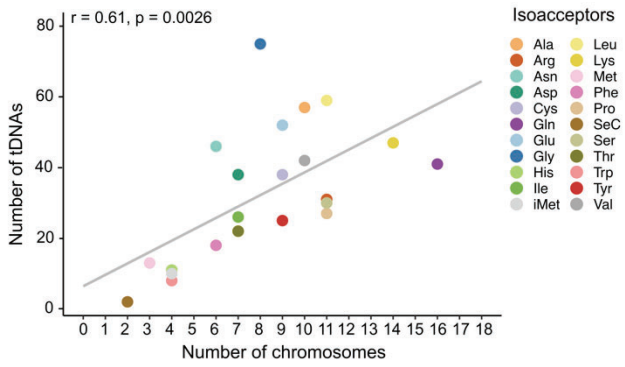


Pro

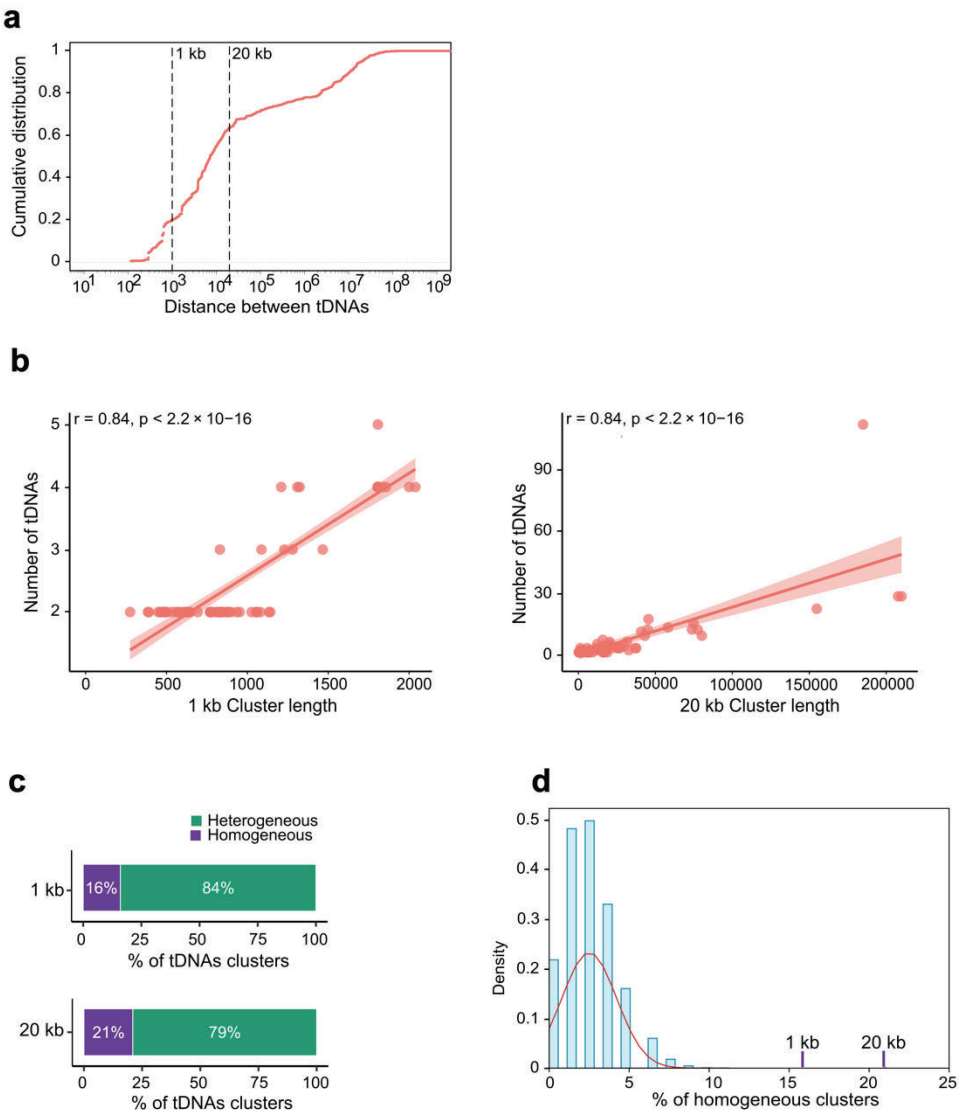




Supplementary Figure 2. Distribution of each tDNA isoacceptor in the human genome. The graphical representation is based on ideograms that indicate the chromosomal location of each tDNA for each isoacceptor across the human genome, using the T2T-CHM13 assembly as a reference. Prediction confidence from the T2T-CHM13 annotation obtained with tRNAscan-SE is described: high-confidence genes (green), pseudogenes (light purple), and genes of uncertain function (dark purple).

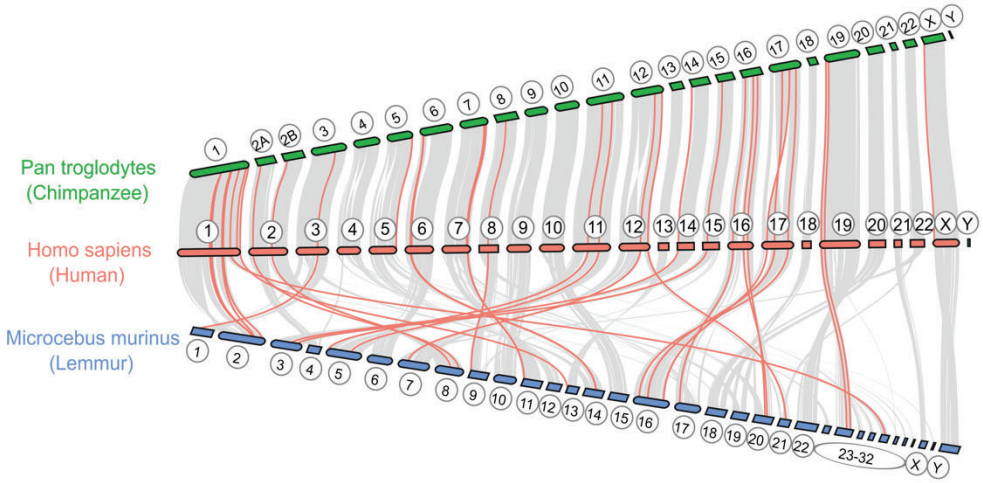


Supplementary Figure 3. Correlation between tDNA isoacceptor copy number and chromosomal occurrence. The plots display the Spearman correlation coefficient (r) and associated *P*-value (p).

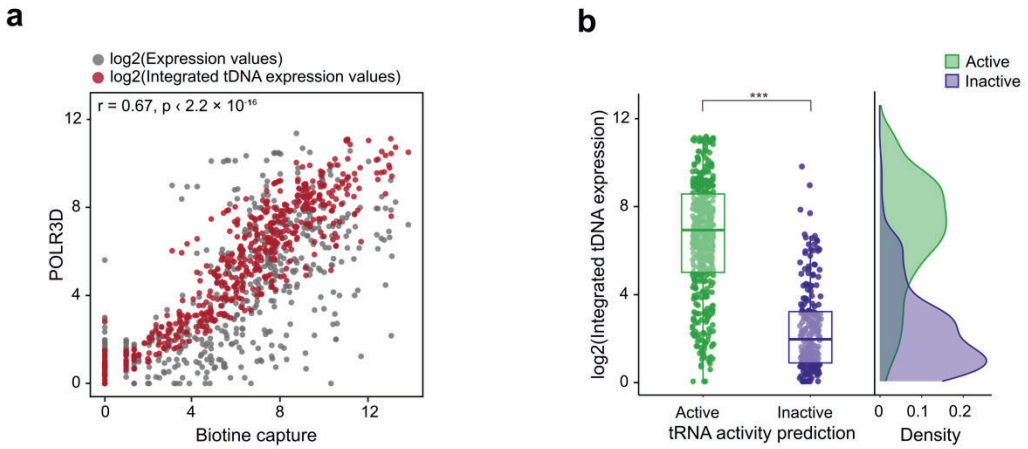


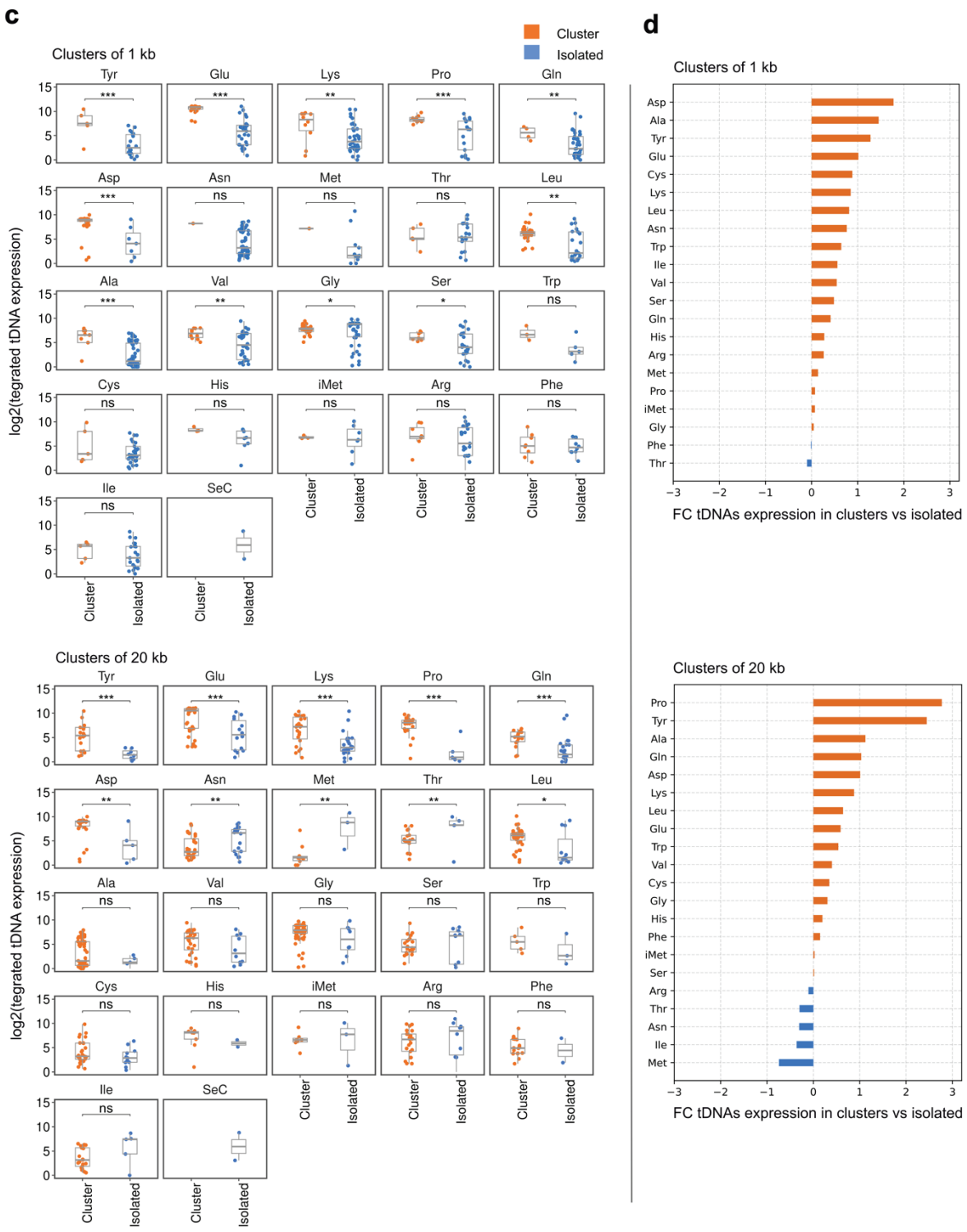
Supplementary Figure 4. Characterization of tDNAs clusters in the T2T-CHM13 genome. (a) Cumulative distribution of tDNA pair distances for *Homo sapiens* (human). The selected distances for cluster definition were 1 and 20 kb. **(b)** Correlation between the number of tDNAs in each cluster and

cluster length used to assess tDNA density. On the right, results for 1 kb clusters, on the left, for 20 kb clusters. The plots include the Spearman correlation coefficient (r) and associated *P-values* (p). **(c)** Observed percentage of heterogeneous clusters (containing more than one isoacceptor) and homogeneous clusters (composed of the same isoacceptor). **(d)** Results of a randomization test performed 1000 times: tDNAs were randomly reassigned to clusters, and the percentage of homogeneous clusters was calculated for each test to evaluate whether the observed distribution deviated from random chance. The purple bars show the results obtained for the observed percentage of tDNAs in the clusters.



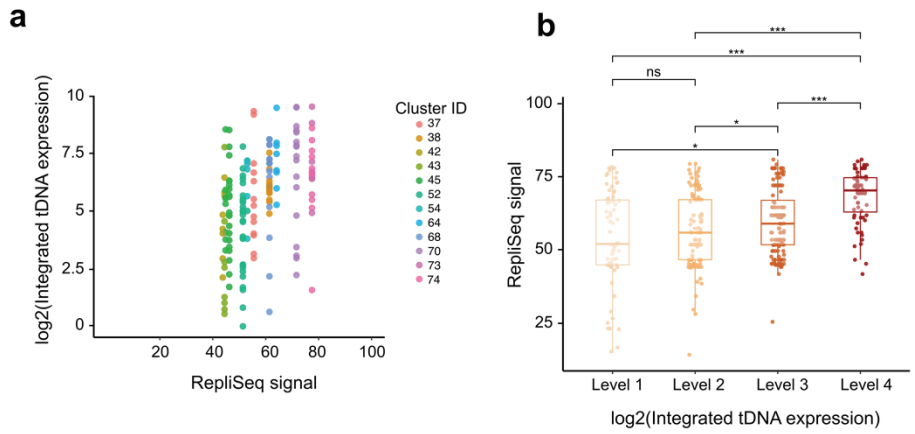
Supplementary Figure 5. Synteny analysis of tDNAs clusters between different primate species. Synteny analysis highlighting the conservation of tDNA 20kb clusters between species *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), and *Microcebus murinus* (lemur). Red lines denote conserved syntenic blocks of tDNA clusters, and gray lines indicate other homologous relationships. Chromosomal numbers are provided for each species.



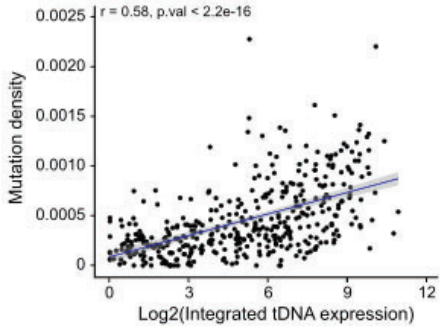


Supplementary Figure 6. Integrated tDNA expression analysis in cluster and isolated tDNAs. (a) Correlation between POLR3D and biotin-capture data, both representing nascent tRNA transcript levels. The Spearman correlation coefficient (r) and associated *P*-value (p) are shown in the plot. Integrated tDNA expression values reflect a combination of POLR3D and biotin capture signals (see Methods). **(b)** Comparison of integrated tDNA expression values between tDNAs classified as active and inactive based on activity predictions from tRAP. Wilcoxon rank-sum test (one-sided with alternative hypothesis: greater; ***: *P*-value ≤ 0.001). **(c)** Box plots of log2 integrated tDNA expression values, comparing clustered and isolated tDNAs by isoacceptor; clusters of 1 and 20 kb are shown on the left and right, respectively. Wilcoxon rank-sum test (one-sided with alternative hypothesis: greater) with Benjamini–Hochberg adjusted P-values (***: *P*-value ≤ 0.001; **: *P*-value ≤ 0.01; *: *P*-value ≤ 0.05; ns: P-value >

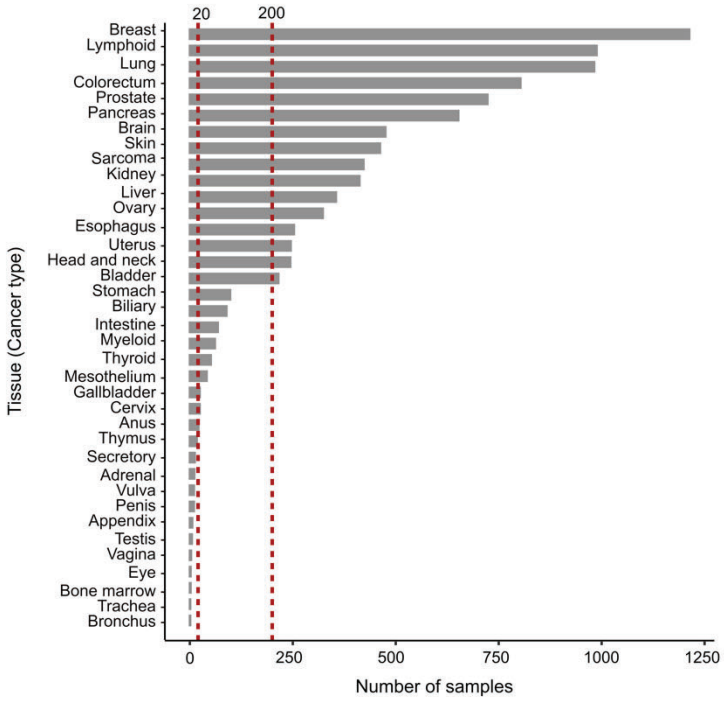
0.05). **(d)** Fold change (FC) of expression values between clustered and isolated tDNAs by isoacceptor. Orange bars indicate when expression levels are higher in clustered tDNAs than in isolated tDNAs.



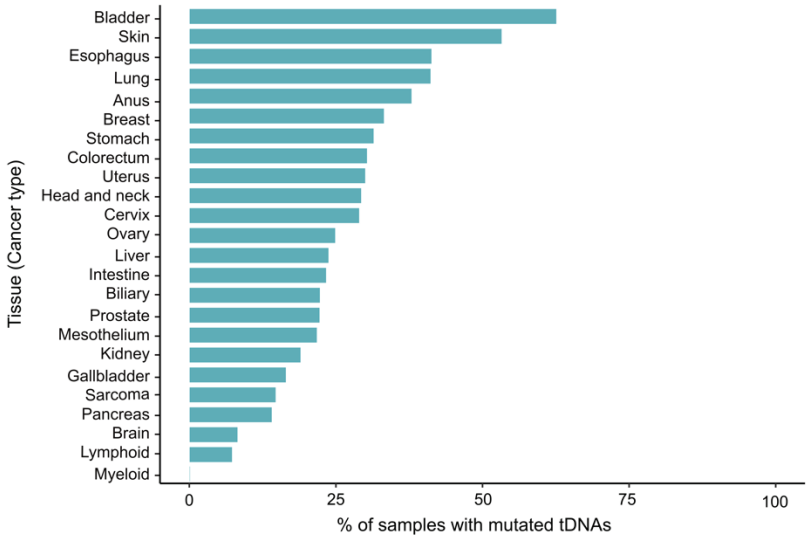
Supplementary Figure 7. Relationship between tDNA replication timing, clustering, and expression. (a) Repli-Seq values and log2 integrated tDNA expression values across tDNAs by cluster. **(b)** Repli-Seq values across tDNA expression levels, from the lowest (Level 1) to the highest (Level 4) (see Methods). Statistical significance was assessed using the two-sided Wilcoxon rank-sum test (one-sided with alternative hypothesis: greater) with Benjamini–Hochberg-adjusted *P*-values (***: *P*-value ≤ 0.001; *: *P*-value ≤ 0.05; ns: *P*-value > 0.05).



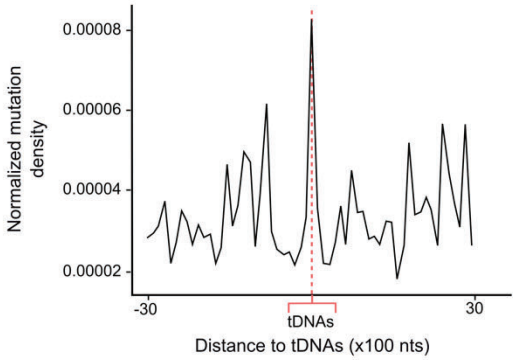
Supplementary Figure 8. Correlation between tDNA mutagenesis and tRNA expression. The Spearman correlation coefficient (r) and associated *P*-value (p) are shown in the plot. Integrated tDNA expression values reflect a combination of POLR3D and biotin-capture signals.



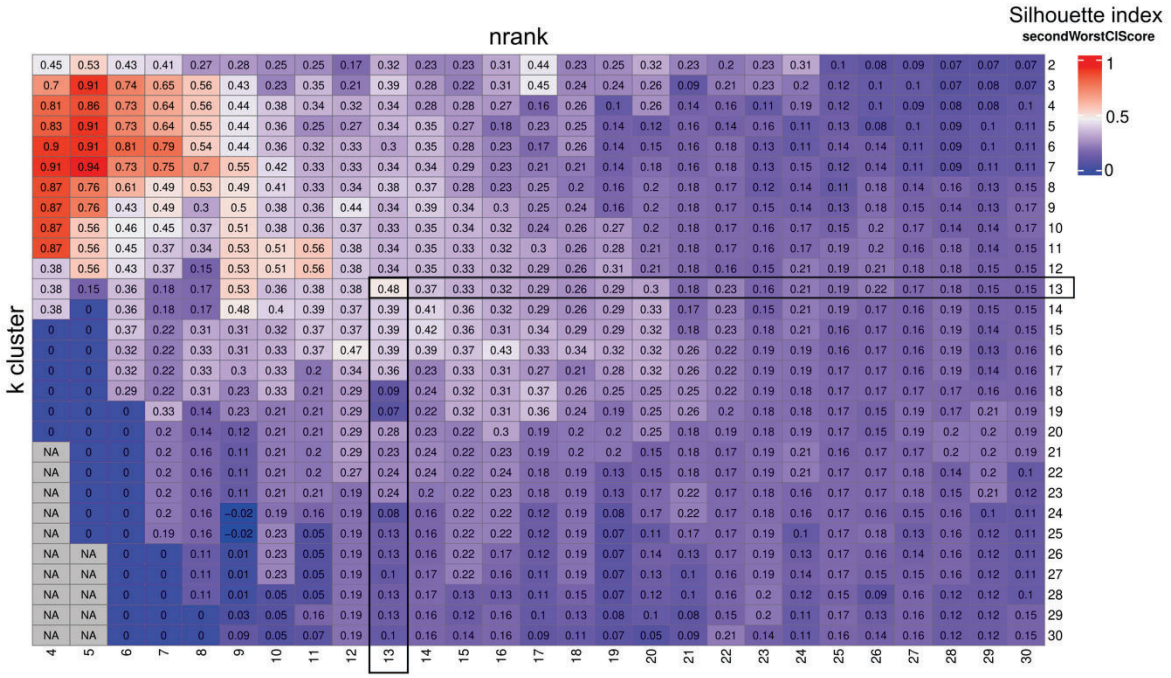
Supplementary Figure 9. Cancer samples metadata. Metadata showing the number of cancer samples analyzed according to tissue type. For the analysis of the mutagenesis profile in tDNAs by cancer type, we filtered the data and selected only those cancer types with more than 20 samples. For the tissue-age analysis, we used cancer types with at least 200 samples per type.



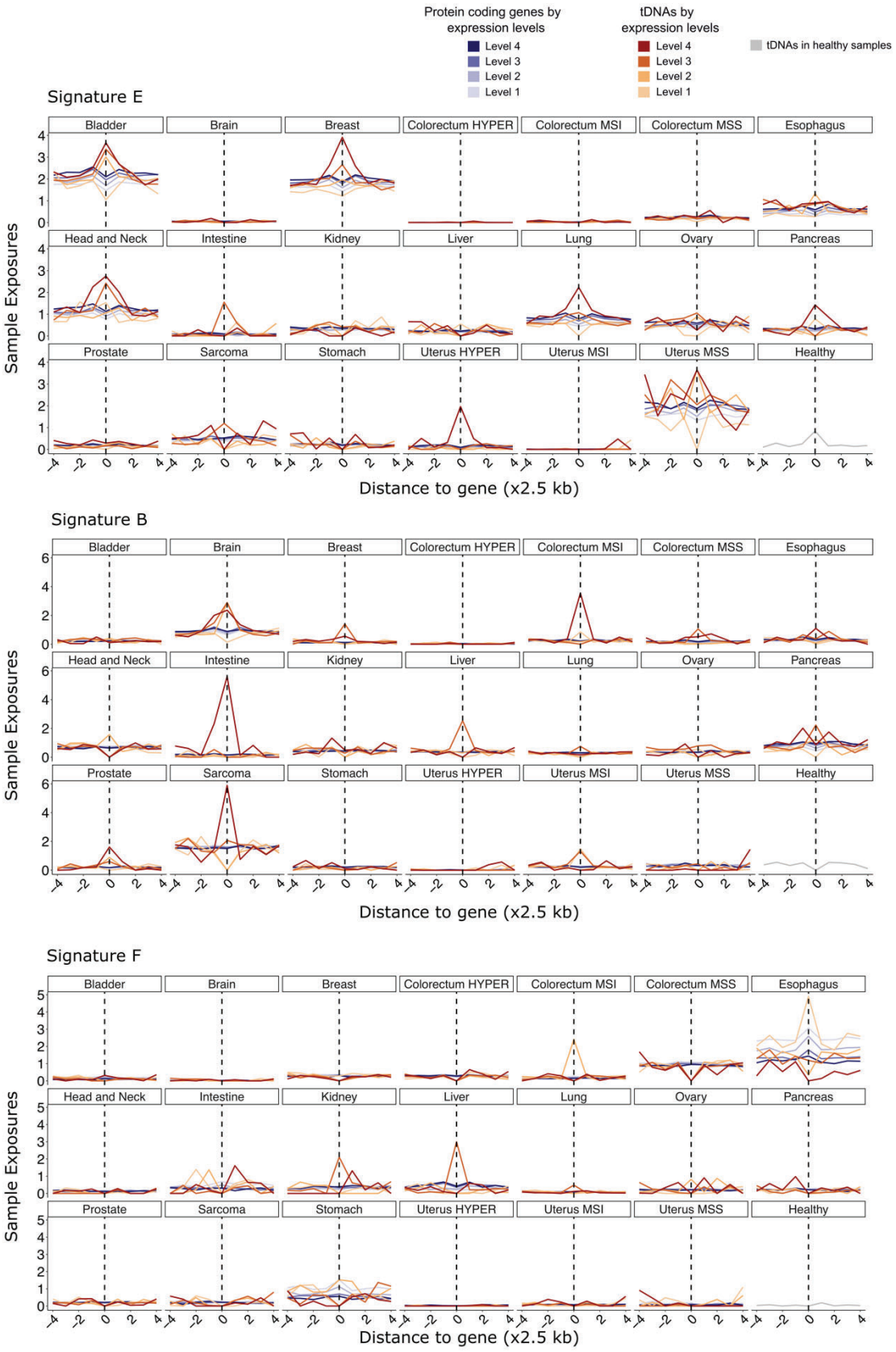
Supplementary Figure 10. Somatic mutational profile in tDNAs by cancer type. Percentage of samples that at least have one tDNA mutated for each cancer type. Only cancer types with more than 20 samples were analyzed.

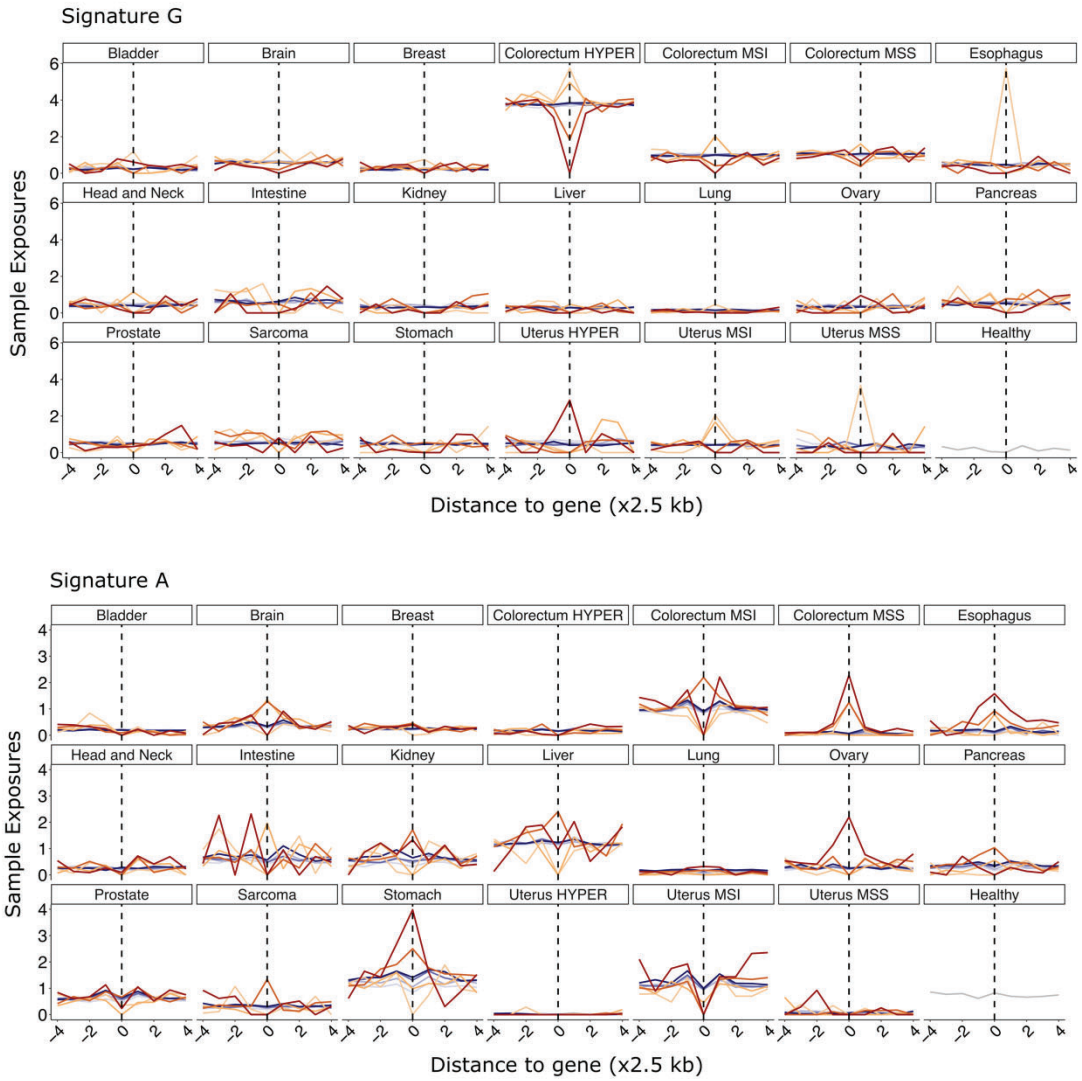


Supplementary Figure 11. Somatic mutagenesis analysis in tDNAs of healthy tissues. Normalized mutation density in tDNAs from healthy non-cancerous tissues (flanking regions of 30 kb downstream and upstream divided into windows of 100 nt).

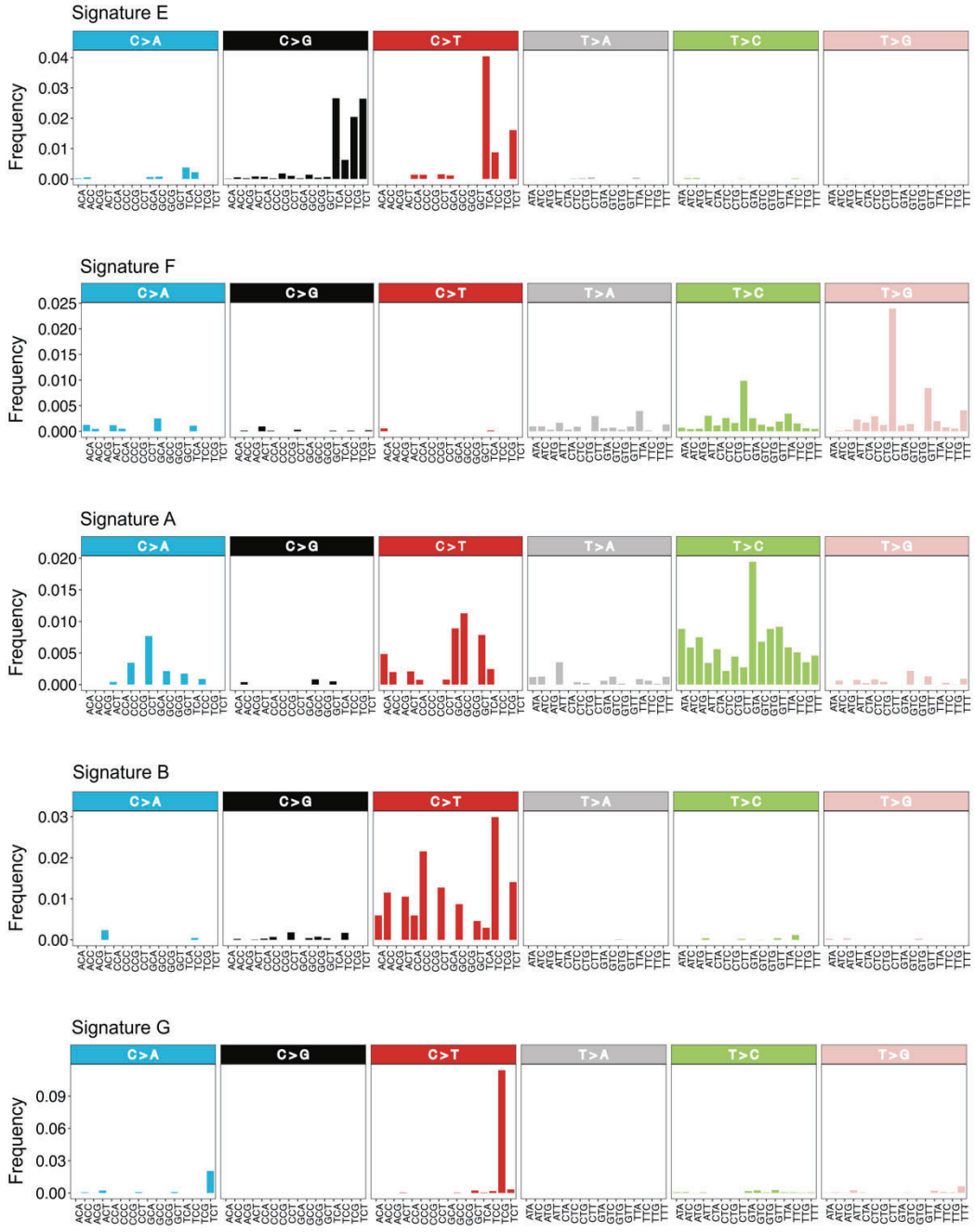


Supplementary Figure 12. De novo analysis of mutation signatures using NMF analysis. Silhouette Index (SI) scores used to evaluate the quality of clustering in NMF analysis. The selected case (nFact = 13, k = 13).

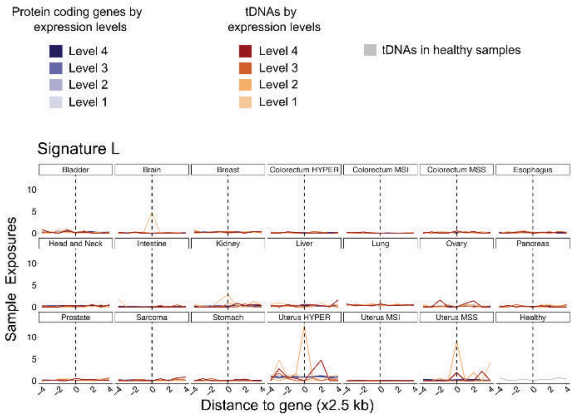

Protein coding genes by expression levels
Level 4
Level 3
Level 2
Level 1
tDNAs by expression levels
Level 4
Level 3
Level 2
Level 1
tDNAs in healthy samples
Signature E
Bladder
Brain
Breast
Colorectum HYPER
Colorectum MSI
Colorectum MSS
Esophagus
Head and Neck
Intestine
Kidney
Liver
Lung
Ovary
Pancreas
Prostate
Sarcoma
Stomach
Uterus HYPER
Uterus MSI
Uterus MSS
Healthy
Sample Exposures
Distance to gene (x2.5 kb)
Signature B
Sample Exposures
Distance to gene (x2.5 kb)
Signature F
Sample Exposures
Distance to gene (x2.5 kb)



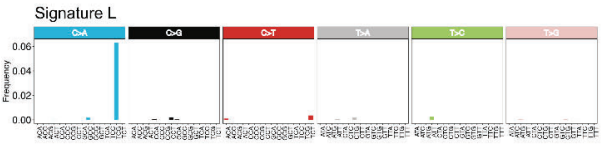
Supplementary Figure 13. Sample exposures for the selected NMF-extracted signatures. The plots include exposures for each gene type (tDNAs and protein-coding genes) and their flanking regions (10 kb upstream and downstream, split into four chunks of 2.5 kb). The exposures for each NMF-extracted signature were obtained for tDNAs and protein-coding genes by expression levels in cancer samples and for tDNAs in healthy samples.

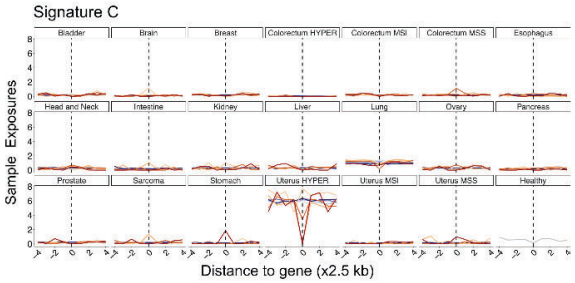


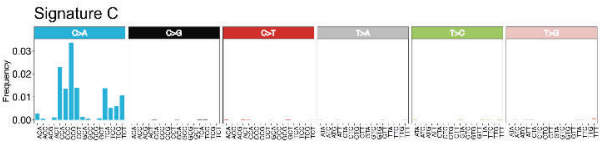
Supplementary Figure 14. Mutational spectrum profile for the selected NMF-extracted signatures. Substitution types: C>A, C>G, C>T, T>A, T>C, and T>G. And their context classification of 96 distinct mutation types.

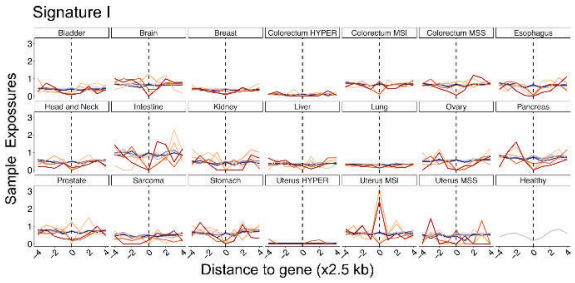

Protein coding genes by expression levels
Level 4
Level 3
Level 2
Level 1
tDNAs by expression levels
Level 4
Level 3
Level 2
Level 1
tDNAs in healthy samples

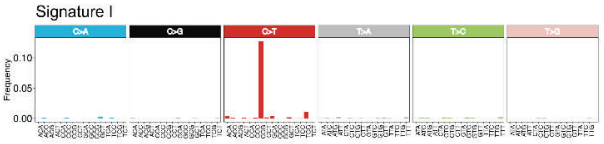
Signature L
Bladder
Brain
Breast
Colorectum HYPER
Colorectum MSI
Colorectum MSS
Esophagus
Head and Neck
Intestine
Kidney
Liver
Lung
Ovary
Pancreas
Prostate
Sarcoma
Stomach
Uterus HYPER
Uterus MSI
Uterus MSS
Healthy
Sample Exposures
Distance to gene (x2.5 kb)

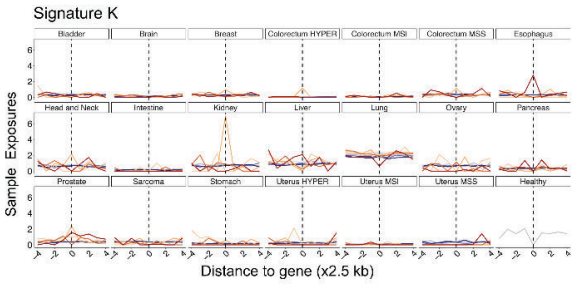

Signature L
C>A
C>G
C>T
T>A
T>C
T>G
Frequency

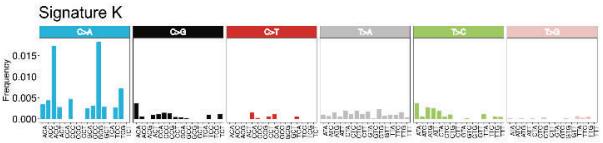

Signature C
Bladder
Brain
Breast
Colorectum HYPER
Colorectum MSI
Colorectum MSS
Esophagus
Head and Neck
Intestine
Kidney
Liver
Lung
Ovary
Pancreas
Prostate
Sarcoma
Stomach
Uterus HYPER
Uterus MSI
Uterus MSS
Healthy
Sample Exposures
Distance to gene (x2.5 kb)

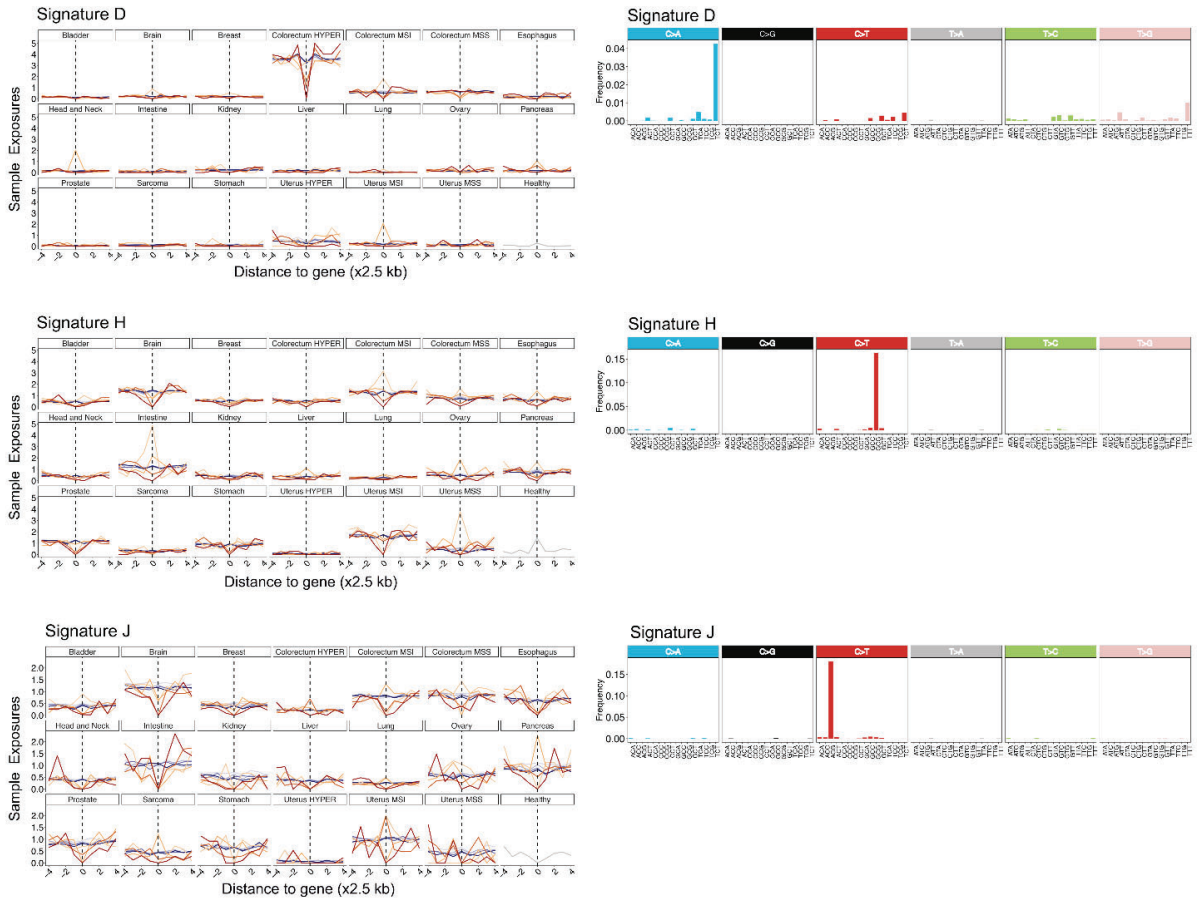

Signature C
C>A
C>G
C>T
T>A
T>C
T>G
Frequency


Signature I
Bladder
Brain
Breast
Colorectum HYPER
Colorectum MSI
Colorectum MSS
Esophagus
Head and Neck
Intestine
Kidney
Liver
Lung
Ovary
Pancreas
Prostate
Sarcoma
Stomach
Uterus HYPER
Uterus MSI
Uterus MSS
Healthy
Sample Exposures
Distance to gene (x2.5 kb)

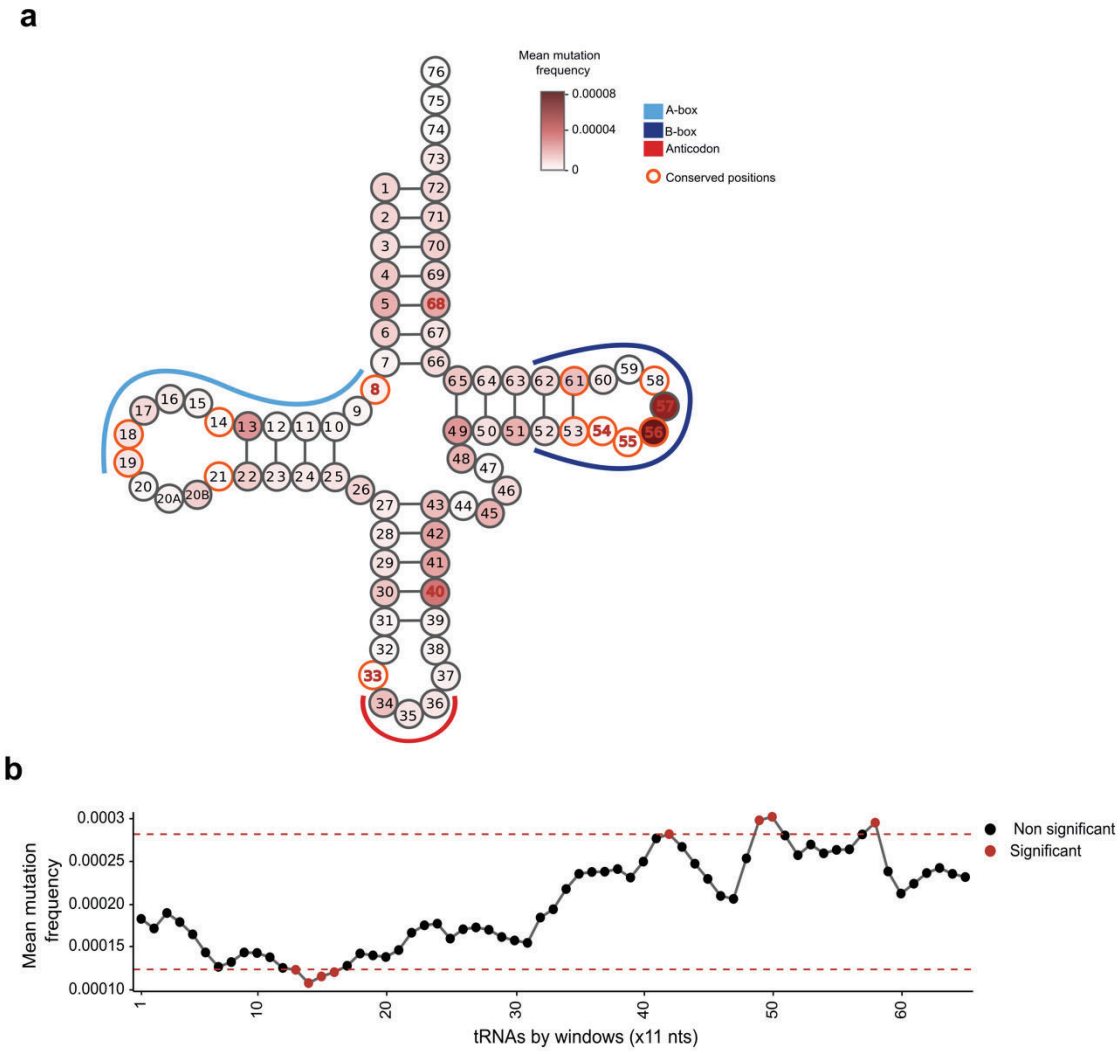

Signature I
C>A
C>G
C>T
T>A
T>C
T>G
Frequency


Signature K
Bladder
Brain
Breast
Colorectum HYPER
Colorectum MSI
Colorectum MSS
Esophagus
Head and Neck
Intestine
Kidney
Liver
Lung
Ovary
Pancreas
Prostate
Sarcoma
Stomach
Uterus HYPER
Uterus MSI
Uterus MSS
Healthy
Sample Exposures
Distance to gene (x2.5 kb)


Signature K
C>A
C>G
C>T
T>A
T>C
T>G
Frequency



Supplementary Figure 15. Sample exposures and mutational spectrum for the discarded NMF-extracted signatures. The plots include exposures for each gene type (tDNAs and protein-coding genes) and their flanking regions (10 kb upstream and downstream, split into four chunks of 2.5 kb). The exposures for each NMF-extracted signature were obtained for tDNAs and protein-coding genes by expression levels in cancer samples and for tDNAs in healthy samples. Mutational spectra for the selected NMF-extracted signatures. Substitution types: C>A, C>G, C>T, T>A, T>C, and T> G, and their context classification comprised 96 distinct mutation types.



Supplementary Figure 16. Distribution of somatic mutations across the tRNA sequence (a) Somatic tDNA mutations at single-base resolution represented in the tRNA cloverleaf secondary structure. Mean somatic mutation frequency is indicated for each position. Highlighting universally conserved positions (U8, A14, G18, G19, A21, U33, G53, T54, U55, C56, A58, C61, C74, C75), internal promoter regions such as the A-box (positions 8-19), the B-box (positions 52-62), and the anticodon (positions 34-36). Statistically significant positions are indicated with red numbers. Specific tDNA positions are considered significant when adj. p ≤ 0.01 and if its mean mutation density exceeds the 0.95 quantile or falls below the 0.05 quantile of the overall distribution. **(b)** Mutation rates computed using a sliding window of 11 nucleotides across the tDNA sequences. Statistically significant positions are indicated with red dots. Specific tDNA positions are considered significant when adj. p ≤ 0.01 and if its mean mutation density exceeds the 0.95 quantile or falls below the 0.05 quantile of the overall distribution (quantile thresholds are indicated by red dashed lines).

References

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., PCAWG Mutational Signatures Working Group, … PCAWG Consortium. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101.

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE blacklist: Identification of problematic regions of the genome. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-45839-z

Anisimova, A. S., Alexandrov, A. I., Makarova, N. E., Gladyshev, V. N., & Dmitriev, S. E. (2018). Protein synthesis and quality control in aging. *Aging*, *10*(12), 4269–4288.

Berg, M. D., Giguere, D. J., Dron, J. S., Lant, J. T., Genereaux, J., Liao, C., Wang, J., Robinson, J. F., Gloor, G. B., Hegele, R. A., O'Donoghue, P., & Brandl, C. J. (2019). Targeted sequencing reveals expanded genetic diversity of human transfer RNAs. *RNA Biology*, *16*(11), 1574–1585.

Bermudez-Santana, C., Attolini, C. S.-O., Kirsten, T., Engelhardt, J., Prohaska, S. J., Steigele, S., & Stadler, P. F. (2010). Genomic organization of eukaryotic tRNAs. *BMC Genomics*, *11*(1), 270.

Besedina, E., & Supek, F. (2024). Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver mutations. *Nature Communications*, *15*(1), 6139.

Beuning, P. J., & Musier-Forsyth, K. (1999). Transfer RNA recognition by aminoacyl-tRNA synthetases. *Biopolymers*, *52*(1), 1–28.

Bezerra, A. R., Simões, J., Lee, W., Rung, J., Weil, T., Gut, I. G., Gut, M., Bayés, M., Rizzetto, L., Cavalieri, D., Giovannini, G., Bozza, S., Romani, L., Kapushesky, M., Moura, G. R., & Santos, M. A. S. (2013). Reversion of a fungal genetic code alteration links proteome instability with genomic and phenotypic diversification. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(27), 11079–11084.

Biela, A., Hammermeister, A., Kaczmarczyk, I., Walczak, M., Koziej, L., Lin, T.-Y., & Glatt, S. (2023). The diverse structural modes of tRNA binding and recognition. *The Journal of Biological Chemistry*, *299*(8), 104966.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., Nijman, I. J., Martincorena, I., Mokry, M., Wiegerinck, C. L., Middendorp, S., Sato, T., Schwank, G., Nieuwenhuis, E. E. S., Verstegen, M. M. A., … van Boxtel, R. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, *538*(7624), 260–264.

Brunner, S. F., Roberts, N. D., Wylie, L. A., Moore, L., Aitken, S. J., Davies, S. E., Sanders, M. A., Ellis, P., Alder, C., Hooks, Y., Abascal, F., Stratton, M. R., Martincorena, I., Hoare, M., & Campbell, P. J. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature*, *574*(7779), 538–542.

Buisson, R., Langenbucher, A., Bowen, D., Kwan, E. E., Benes, C. H., Zou, L., & Lawrence, M. S. (2019). Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science (New York, N.Y.)*, *364*(6447), eaaw2872.

Burns, M. B., Temiz, N. A., & Harris, R. S. (2013). Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nature Genetics*, *45*(9), 977–983.

Cabrelle, C., Giorgi, F. M., & Mercatelli, D. (2024). Quantitative and qualitative detection of tRNAs, tRNA halves and tRFs in human cancer samples: Molecular grounds for biomarker development and clinical perspectives. *Gene*, *898*, 148097.

Cagan, A., Baez-Ortega, A., Brzozowska, N., Abascal, F., Coorens, T. H. H., Sanders, M. A., Lawson,

A. R. J., Harvey, L. M. R., Bhosle, S., Jones, D., Alcantara, R. E., Butler, T. M., Hooks, Y., Roberts, K., Anderson, E., Lunn, S., Flach, E., Spiro, S., Januszczak, I., … Martincorena, I. (2022). Somatic mutation rates scale with lifespan across mammals. *Nature*, *604*(7906), 517–524.

Cai, L., Xuan, J., Lin, Q., Wang, J., Liu, S., Xie, F., Zheng, L., Li, B., Qu, L., & Yang, J. (2022). Pol3Base: a resource for decoding the interactome, expression, evolution, epitranscriptome and disease variations of Pol III-transcribed ncRNAs. *Nucleic Acids Research*, *50*(D1), D279–D286.

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096.

Chan, P. P., & Lowe, T. M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Research*, *44*(D1), D184-9.

Christensen, S., Van der Roest, B., Besselink, N., Janssen, R., Boymans, S., Martens, J., Yaspo, M.-L., Priestley, P., Kuijk, E., Cuppen, E., Van Hoeck, A., & Center for Personalized Cancer Treatment. (2019). 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. In *bioRxiv*. https://doi.org/10.1101/681262

Dammann, R., & Pfeifer, G. P. (1997). Lack of gene- and strand-specific DNA repair in RNA polymerase III-transcribed human tRNA genes. *Molecular and Cellular Biology*, *17*(1), 219–229.

Degasperi, A., Zou, X., Amarante, T. D., Martinez-Martinez, A., Koh, G. C. C., Dias, J. M. L., Heskin, L., Chmelova, L., Rinaldi, G., Wang, V. Y. W., Nanda, A. S., Bernstein, A., Momen, S. E., Young, J., Perez-Gil, D., Memari, Y., Badja, C., Shooter, S., Czarnecki, J., … Nik-Zainal, S. (2022). Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science (New York, N.Y.)*, *376*(6591). https://doi.org/10.1126/science.abl9283

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome mappability. *PloS One*, *7*(1), e30377.

Dittmar, K. A., Goodenbour, J. M., & Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. *PLoS Genetics*, *2*(12), e221.

Earnest-Noble, L. B., Hsu, D., Chen, S., Asgharian, H., Nandan, M., Passarelli, M. C., Goodarzi, H., & Tavazoie, S. F. (2022). Two isoleucyl tRNAs that decode synonymous codons divergently regulate breast cancer metastatic growth by controlling translation of proliferation-regulating genes. *Nature Cancer*, *3*(12), 1484–1497.

Gao, L., Behrens, A., Rodschinka, G., Forcelloni, S., Wani, S., Strasser, K., & Nedialkova, D. D. (2024). Selective gene expression maintains human tRNA anticodon pools during differentiation. *Nature Cell Biology*, *26*(1), 100–112.

García-Vílchez, R., Añazco-Guenkova, A. M., López, J., Dietmann, S., Tomé, M., Jimeno, S., Azkargorta, M., Elortza, F., Bárcena, L., Gonzalez-Lopez, M., Aransay, A. M., Sánchez-Martín, M. A., Huertas, P., Durán, R. V., & Blanco, S. (2023). N7-methylguanosine methylation of tRNAs regulates survival to stress in cancer. *Oncogene*, *42*(43), 3169–3181.

Geslain, R., Cubells, L., Bori-Sanz, T., Alvarez-Medina, R., Rossell, D., Martí, E., & Ribas de Pouplana, L. (2010). Chimeric tRNAs as tools to induce proteome damage and identify components of stress responses. *Nucleic Acids Research*, *38*(5), e30.

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

Giegé, Richard, & Eriani, G. (2023). The tRNA identity landscape for aminoacylation and beyond. *Nucleic Acids Research*, *51*(4), 1528–1570.

Giegé, Richard, Jühling, F., Pütz, J., Stadler, P., Sauter, C., & Florentz, C. (2012). Structure of transfer RNAs: similarity and variability. *Wiley Interdisciplinary Reviews. RNA*, *3*(1), 37–61.

Gingold, H., Tehler, D., Christoffersen, N. R., Nielsen, M. M., Asmar, F., Kooistra, S. M., Christophersen, N. S., Christensen, L. L., Borre, M., Sørensen, K. D., Andersen, L. D., Andersen, C. L., Hulleman, E., Wurdinger, T., Ralfkiær, E., Helin, K., Grønbæk, K., Ørntoft, T., Waszak, S. M., … Pilpel, Y. (2014). A dual program for translation regulation in cellular proliferation and differentiation. *Cell*, *158*(6), 1281–1292.

Gonzalez-Perez, A., Sabarinathan, R., & Lopez-Bigas, N. (2019). Local determinants of the mutational landscape of the human genome. *Cell*, *177*(1), 101–114.

Goodarzi, H., Nguyen, H. C. B., Zhang, S., Dill, B. D., Molina, H., & Tavazoie, S. F. (2016). Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell*, *165*(6), 1416–1427.

Gupta, T., Malkin, M. G., & Huang, S. (2022). TRNA function and dysregulation in cancer. *Frontiers in Cell and Developmental Biology*, *10*, 886642.

Hanawalt, P. C., & Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. *Nature Reviews. Molecular Cell Biology*, *9*(12), 958–970.

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., & Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(1), 139–144.

Haradhvala, N. J., Kim, J., Maruvka, Y. E., Polak, P., Rosebrock, D., Livitz, D., Hess, J. M., Leshchiner, I., Kamburov, A., Mouw, K. W., Lawrence, M. S., & Getz, G. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nature Communications*, *9*(1), 1746.

Hou, Y. M., & Schimmel, P. (1988). A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature*, *333*(6169), 140–145.

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., de Lima, L. G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., Altemose, N., Rhie, A., Core, L. J., Gerton, J. L., Makalowski, W., Olson, D., Rosen, J., Smit, A. F. A., Straight, A. F., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science (New York, N.Y.)*, *376*(6588), eabk3112.

Huang, S.-Q., Sun, B., Xiong, Z.-P., Shu, Y., Zhou, H.-H., Zhang, W., Xiong, J., & Li, Q. (2018). The dysregulation of tRNAs and tRNA derivatives in cancer. *Journal of Experimental & Clinical Cancer Research: CR*, *37*(1). https://doi.org/10.1186/s13046-018-0745-z

Hummel, G., Warren, J., & Drouard, L. (2019). The multi-faceted regulation of nuclear tRNA gene transcription. *IUBMB Life*, *71*(8), 1099–1108.

Ishimura, R., Nagy, G., Dotu, I., Zhou, H., Yang, X.-L., Schimmel, P., Senju, S., Nishimura, Y., Chuang, J. H., & Ackerman, S. L. (2014). RNA function. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science (New York, N.Y.)*, *345*(6195), 455–459.

Jang, G. M., Sudarsan, A. K. A., Shayeganmehr, A., Munhoz, E. P., Lao, R., Gaba, A., Rodríguez, M. G., Love, R. P., Polacco, B. J., Zhou, Y., Krogan, N. J., Kaake, R. M., & Chelico, L. (2024). Protein interaction map of APOBEC3 enzyme family reveals deamination-independent role in cellular function. In *bioRxivorg*. https://doi.org/10.1101/2024.02.06.579137

Killcoyne, S., & Fitzgerald, R. C. (2021). Evolution and progression of Barrett's oesophagus to oesophageal cancer. *Nature Reviews. Cancer*, *21*(11), 731–741.

Kim, T.-M., Laird, P. W., & Park, P. J. (2013). The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell*, *155*(4), 858–868.

Kochavi, A., Nagel, R., Körner, P.-R., Bleijerveld, O. B., Lin, C.-P., Huinen, Z., Malka, Y., Proost, N., van de Ven, M., Feng, X., Navarro, J. M., Pataskar, A., Peeper, D. S., Champagne, J., & Agami, R. (2024). Chemotherapeutic agents and leucine deprivation induce codon-biased aberrant protein production in cancer. *Nucleic Acids Research*, *52*(22), 13964–13979.

Langenbucher, A., Bowen, D., Sakhtemani, R., Bournique, E., Wise, J. F., Zou, L., Bhagwat, A. S., Buisson, R., & Lawrence, M. S. (2021). An extended APOBEC3A mutation signature in cancer. *Nature Communications*, *12*(1), 1602.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930.

Lindskrog, S. V., Prip, F., Lamy, P., Taber, A., Groeneveld, C. S., Birkenkamp-Demtröder, K., Jensen, J. B., Strandgaard, T., Nordentoft, I., Christensen, E., Sokac, M., Birkbak, N. J., Maretty, L., Hermann, G. G., Petersen, A. C., Weyerer, V., Grimm, M.-O., Horstmann, M., Sjödahl, G., … Dyrskjøt, L. (2021). An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nature Communications*, *12*(1), 2301.

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., Lee, S., Chittenden, T. W., D'Gama, A. M., Cai, X., Luquette, L. J., Lee, E., Park, P. J., & Walsh, C. A. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science (New York, N.Y.)*, *350*(6256), 94–98.

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2023). Hallmarks of aging: An expanding universe. *Cell*, *186*(2), 243–278.

Maric, C., & Prioleau, M.-N. (2010). Interplay between DNA replication and gene expression: a harmonious coexistence. *Current Opinion in Cell Biology*, *22*(3), 277–283.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., & Campbell, P. J. (2018). Universal patterns of selection in cancer and somatic tissues. *Cell*, *173*(7), 1823.

Mas-Ponte, D., & Supek, F. (2020). DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nature Genetics*, *52*(9), 958–968.

McCann, J. L., Cristini, A., Law, E. K., Lee, S. Y., Tellier, M., Carpenter, M. A., Beghè, C., Kim, J. J., Sanchez, A., Jarvis, M. C., Stefanovska, B., Temiz, N. A., Bergstrom, E. N., Salamango, D. J., Brown, M. R., Murphy, S., Alexandrov, L. B., Miller, K. M., Gromak, N., & Harris, R. S. (2023). APOBEC3B regulates R-loops and promotes transcription-associated mutagenesis in cancer. *Nature Genetics*, *55*(10), 1721–1734.

Mohler, K., & Ibba, M. (2017). Translational fidelity and mistranslation in the cellular response to stress. *Nature Microbiology*, *2*, 17117.

Müller, C. A., & Nieduszynski, C. A. (2017). DNA replication timing influences gene expression level. *The Journal of Cell Biology*, *216*(7), 1907–1914.

Mungall, A. J., Palmer, S. A., Sims, S. K., Edwards, C. A., Ashurst, J. L., Wilming, L., Jones, M. C., Horton, R., Hunt, S. E., Scott, C. E., Gilbert, J. G. R., Clamp, M. E., Bethel, G., Milne, S., Ainscough, R., Almeida, J. P., Ambrose, K. D., Andrews, T. D., Ashwell, R. I. S., … Beck, S. (2003). The DNA sequence and analysis of human chromosome 6. *Nature*, *425*(6960), 805–811.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose,

N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science (New York, N.Y.)*, *376*(6588), 44–53.

Orellana, E. A., Siegal, E., & Gregory, R. I. (2022). tRNA dysregulation and disease. *Nature Reviews. Genetics*, *23*(11), 651–664.

Ormond, C., Ryan, N. M., Corvin, A., & Heron, E. A. (2021). Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Briefings in Bioinformatics*, *22*(5). https://doi.org/10.1093/bib/bbab069

Ortiz, R., Perazzoli, G., Cabeza, L., Jiménez-Luna, C., Luque, R., Prados, J., & Melguizo, C. (2021). Temozolomide: An updated overview of resistance mechanisms, nanotechnology advances and clinical applications. *Current Neuropharmacology*, *19*(4), 513–537.

Papadimitriou, M.-A., Avgeris, M., Levis, P., Papasotiriou, E. C., Kotronopoulos, G., Stravodimos, K., & Scorilas, A. (2020). TRNA-derived fragments (tRFs) in bladder cancer: Increased 5'-tRF-LysCTT results in disease early progression and patients' poor treatment outcome. *Cancers*, *12*(12), 3661.

Parisien, M., Wang, X., & Pan, T. (2013). Diversity of human tRNA genes from the 1000-genomes project. *RNA Biology*, *10*(12), 1853–1867.

Park, S. J., & Schimmel, P. (1988). Evidence for interaction of an aminoacyl transfer RNA synthetase with a region important for the identity of its cognate transfer RNA. *The Journal of Biological Chemistry*, *263*(32), 16527–16530.

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

Petljak, M., Dananberg, A., Chu, K., Bergstrom, E. N., Striepen, J., von Morgen, P., Chen, Y., Shah, H., Sale, J. E., Alexandrov, L. B., Stratton, M. R., & Maciejowski, J. (2022). Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature*, *607*(7920), 799–807.

Pinzaru, A. M., & Tavazoie, S. F. (2023). Transfer RNAs as dynamic and critical regulators of cancer progression. *Nature Reviews. Cancer*, *23*(11), 746–761.

Polak, P., Lawrence, M. S., Haugen, E., Stoletzki, N., Stojanov, P., Thurman, R. E., Garraway, L. A., Mirkin, S., Getz, G., Stamatoyannopoulos, J. A., & Sunyaev, S. R. (2014). Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature Biotechnology*, *32*(1), 71–75.

Priestley, P., Baber, J., Lolkema, M. P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., Roepman, P., Voda, M., Bloemendal, H. J., Tjan-Heijnen, V. C. G., van Herpen, C. M. L., Labots, M., Witteveen, P. O., Smit, E. F., Sleijfer, S., … Cuppen, E. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, *575*(7781), 210–216.

Reverendo, M., Soares, A. R., Pereira, P. M., Carreto, L., Ferreira, V., Gatti, E., Pierre, P., Moura, G. R., & Santos, M. A. S. (2014). TRNA mutations that affect decoding fidelity deregulate development and the proteostasis network in zebrafish. *RNA Biology*, *11*(9), 1199–1213.

Rhind, N., & Gilbert, D. M. (2013). DNA replication timing. *Cold Spring Harbor Perspectives in Biology*, *5*(8), a010132.

Ribas de Pouplana, L., Santos, M. A. S., Zhu, J.-H., Farabaugh, P. J., & Javid, B. (2014). Protein mistranslation: friend or foe? *Trends in Biochemical Sciences*, *39*(8), 355–362.

Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G. V., Carter, S. L., Saksena, G., Harris, S., Shah, R. R., Resnick, M. A., Getz, G., &

Gordenin, D. A. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics*, *45*(9), 970–976.

Roszkowska, M. (2024). Multilevel mechanisms of cancer drug resistance. *International Journal of Molecular Sciences*, *25*(22), 12402.

Rubio Gomez, M. A., & Ibba, M. (2020). Aminoacyl-tRNA synthetases. *RNA (New York, N.Y.)*, *26*(8), 910–936.

Saini, N., Roberts, S. A., Sterling, J. F., Malc, E. P., Mieczkowski, P. A., & Gordenin, D. A. (2017). APOBEC3B cytidine deaminase targets the non-transcribed strand of tRNA genes in yeast. *DNA Repair*, *53*, 4–14.

Sakhtemani, R., Senevirathne, V., Stewart, J., Perera, M. L. W., Pique-Regi, R., Lawrence, M. S., & Bhagwat, A. S. (2019). Genome-wide mapping of regions preferentially targeted by the human DNA-cytosine deaminase APOBEC3A using uracil-DNA pulldown and sequencing. *The Journal of Biological Chemistry*, *294*(41), 15037–15051.

Sanchez, A., Ortega, P., Sakhtemani, R., Manjunath, L., Oh, S., Bournique, E., Becker, A., Kim, K., Durfee, C., Temiz, N. A., Chen, X. S., Harris, R. S., Lawrence, M. S., & Buisson, R. (2024). Mesoscale DNA features impact APOBEC3A and APOBEC3B deaminase activity and shape tumor mutational landscapes. *Nature Communications*, *15*(1), 2370.

Santos, M., Fidalgo, A., Varanda, A. S., Oliveira, C., & Santos, M. A. S. (2019). TRNA deregulation and its consequences in cancer. *Trends in Molecular Medicine*, *25*(10), 853–865.

Santos, M., Pereira, P. M., Varanda, A. S., Carvalho, J., Azevedo, M., Mateus, D. D., Mendes, N., Oliveira, P., Trindade, F., Pinto, M. T., Bordeira-Carriço, R., Carneiro, F., Vitorino, R., Oliveira, C., & Santos, M. A. S. (2018). Codon misreading tRNAs promote tumor growth in mice. *RNA Biology*, *15*(6), 773–786.

Schmidt, E., & Schimmel, P. (1993). Dominant lethality by expression of a catalytically inactive class I tRNA synthetase. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(15), 6919–6923.

Schuntermann, D. B., Jaskolowski, M., Reynolds, N. M., & Vargas-Rodriguez, O. (2024). The central role of transfer RNAs in mistranslation. *The Journal of Biological Chemistry*, *300*(9), 107679.

Seplyarskiy, V., Koch, E. M., Lee, D. J., Lichtman, J. S., Luan, H. H., & Sunyaev, S. R. (2023). A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription. *Nature Genetics*, *55*(12), 2235–2242.

Seplyarskiy, V., Soldatov, R. A., Popadin, K. Y., Antonarakis, S. E., Bazykin, G. A., & Nikolaev, S. I. (2016). APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Research*, *26*(2), 174–182.

Shi, M.-J., Meng, X.-Y., Fontugne, J., Chen, C.-L., Radvanyi, F., & Bernard-Pierrot, I. (2020). Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer. *Genome Medicine*, *12*(1), 85.

Sondka, Z., Dhir, N. B., Carvalho-Silva, D., Jupe, S., Madhumita, McLaren, K., Starkey, M., Ward, S., Wilding, J., Ahmed, M., Argasinska, J., Beare, D., Chawla, M. S., Duke, S., Fasanella, I., Neogi, A. G., Haller, S., Hetenyi, B., Hodges, L., … Teague, J. (2024). COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research*, *52*(D1), D1210–D1217.

Sui, Y., Qi, L., Zhang, K., Saini, N., Klimczak, L. J., Sakofsky, C. J., Gordenin, D. A., Petes, T. D., & Zheng, D.-Q. (2020). Analysis of APOBEC-induced mutations in yeast strains with low levels of replicative DNA polymerases. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9440–9450.

Supek, F., & Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, *521*(7550), 81–84.

Supek, F., & Lehner, B. (2017). Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell*, *170*(3), 534-547.e23.

Taylor, B. J., Nik-Zainal, S., Wu, Y. L., Stebbings, L. A., Raine, K., Campbell, P. J., Rada, C., Stratton, M. R., & Neuberger, M. S. (2013). DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *ELife*, *2*, e00534.

Thornlow, B. P., Armstrong, J., Holmes, A. D., Howard, J. M., Corbett-Detig, R. B., & Lowe, T. M. (2020). Predicting transfer RNA gene activity from sequence and genome context. *Genome Research*, *30*(1), 85–94.

Thornlow, B. P., Hough, J., Roger, J. M., Gong, H., Lowe, T. M., & Corbett-Detig, R. B. (2018). Transfer RNA genes experience exceptionally elevated mutation rates. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(36), 8996–9001.

Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A., & Madan Babu, M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science Signaling*, *11*(546), eaat6409.

Torres, A. G., Reina, O., Stephan-Otto Attolini, C., & Ribas de Pouplana, L. (2019). Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(17), 8451–8456.

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

Vijg, J., & Dong, X. (2020). Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell*, *182*(1), 12–23.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7), e49.

Wernaart, D., Fumagalli, A., & Agami, R. (2024). Molecular mechanisms of non-genetic aberrant peptide production in cancer. *Oncogene*, *43*(27), 2053–2062.

Yang, C., Pataskar, A., Feng, X., Montenegro Navarro, J., Paniagua, I., Jacobs, J. J. L., Zaal, E. A., Berkers, C. R., Bleijerveld, O. B., & Agami, R. (2024). Arginine deprivation enriches lung cancer proteomes with cysteine by inducing arginine-to-cysteine substitutants. *Molecular Cell*, *84*(10), 1904-1916.e7.

Yoshida, K., Gowers, K. H. C., Lee-Six, H., Chandrasekharan, D. P., Coorens, T., Maughan, E. F., Beal, K., Menzies, A., Millar, F. R., Anderson, E., Clarke, S. E., Pennycuick, A., Thakrar, R. M., Butler, C. R., Kakiuchi, N., Hirano, T., Hynds, R. E., Stratton, M. R., Martincorena, I., … Campbell, P. J. (2020). Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*, *578*(7794), 266–272.

Zhang, Z., Ye, Y., Gong, J., Ruan, H., Liu, C.-J., Xiang, Y., Cai, C., Guo, A.-Y., Ling, J., Diao, L., Weinstein, J. N., & Han, L. (2018). Global analysis of tRNA and translation factor expression reveals a dynamic landscape of translational regulation in human cancers. *Communications Biology*, *1*(1), 234.

4 GENERAL DISCUSSION

Given the complexity of tRNA biology, the study of the tRNAome remains a challenge for the scientific community. Therefore, the overall aim of this thesis was to develop and apply bioinformatic approaches to investigate previously unexplored aspects of tRNA biology, that could further provide insight into their multifaceted roles in disease (P. Anderson & Ivanov, 2014; Orellana et al., 2022; Torres et al., 2014a).

First, we focused on the development of tRNAstudio, a pipeline that addresses the challenges associated with studying tRNA-Seq datasets, including differential expression analysis and processing of tRNAs (Padhiar et al., 2024; Telonis et al., 2016). tRNAstudio allows for the characterization of the tRNA pool from different perspectives, including the report and quantification of tRNA modifications. This functionality is extremely useful because alterations in tRNA modifications are associated with multiple human diseases (Torres et al., 2014a). For instance, mutations in the subunits of human ADAT, which catalyzes the modification of A-to-I at position 34, have been associated with neuropathies (Alazami et al., 2013; Salehi Chaleshtori et al., 2018; Thomas et al., 2019; Torres et al., 2014a). In this regard, tRNAstudio facilitated our research into understanding the biological relevance of I34-tRNAs to elucidate how alterations within this context can be related to human diseases.

Second, we addressed several questions related to unexplored areas of tRNA biology, focusing on the field of tDNA genomics. Evidence suggests that the local and long-range proximity of tDNAs may impact their transcriptional activity (Gao et al., 2024; Van Bortle et al., 2017). Furthermore, studies have reported that human tDNAs accumulate somatic mutations at high rates, and that APOBEC enzymes are a major contributor to this event (Sakhtemani et al., 2019). However, the relationship between the genomic proximity of tDNAs and the coordination of their transcriptional activity remains unclear. In addition, a detailed analysis of tDNA somatic mutagenesis is still lacking, including an understanding of the mutational mechanism and tissue-specific analyses. Therefore, in this thesis, we aimed to contribute to expanding the knowledge in these unexplored areas.

The computational analysis of tRNA-Seq data presents several challenges due to the inherent complexity of tRNA biology (Padhiar et al., 2024; Telonis et al., 2016) **(see 1.7)**. To overcome these challenges, we successfully developed tRNAstudio, an integrated pipeline for the study of tRNA-Seq data **(see 3.1)**. tRNAstudio allows the exploration of tRNA biology from multiple perspectives, including tRNA processing by the identification and classification of pre-tRNA and mature tRNA sequences, the use of a coverage profile to identify tsRNAs, mismatch analysis for the detection of post-transcriptional chemical modifications, as well as the quantification of individual tRNA sequences to perform differential expression analysis. Although there are many different variants within the pipelines developed for tRNA analysis, we consider tRNAstudio an integrated pipeline capable of performing an in-depth tRNA analysis from different points of view.

Unlike tRNAstudio, other pipelines developed for tRNA analysis focus on specific outcomes rather than a general exploration of the tRNA pool **(see 1.7)**. Examining the tRNA pool from multiple perspectives, as tRNAstudio does, is crucial because phenotypes produced by changes in the tRNA pool can stem from alterations at any stage of its biogenesis. In contrast, performing tRNA-Seq analysis from a narrow perspective (e.g., only analyzing mature sequences) risks overlooking valuable insights about tRNA processing. Furthermore, tRNAstudio is designed to be accessible to non-computational users, while most available pipelines require advanced computational skills. The combination of its accessibility and capacity to perform an extended tRNA pool analysis shows that it can be used as an exploratory tool by researchers who want to explore and analyze tRNA-Seq datasets but do not have their own resources to do so.

Different approaches have been tested to perform the mapping of tRNA-related reads. Some methods involve mapping first against the native genome, further extended with mature tRNA sequences (P. P. Chan et al., 2025; Clark et al., 2016; Cozen et al., 2015; Erber et al., 2020; Hauenschild et al., 2015; A. Hoffmann et al., 2018), whereas others perform the mapping against the tRNA space alone, with only tRNA sequences (Behrens et al., 2021; Selitsky & Sethupathy, 2015). In tRNAstudio, we use the strategy of mapping in different steps against different genomes to ensure reliable detection and classification of tRNA reads **(see. 3.1 Supp. Fig. S1)**. Although some pipelines map exclusively against the tRNA space alone, in tRNAstudio, we first align against the complete native human genome. This strategy allows for the removal of non-tRNA reads. This is crucial because the human genome contains both nuclear and mitochondrial 'tRNA-lookalike' sequences (Telonis et al., 2014). If mapping is performed exclusively against tRNA sequences, it can force the alignment of these non-tRNA reads onto tRNA sequences, resulting in false positives (Loher et al., 2017; Padhiar et al., 2024; Telonis et al., 2014, 2016). Following this first alignment, tRNAstudio extracts, classifies and remaps tRNA reads against customized tRNA reference genomes (e.g., a collection of unique mt-tRNAs sequences, pre-tRNA, or mature tRNA sequences). This multi-step mapping approach allows tRNAstudio to avoid human genome encoded tRNA-lookalikes, to classify tRNA sequences and, to capture tRNA-related reads that are highly modified and to control for multimapping.

Once we evaluated the performance of tRNAstudio **(see 3.1)**, we used the pipeline to provide analytical support for a project in our group focused on the evaluation of the biological relevance of I34-tRNAs **(see 3.2)**. It is worth noting that this study was carried out before tRNAstudio was published, although the pipeline used is the one that tRNAstudio is based on. Previous experimental and computational analyses conducted by former members of the lab showed that the reduction of I34-tRNAs impaired the translation of extracellular matrix (ECM) proteins. These proteins are enriched in low-complexity TAPSLVR-rich regions that rely on ADAT-dependent codons for their translation (Rafels-Ybern et al., 2015; Rodríguez-Escribà, 2020). The ECM provides structural support and enables cell-to-cell communication, which is essential for multicellular organisms (Karamanos et al., 2021). By applying tRNAstudio pipeline, we reported that knocking down ADAT2 results in a reduction of I34-tRNAs **(see 3.2 Fig. 2B and Supp. Fig. 2B)**. We also showed that the expression levels of the rest of the tRNAs do not change **(see 3.2 Fig. 2C and Supp. Fig. 2C)**. These results indicate that eukaryotic cells are unable to compensate for I34-tRNA deficiency through overexpression of alternative tRNA, highlighting how essential and indispensable I34-tRNAs are. Given that the expansion of I34 to TAPSLIVR tRNAs is a hallmark of eukaryotes, that I34-tRNAs are essential to efficiently translate ECM proteins required for multicellularity, and that eukaryotic cells cannot compensate for the lack of I34-tRNAs with other tRNAs to produce such proteins; it is possible that, among other crucial functions, the emergence of I34-tRNAs was fundamental in the evolution of multicellularity.

Interestingly, some of the ECM components identified (Rodríguez-Escribà, 2020), such as Syndecan 3, have been associated with neurological development (Bespalov et al., 2011; Hienola et al., 2006; Hudák et al., 2022), while others components like mucins, such as MUC5AC, have been associated with respiratory diseases, including asthma, chronic bronchitis (CB), cystic fibrosis (CF), and obstructive pulmonary disease (COPD) (Carpenter et al., 2021; Ma et al., 2018). This connection provides a possible explanation for the phenotypes observed in individuals with mutations in ADAT genes (Alazami et al., 2013; Salehi Chaleshtori et al., 2018; Thomas et al., 2019; Torres et al., 2014a). Thus, it would be interesting to address whether the modulation of the levels of functional ADAT could be used as a strategy in biomedicine, either overexpressing it to compensate for the effect of such mutations in neurological diseases or downregulating its expression to reduce mucin production in diseases like COPD.

This analysis demonstrated that by applying tRNAstudio we can successfully analyze tRNA-Seq data, including the quantification of tRNA modifications and tRNA differential expression analysis. However, now we will discuss some limitations or improvements that would enhance the performance of tRNAstudio.

All mapping steps in tRNAstudio are performed using Bowtie2 (Langmead & Salzberg, 2012). While Bowtie2 remains widely used (P. P. Chan et al., 2025; Pinkard et al., 2021; Smith et al., 2024), there are other aligners that also enable short-read mapping and have been employed for tRNA analysis, such as Segemehl (A. Hoffmann et al., 2018; S. Hoffmann et al., 2009), SHRiMP2 (David et al., 2011; Shigematsu et al., 2017), GSNAP (Behrens et al., 2021; T. D. Wu et al., 2016) or Burrows–Wheeler

aligner (BWA) (Erber et al., 2020; Gogakos et al., 2017; Scheepbouwer et al., 2023). Furthermore, alignment algorithms developed specifically for tRNA-Seq analysis have been developed, like tDRmapper (Selitsky & Sethupathy, 2015). Additionally, the parameters used to run the aligners are usually adjusted to accommodate for mismatches, which allows for a broader detection of tRNA modifications (Behrens et al., 2021). In fact, this methodology is implemented in the final step of tRNAstudio pipeline by relaxing Bowtie2 parameters to enable additional mismatches in the seed (i.e., the matching segment between a sequencing read and a reference). Some pipelines incorporate the information from curated databases that contain known RNA modifications (Behrens et al., 2021; Cappannini et al., 2024; Dunin-Horkawicz et al., 2006). An example of such databases is MODOMICS, which includes sequences with tRNA modifications (Cappannini et al., 2024; Dunin-Horkawicz et al., 2006). The addition of MODOMICS to the pipeline helps to distinguish which mismatches or sequence truncations observed are likely due to known modifications rather than sequencing errors (Behrens et al., 2021). While tRNAstudio already reports tRNA modifications using as reference the consensus tRNA sequence (Sprinzl et al., 1998), updating tRNAstudio with a similar strategy would be useful to facilitate the report of known modifications to ease the biological interpretation of the results.

Although a comparison of all available aligners for tRNA-Seq analysis has not been performed, there are recent benchmarking studies that compared the performance of Bowtie2, SHRiMP2, and GSNAP (Smith et al., 2024). Based on simulation data mimicking real tRNA-Seq reads, Bowtie2 and SHRiMP2 were found to align the highest proportion of reads (98.5% for Bowtie2, 93.0% for SHRiMP2), whereas GSNAP had a lower mapping rate (71.3%) (Smith et al., 2024). However, both Bowtie2 and SHRiMP2 produced up to 4.0% and 3.3% fewer correct alignments at the anticodon level, compared to the GSNAP approach. Therefore, some aligners like GSNAP lose information but are more accurate, whereas aligners like Bowtie2 and SHRiMP2 allow for the detection of more reads but are subject to a slightly higher error rate than GSNAP. These results suggest that when analyzing tRNA-Seq data it may be beneficial to integrate the results from multiple aligners to enhance the accuracy of the final output.

A current limitation of tRNAstudio is its exclusive design for the analysis of human datasets, which is currently restricted to the GRCh38/hg38 genome. In contrast, other pipelines, such as tRAX (Holmes et al., 2022), support the analysis of multiple species by providing a custom reference genome for each of them. But, the number of available species is still limited. Collaborations within our group were established to implement tRNAstudio for mouse and yeast models. This effort was successful, and the project is currently ongoing, led by these laboratories. In each case, however, tRNAstudio had to be adapted individually for the specific genome of each organism. For that reason, it would be interesting to expand tRNAstudio with a feature that allows the automated generation of reference genomes. This would allow the analysis of a broader range of species and cell lines. Furthermore, this improvement would make the pipeline more accessible to non-computational users, since they could provide the desired reference genome and have all the necessary custom genomes generated automatically. Such functionality could be achieved by integrating tRNAscan-SE to predict the tDNAs loci and use this information to automatically build custom reference genomes (P. P. Chan et al., 2021).

A new version of the assembly of the human genome has been published (T2T-CH13), which is considered a complete, gapless human reference genome (except for some ribosomal DNA arrays that remain unresolved) generated using long-read sequencing technologies (Nurk et al., 2022). This reinforces the need to update tRNAstudio with functionalities that allow the pipeline to be adapted to other, more recent reference genomes. Nevertheless, most references for tRNA analysis studies are still based on GRCh38/hg38; therefore, tRNAstudio remains a competent tool that allows for exhaustive analysis of tRNA sequences.

We used the knowledge gained during the development of tRNAstudio to use other computational approaches that enable additional analyses to explore other aspects of tRNA biology. Knowing that a more complete assembly of the human genome was available, we analyzed the distribution of tDNAs in the T2T-CHM13 human genome assembly using tRNAscan-SE (P. P. Chan & Lowe, 2019). We verified that tDNA localization is non-random and that tDNAs can be linearly localized in clusters or pairs within the genome (Bermudez-Santana et al., 2010; Van Bortle et al., 2017) **(see 3.3 Fig. 2a)**. Similar analysis in other species have also reported that there are more tDNA pairs than expected (Bermudez-Santana et al., 2010). These findings suggest that tDNA evolution may favor the formation of tDNA clusters. The fact that tDNA localization is not random provides insight into the evolutionary history of tDNAs, as the close proximity of many tDNAs suggests that local duplication events have played a major role in the expansion of tDNA copy number and the generation of identical tDNA copies within genomes (Ayan et al., 2020; Bermudez-Santana et al., 2010). Importantly, we also identified tDNA copies that are located farther apart, indicating that tDNAs can be dispersed throughout the genome by additional mechanisms (Kramerov & Vassetzky, 2011).

Our identification of tDNAs using the T2T-CHM13 assembly represented a significant novelty. Such characterization of tDNAs was previously conducted using reference genomes that contained incomplete assembled regions particularly those with repetitive sequences that could not be resolved (e.g., hg19 and hg38) (P. P. Chan & Lowe, 2016). Those caveats are now resolved in the T2T-CHM13 assembly. This is particularly relevant when characterizing tDNAs, as they often reside within repetitive regions which are now fully resolved in the T2T-CHM13 assembly (Hoyt et al., 2022). Whereas with T2T-CHM13 we reported a total of 733 tDNAs with around 521 classified as high-confidence, with hg19 and hg38 we reported a total of 614 and 637 tDNAs respectively with around 400 of them being high-confidence for both assemblies **(see 3.3 Fig. 2b)**. This suggests that the number of tDNAs have been underestimated in older assemblies.

The difference in the number of detected tDNAs between assemblies comes mainly from a variation of a tandem repetition inside a cluster located in Chr 1 (in the T2T Chr1:160,577,243-160,733,822) **(see 3.3 Fig. 4a)**. While in the T2T-CHM13 assembly, we defined it as a 21-copy tandem repeat, in both hg19 and hg38, this region was defined as a 4-copy tandem repeat (Gao et al., 2024; Iben & Maraia, 2014). Interestingly, this tandem repetition has been previously reported to present different number of copies between individuals, leading to tgCNV (Iben & Maraia, 2014). Polymorphisms involving copy number variations have also been reported in the homologous mouse region, also located in Chr 1 (with a

variation of 9-43 repeat units) (Darrow & Chadwick, 2014). Altogether, these findings suggest that the tandem repeat region on Chr 1 is more prone to genetic variability through duplication events than other tDNA-enriched regions. Furthermore, variation of individual tDNAs have also been reported in human genomes (Berg et al., 2019; Iben & Maraia, 2012, 2014; Lant et al., 2019; Parisien et al., 2013). This highlights how tDNA copy number can be individual-specific and shows that relying on general reference genomes may lead to underestimation or overestimation of the tDNA number.

Nevertheless, T2T-CHM13 does not represent genetic diversity between individuals (Liao et al., 2023; Nurk et al., 2022). This is because the T2T-CHM13 assembly was primarily obtained only from the CHM13 human cell line (Schneider et al., 2017). Moreover, both hg19 and hg38 are 'mosaic genomes', meaning that they are composed of sequences from diverse individuals but do not take into account human genomic variability either (Liao et al., 2023; Schneider et al., 2017). Therefore, the differences in assembly origins can lead to variations in the number of tDNAs produced by both technical factors and the genetic diversity of the donor samples used to obtain the assemblies. This implies that we cannot know if the T2T-CHM13 represents most of the population, especially for tDNA-regions that are highly variable between individuals. In this regard, any of the developed tools to analyze tRNA-Seq data, including tRNAstudio, do not account for potential sites of tgCNV. This underscores the importance of developing inclusive reference genomes to better capture human tDNA genetic variation, like the Human Pangenome Reference Consortium (HPRC) that works towards the generation of a reference that includes assemblies from genetically diverse individuals (Liao et al., 2023). However, the best-case scenario will be to determine the specific tDNA copy number for each cell type or individual under study in order to assigning the possible genomic origins of tRNA sequences.

Each repetition unit in the previously described tandem repeat found in Chr 1 is composed of five tDNAs: Gly-TCC, Asp-GTC, Leu-CAG, Gly-GCC, and Glu-CTC **(see 3.3 Fig. 4a)**. This implies that individuals may differ in the tDNA copy number for these specific isodecoders, depending on how many tandem units their genome contains. For instance, in the T2T-CHM13 reference genome, tDNAs for Gly are the most abundant, with more than 70 copies. In contrast, hg19 and hg38 contain only about 40 copies each, and tDNAs for Ala are the most abundant. While it remains unclear whether these differences arise from technical or biological factors, these observations emphasize how the choice of the reference genome can significantly influence the perceived composition of tDNAs.

Consequently, tgCNV could directly influence the interpretation of tDNA expression data, as observed differences in tRNA sequences might reflect gene dosage variation rather than changes in tDNA transcriptional activity. Hence, the selection of the human reference genome for the analysis of tRNA-Seq datasets could influence the interpretation of the results, as we found differences in the number of tDNAs reported for each analyzed human reference genome. Nevertheless, our studies exploring tRNA gene expression are still informative, as regardless of whether the overexpression of a given tRNA transcript comes from additional tDNA copy numbers or enhanced transcription of such tRNA genes, the overall result is a physiological change in tRNA abundance resulting in observed phenotypes.

tgCNVs could lead to gene dosage imbalances, which could alter the cellular abundance of specific tRNA molecules. This may influence the efficiency and fidelity of protein translation, in particular for proteins with a codon usage biased toward codons decoded by the specific tRNAs involved in tgCNV. In a worst-case scenario such alterations could lead to disease. For example, the loss of only one gene for tRNA-Phe has been previously described to affect neuronal function (Hughes et al., 2023). Whereas other reported tgCNVs are not attributed to a specific phenotype or an evident contribution to disease (Berg et al., 2019; Iben & Maraia, 2014; Lant et al., 2019; P. Yang et al., 2019). Furthermore, changes in tDNA copy number could affect the production of tsRNAs or even have implications in their role of organization of the eukaryotic genome (McFarlane & Whitehall, 2009). However, the impact of tgCNV on cellular function and disease is still underexplored. In this regard, it would be interesting to expand our knowledge of tDNA copy number variation by analyzing human genomes from different individuals using long-read sequencing strategies, as those used to obtain the T2T-CHM13 genome.

Gene copy-number changes that result from genomic instability can be used as an adaptive mechanism by tumoral cells to improve their fitness (Ippolito et al., 2021; Mishra & Whetstine, 2016). With this in mind, we hypothesize that cancer-associated genomic instability could generate tgCNV in somatic cells, which in certain cases could provide selective advantages. For instance, an increase in tRNA transcripts is frequently observed in cancer cells in order to compensate for the high demand in protein synthesis that results from their high proliferative state (Goodarzi et al., 2016; T. Gupta et al., 2022; Hernandez-Alias et al., 2020; Orellana et al., 2022; Pinzaru & Tavazoie, 2023; Z. Zhang et al., 2018). According to our hypothesis, such increases might not only result from overexpression of specific tDNAs but also from an increase in tDNA copy number. Nevertheless, this remains a hypothesis and further analysis will be needed to test it, as somatic alterations in tDNA copy number have not been investigated in the context of cancer.

This hypothesis could be supported by several studies that have shown how tgCNV undergo adaptive changes in populations of bacteria in response to altered translational demands (Ayan et al., 2020; Khomarbaghi et al., 2024; Nilsson et al., 2006). These changes are often driven by spontaneous large-scale tandem duplications or amplifications involving tDNAs. A relevant example involves a large-scale duplication containing Gly-GCC in *Pseudomonas fluorescens,* that occurs in order to genetically compensate the loss of other Gly-GCC genes (Khomarbaghi et al., 2024). Surprisingly, Gly-GCC is one of the tDNAs included in the aforementioned tandem repeat found on Chr 1 in human and mice (Darrow & Chadwick, 2014; Gao et al., 2024; Iben & Maraia, 2014). It is worth highlighting that high amounts of tsRNAs coming from tRNA-Gly-GCC and tRNA-Glu-CTC have been previously reported in specific human biological contexts. For example, an upregulation of 5′-tRFs derived from Gly-GCC tRNAs has been observed during stress responses (Jin et al., 2024)**.** Furthermore, studies have shown that tsRNAs derived from Gly-GCC tRNAs are upregulated and increase the malignancy of several types of tumors, including ovarian, colorectal and bladder (Panoutsopoulou et al., 2021; Qin et al., 2022; Y. Wu et al., 2021). Also, 5′-tRFs derived from Glu-TCT have been involved with increased proliferation in ovarian cancer (Zhou et al., 2017)**.** Following our hypothesis, this suggests that alterations in the genomic regions encoding tDNAs for those specific tsRNAs could influence the abundance of such transcripts,

contributing to the observed phenotype. To test this hypothesis further analysis on WGS data of cancer tissues would be needed to analyze tgCNV.

Next, we shifted our focus towards understanding how tDNA transcription could be orchestrated by the genomic organization of tDNAs as previous studies have noted how proximity between tDNAs could favor their transcription (Gao et al., 2024; Van Bortle et al., 2017). We found that tDNA clustering can enhance tDNA expression and can even coordinate the expression levels within tDNAs in a cluster **(see 3.3 Fig. 5)**. This phenomenon may occur because such proximity facilitates the recruitment of Pol III and its corresponding transcription factors to multiple tDNAs at a time, into what are called 'tDNA transcription factories' (Gao et al., 2024). Importantly, as mentioned previously, tDNA proximity has been noted in different eukaryotic species, suggesting that evolution favors tDNA clustering (Bermudez-Santana et al., 2010). This implies that natural selection may favor tDNA clustering since it could optimize and coordinate tDNA expression. Moreover, the linear arrangement of tDNAs within the genome can facilitate the formation of TADs and chromatin loop structures, both known to also influence tDNA expression (Van Bortle et al., 2017). However, it is important to note that some isolated tDNAs also exhibit high expression levels, while genes within tDNA clusters can display varying or even low levels of expression. This underscores the complexity and multiple layers of regulation involved in tDNA transcription.

We have identified that most of the tDNAs are located in early-replicating regions of the genome, which correspond to regions that in general, are highly transcribed (Maric & Prioleau, 2010; Müller & Nieduszynski, 2017; Rhind & Gilbert, 2013) **(see 3.3 Fig. 6)**. This matches with previous studies that reported that most tDNAs co-localize with predicted transcriptional hotspots (Mungall et al., 2003). However, even if most tDNAs are located in those highly active regions and contain suitable tDNA promoter sequences, some tDNAs are still transcriptionally silent (Thornlow et al., 2020; Torres, 2019). Again, this highlights how tDNA expression is regulated by multiple factors and involves several layers of control. Moreover, the presence of constitutively silent tDNAs could suggest that they are conserved in the genome due to non-canonical function attributed to tDNAs, such as their role in genome organization (Guimarães et al., 2021; Hamdani et al., 2019; Iwasaki et al., 2020; McFarlane & Whitehall, 2009; Raab et al., 2012; Sizer et al., 2022; Van Bortle & Corces, 2012; Van Bortle et al., 2017).

Using a large sample size, we observed that tDNAs are hotspots of somatic mutagenesis in cancer samples **(see 3.3 Fig. 7)**. This is consistent with previous reports (Sakhtemani et al., 2019). Furthermore, we found that mutation rates in tDNAs increased with transcriptional activity, supporting the idea that tDNAs are subjected to TAM (Thornlow et al., 2018). Interestingly, this contrasts with what is observed in protein-coding genes, where higher transcriptional activity is associated with lower mutation rates **(see 3.3 Fig. 8)**. The observed differences in mutational patterns between these two gene types could be attributed to their distinct transcriptional machinery, whereas protein-coding genes are transcribed by Pol II, tDNAs are transcribed by Pol III. This distinction is crucial because Pol II can efficiently recruit TCR machinery, which is highly efficient in repairing lesions on the transcribed strand of active genes, mitigating the accumulation of mutations in highly expressed protein-coding genes

(Spivak, 2015). In contrast, Pol III does not recruit TCR machinery (Dammann & Pfeifer, 1997; Seplyarskiy et al., 2023), resulting in a higher accumulation of mutations at highly expressed tDNAs. Although general repair pathways that are Pol II-independent, such as base excision repair (BER) and global genome nucleotide excision repair (GG-NER), are active throughout the genome, they do not provide the same protection as TCR (Krokan & Bjørås, 2013; Saini et al., 2017; Spivak, 2015). Altogether this highlights that while TAM can have an impact on both protein-coding genes and tDNAs, their distinct repair mechanisms may contribute to explaining the susceptibility of tDNAs to accumulate somatic mutations.

With this in mind, we would expect similar mutational loads across all Pol III-transcribed genes. Surprisingly, our analysis of other Pol III-transcribed genes (including 5S rDNA, RNUs and miscRNA genes) shows that while these genes, especially RNUs, can accumulate mutations (Seplyarskiy et al., 2023), tDNAs are exceptionally prone to accumulate mutations **(see 3.3 Fig. 9)**. Hence, the absence of TCR is insufficient to explain the observed increase in mutational density within tDNAs. This suggests that among the genes transcribed by Pol III, tDNAs could have specific characteristics that make them particularly vulnerable to mutagenesis.

In concordance with previous results, the mutational signature analysis suggests that members of the APOBEC3 family are the main contributors to tDNA mutagenesis (Butt et al., 2024; Saini et al., 2017; Sakhtemani et al., 2019; Sui et al., 2020). From the APOBEC3 family, APOBEC3A (A3A) and APOBEC3B (A3B) are responsible for most mutations in human somatic cells and have been associated with cancer development and progression (Alexandrov et al., 2020; Butler & Banday, 2023; Petljak et al., 2022). Both A3A and A3B have been previously linked with tDNA mutagenesis in bacteria, yeast, and human (Butt et al., 2024; Saini et al., 2017; Sakhtemani et al., 2022, 2019). Knowing that APOBEC3 activity increased with tDNA transcription levels and that APOBEC3 enzymes act on ssDNA that is exposed during transcription (Langenbucher et al., 2021; Roberts et al., 2013), we propose that the observed TAM could be mainly driven by the activity of APOBEC3 enzymes.

As other genes transcribed by Pol III showed different mutational profiles, this suggests that APOBEC3-driven mutagenesis at tDNAs is not caused only by high levels of transcription but requires additional factors specific to tDNAs. In this regard, we explored which factors could explain the high mutagenic rate observed in tDNAs. We suspected that tDNAs may be prone to produce secondary DNA structures that arise in ssDNA and that favor APOBEC3-driven mutagenesis, such as R-loops and DNA hairpins **(see 1.2.2).** This idea was driven by the characteristic cloverleaf structure of tRNAs based on stem-loops conformations that reassemble hairpin structures, that prompted us to consider that tDNAs can produce similar configurations due to intramolecular complementary regions. By analyzing APOBEC3-preferred hairpin DNA sites prediction (Buisson et al., 2019), we found a significant enrichment of hairpin formations in tDNAs compared to the rest of the genome and compared with other Pol III-transcribed genes. Moreover, tDNAs have been previously reported to be R-loops hotspots in human, yeast and plants (Chen et al., 2017; El Hage et al., 2014; Elsakrmy & Cui, 2023; K. Liu & Sun, 2021; Meng & Zou, 2025; Santos-Pereira & Aguilera, 2015). Therefore, these findings highlight that tDNAs have specific

sequence characteristics that can contribute to explaining why tDNAs accumulate such high levels of mutagenesis in comparison to other regions of the genome.

For instance, when analyzing the mutational profile of tDNAs at single base resolution, we detected that positions that are hotspots for mutagenesis contain TCN motifs preferred by APOBEC3 **(see 3.3 Fig. 12a)**. Notably, position 56 has the highest mutation rate and the highest percentage of tDNA sequences containing TCN motifs at this position **(see 3.3 Supp. Fig. 16)**. This position is located within the T-loop of the tRNA, suggesting it may be part of a region capable of producing a hairpin-like structure in tDNAs, which could explain the elevated mutagenic activity observed at this position. However, other positions also display elevated mutation rates that do not overlap with APOBEC3 motifs, indicating that distinct mutagenic mechanisms may be responsible for these mutations.

tDNAs are well-known sites of replication-fork pausing. Several studies have reported that the high transcriptional activity of tDNAs can pause replication fork progression, due to the transcription complex itself (Pol III and transcription factors) and the condensin-mediated clustering in tDNAs, which generate obstacles for the replication fork, leading to the ssDNA being exposed for longer time (Bermudez-Santana et al., 2010; McFarlane & Whitehall, 2009; Yeung & Smith, 2020). We hypothesize that these replication-transcription conflicts could create vulnerable ssDNA regions in tDNAs, increasing their susceptibility to mutagens, such as APOBEC3 enzymes.

Another factor that could contribute to the increase in the likelihood of mutagenesis at tDNAs is the interaction between tRNA processing enzymes and members from the APOBEC3-family. This idea arises from the reconstruction of the protein interaction map of the APOBEC3 subfamily, which reported that APOBEC3G (A3G) and APOBEC3H (A3H) can interact with enzymes involved in tRNA processing and methylation (Jang et al., 2024). Although direct interactions between the tRNA processing machinery and other APOBEC3 enzymes, such as A3A or A3B, have not yet been described, their structural and functional similarities to A3G and A3H raise the possibility that analogous interactions may exist but remain undetected. Given that tRNA processing is initiated in proximity to tDNA loci immediately after transcription (Hopper et al., 2010), such proposed interactions could position APOBEC3 enzymes near sites of active tDNA transcription, increasing the chances of APOBEC3-driven mutagenesis. Future studies are needed to investigate whether A3A and A3B are similarly recruited by tRNA processing enzymes, which could contribute to elucidating the molecular basis for the unique vulnerability of tDNAs to mutagenesis.

Among the mutational signatures detected, APOBEC signature exhibited the most robust profile, whereas other signatures were in some cases ambiguous. These signatures included DNA mismatch repair alterations, oxidative damage, mutations in the POLE gene, or specific chemotherapy treatments like temozolomide. Our analysis of mutational signatures was performed using the NMF method. We selected NMF as a first approach because it enables the discovery of novel signatures, in contrast to methods like SigProfilerAssignment, which are limited to identifying signatures already reported and annotated in the COSMIC catalogue (Alexandrov et al., 2020). However, the ambiguity observed in

some signatures indicates that complementary analytical approaches are necessary to validate and reinforce the robustness of the other detected signatures with NMF.

In summary, in our study we reported different mechanisms that could contribute to answer why tDNAs are hotspots of somatic mutagenesis. These mechanisms include: i) Transcription by Pol III and APOBEC3-driven mutagenesis: tDNAs are transcribed by Pol III, which does not recruit the TCR machinery, which makes tDNAs more vulnerable to TAM induced by APOBEC3 enzymes. ii) Enrichment in DNA motifs and secondary DNA structures: tDNAs contain APOBEC3 motifs and are enriched in secondary DNA structures, such as R-loops and hairpins, that promote APOBEC3-mediated mutagenesis. Moreover, we propose additional mechanisms that could contribute to explaining tDNA mutagenesis: iii) Replication-transcription conflicts at tDNAs that increase the occurrence of ssDNA and could increase their susceptibility to mutagenesis. iv) APOBEC3-driven mutagenesis could be promoted by the interaction of APOBEC3 enzymes with tRNA processing enzymes.

Next, we found that the mutational load of tDNAs is cancer-type and tissue-dependent. Thus, tDNA mutagenesis is influenced by the specific biological context of each tissue **(see 3.3 Fig. 10a and Supp. Fig. 10)**. For instance, we reported APOBEC3 activity in an specific set of tissues including bladder, breast, head and neck, uterus and lung, which aligns with previous findings where overexpression of APOBEC3 activity was reported for these cancer types (Alexandrov et al., 2020; Burns et al., 2013a). Therefore, the differences in the mutational load for each cancer type can be driven by differences in the mutagenic agents acting in each cancer type, such as the different activity levels of APOBEC3.

Indeed, we have found a correlation between APOBEC3 expression levels and rates of tDNA mutagenesis. For example, we reported that bladder cancer (BLCA) exhibits the highest rates of tDNAs mutagenesis, which increases with APOBEC3 activity. Moreover BLCA has been previously reported to usually present overexpression of APOBEC3 (Lindskrog et al., 2021; M.-J. Shi et al., 2020). Interestingly, in contrast to BLCA, we found that brain tumors show almost no tDNA somatic mutations. The brain is described as an immune-privileged organ, which implies a lower baseline mutation rate regarding the types of mutations produced by APOBEC3 enzymes (Benhar et al., 2012).

We also reported mutagenesis of tDNAs in healthy, non-cancerous human samples **(see 3.3 Supp. Fig. 11)**. While genome-wide signatures of APOBEC3 mutagenesis are rare in healthy somatic cells, in contrast to cancer genomes (Franco et al., 2019), we do observe this mutagenesis signature in tDNAs from healthy tissues. We believe that this observation could be attributed to the previously described unique characteristics of tDNAs, which make tDNAs more susceptible to APOBEC3-induced somatic mutagenesis, even in healthy samples. In healthy cells, APOBEC3 overexpression can be triggered by many different factors, including viral infections and inflammation (Butler & Banday, 2023). In most cases, APOBEC3 returns to basal expression levels once the viral infection or the inflammatory process is resolved (Ferreira et al., 2021). Since the datasets analyzed included data from lung and colon, that are usually exposed to infection and inflammatory events over time (Kawalec, 2016; Kombe Kombe et

al., 2024), this suggests that the observed somatic mutations in tDNAs from healthy samples may result from transient overexpression of APOBEC3 occurring specifically in these tissues.

Unfortunately, the number of healthy samples was insufficient to perform a tissue-specific analysis. Therefore, we were unable to determine in which specific tissues was mutagenesis occurring. Specifically, our dataset included only 1,192 healthy samples (including brain, colon, liver, and lung), compared to 9,596 cancer samples that were analyzed. To validate our observations and achieve a more comprehensive view of the landscape of tDNA mutagenesis in healthy tissues, future studies should increase both the sample size and the diversity of healthy tissue types analyzed.

The impact of tDNA somatic mutagenesis in tRNA biogenesis and function remains to be determined. It is important to note that the consequences of tDNA mutagenesis will depend on many different factors, such as which tDNAs are mutated or the localization of the mutations within the tDNA sequence. On the one hand, mutations in tDNAs may be functionally neutral, or their effect could be compensated by the multicopy nature of tDNAs, which could allow non-mutated copies to compensate for the defective ones. On the other hand, if these mutations are not neutral, they could result in either detrimental or beneficial effects according to the specific cellular context. For example, we reported that somatic mutations at the anticodon of actively transcribed tDNAs could produce chimeric tRNAs known to produce mistranslation by introducing non-cognate amino acids in the human proteome (Geslain et al., 2010; M. Santos et al., 2018). In the context of cancer, experimental models have demonstrated that the expression of chimeric tRNAs can enhance cell transformation, stimulate angiogenesis, and accelerate tumor growth in mice (M. Santos et al., 2018). Therefore, if chimeric tRNAs that result from tDNA mutagenesis are still active, they can produce widespread misincorporations in the proteome. Additionally, if mutations occur in the anticodon of tRNAs whose aminoacyl-tRNA synthetases require the anticodon for recognition, this could result in the presence of inactive uncharged tRNAs.

Besides mutations in the anticodon, mutations elsewhere in the tRNA sequence can alter tRNA biogenesis and maturation, which could lead to altered tRNA pool. Such alterations could also increase the risk of amino acid misincorporation during protein synthesis. For example, mutations could alter tDNA transcription. This could occur due to mutations that fall within the internal promoter regions or within the immediate flaking regions, which have been described to influence tRNA transcription (Thornlow et al., 2020, 2018). In addition, mutations in the flanking regions could disrupt the processing of pre-tRNAs since 5’ leader and the 3’ trailer sequences of the pre-tRNA are encoded within those regions (Gogakos et al., 2017), and mutations in those sequences. In future analyses it will be interesting to characterize the mutations falling within the immediate flaking regions of tDNAs at the single-nucleotide level to address their potential impact on tDNA transcription and processing.

Moreover, mutant tRNAs could lead to aberrant tRNA post-transcriptional modification patterns. As mentioned in the introduction, post-transcriptional modifications are essential for tRNA stability and function (Pan, 2018; Suzuki, 2021; Torres et al., 2014a). In our results we can identify several modification sites that coincided with hotspots of somatic mutagenesis **(see 1.3.3 Fig. 10 and 3.3 Supp.**

Fig. 16a). For example, position 13 within the tRNA sequence is highly mutated and frequently modified. This position is a well-known modification site crucial for tRNA folding and stability (Lorenz et al., 2017). Another example is the position 34, if mutations occur in tDNAs of A34-tRNAs, it could reduce the levels of functional I34-tRNAs. Taken together, our results suggest that somatic mutations in tDNAs could interfere with key tRNA modifications and consequently, compromise tRNA stability, decoding capacity, aminoacylation and ribosome interactions, among other critical processes.

Mutant tRNAs could impair the production of tsRNAs. Alterations in the sequences of tRNAs can hinder the recognition of the enzymes responsible for the cleavage of tRNAs into tsRNA, including RNase Z/ELAC2, Dicer, and ANG (Ivanov et al., 2011; Kumar et al., 2014; Y. S. Lee et al., 2009). The effect of such mutations may not be related with protein synthesis but rather with non-canonical functions. Alterations in tsRNAs levels have been reported in different cancer types (P. Anderson & Ivanov, 2014; Pekarsky et al., 2023; Soares & Santos, 2017). For example, some studies have reported that a decrease of specific tsRNAs promotes cellular proliferation in lung cancer (T. Gupta et al., 2022). Interestingly, in BLCA it has been described a link between the dysregulation in the levels of specific tsRNAs and the aggressiveness of tumors (Papadimitriou et al., 2020). This observation together with the fact that BLCA shows high levels of tDNA mutagenesis opens the possibility of a functional relationship between mutated tDNAs and BLCA development.

Further analysis will be needed to verify the presence of potentially active mutated tRNAs within the cell. One approach to identify such mutated tDNAs will be to combine WGS (detecting mutations in tDNAs) with tRNA-Seq (quantifying and characterizing tRNAs). This integrated approach will help to identify if mutated tDNAs are transcriptionally active and if the resulting tRNAs, carrying these mutations, are present within the cell rather than being degraded. For this purpose, tRNAstudio could be used to identify active mutated tRNAs from tRNA-Seq data. This task will be difficulted by the presence of post-transcriptional chemical modifications in tRNAs. Both somatic mutations and chemical modifications will be detected as mismatches during sequence alignment. We propose a possible solution that consists of comparing the reported base changes with well-known positions of tRNA modifications. This could be achieved by using MODOMICS, a comprehensive database that catalogs tRNA sequences and their known chemical modifications (Cappannini et al., 2024). In any case, this characterization will be extremely useful to later determine the effect of such mutated tRNAs and to aid on the design of experimental approaches to elucidate the effect of such mutations.

The proposed scenarios evidence how tDNA mutagenesis could disturb tRNA biogenesis in many different aspects, including transcription, processing, addition of chemical modification or tsRNAs production (Cabrelle et al., 2024; Earnest-Noble et al., 2022; Goodarzi et al., 2016; T. Gupta et al., 2022; Z. Zhang et al., 2018). Notably, such alterations have been reported in cancer and are used by oncogenic programs to promote tumor progression and enhance resistance to chemotherapy (Earnest-Noble et al., 2022; García-Vílchez et al., 2023; T. Gupta et al., 2022; Kochavi et al., 2024; Wernaart et al., 2024; C. Yang et al., 2024). However, it is still unexplored whether these alterations result from mutated tDNAs.

Alterations in tRNA biogenesis produced by mutations in tDNAs could lead to an imbalance of the tRNA pool which can promote the addition of non-cognate amino acids by the ribosome during protein synthesis (Kochavi et al., 2024). This suggests a potential link between tDNA mutagenesis and cancer development and/or progression given that translational fidelity is often compromised in cancer cells (Silvera et al., 2010; Weller et al., 2025). Mistranslation can be beneficial for cancer cells as it allows them to reshape the proteome and acquire aberrant peptides by increasing proteome diversity produced by non-genetic alterations of protein-coding genes (Weller et al., 2025; Wernaart et al., 2024). Such mistranslation could compromise the activity of tumor suppressors or generate proteins that favor tumor growth. Therefore, tDNA mutagenesis could be used by cancer cells as a mechanisms of adaptive mistranslation (Pan, 2013), which may support their readjustment to environmental changes, to allow tumor progression under harsh conditions or to develop drug resistance (Kochavi et al., 2024).

Aberrant peptides produced by errors during protein synthesis can act as tumor-specific antigens, known as neoantigens (Kochavi et al., 2024; Weller et al., 2025). These neoantigens are presented on the cell surface triggering immunological responses and can be used to boost immunotherapy treatments (Gubin et al., 2014). Remarkably, a recent study reported that cells with impaired translation produced by Phe-tRNAs that lack an essential modification to stabilize codon-anticodon interactions, showed an increase in presented neoantigens and consequently a superior response to immunotherapy treatments (Weller et al., 2025). Altogether, these findings suggest that similar phenotypes may be observed when mutations in tDNAs disrupt tRNA biogenesis and function, thereby altering translation and contributing to the production of targetable neoantigens. Therefore, we hypothesize that these mutations could predict tumor sensitivity to immunotherapy, indicating that patients with tDNA-mutated tumors may benefit from immunotherapy treatments.

We reported an apparent age-dependent accumulation of somatic mutagenesis in cancer tissues. The presence of somatic mutagenesis in healthy tissues suggests that age-related accumulation of somatic mutations could not be limited to cancer but could also be happening in normal tissues. However, as mentioned previously the sample size of healthy samples was limited and additional analyses are required to confirm this hypothesis. Exploring this age-related increase in healthy samples would be interesting because the accumulation of mutations in tDNAs could also compromise proteostasis, a recognized hallmark of aging (Anisimova et al., 2018; López-Otín et al., 2023; E. Schmidt & Schimmel, 1993). In this regard, age-associated hypermethylation of certain tDNAs has been observed, leading to the transcriptional silencing of certain tRNAs and decreasing their availability during translation (Acton et al., 2021). Hence, mutagenesis in tDNAs, could act similarly and contribute to tDNA impairment during aging.

Overall, the results and hypotheses presented here open exciting new fields to explore. This work raises many interesting questions and suggests promising links between tRNA biology, cancer, and aging.

5 CONCLUSIONS

- tRNAstudio provides a user-friendly platform for tRNA-Seq data analysis, designed for non-computational users.

- In comparison with other tools, the mapping strategy of tRNAstudio allows the exploration of tRNA biology from different perspectives:
    - Classifies tRNA-derived reads from mt-tRNAs and cytosolic pre-tRNAs or mature tRNAs.
    - Reports on tRNA gene sequence coverages, which can aid in the identification of *bona fide* tsRNAs.
    - Controls for multimapping to quantify individual tRNA sequences. This allows for differential expression analysis using two complementary methods, DESeq2 and iso-tRNA-CP, which account for the inherent biases specific to tRNA-Seq.
    - Performs mismatch analysis for the detection of post-transcriptional chemical modifications, including inosine and methylation.
- I34-tRNAs deficiency cannot be compensated by upregulating the expression of alternative tRNAs. This highlights the essential role of I34-tRNAs to efficiently translate eukarya-specific low-complexity proteins, such as ECM components.
- tDNA copy number can vary between human genome assemblies, with the T2T-CHM13 assembly containing the larger number of tDNAs.
- tDNAs organized in clusters exhibit higher expression levels than isolated tDNAs, and tDNA clustering can coordinate tDNA expression.
- tDNAs are hotspots of somatic mutagenesis in human cells, both in tumor and healthy tissues.
- tDNAs exhibit higher somatic mutation rates than protein-coding genes and other Pol III-transcribed genes.
- Mutation rates in tDNAs increase with transcriptional activity, providing evidence that tDNAs are especially prone to TAM.
- APOBEC3 is one of the main contributors to tDNAs mutagenesis. This mutagenesis is promoted by ssDNA secondary structures, such as DNA hairpins.
- tDNAs mutation rates vary by cancer type, with BLCA exhibiting the highest accumulation of such mutations.
- Mutation rate in tDNAs increases with age across multiple cancer types.
- In general, conserved positions within tDNAs show significantly lower mutation rates than the rest of the nucleotides in tDNAs. Other non-conserved tRNA positions are highly mutated and still may disrupt tRNA biogenesis and function, for example, by generating chimeric tRNAs that could promote mistranslation.

6 REFERENCES

Acton, R. J., Yuan, W., Gao, F., Xia, Y., Bourne, E., Wozniak, E., Bell, J., Lillycrop, K., Wang, J., Dennison, E., Harvey, N. C., Mein, C. A., Spector, T. D., Hysi, P. G., Cooper, C., & Bell, C. G. (2021). The genomic loci of specific human tRNA genes exhibit ageing-related DNA hypermethylation. *Nature Communications*, *12*(1), 2655.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, *136*(3), 927–935.

Alazami, A. M., Hijazi, H., Al-Dosari, M. S., Shaheen, R., Hashem, A., Aldahmesh, M. A., Mohamed, J. Y., Kentab, A., Salih, M. A., Awaji, A., Masoodi, T. A., & Alkuraya, F. S. (2013). Mutation in ADAT3, encoding adenosine deaminase acting on transfer RNA, causes intellectual disability and strabismus. *Journal of Medical Genetics*, *50*(7), 425–430.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *From RNA to Protein*. Garland Science.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., PCAWG Mutational Signatures Working Group, … PCAWG Consortium. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101.

Anderson, P., & Ivanov, P. (2014). tRNA fragments in human health and disease. *FEBS Letters*, *588*(23), 4297–4304.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., & Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, *290*(5806), 457–465.

Anisimova, A. S., Alexandrov, A. I., Makarova, N. E., Gladyshev, V. N., & Dmitriev, S. E. (2018). Protein synthesis and quality control in aging. *Aging*, *10*(12), 4269–4288.

Avcilar-Kucukgoze, I., & Kashina, A. (2020). Hijacking tRNAs from translation: Regulatory functions of tRNAs in mammalian cell physiology. *Frontiers in Molecular Biosciences*, *7*, 610617.

Ayan, G. B., Park, H. J., & Gallie, J. (2020). The birth of a bacterial tRNA gene by large-scale, tandem duplication events. *ELife*, *9*. https://doi.org/10.7554/eLife.57947

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D. C., Tamuri, A. U., Martincorena, I., Petljak, M., Alexandrov, L. B., Gundem, G., Tarpey, P. S., Roerink, S., Blokker, J., Maddison, M., Mudie, L., Robinson, B., Nik-Zainal, S., Campbell, P., Goldman, N., … Stratton, M. R. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, *513*(7518), 422–425.

Behrens, A., Rodschinka, G., & Nedialkova, D. D. (2021). High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Molecular Cell*, *81*(8), 1802-1815.e7.

Benhar, I., London, A., & Schwartz, M. (2012). The privileged immunity of immune privileged organs: the case of the eye. *Frontiers in Immunology*, *3*, 296.

Berg, M. D., & Brandl, C. J. (2021). Transfer RNAs: diversity in form and function. *RNA Biology*, *18*(3), 316–339.

Berg, M. D., Giguere, D. J., Dron, J. S., Lant, J. T., Genereaux, J., Liao, C., Wang, J., Robinson, J. F., Gloor, G. B., Hegele, R. A., O'Donoghue, P., & Brandl, C. J. (2019). Targeted sequencing reveals expanded genetic diversity of human transfer RNAs. *RNA Biology*, *16*(11), 1574–1585.

Beringer, M., & Rodnina, M. V. (2007). The ribosomal peptidyl transferase. *Molecular Cell*, *26*(3), 311–321.

Bermudez-Santana, C., Attolini, C. S.-O., Kirsten, T., Engelhardt, J., Prohaska, S. J., Steigele, S., & Stadler, P. F. (2010). Genomic organization of eukaryotic tRNAs. *BMC Genomics*, *11*(1), 270.

Bespalov, M. M., Sidorova, Y. A., Tumova, S., Ahonen-Bishopp, A., Magalhães, A. C., Kulesskiy, E., Paveliev, M., Rivera, C., Rauvala, H., & Saarma, M. (2011). Heparan sulfate proteoglycan syndecan-3 is a novel receptor for GDNF, neurturin, and artemin. *The Journal of Cell Biology*, *192*(1), 153–169.

Betat, H., & Mörl, M. (2015). The CCA-adding enzyme: A central scrutinizer in tRNA quality control. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *37*(9), 975–982.

Bezerra, A. R., Simões, J., Lee, W., Rung, J., Weil, T., Gut, I. G., Gut, M., Bayés, M., Rizzetto, L., Cavalieri, D., Giovannini, G., Bozza, S., Romani, L., Kapushesky, M., Moura, G. R., & Santos, M. A. S. (2013). Reversion of a fungal genetic code alteration links proteome instability with genomic and phenotypic diversification. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(27), 11079–11084.

Biela, A., Hammermeister, A., Kaczmarczyk, I., Walczak, M., Koziej, L., Lin, T.-Y., & Glatt, S. (2023). The diverse structural modes of tRNA binding and recognition. *The Journal of Biological Chemistry*, *299*(8), 104966.

Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., Kurkowska, M., Shirvanizadeh, N., Destefanis, E., Groza, P., Avşar, G., Romitelli, A., Pir, P., Dassi, E., Conticello, S. G., Aguilo, F., & Bujnicki, J. M. (2022). MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Research*, *50*(D1), D231–D235.

Buisson, R., Langenbucher, A., Bowen, D., Kwan, E. E., Benes, C. H., Zou, L., & Lawrence, M. S. (2019). Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science (New York, N.Y.)*, *364*(6447), eaaw2872.

Burns, M. B., Lackey, L., Carpenter, M. A., Rathore, A., Land, A. M., Leonard, B., Refsland, E. W., Kotandeniya, D., Tretyakova, N., Nikas, J. B., Yee, D., Temiz, N. A., Donohue, D. E., McDougle, R. M., Brown, W. L., Law, E. K., & Harris, R. S. (2013b). APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*, *494*(7437), 366–370.

Burns, M. B., Temiz, N. A., & Harris, R. S. (2013a). Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nature Genetics*, *45*(9), 977–983.

Butler, K., & Banday, A. R. (2023). APOBEC3-mediated mutagenesis in cancer: causes, clinical significance and therapeutic potential. *Journal of Hematology & Oncology*, *16*(1), 31.

Butt, Y., Sakhtemani, R., Mohamad-Ramshan, R., Lawrence, M. S., & Bhagwat, A. S. (2024). Distinguishing preferences of human APOBEC3A and APOBEC3B for cytosines in hairpin loops, and reflection of these preferences in APOBEC-signature cancer genome mutations. *Nature Communications*, *15*(1), 2369.

Cabrelle, C., Giorgi, F. M., & Mercatelli, D. (2024). Quantitative and qualitative detection of tRNAs, tRNA halves and tRFs in human cancer samples: Molecular grounds for biomarker development and clinical perspectives. *Gene*, *898*, 148097.

Cappannini, A., Ray, A., Purta, E., Mukherjee, S., Boccaletto, P., Moafinejad, S. N., Lechner, A., Barchet, C., Klaholz, B. P., Stefaniak, F., & Bujnicki, J. M. (2024). MODOMICS: a database of RNA modifications and related information. 2023 update. *Nucleic Acids Research*, *52*(D1), D239–D244.

Carpenter, J., Wang, Y., Gupta, R., Li, Y., Haridass, P., Subramani, D. B., Reidel, B., Morton, L., Ridley,

C., O'Neal, W. K., Buisine, M.-P., Ehre, C., Thornton, D. J., & Kesimer, M. (2021). Assembly and organization of the N-terminal region of mucin MUC5AC: Indications for structural and functional distinction from MUC5B. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(39), e2104490118.

Chan, C. T. Y., Dyavaiah, M., DeMott, M. S., Taghizadeh, K., Dedon, P. C., & Begley, T. J. (2010). A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genetics*, *6*(12), e1001247.

Chan, P. P., Holmes, A. D., & Lowe, T. M. (2025). Analyzing, visualizing, and annotating tRNA-derived RNAs using tRAX and tDRnamer. *Methods in Enzymology*, *711*, 103–133.

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096.

Chan, P. P., & Lowe, T. M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Research*, *44*(D1), D184-9.

Chan, P. P., & Lowe, T. M. (2019). TRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods in Molecular Biology (Clifton, N.J.)*, *1962*, 1–14.

Chen, L., Chen, J.-Y., Zhang, X., Gu, Y., Xiao, R., Shao, C., Tang, P., Qian, H., Luo, D., Li, H., Zhou, Y., Zhang, D.-E., & Fu, X.-D. (2017). R-ChIP using inactive RNase H reveals dynamic coupling of R-loops with transcriptional pausing at gene promoters. *Molecular Cell*, *68*(4), 745-757.e5.

Clark, W. C., Evans, M. E., Dominissini, D., Zheng, G., & Pan, T. (2016). tRNA base methylation identification and quantification via high-throughput sequencing. *RNA (New York, N.Y.)*, *22*(11), 1771–1784.

Correia, I., Oliveira, C., Reis, A., Guimarães, A. R., Aveiro, S., Domingues, P., Bezerra, A. R., Vitorino, R., Moura, G., & Santos, M. A. S. (2024). A proteogenomic pipeline for the analysis of protein biosynthesis errors in the human pathogen candida albicans. *Molecular & Cellular Proteomics: MCP*, *23*(9), 100818.

Cozen, A. E., Quartley, E., Holmes, A. D., Hrabeta-Robinson, E., Phizicky, E. M., & Lowe, T. M. (2015). ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nature Methods*, *12*(9), 879–884.

Cozma, E., Rao, M., Dusick, M., Genereaux, J., Rodriguez-Mias, R. A., Villén, J., Brandl, C. J., & Berg, M. D. (2023). Anticodon sequence determines the impact of mistranslating tRNAAla variants. *RNA Biology*, *20*(1), 791–804.

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

Crick, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, *12*, 138–163.

Crick, F. H. (1966). Codon--anticodon pairing: the wobble hypothesis. *Journal of Molecular Biology*, *19*(2), 548–555.

Crick, F. H., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, *192*(4809), 1227–1232.

Cui, J., Liu, Q., Sendinc, E., Shi, Y., & Gregory, R. I. (2021). Nucleotide resolution profiling of m3C RNA modification by HAC-seq. *Nucleic Acids Research*, *49*(5), e27.

Dammann, R., & Pfeifer, G. P. (1997). Lack of gene- and strand-specific DNA repair in RNA polymerase III-transcribed human tRNA genes. *Molecular and Cellular Biology*, *17*(1), 219–229.

Darrow, E. M., & Chadwick, B. P. (2014). A novel tRNA variable number tandem repeat at human chromosome 1q23.3 is implicated as a boundary element based on conservation of a CTCF motif in mouse. *Nucleic Acids Research*, *42*(10), 6421–6435.

David, M., Dzamba, M., Lister, D., Ilie, L., & Brudno, M. (2011). SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics (Oxford, England)*, *27*(7), 1011–1012.

Davies, J. E., & Rubinsztein, D. C. (2006). Polyalanine and polyserine frameshift products in Huntington's disease. *Journal of Medical Genetics*, *43*(11), 893–896.

de Crécy-Lagard, V., Boccaletto, P., Mangleburg, C. G., Sharma, P., Lowe, T. M., Leidel, S. A., & Bujnicki, J. M. (2019). Matching tRNA modifications in humans to their known and predicted enzymes. *Nucleic Acids Research*, *47*(5), 2143–2159.

de Crécy-Lagard, V., & Jaroch, M. (2021). Functions of bacterial tRNA modifications: From ubiquity to diversity. *Trends in Microbiology*, *29*(1), 41–53.

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

Dittmar, K. A., Goodenbour, J. M., & Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. *PLoS Genetics*, *2*(12), e221.

Dittmar, K. A., Sørensen, M. A., Elf, J., Ehrenberg, M., & Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Reports*, *6*(2), 151–157.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380.

Donovan, P. D., McHale, N. M., Venø, M. T., & Prehn, J. H. M. (2021). tsRNAsearch: a pipeline for the identification of tRNA and ncRNA fragments from small RNA-sequencing data. *Bioinformatics (Oxford, England)*, *37*(23), 4424–4430.

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, *134*(2), 341–352.

Dunin-Horkawicz, S., Czerwoniec, A., Gajda, M. J., Feder, M., Grosjean, H., & Bujnicki, J. M. (2006). MODOMICS: a database of RNA modification pathways. *Nucleic Acids Research*, *34*(Database issue), D145-9.

Earnest-Noble, L. B., Hsu, D., Chen, S., Asgharian, H., Nandan, M., Passarelli, M. C., Goodarzi, H., & Tavazoie, S. F. (2022). Two isoleucyl tRNAs that decode synonymous codons divergently regulate breast cancer metastatic growth by controlling translation of proliferation-regulating genes. *Nature Cancer*, *3*(12), 1484–1497.

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

El Hage, A., Webb, S., Kerr, A., & Tollervey, D. (2014). Genome-wide distribution of RNA-DNA hybrids identifies RNase H targets in tRNA genes, retrotransposons and mitochondria. *PLoS Genetics*, *10*(10), e1004716.

Elsakrmy, N., & Cui, H. (2023). R-loops and R-loop-binding proteins in cancer progression and drug resistance. *International Journal of Molecular Sciences*, *24*(8). https://doi.org/10.3390/ijms24087064

Erber, L., Hoffmann, A., Fallmann, J., Betat, H., Stadler, P. F., & Mörl, M. (2020). LOTTE-seq (Long

hairpin oligonucleotide based tRNA high-throughput sequencing): specific selection of tRNAs with 3'-CCA end for high-throughput sequencing. *RNA Biology*, *17*(1), 23–32.

Feinberg, J. S., & Joseph, S. (2001). Identification of molecular interactions between P-site tRNA and the ribosome essential for translocation. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11120–11125.

Ferreira, D. A., Tayyar, Y., Idris, A., & McMillan, N. A. J. (2021). A "hit-and-run" affair - A possible link for cancer progression in virally driven cancers. *Biochimica et Biophysica Acta. Reviews on Cancer*, *1875*(1), 188476.

Fichant, G. A., & Burks, C. (1991). Identifying potential tRNA genes in genomic DNA sequences. *Journal of Molecular Biology*, *220*(3), 659–671.

Franco, I., Helgadottir, H. T., Moggio, A., Larsson, M., Vrtačnik, P., Johansson, A., Norgren, N., Lundin, P., Mas-Ponte, D., Nordström, J., Lundgren, T., Stenvinkel, P., Wennberg, L., Supek, F., & Eriksson, M. (2019). Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. *Genome Biology*, *20*(1), 285.

Frugier, M., Bour, T., Ayach, M., Santos, M. A. S., Rudinger-Thirion, J., Théobald-Dietrich, A., & Pizzi, E. (2010). Low Complexity Regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. *FEBS Letters*, *584*(2), 448–454.

Gao, L., Behrens, A., Rodschinka, G., Forcelloni, S., Wani, S., Strasser, K., & Nedialkova, D. D. (2024). Selective gene expression maintains human tRNA anticodon pools during differentiation. *Nature Cell Biology*, *26*(1), 100–112.

García-Vílchez, R., Añazco-Guenkova, A. M., López, J., Dietmann, S., Tomé, M., Jimeno, S., Azkargorta, M., Elortza, F., Bárcena, L., Gonzalez-Lopez, M., Aransay, A. M., Sánchez-Martín, M. A., Huertas, P., Durán, R. V., & Blanco, S. (2023). N7-methylguanosine methylation of tRNAs regulates survival to stress in cancer. *Oncogene*, *42*(43), 3169–3181.

Gerber, A. P., & Keller, W. (1999). An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science (New York, N.Y.)*, *286*(5442), 1146–1149.

Geslain, R., Cubells, L., Bori-Sanz, T., Álvarez-Medina, R., Rossell, D., Martí, E., & de Pouplana, L. R. (2010). Chimeric tRNAs as tools to induce proteome damage and identify components of stress responses. *Nucleic Acids Research*, *38*(5), e30–e30.

Geslain, R., & Pan, T. (2010). Functional analysis of human tRNA isodecoders. *Journal of Molecular Biology*, *396*(3), 821–831.

Giegé, R., & Eriani, G. (2023). The tRNA identity landscape for aminoacylation and beyond. *Nucleic Acids Research*, *51*(4), 1528–1570.

Giegé, R., Jühling, F., Pütz, J., Stadler, P., Sauter, C., & Florentz, C. (2012). Structure of transfer RNAs: similarity and variability. *Wiley Interdisciplinary Reviews. RNA*, *3*(1), 37–61.

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

Gingold, H., Tehler, D., Christoffersen, N. R., Nielsen, M. M., Asmar, F., Kooistra, S. M., Christophersen, N. S., Christensen, L. L., Borre, M., Sørensen, K. D., Andersen, L. D., Andersen, C. L., Hulleman, E., Wurdinger, T., Ralfkiær, E., Helin, K., Grønbæk, K., Ørntoft, T., Waszak, S. M., … Pilpel, Y. (2014). A dual program for translation regulation in cellular proliferation and differentiation. *Cell*, *158*(6), 1281–1292.

Gogakos, T., Brown, M., Garzia, A., Meyer, C., Hafner, M., & Tuschl, T. (2017). Characterizing expression and processing of precursor and mature human tRNAs by hydro-tRNAseq and PAR-CLIP. *Cell Reports*, *20*(6), 1463–1475.

Goldkamp, A. K., Li, Y., Rivera, R. M., & Hagen, D. E. (2022). Characterization of tRNA expression profiles in large offspring syndrome. *BMC Genomics*, *23*(1), 273.

Good, P. D., Kendall, A., Ignatz-Hoover, J., Miller, E. L., Pai, D. A., Rivera, S. R., Carrick, B., & Engelke, D. R. (2013). Silencing near tRNA genes is nucleosome-mediated and distinct from boundary element function. *Gene*, *526*(1), 7–15.

Goodarzi, H., Nguyen, H. C. B., Zhang, S., Dill, B. D., Molina, H., & Tavazoie, S. F. (2016). Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell*, *165*(6), 1416–1427.

Goodenbour, J. M., & Pan, T. (2006). Diversity of tRNA genes in eukaryotes. *Nucleic Acids Research*, *34*(21), 6137–6146.

Gubin, M. M., Zhang, X., Schuster, H., Caron, E., Ward, J. P., Noguchi, T., Ivanova, Y., Hundal, J., Arthur, C. D., Krebber, W.-J., Mulder, G. E., Toebes, M., Vesely, M. D., Lam, S. S. K., Korman, A. J., Allison, J. P., Freeman, G. J., Sharpe, A. H., Pearce, E. L., … Schreiber, R. D. (2014). Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*, *515*(7528), 577–581.

Guimarães, A. R., Correia, I., Sousa, I., Oliveira, C., Moura, G., Bezerra, A. R., & Santos, M. A. S. (2021). TRNAs as a driving force of genome evolution in yeast. *Frontiers in Microbiology*, *12*, 634004.

Gupta, R., & Laxman, S. (2020). tRNA wobble-uridine modifications as amino acid sensors and regulators of cellular metabolic state. *Current Genetics*, *66*(3), 475–480.

Gupta, T., Malkin, M. G., & Huang, S. (2022). TRNA function and dysregulation in cancer. *Frontiers in Cell and Developmental Biology*, *10*, 886642.

Hamdani, O., Dhillon, N., Hsieh, T.-H. S., Fujita, T., Ocampo, J., Kirkland, J. G., Lawrimore, J., Kobayashi, T. J., Friedman, B., Fulton, D., Wu, K. Y., Chereji, R. V., Oki, M., Bloom, K., Clark, D. J., Rando, O. J., & Kamakaka, R. T. (2019). TRNA genes affect chromosome structure and function via local effects. *Molecular and Cellular Biology*, *39*(8), 1–26.

Han, L., & Phizicky, E. M. (2018). A rationale for tRNA modification circuits in the anticodon loop. *RNA (New York, N.Y.)*, *24*(10), 1277–1284.

Hanson, G., & Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews. Molecular Cell Biology*, *19*(1), 20–30.

Hauenschild, R., Tserovski, L., Schmid, K., Thüring, K., Winz, M.-L., Sharma, S., Entian, K.-D., Wacheul, L., Lafontaine, D. L. J., Anderson, J., Alfonzo, J., Hildebrandt, A., Jäschke, A., Motorin, Y., & Helm, M. (2015). The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Research*, *43*(20), 9950–9964.

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

Hienola, A., Tumova, S., Kulesskiy, E., & Rauvala, H. (2006). N-syndecan deficiency impairs neural migration in brain. *The Journal of Cell Biology*, *174*(4), 569–580.

Hoffmann, A., Fallmann, J., Vilardo, E., Mörl, M., Stadler, P. F., & Amman, F. (2018). Accurate mapping of tRNA reads. *Bioinformatics (Oxford, England)*, *34*(13), 2339.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F., & Hackermüller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Computational Biology*, *5*(9), e1000502.

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

Holmes, A. D., Howard, J. M., Chan, P. P., & Lowe, T. M. (2022). tRNA Analysis of eXpression (tRAX): A tool for integrating analysis of tRNAs, tRNA-derived small RNAs, and tRNA modifications. In *bioRxiv*. https://doi.org/10.1101/2022.07.02.498565

Honda, S., Loher, P., Shigematsu, M., Palazzo, J. P., Suzuki, R., Imoto, I., Rigoutsos, I., & Kirino, Y. (2015). Sex hormone-dependent tRNA halves enhance cell proliferation in breast and prostate cancers. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(29), E3816-25.

Hopper, A. K., Pai, D. A., & Engelke, D. R. (2010). Cellular dynamics of tRNAs and their genes. *FEBS Letters*, *584*(2), 310–317.

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., de Lima, L. G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., Altemose, N., Rhie, A., Core, L. J., Gerton, J. L., Makalowski, W., Olson, D., Rosen, J., Smit, A. F. A., Straight, A. F., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science (New York, N.Y.)*, *376*(6588), eabk3112.

Hu, J. F., Yim, D., Ma, D., Huber, S. M., Davis, N., Bacusmo, J. M., Vermeulen, S., Zhou, J., Begley, T. J., DeMott, M. S., Levine, S. S., de Crécy-Lagard, V., Dedon, P. C., & Cao, B. (2021). Quantitative mapping of the cellular small RNA landscape with AQRNA-seq. *Nature Biotechnology*, *39*(8), 978–988.

Huang, S.-Q., Sun, B., Xiong, Z.-P., Shu, Y., Zhou, H.-H., Zhang, W., Xiong, J., & Li, Q. (2018). The dysregulation of tRNAs and tRNA derivatives in cancer. *Journal of Experimental & Clinical Cancer Research: CR*, *37*(1). https://doi.org/10.1186/s13046-018-0745-z

Hudák, A., Letoha, A., Vizler, C., & Letoha, T. (2022). Syndecan-3 as a novel biomarker in Alzheimer's disease. *International Journal of Molecular Sciences*, *23*(6), 3407.

Hughes, L. A., Rudler, D. L., Siira, S. J., McCubbin, T., Raven, S. A., Browne, J. M., Ermer, J. A., Rientjes, J., Rodger, J., Marcellin, E., Rackham, O., & Filipovska, A. (2023). Copy number variation in tRNA isodecoder genes impairs mammalian development and balanced translation. *Nature Communications*, *14*(1), 2210.

Hummel, G., Warren, J., & Drouard, L. (2019). The multi-faceted regulation of nuclear tRNA gene transcription. *IUBMB Life*, *71*(8), 1099–1108.

Ibba, M., & Söll, D. (2000). Aminoacyl-tRNA synthesis. *Annual Review of Biochemistry*, *69*(1), 617–650.

Iben, J. R., & Maraia, R. J. (2012). tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon:codon wobble pair in a eukaryote. *RNA (New York, N.Y.)*, *18*(7), 1358–1372.

Iben, J. R., & Maraia, R. J. (2014). tRNA gene copy number variation in humans. *Gene*, *536*(2), 376–384.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, *2*(1), 13–34.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, *324*(5924), 218–223.

Ippolito, M. R., Martis, V., Martin, S., Tijhuis, A. E., Hong, C., Wardenaar, R., Dumont, M., Zerbib, J., Spierings, D. C. J., Fachinetti, D., Ben-David, U., Foijer, F., & Santaguida, S. (2021). Gene copy-

number changes and chromosomal instability induced by aneuploidy confer resistance to chemotherapy. *Developmental Cell*, *56*(17), 2440-2454.e6.

Ishimura, R., Nagy, G., Dotu, I., Zhou, H., Yang, X.-L., Schimmel, P., Senju, S., Nishimura, Y., Chuang, J. H., & Ackerman, S. L. (2014). RNA function. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science (New York, N.Y.)*, *345*(6195), 455–459.

Ivanov, P., Emara, M. M., Villen, J., Gygi, S. P., & Anderson, P. (2011). Angiogenin-induced tRNA fragments inhibit translation initiation. *Molecular Cell*, *43*(4), 613–623.

Iwasaki, Y., Ikemura, T., Kurokawa, K., & Okada, N. (2020). Implication of a new function of human tDNAs in chromatin organization. *Scientific Reports*, *10*(1), 17440.

Jang, G. M., Annan Sudarsan, A. K., Shayeganmehr, A., Prando Munhoz, E., Lao, R., Gaba, A., Granadillo Rodríguez, M., Love, R. P., Polacco, B. J., Zhou, Y., Krogan, N. J., Kaake, R. M., & Chelico, L. (2024). Protein interaction map of APOBEC3 enzyme family reveals deamination-independent role in cellular function. *Molecular & Cellular Proteomics: MCP*, *23*(5), 100755.

Jin, H., Yeom, J.-H., Shin, E., Ha, Y., Liu, H., Kim, D., Joo, M., Kim, Y.-H., Kim, H. K., Ryu, M., Kim, H.-M., Kim, J., Kim, K. P., Hahn, Y., Bae, J., & Lee, K. (2024). 5'-tRNAGly(GCC) halves generated by IRE1α are linked to the ER stress response. *Nature Communications*, *15*(1), 9273.

Karamanos, N. K., Theocharis, A. D., Piperigkou, Z., Manou, D., Passi, A., Skandalis, S. S., Vynios, D. H., Orian-Rousseau, V., Ricard-Blum, S., Schmelzer, C. E. H., Duca, L., Durbeej, M., Afratis, N. A., Troeberg, L., Franchi, M., Masola, V., & Onisto, M. (2021). A guide to the composition and functions of the extracellular matrix. *The FEBS Journal*, *288*(24), 6850–6912.

Kawalec, P. (2016). Indirect costs of inflammatory bowel diseases: Crohn's disease and ulcerative colitis. A systematic review. *Archives of Medical Science: AMS*, *12*(2), 295–302.

Khomarbaghi, Z., Ngan, W. Y., Ayan, G. B., Lim, S., Dechow-Seligmann, G., Nandy, P., & Gallie, J. (2024). Large-scale duplication events underpin population-level flexibility in tRNA gene copy number in Pseudomonas fluorescens SBW25. *Nucleic Acids Research*, *52*(5), 2446–2462.

Kim, S. H., Sussman, J. L., Suddath, F. L., Quigley, G. J., McPherson, A., Wang, A. H., Seeman, N. C., & Rich, A. (1974). The general structure of transfer RNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, *71*(12), 4970–4974.

Kochavi, A., Nagel, R., Körner, P.-R., Bleijerveld, O. B., Lin, C.-P., Huinen, Z., Malka, Y., Proost, N., van de Ven, M., Feng, X., Navarro, J. M., Pataskar, A., Peeper, D. S., Champagne, J., & Agami, R. (2024). Chemotherapeutic agents and leucine deprivation induce codon-biased aberrant protein production in cancer. *Nucleic Acids Research*, *52*(22), 13964–13979.

Kombe Kombe, A. J., Fotoohabadi, L., Gerasimova, Y., Nanduri, R., Lama Tamang, P., Kandala, M., & Kelesidis, T. (2024). The role of inflammation in the pathogenesis of viral respiratory infections. *Microorganisms*, *12*(12), 2526.

Kramerov, D. A., & Vassetzky, N. S. (2011). Origin and evolution of SINEs in eukaryotic genomes. *Heredity*, *107*(6), 487–495.

Krokan, H. E., & Bjørås, M. (2013). Base excision repair. *Cold Spring Harbor Perspectives in Biology*, *5*(4), a012583.

Kumar, P., Anaya, J., Mudunuri, S. B., & Dutta, A. (2014). Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biology*, *12*(1), 78.

Kumar, P., Kuscu, C., & Dutta, A. (2016). Biogenesis and function of transfer RNA-related fragments (tRFs). *Trends in Biochemical Sciences*, *41*(8), 679–689.

Kuscu, C., Kumar, P., Kiran, M., Su, Z., Malik, A., & Dutta, A. (2018). tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA (New York, N.Y.)*, *24*(8), 1093–1105.

La Ferlita, A., Nigita, G., Ferro, A., Alaimo, S., & Pulvirenti, A. (2025). Protocol for analyzing tRNA-derived ncRNAs from small RNA-seq data using tRFUniverse functional analyses. *STAR Protocols*, *6*(2), 103884.

Langenbucher, A., Bowen, D., Sakhtemani, R., Bournique, E., Wise, J. F., Zou, L., Bhagwat, A. S., Buisson, R., & Lawrence, M. S. (2021). An extended APOBEC3A mutation signature in cancer. *Nature Communications*, *12*(1), 1602.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Lant, J. T., Berg, M. D., Heinemann, I. U., Brandl, C. J., & O'Donoghue, P. (2019). Pathways to disease from natural variations in human cytoplasmic tRNAs. *The Journal of Biological Chemistry*, *294*(14), 5294–5308.

Lassak, J., Wilson, D. N., & Jung, K. (2016). Stall no more at polyproline stretches with the translation elongation factors EF-P and IF-5A: Stall no more. *Molecular Microbiology*, *99*(2), 219–235.

Lee, J. W., Beebe, K., Nangle, L. A., Jang, J., Longo-Guess, C. M., Cook, S. A., Davisson, M. T., Sundberg, J. P., Schimmel, P., & Ackerman, S. L. (2006). Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nature*, *443*(7107), 50–55.

Lee, J.-O., Chu, J., Jang, G., Lee, M., & Chung, Y.-J. (2022). tReasure: R-based GUI package analyzing tRNA expression profiles from small RNA sequencing data. *BMC Bioinformatics*, *23*(1), 155.

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

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324.

Lindskrog, S. V., Prip, F., Lamy, P., Taber, A., Groeneveld, C. S., Birkenkamp-Demtröder, K., Jensen, J. B., Strandgaard, T., Nordentoft, I., Christensen, E., Sokac, M., Birkbak, N. J., Maretty, L., Hermann, G. G., Petersen, A. C., Weyerer, V., Grimm, M.-O., Horstmann, M., Sjödahl, G., … Dyrskjøt, L. (2021). An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nature Communications*, *12*(1), 2301.

Liu, K., & Sun, Q. (2021). Intragenic tRNA-promoted R-loops orchestrate transcription interference for plant oxidative stress responses. *The Plant Cell*, *33*(11), 3574–3591.

Liu, Y., Satz, J. S., Vo, M.-N., Nangle, L. A., Schimmel, P., & Ackerman, S. L. (2014). Deficiencies in tRNA synthetase editing activity cause cardioproteinopathy. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(49), 17570–17575.

Loher, P., Telonis, A. G., & Rigoutsos, I. (2017). MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Scientific Reports*, *7*(1), 41184.

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2023). Hallmarks of aging: An expanding universe. *Cell*, *186*(2), 243–278.

Lorenz, C., Lünse, C. E., & Mörl, M. (2017). TRNA modifications: Impact on structure and thermal adaptation. *Biomolecules*, *7*(2). https://doi.org/10.3390/biom7020035

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

Lucas, M. C., Pryszcz, L. P., Medina, R., Milenkovic, I., Camacho, N., Marchand, V., Motorin, Y., Ribas de Pouplana, L., & Novoa, E. M. (2024). Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. *Nature Biotechnology*, *42*(1), 72–86.

Ma, J., Rubin, B. K., & Voynow, J. A. (2018). Mucins, mucus, and goblet cells. *Chest*, *154*(1), 169–176.

Machnicka, M. A., Olchowik, A., Grosjean, H., & Bujnicki, J. M. (2014). Distribution and frequencies of post-transcriptional modifications in tRNAs. *RNA Biology*, *11*(12), 1619–1629.

Male, G., von Appen, A., Glatt, S., Taylor, N. M. I., Cristovao, M., Groetsch, H., Beck, M., & Müller, C. W. (2015). Architecture of TFIIIC and its role in RNA polymerase III pre-initiation complex assembly. *Nature Communications*, *6*(1), 7387.

Maraia, R. J., & Arimbasseri, A. G. (2017). Factors that shape eukaryotic tRNAomes: Processing, modification and anticodon-Codon use. *Biomolecules*, *7*(1). https://doi.org/10.3390/biom7010026

Maric, C., & Prioleau, M.-N. (2010). Interplay between DNA replication and gene expression: a harmonious coexistence. *Current Opinion in Cell Biology*, *22*(3), 277–283.

McCann, J. L., Cristini, A., Law, E. K., Lee, S. Y., Tellier, M., Carpenter, M. A., Beghè, C., Kim, J. J., Sanchez, A., Jarvis, M. C., Stefanovska, B., Temiz, N. A., Bergstrom, E. N., Salamango, D. J., Brown, M. R., Murphy, S., Alexandrov, L. B., Miller, K. M., Gromak, N., & Harris, R. S. (2023). APOBEC3B regulates R-loops and promotes transcription-associated mutagenesis in cancer. *Nature Genetics*, *55*(10), 1721–1734.

McFarlane, R. J., & Whitehall, S. K. (2009). tRNA genes in eukaryotic genome organization and reorganization. *Cell Cycle (Georgetown, Tex.)*, *8*(19), 3102–3106.

Meng, Y., & Zou, L. (2025). Building an integrated view of R-loops, transcription, and chromatin. *DNA Repair*, *149*(103832), 103832.

Mier, P., Paladin, L., Tamana, S., Petrosian, S., Hajdu-Soltész, B., Urbanek, A., Gruca, A., Plewczynski, D., Grynberg, M., Bernadó, P., Gáspári, Z., Ouzounis, C. A., Promponas, V. J., Kajava, A. V., Hancock, J. M., Tosatto, S. C. E., Dosztanyi, Z., & Andrade-Navarro, M. A. (2020). Disentangling the complexity of low complexity proteins. *Briefings in Bioinformatics*, *21*(2), 458–472.

Miranda, I., Rocha, R., Santos, M. C., Mateus, D. D., Moura, G. R., Carreto, L., & Santos, M. A. S. (2007). A genetic code alteration is a phenotype diversity generator in the human pathogen Candida albicans. *PloS One*, *2*(10), e996.

Miranda, I., Silva-Dias, A., Rocha, R., Teixeira-Santos, R., Coelho, C., Gonçalves, T., Santos, M. A. S., Pina-Vaz, C., Solis, N. V., Filler, S. G., & Rodrigues, A. G. (2013). Candida albicans CUG mistranslation is a mechanism to create cell surface variation. *MBio*, *4*(4). https://doi.org/10.1128/mBio.00285-13

Mishra, S., & Whetstine, J. R. (2016). Different facets of copy number changes: Permanent, transient, and adaptive. *Molecular and Cellular Biology*, *36*(7), 1050–1063.

Mohler, K., & Ibba, M. (2017). Translational fidelity and mistranslation in the cellular response to stress. *Nature Microbiology*, *2*, 17117.

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *38*(1), 23–38.

Morgado, S. M., & Vicente, A. C. P. (2019). Exploring tRNA gene cluster in archaea. *Memorias Do*

Instituto Oswaldo Cruz, *114*(0), e180348.

Motorin, Y., & Helm, M. (2019). Methods for RNA modification mapping using deep sequencing: Established and new emerging technologies. *Genes*, *10*(1), 35.

Motorin, Y., & Helm, M. (2024). General principles and limitations for detection of RNA modifications by sequencing. *Accounts of Chemical Research*, *57*(3), 275–288.

Müller, C. A., & Nieduszynski, C. A. (2017). DNA replication timing influences gene expression level. *The Journal of Cell Biology*, *216*(7), 1907–1914.

Mungall, A. J., Palmer, S. A., Sims, S. K., Edwards, C. A., Ashurst, J. L., Wilming, L., Jones, M. C., Horton, R., Hunt, S. E., Scott, C. E., Gilbert, J. G. R., Clamp, M. E., Bethel, G., Milne, S., Ainscough, R., Almeida, J. P., Ambrose, K. D., Andrews, T. D., Ashwell, R. I. S., … Beck, S. (2003). The DNA sequence and analysis of human chromosome 6. *Nature*, *425*(6960), 805–811.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics (Oxford, England)*, *29*(22), 2933–2935.

Netzer, N., Goodenbour, J. M., David, A., Dittmar, K. A., Jones, R. B., Schneider, J. R., Boone, D., Eves, E. M., Rosner, M. R., Gibbs, J. S., Embry, A., Dolan, B., Das, S., Hickman, H. D., Berglund, P., Bennink, J. R., Yewdell, J. W., & Pan, T. (2009). Innate immune and chemically triggered oxidative stress modifies translational fidelity. *Nature*, *462*(7272), 522–526.

Nilsson, A. I., Zorzet, A., Kanth, A., Dahlström, S., Berg, O. G., & Andersson, D. I. (2006). Reducing the fitness cost of antibiotic resistance by amplification of initiator tRNA genes. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(18), 6976–6981.

Nirenberg, M. W., & Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, *47*(10), 1588–1602.

Novoa, E. M., Pavon-Eternod, M., Pan, T., & Ribas de Pouplana, L. (2012). A role for tRNA modifications in genome structure and codon usage. *Cell*, *149*(1), 202–213.

Novoa, E. M., & Ribas de Pouplana, L. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. *Trends in Genetics: TIG*, *28*(11), 574–581.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science (New York, N.Y.)*, *376*(6588), 44–53.

Oberbauer, V., & Schaefer, M. R. (2018). TRNA-derived small RNAs: Biogenesis, modification, function and potential impact on human disease development. *Genes*, *9*(12), 607.

Ojala, D., Montoya, J., & Attardi, G. (1981). tRNA punctuation model of RNA processing in human mitochondria. *Nature*, *290*(5806), 470–474.

Orellana, E. A., Siegal, E., & Gregory, R. I. (2022). tRNA dysregulation and disease. *Nature Reviews. Genetics*, *23*(11), 651–664.

Padhiar, N. H., Katneni, U., Komar, A. A., Motorin, Y., & Kimchi-Sarfaty, C. (2024). Advances in methods for tRNA sequencing and quantification. *Trends in Genetics: TIG*, *40*(3), 276–290.

Pan, T. (2013). Adaptive translation as a mechanism of stress response and adaptation. *Annual Review of Genetics*, *47*(1), 121–137.

Pan, T. (2018). Modifications and functional genomics of human transfer RNA. *Cell Research*, *28*(4),

395–404.

Pandey, K. K., Madhry, D., Ravi Kumar, Y. S., Malvankar, S., Sapra, L., Srivastava, R. K., Bhattacharyya, S., & Verma, B. (2021). Regulatory roles of tRNA-derived RNA fragments in human pathophysiology. *Molecular Therapy. Nucleic Acids*, *26*, 161–173.

Panoutsopoulou, K., Dreyer, T., Dorn, J., Obermayr, E., Mahner, S., van Gorp, T., Braicu, I., Zeillinger, R., Magdolen, V., Avgeris, M., & Scorilas, A. (2021). TRNAGlyGCC-derived internal fragment (i-tRF-GlyGCC) in ovarian cancer treatment outcome and progression. *Cancers*, *14*(1), 24.

Papadimitriou, M.-A., Avgeris, M., Levis, P., Papasotiriou, E. C., Kotronopoulos, G., Stravodimos, K., & Scorilas, A. (2020). TRNA-derived fragments (tRFs) in bladder cancer: Increased 5'-tRF-LysCTT results in disease early progression and patients' poor treatment outcome. *Cancers*, *12*(12), 3661.

Parisien, M., Wang, X., & Pan, T. (2013). Diversity of human tRNA genes from the 1000-genomes project. *RNA Biology*, *10*(12), 1853–1867.

Parvathy, S. T., Udayasuriyan, V., & Bhadana, V. (2022). Codon usage bias. *Molecular Biology Reports*, *49*(1), 539–565.

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

Pavon-Eternod, M., Wei, M., Pan, T., & Kleiman, L. (2010). Profiling non-lysyl tRNAs in HIV-1. *RNA (New York, N.Y.)*, *16*(2), 267–273.

Pekarsky, Y., Balatti, V., & Croce, C. M. (2023). tRNA-derived fragments (tRFs) in cancer. *Journal of Cell Communication and Signaling*, *17*(1), 47–54.

Petljak, M., Dananberg, A., Chu, K., Bergstrom, E. N., Striepen, J., von Morgen, P., Chen, Y., Shah, H., Sale, J. E., Alexandrov, L. B., Stratton, M. R., & Maciejowski, J. (2022). Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature*, *607*(7920), 799–807.

Pinkard, O., Mcfarland, S., Sweet, T., & Coller, J. (2021). Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. In *Research Square*. Research Square. https://doi.org/10.21203/rs.3.pex-1031/v1

Pinzaru, A. M., & Tavazoie, S. F. (2023). Transfer RNAs as dynamic and critical regulators of cancer progression. *Nature Reviews. Cancer*, *23*(11), 746–761.

Prehn, J. H. M., & Jirström, E. (2020). Angiogenin and tRNA fragments in Parkinson's disease and neurodegeneration. *Acta Pharmacologica Sinica*, *41*(4), 442–446.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., & Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, *160*(6), 1111–1124.

Qin, C., Chen, Z.-H., Cao, R., Shi, M.-J., & Tian, Y. (2022). A novel tiRNA-Gly-GCC-1 promotes progression of urothelial bladder carcinoma and directly targets TLR4. *Cancers*, *14*(19), 4555.

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Molecular Cell*, *59*(2), 149–161.

Raab, J. R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P. A., Haussler, D., & Kamakaka, R. T. (2012). Human tRNA genes function as chromatin insulators. *The EMBO Journal*, *31*(2), 330–350.

Rafels-Ybern, À., Attolini, C. S.-O., & Ribas de Pouplana, L. (2015). Distribution of ADAT-dependent

codons in the human transcriptome. *International Journal of Molecular Sciences*, *16*(8), 17303–17314.

Rafels-Ybern, À., Torres, A. G., Camacho, N., Herencia-Ropero, A., Roura Frigolé, H., Wulff, T. F., Raboteg, M., Bordons, A., Grau-Bove, X., Ruiz-Trillo, I., & Ribas de Pouplana, L. (2019). The expansion of inosine at the wobble position of tRNAs, and its role in the evolution of proteomes. *Molecular Biology and Evolution*, *36*(4), 650–662.

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

Raina, M., & Ibba, M. (2014). tRNAs as regulators of biological processes. *Frontiers in Genetics*, *5*, 171.

Reverendo, M., Soares, A. R., Pereira, P. M., Carreto, L., Ferreira, V., Gatti, E., Pierre, P., Moura, G. R., & Santos, M. A. S. (2014). TRNA mutations that affect decoding fidelity deregulate development and the proteostasis network in zebrafish. *RNA Biology*, *11*(9), 1199–1213.

Rhind, N., & Gilbert, D. M. (2013). DNA replication timing. *Cold Spring Harbor Perspectives in Biology*, *5*(8), a010132.

Ribas de Pouplana, L., Santos, M. A. S., Zhu, J.-H., Farabaugh, P. J., & Javid, B. (2014). Protein mistranslation: friend or foe? *Trends in Biochemical Sciences*, *39*(8), 355–362.

Richter, U., McFarland, R., Taylor, R. W., & Pickett, S. J. (2021). The molecular pathology of pathogenic mitochondrial tRNA variants. *FEBS Letters*, *595*(8), 1003–1024.

Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G. V., Carter, S. L., Saksena, G., Harris, S., Shah, R. R., Resnick, M. A., Getz, G., & Gordenin, D. A. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics*, *45*(9), 970–976.

Rodnina, M. V., & Wintermeyer, W. (2016). Protein elongation, co-translational folding and targeting. *Journal of Molecular Biology*, *428*(10), 2165–2185.

Rodríguez-Escribà, M. (2020). Role of tRNA modifications in the synthesis of the extracellular matrix. *University of Barcelona*. https://diposit.ub.edu/dspace/bitstream/2445/149477/1/MRE_PhD_THESIS.pdf

Rosselló-Tortella, M., Bueno-Costa, A., Martínez-Verbo, L., Villanueva, L., & Esteller, M. (2022). DNA methylation-associated dysregulation of transfer RNA expression in human cancer. *Molecular Cancer*, *21*(1), 48.

Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nature Reviews. Genetics*, *19*(12), 789–800.

Rubio Gomez, M. A., & Ibba, M. (2020). Aminoacyl-tRNA synthetases. *RNA (New York, N.Y.)*, *26*(8), 910–936.

Rudinger, J., Florentz, C., & Giegé, R. (1994). Histidylation by yeast HisRS of tRNA or tRNA-like structure relies on residues -1 and 73 but is dependent on the RNA context. *Nucleic Acids Research*, *22*(23), 5031–5037.

Ryvkin, P., Leung, Y. Y., Silverman, I. M., Childress, M., Valladares, O., Dragomir, I., Gregory, B. D., & Wang, L.-S. (2013). HAMR: high-throughput annotation of modified ribonucleotides. *RNA (New York, N.Y.)*, *19*(12), 1684–1692.

Sabi, R., & Tuller, T. (2014). Modelling the efficiency of codon-tRNA interactions based on codon usage bias. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and*

Genomes, *21*(5), 511–526.

Saini, N., Roberts, S. A., Sterling, J. F., Malc, E. P., Mieczkowski, P. A., & Gordenin, D. A. (2017). APOBEC3B cytidine deaminase targets the non-transcribed strand of tRNA genes in yeast. *DNA Repair*, *53*, 4–14.

Sakhtemani, R., Perera, M. L. W., Hübschmann, D., Siebert, R., Lawrence, M. S., & Bhagwat, A. S. (2022). Human activation-induced deaminase lacks strong replicative strand bias or preference for cytosines in hairpin loops. *Nucleic Acids Research*, *50*(9), 5145–5157.

Sakhtemani, R., Senevirathne, V., Stewart, J., Perera, M. L. W., Pique-Regi, R., Lawrence, M. S., & Bhagwat, A. S. (2019). Genome-wide mapping of regions preferentially targeted by the human DNA-cytosine deaminase APOBEC3A using uracil-DNA pulldown and sequencing. *The Journal of Biological Chemistry*, *294*(41), 15037–15051.

Salehi Chaleshtori, A. R., Miyake, N., Ahmadvand, M., Bashti, O., Matsumoto, N., & Noruzinia, M. (2018). A novel 8-bp duplication in ADAT3 causes mild intellectual disability. *Human Genome Variation*, *5*(1). https://doi.org/10.1038/s41439-018-0007-9

Santos, M. A. S., & Tuite, M. F. (1995). The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. *Nucleic Acids Research*, *23*(9), 1481–1486.

Santos, M., Fidalgo, A., Varanda, A. S., Oliveira, C., & Santos, M. A. S. (2019). TRNA deregulation and its consequences in cancer. *Trends in Molecular Medicine*, *25*(10), 853–865.

Santos, M., Pereira, P. M., Varanda, A. S., Carvalho, J., Azevedo, M., Mateus, D. D., Mendes, N., Oliveira, P., Trindade, F., Pinto, M. T., Bordeira-Carriço, R., Carneiro, F., Vitorino, R., Oliveira, C., & Santos, M. A. S. (2018). Codon misreading tRNAs promote tumor growth in mice. *RNA Biology*, 1–14.

Santos-Pereira, J. M., & Aguilera, A. (2015). R loops: new modulators of genome dynamics and function. *Nature Reviews. Genetics*, *16*(10), 583–597.

Schaefer, M., Pollex, T., Hanna, K., & Lyko, F. (2009). RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Research*, *37*(2), e12.

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., & Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes & Development*, *24*(15), 1590–1595.

Scheepbouwer, C., Aparicio-Puerta, E., Gomez-Martin, C., Verschueren, H., van Eijndhoven, M., Wedekind, L. E., Giannoukakos, S., Hijmering, N., Gasparotto, L., van der Galien, H. T., van Rijn, R. S., Aronica, E., Kibbelaar, R., Heine, V. M., Wesseling, P., Noske, D. P., Vandertop, W. P., de Jong, D., Pegtel, D. M., … Koppers-Lalic, D. (2023). ALL-tRNAseq enables robust tRNA profiling in tissue samples. *Genes & Development*, *37*(5–6), 243–257.

Schimmel, P. (2018). The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. *Nature Reviews. Molecular Cell Biology*, *19*(1), 45–58.

Schmidt, C. A., & Matera, A. G. (2020). tRNA introns: Presence, processing, and purpose. *Wiley Interdisciplinary Reviews. RNA*, *11*(3), e1583.

Schmidt, E., & Schimmel, P. (1993). Dominant lethality by expression of a catalytically inactive class I tRNA synthetase. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(15), 6919–6923.

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring

quality of the reference assembly. *Genome Research*, *27*(5), 849–864.

Schoenmakers, E., Carlson, B., Agostini, M., Moran, C., Rajanayagam, O., Bochukova, E., Tobe, R., Peat, R., Gevers, E., Muntoni, F., Guicheney, P., Schoenmakers, N., Farooqi, S., Lyons, G., Hatfield, D., & Chatterjee, K. (2016). Mutation in human selenocysteine transfer RNA selectively disrupts selenoprotein synthesis. *The Journal of Clinical Investigation*, *126*(3), 992–996.

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

Schuller, A. P., & Green, R. (2018). Roadblocks and resolutions in eukaryotic translation. *Nature Reviews. Molecular Cell Biology*, *19*(8), 526–541.

Schuntermann, D. B., Fischer, J. T., Bile, J., Gaier, S. A., Shelley, B. A., Awawdeh, A., Jahn, M., Hoffman, K. S., Westhof, E., Söll, D., Clarke, C. R., & Vargas-Rodriguez, O. (2023). Mistranslation of the genetic code by a new family of bacterial transfer RNAs. *The Journal of Biological Chemistry*, *299*(7), 104852.

Schuntermann, D. B., Jaskolowski, M., Reynolds, N. M., & Vargas-Rodriguez, O. (2024). The central role of transfer RNAs in mistranslation. *The Journal of Biological Chemistry*, *300*(9), 107679.

Selitsky, S. R., & Sethupathy, P. (2015). tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. *BMC Bioinformatics*, *16*(1), 354.

Seplyarskiy, V., Koch, E. M., Lee, D. J., Lichtman, J. S., Luan, H. H., & Sunyaev, S. R. (2023). A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription. *Nature Genetics*, *55*(12), 2235–2242.

Sheppard, K., Yuan, J., Hohn, M. J., Jester, B., Devine, K. M., & Söll, D. (2008). From one amino acid to another: tRNA-dependent amino acid biosynthesis. *Nucleic Acids Research*, *36*(6), 1813–1825.

Shi, J., Zhang, Y., Tan, D., Zhang, X., Yan, M., Zhang, Y., Franklin, R., Shahbazi, M., Mackinlay, K., Liu, S., Kuhle, B., James, E. R., Zhang, L., Qu, Y., Zhai, Q., Zhao, W., Zhao, L., Zhou, C., Gu, W., … Chen, Q. (2021). PANDORA-seq expands the repertoire of regulatory small RNAs by overcoming RNA modifications. *Nature Cell Biology*, *23*(4), 424–436.

Shi, M.-J., Meng, X.-Y., Fontugne, J., Chen, C.-L., Radvanyi, F., & Bernard-Pierrot, I. (2020). Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer. *Genome Medicine*, *12*(1), 85.

Shigematsu, M., Honda, S., Loher, P., Telonis, A. G., Rigoutsos, I., & Kirino, Y. (2017). YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs. *Nucleic Acids Research*, gkx005.

Silvera, D., Formenti, S. C., & Schneider, R. J. (2010). Translational control in cancer. *Nature Reviews. Cancer*, *10*(4), 254–266.

Sizer, R. E., Chahid, N., Butterfield, S. P., Donze, D., Bryant, N. J., & White, R. J. (2022). TFIIIC-based chromatin insulators through eukaryotic evolution. *Gene*, *835*(146533), 146533.

Smith, T., Monti, M., Willis, A. E., & Kalmár, L. (2024). *Benchmarking tRNA-Seq quantification approaches by realistic tRNA-Seq data simulation identifies two novel approaches with higher accuracy*. https://doi.org/10.7554/elife.96955.1

Soares, A. R., & Santos, M. A. S. (2017). Discovery and function of transfer RNA-derived fragments and their role in disease. *Wiley Interdisciplinary Reviews. RNA*, *8*(5), e1423.

Sonenberg, N., & Hinnebusch, A. G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, *136*(4), 731–745.

Song, J., Zhuang, Y., Zhu, C., Meng, H., Lu, B., Xie, B., Peng, J., Li, M., & Yi, C. (2020). Differential roles of human PUS10 in miRNA processing and tRNA pseudouridylation. *Nature Chemical Biology*, *16*(2), 160–169.

Spencer, P. S., Siller, E., Anderson, J. F., & Barral, J. M. (2012). Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *Journal of Molecular Biology*, *422*(3), 328–335.

Spivak, G. (2015). Nucleotide excision repair in humans. *DNA Repair*, *36*, 13–18.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., & Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, *26*(1), 148–153.

Srinivasan, S., Torres, A. G., & Ribas de Pouplana, L. (2021). Inosine in biology and disease. *Genes*, *12*(4), 600.

Su, Z., Kuscu, C., Malik, A., Shibata, E., & Dutta, A. (2019). Angiogenin generates specific stress-induced tRNA halves and is not involved in tRF-3-mediated gene silencing. *The Journal of Biological Chemistry*, *294*(45), 16930–16941.

Su, Z., Wilson, B., Kumar, P., & Dutta, A. (2020). Noncanonical roles of tRNAs: tRNA fragments and beyond. *Annual Review of Genetics*, *54*(1), 47–69.

Sui, Y., Qi, L., Zhang, K., Saini, N., Klimczak, L. J., Sakofsky, C. J., Gordenin, D. A., Petes, T. D., & Zheng, D.-Q. (2020). Analysis of APOBEC-induced mutations in yeast strains with low levels of replicative DNA polymerases. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9440–9450.

Supek, F., & Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, *521*(7550), 81–84.

Supek, F., & Lehner, B. (2017). Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell*, *170*(3), 534-547.e23.

Suzuki, T. (2021). The expanding world of tRNA modifications and their disease relevance. *Nature Reviews. Molecular Cell Biology*, *22*(6), 375–392.

Suzuki, T., Nagao, A., & Suzuki, T. (2011). Human mitochondrial tRNAs: biogenesis, function, structural aspects, and diseases. *Annual Review of Genetics*, *45*(1), 299–329.

Telonis, A. G., Loher, P., Honda, S., Jing, Y., Palazzo, J., Kirino, Y., & Rigoutsos, I. (2015). Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget*, *6*(28), 24797–24822.

Telonis, A. G., Loher, P., Kirino, Y., & Rigoutsos, I. (2014). Nuclear and mitochondrial tRNA-lookalikes in the human genome. *Frontiers in Genetics*, *5*, 344.

Telonis, A. G., Loher, P., Kirino, Y., & Rigoutsos, I. (2016). Consequential considerations when mapping tRNA fragments. *BMC Bioinformatics*, *17*(1), 123.

Thomas, E., Lewis, A. M., Yang, Y., Chanprasert, S., Potocki, L., & Scott, D. A. (2019). Novel missense variants in ADAT3 as a cause of syndromic intellectual disability. *Journal of Pediatric Genetics*, *8*(4), 244–251.

Thompson, M., Haeusler, R. A., Good, P. D., & Engelke, D. R. (2003). Nucleolar clustering of dispersed tRNA genes. *Science (New York, N.Y.)*, *302*(5649), 1399–1401.

Thornlow, B. P., Armstrong, J., Holmes, A. D., Howard, J. M., Corbett-Detig, R. B., & Lowe, T. M. (2020). Predicting transfer RNA gene activity from sequence and genome context. *Genome Research*, *30*(1), 85–94.

Thornlow, B. P., Hough, J., Roger, J. M., Gong, H., Lowe, T. M., & Corbett-Detig, R. B. (2018). Transfer RNA genes experience exceptionally elevated mutation rates. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(36), 8996–9001.

Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A., & Madan Babu, M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science Signaling*, *11*(546), eaat6409.

Torres, A. G. (2019). Enjoy the silence: Nearly half of human tRNA genes are silent [Review of *Enjoy the silence: Nearly half of human tRNA genes are silent*]. *Bioinformatics and Biology Insights*, *13*, 1177932219868454. SAGE Publications.

Torres, A. G., Batlle, E., & Ribas de Pouplana, L. (2014a). Role of tRNA modifications in human diseases. *Trends in Molecular Medicine*, *20*(6), 306–314.

Torres, A. G., Piñeyro, D., Filonava, L., Stracker, T. H., Batlle, E., & Ribas de Pouplana, L. (2014b). A-to-I editing on tRNAs: biochemical, biological and evolutionary implications. *FEBS Letters*, *588*(23), 4279–4286.

Torres, A. G., Reina, O., Stephan-Otto Attolini, C., & Ribas de Pouplana, L. (2019). Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(17), 8451–8456.

Turowski, T. W., & Tollervey, D. (2016). Transcription by RNA polymerase III: insights into mechanism and regulation. *Biochemical Society Transactions*, *44*(5), 1367–1375.

Van Bortle, K., & Corces, V. G. (2012). tDNA insulators and the emerging role of TFIIIC in genome organization. *Transcription*, *3*(6), 277–284.

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

Vetsigian, K., Woese, C., & Goldenfeld, N. (2006). Collective evolution and the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(28), 10696–10701.

Weil, T., Santamaría, R., Lee, W., Rung, J., Tocci, N., Abbey, D., Bezerra, A. R., Carreto, L., Moura, G. R., Bayés, M., Gut, I. G., Csikasz-Nagy, A., Cavalieri, D., Berman, J., & Santos, M. A. S. (2017). Adaptive mistranslation accelerates the evolution of fluconazole resistance and induces major genomic and gene expression alterations in Candida albicans. *MSphere*, *2*(4). https://doi.org/10.1128/mSphere.00167-17

Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., & Bartel, D. P. (2016). Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Reports*, *14*(7), 1787–1799.

Weller, C., Bartok, O., McGinnis, C. S., Palashati, H., Chang, T.-G., Malko, D., Shmueli, M. D., Nagao, A., Hayoun, D., Murayama, A., Sakaguchi, Y., Poulis, P., Khatib, A., Erlanger Avigdor, B., Gordon, S., Cohen Shvefel, S., Zemanek, M. J., Nielsen, M. M., Boura-Halfon, S., … Samuels, Y. (2025). Translation dysregulation in cancer as a source for targetable antigens. *Cancer Cell*, *43*(5), 823-840.e18.

Wernaart, D., Fumagalli, A., & Agami, R. (2024). Molecular mechanisms of non-genetic aberrant peptide production in cancer. *Oncogene*, *43*(27), 2053–2062.

Wolf, J., Gerber, A. P., & Keller, W. (2002). tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. *The EMBO Journal*, *21*(14), 3841–3851.

Wu, T. D., Reeder, J., Lawrence, M., Becker, G., & Brauer, M. J. (2016). GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. *Methods in Molecular Biology (Clifton, N.J.)*, *1418*, 283–334.

Wu, Y., Yang, X., Jiang, G., Zhang, H., Ge, L., Chen, F., Li, J., Liu, H., & Wang, H. (2021). 5'-tRF-GlyGCC: a tRNA-derived small RNA as a novel biomarker for colorectal cancer diagnosis. *Genome Medicine*, *13*(1), 20.

Wulff, T. F., Hahnke, K., Lécrivain, A.-L., Schmidt, K., Ahmed-Begrich, R., Finstermeier, K., & Charpentier, E. (2024). Dynamics of diversified A-to-I editing in Streptococcus pyogenes is governed by changes in mRNA stability. *Nucleic Acids Research*, *52*(18), 11234–11253.

Xue, C., Tian, J., Chen, Y., & Liu, Z. (2024). Structural insights into human ELAC2 as a tRNA 3' processing enzyme. *Nucleic Acids Research*, *52*(21), 13434–13446.

Yang, C., Pataskar, A., Feng, X., Montenegro Navarro, J., Paniagua, I., Jacobs, J. J. L., Zaal, E. A., Berkers, C. R., Bleijerveld, O. B., & Agami, R. (2024). Arginine deprivation enriches lung cancer proteomes with cysteine by inducing arginine-to-cysteine substitutants. *Molecular Cell*, *84*(10), 1904-1916.e7.

Yang, P., Beltramo, D. M., Ribas de Pouplana, L., Soria, N. W., & Torres, A. G. (2019). Loss of the tRNALysCUU encoding gene, Chr-11 tRNA-Lys-CUU, is not associated with Type 2 diabetes mellitus. *Biomarkers in Medicine*, *13*(4), 259–266.

Yeung, R., & Smith, D. J. (2020). Determinants of Replication-Fork Pausing at tRNA Genes in Saccharomyces cerevisiae. *Genetics*, *214*(4), 825–838.

Yoshihisa, T. (2014). Handling tRNA introns, archaeal way and eukaryotic way. *Frontiers in Genetics*, *5*, 213.

Yuan, L., Han, Y., Zhao, J., Zhang, Y., & Sun, Y. (2023). Recognition and cleavage mechanism of intron-containing pre-tRNA by human TSEN endonuclease complex. *Nature Communications*, *14*(1), 6071.

Zahra, S., Singh, A., & Kumar, S. (2023). tncRNA Toolkit: A pipeline for convenient identification of RNA (tRNA)-derived non-coding RNAs. *MethodsX*, *10*(101991), 101991.

Zhang, W., Foo, M., Eren, A. M., & Pan, T. (2022). tRNA modification dynamics from individual organisms to metaepitranscriptomics of microbiomes. *Molecular Cell*, *82*(5), 891–906.

Zhang, Z., Ye, Y., Gong, J., Ruan, H., Liu, C.-J., Xiang, Y., Cai, C., Guo, A.-Y., Ling, J., Diao, L., Weinstein, J. N., & Han, L. (2018). Global analysis of tRNA and translation factor expression reveals a dynamic landscape of translational regulation in human cancers. *Communications Biology*, *1*(1), 234.

Zheng, G., Qin, Y., Clark, W. C., Dai, Q., Yi, C., He, C., Lambowitz, A. M., & Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. *Nature Methods*, *12*(9), 835–837.

Zhou, K., Diebel, K. W., Holy, J., Skildum, A., Odean, E., Hicks, D. A., Schotl, B., Abrahante, J. E., Spillman, M. A., & Bemis, L. T. (2017). A tRNA fragment, tRF5-Glu, regulates BCAR3 expression and proliferation in ovarian cancer cells. *Oncotarget*, *8*(56), 95377–95391.