



UNIVERSITAT<sup>DE</sup>  
BARCELONA

## Evolution of I34 modifications in tRNAs and their role in proteome composition

Àlbert Rafels Ybern



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement 3.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento 3.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution 3.0. Spain License.**

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

# **Evolution of I34 modifications in tRNAs and their role in proteome composition**

Àlbert Rafels Ybern  
2018





UNIVERSITAT DE  
BARCELONA



INSTITUT  
DE RECERCA  
BIOMÈDICA

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

PROGRAMA DE DOCTORAT EN BIOMEDICINA

INSTITUT DE RECERCA BIOMÈDICA

# **Evolution of I34 modifications in tRNAs and their role in proteome composition**

Memòria presentada per Àlbert Rafels Ybern per optar al títol de doctor per la Universitat  
de Barcelona

Director

Tutor

Doctorand

Lluís Ribas de Pouplana

Gemma Marfany

Àlbert Rafels Ybern



# 1. Acknowledgements

I would like to sincerely thank the members from my laboratory, from IRB, UB and other institutions, and my family, who helped me during these years. They made possible that this thesis has been carried out. I would like to personally express my gratefulness to Lluís Ribas (PI), Gemma Marfany (UB), David Piñeiro (Postdoc), Federica Lombardi (Postdoc), Adrian Torres (Postdoc), Liudmilla Filonava (Postdoc), Daria Picchoni (PhD student), Albert Antolín (PhD student), Marta Rodríguez (PhD student), Helena Roura (PhD student), Alba Pons (PhD student), Enric Ros (PhD student), Noelia Camacho (Lab technician), Ana Tort (Lab technician), Glenn Hauquier (Lab technician), Andrea Herencia, (visiting student), Marina Raboteg (visiting student), Maria Carretero (visiting student), Maria Noguera (visiting student), Thomas Wulff (visiting student), Camille Stephan-Otto (biostatistics unit), Oscar Reina (biostatistics unit), Adria Caballe, (biostatistics unit), Toni Berenguer (biostatistics unit), José Ignacio (functional genomics unit), David Fernandez (functional genomics unit), Iñaki Ruiz (Multicellgenome lab), Xavier Grau (Multicellgenome lab), Lorena Valverde (IRB), Salvador Guardiola (IRB) and Jesus Herraiz (IRB).



# Table of contents

1. Acknowledgements.....	v
2. Abstract.....	9
3. Abbreviations.....	11
4. Introduction.....	13
4.1. General Overview.....	13
4.1.1. The genome.....	14
• Protein coding genes.....	14
• RNA-specifying genes.....	16
• Untranscribed genes.....	17
• Pseudogenes.....	17
4.1.2. The process of translation.....	17
• aaRSs.....	18
• The ribosome.....	19
4.1.3. The genetic code.....	22
4.1.4. Proteins.....	23
4.2. Transfer RNA (tRNA).....	26
4.2.1. tRNA structure.....	25
4.2.2. tRNA gene copy number and codon usage bias.....	28
4.2.3. tRNA wobbling.....	31
4.2.4. tRNA modifications.....	32
4.3. Molecular evolution and phylogeny.....	37
4.3.1. A brief history.....	34
4.3.2. Mutations.....	38
4.3.3. Molecular phylogenetics.....	40
• Tree terminology.....	40
• Types of trees.....	41
• Tree shape.....	42
• Character change reconstruction.....	43
• Tree-making methods.....	43
• Bootstrapping.....	44
5. Objectives.....	47
6. Publications.....	49
6.1. PhD advisor report.....	49
6.2. Publication 1.....	51
6.3. Publication 2.....	71
6.4. Publication 3.....	89
6.5. Publication 4.....	123
7. Summary of results.....	131
8. Discussion.....	137
9. Conclusions.....	143
10. References.....	145





## 2. Abstract

Inosine is a guanosine analogue that when is found at the wobble position of the tRNAs (I34) expands its codon recognition capability. Inosine can wobble pair with cytosine, adenosine and uridine. Because inosine is not genomically encoded, essential enzymes are responsible for the hydrolytic deamination of adenosine to inosine, specifically at the wobble position of the tRNAs. In Bacteria, the modification is mostly found in tRNA<sup>Arg</sup>, catalysed by the homodimeric tRNA adenosine deaminase A (TadA), with a conserved active site coordinated with an atom of Zn<sup>+2</sup>. In Eukarya, the modification is present in up to eight different tRNAs, catalysed by the heterodimeric enzyme ADAT (ADAT2-ADAT3), which originally evolved from TadA by duplication and divergence. ADAT2 is considered the catalytic subunit because it conserves the active site, whereas ADAT3, which lacks one of the essential catalytic residues, is thought to play a structural role. This substrate expansion, significantly influenced the evolution of eukaryotic genomes in terms of tRNA gene abundance and codon usage. However, the selection pressures driving this process remain unclear.

In this thesis, we characterize the human transcriptome and proteome in terms of frequency and distribution of ADAT-related codons. Human codon usage indicates that I34 modified tRNAs are preferred for the translation of highly repetitive coding sequences, suggesting that I34 is an important modification for the synthesis of proteins of highly skewed amino acid composition. Persuaded by these results we extend the analysis to a series of eukaryotic and bacterial organisms, spanning the whole tree of life. We find that the preference for codons that are recognized by I34-modified tRNAs, in genes with highly biased codon composition, is universal among eukaryotes, and we report that, unexpectedly, the bacterial phylum of Firmicutes shows a similar preference. We experimentally demonstrate that the Firmicute *Oenococcus oeni* presents a functional expansion of I34 modification to other tRNAs other than tRNA<sup>Arg</sup>, and that this process likely starts with the emergence of unmodified A34-containing tRNAs. Our findings also indicate that several ancestral bacterial groups lack both TadA and A34-tRNAs, suggesting that these species never developed the machinery to generate I34-modified tRNAs. On the other hand limited sets of bacterial species have either lost the system secondarily, or expanded it to additional tRNA substrates. In Eukaryotes, we show that a large variability in the use of I34 can be found in protists, while the modification becomes fixed in Metazoa, Fungi and Plant kingdoms.



### 3. Abbreviations

Commonly used abbreviations and definitions			
$\Delta_{1\%}$	Number of codons needed to increase the ADAT enrichment by 1% based on the slope of the linear model.	G34 tRNAs	tRNA with G at the wobble position with cognate amino acids TAPSLIVR
a(t) or A codons	(i.e. ADAT codons). This notation is used in Publication 1.	GDP	Guanosine diphosphate.
a/c	Fraction of ADAT codons. This notation is used in Publication 1.	G-ended codons	Subset of ADAT codons that are not ADAT-sensitive codons, so they are not recognized by I34 tRNAs. Most of them ends in G.
A:T	Canonical base pairing between A and T in DNA.	GTP	Guanosine triphosphate.
A:U	Canonical base pairing between A and U in RNA.	HisRS	HistidinyI-tRNA synthetase.
A34 tRNA diversity	Number of different A34 tRNAs that code for ADAT amino acids. Range from 0 to 8.	indel	Insertion or deletion.
A34 tRNAs	tRNA with A at the position 34 (or wobble position).	IQR	Interquartile range
A34-tRNA ratio	Ratio of A34 tRNA genes compared to C34-, U34-, and G34-tRNA genes.	LIVR	L, I, V and R amino acids.
aaRS	Aminoacyl tRNA-synthetase.	LUCA	Last universal common ancestor
aa-tRNA	aminoacyl-tRNA.	mDNA	Messenger DNA. Portion of DNA that corresponds to mRNA.
ADAT	Adenosine Deaminase Acting on tRNAs	ML	Maximum likelihood.
ADAT amino acids	The eight amino acids T,A,P,S and L, I, V, R. (Publication 3, Figure 1a)	ML method	Maximum likelihood tree inference method.
ADAT codons	The 37 codons that code for ADAT amino acids (TAPSLIVR). See (Publication 3, Figure 1a).	mRNA	Messenger RNA.
ADAT enrichment	Fraction of ADAT-sensitive codons / ADAT codons. See (Publication 3, Figure 1a).	MSA	Multiple sequence alignment
ADAT stretch	Region highly enriched in ADAT codons (and ADAT amino acids).	MUC5b	Mucin 5b
ADAT2-ADAT3	Subunits 2 and 3 from heterodimeric ADAT.	NCBI	National Center of Biotechnology Information ( <a href="https://www.ncbi.nlm.nih.gov/">https://www.ncbi.nlm.nih.gov/</a> )
ADAT-dependent codons	(i.e. ADAT-sensitive codons). This notation is used in Publication 1.	NR protein database	Non-redundant protein database. Compilation of all protein sequences from all the known species to date provided by NCBI.
Adatness	ADAT (or TadA) e-value divided by its corresponding CDA e-value	num-tRNA	tRNA gene copy number.
Adatness	The e-value for ADAT2 (or TadA in bacteria) divided by the corresponding e-value for the CDA superfamily.	OTU	Operational taxonomic unit
ADAT-sensitive codons	Subset of 24 ADAT codons susceptible to be recognized by modified I34 tRNAs. See (Publication 3, Figure 1a).	PAP	PolyA polymerase
BLAST	Basic Local Alignment Search Tool ( <a href="https://blast.ncbi.nlm.nih.gov/Blast.cgi">https://blast.ncbi.nlm.nih.gov/Blast.cgi</a> )	phyloT	Software used to infer standard phylogenies based on NCBI taxonomy.
c(t)	Total number codons in a sequence <i>t</i> . This notation is used in Publication 1.	PolyA	Polyadenylation. Added at 3' of mRNA.
C:G	Canonical base pairing between C and G in DNA and RNA.	Pyl	Pyrolysine
CCDS	Consensus Coding DNA Sequence. <a href="https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi">https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi</a>	Q1 and Q4	First and fourth quartile in a boxplot analysis.
CDA	Family of cytidine deaminases proteins.	$R^2$	R-squared from the corresponding linear model.
CDS	Coding DNA Sequence.	RNA	Ribonucleic acid.
CF	Cystic fibrosis	RNA-pol	RNA-polymerases.
COPD	Chronic obstructive pulmonary disease	rRNA	Ribosomal RNA.
CpG	CG island. Sequence of repetitive CG or GC motifs.	S	Svedberg units of sedimentation. Used to measure the ribosomes size.
CU	codon usage	SAM	S-adenosyl-L-methionine.
d(t) or D codons	Number of ADAT-sensitive codons in a sequence <i>t</i> . This notation is used in Publication 1.	SAR	Monophyletic supergroup that includes Stramenopiles (heterokonts), Alveolates, and Rhizaria. Also called Harosa.
d/a	(i.e. ADAT enrichment). This notation is used in Publication 1.	SDC3	Syndecan 3
DNA	Deoxyribonucleic acid.	TadA	tRNA Adenosine Deaminase A
EF-P	Elongation factor P	TAPS	T,A,P and S amino acids.
eIF5A	Initiation factor 5A	TAPSLIVR	T,A,P,S and L, I, V, R amino acids
Ensembl e-val	Genome browser ( <a href="https://www.ensembl.org/index.html">https://www.ensembl.org/index.html</a> ) e-value	Trm5	tRNA methyltransferase 5 in Eukarya.
FDR	False discovery rate	TrmD	tRNA methyltransferase D in Bacteria.
G:U	Canonical base pairing between G and U in RNA.	tRNA	Transfer RNA.
		tRNA scan-SE	Software used to predict tRNA genes and the num-tRNA for each organism.

## Abbreviations

Nucleotides	
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	A or G (purines)
Y	C or T (pyrimidines)
–	Gap
.	Gap

Amino acids					
A	Ala	alanine		fMet	formyl-methionine
C	Cys	cysteine	N	Asn	asparagine
D	Asp	aspartic acid	P	Pro	proline
E	Glu	glutamic acid	Q	Gln	glutamine
F	Phe	phenylalanine	R	Arg	arginine
G	Gly	glycine	S	Ser	serine
H	His	histidine	T	Thr	threonine
I	Ile	isoleucine	U	Sec	selenocysteine
K	Lys	lysine	V	Val	valine
L	Leu	leucine	W	Trp	tryptophan
M	Met	methionine	Y	Tyr	tyrosine

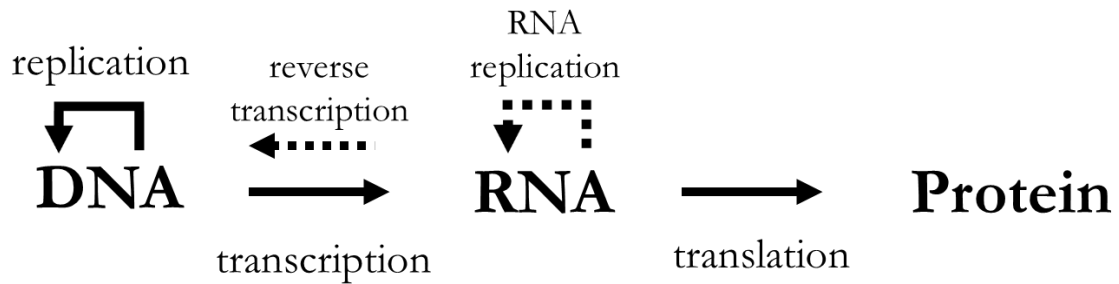
Modified nucleotides			
ac <sup>4</sup> C	N-4 acetylcytidine.	mcm <sup>5</sup> U	C-5 methoxycarbonylmethyl uridine
ac <sup>6</sup> A	N-6 acetyladenosine	mcm <sup>5</sup> Um	5-methoxycarbonylmethyl-2'-O-methyluridine
acp <sup>3</sup> U	3-(3-amino-3-carboxypropyl)uridine	mcm <sup>5</sup> U	uridine 5-oxyacetic acid methyl ester
agm <sup>2</sup> C	agmatidine	mG	Methylated G. Added at 5' of mRNA.
chm <sup>5</sup> U	C-5 carboxyhydroxymethyl uridine	m <sup>n</sup> X	methylation at position n of nucleotide base X
Cm	2'-O-methylcytidine	mnm <sup>5</sup> s <sup>2</sup> U	5-methylaminomethyl-2-thiouridine
cmm <sup>5</sup> U	C-5 carboxymethylaminomethyluridine	mnm <sup>5</sup> U	C-5 methylaminomethyluridine
cmo <sup>5</sup> U	uridine 5-oxyacetic acid	mn,nX	dimethylation at position n of nucleotide base X
D	dihydrouridine	mo <sup>5</sup> U	C-5 methoxyuridine
f <sup>5</sup> C	5-formyl cytosine	ms <sup>2</sup> t <sup>6</sup> A	2-methylthio-N6-threonyl carbamoyladenosine
G+	archaeosine	nm <sup>5</sup> U	C-5 aminomethyluridine
g <sup>6</sup> A	N-6 glycinylcarbamoyladenosine	nmn <sup>5</sup> U	C-5 carbamoylmethyl uridine
hn <sup>6</sup> A	N-6 hydroxynorvalylcarbamoyladenosine	Q	queosine (and related 7-deaza species oQ, preQ1, preQ0, gluQ, galQ, and manQ)
ho <sup>5</sup> U	C-5 hydroxyuridine	s <sup>2</sup> C	2-thiocytidine
I	inosine (from deamination of adenosine)	s4U	4-thiouridine
i <sup>6</sup> A	N-6 isopentenyladenosine	s <sup>n</sup> X	replacement of oxygen with sulfur at position n of nucleotide X
imG	wyosine (and related imG-14, mimG, and imG2 species)	t <sup>6</sup> A	N6-threonylcarbamoyladenosine
io <sup>6</sup> A	N-6 (cis-hydroxyisopentenyl)adenosine	t <sup>6</sup> A	N-6 threonyladenosine
k <sup>2</sup> C	lysidine	tm <sup>5</sup> s <sup>2</sup> U	5-taurinomethyl-2-thiouridine
m <sup>2</sup> <sub>2</sub> G	N2,N2-dimethylguanosine	W or Ψ	Pseudouridine.
m <sup>2</sup> G	N2-methylguanosine	Xm	2'-O methylation of nucleotide X
m <sup>5</sup> C	5-methylcytidine	xm <sup>5</sup> s <sup>2</sup> U	5-methyl-2-thiouridine derivatives with any substitution at carbon 5 of the uracil
m <sup>5</sup> U	5-methyluridine or ribothymidine	xmo <sup>5</sup> U	5-methoxyuridine derivatives with any substitution at carbon 5 of the uracil
m <sup>7</sup> G	7-methylguanosine	Xr(p)	2'-O-ribosyl phosphate derivative of nucleotide X
mcm <sup>5</sup> s <sup>2</sup> U	5-methoxycarbonylmethyl-2-thiouridine	yW	wybutosine (and related OHyW, OHyW*, o2yW, and yW-86 species)

## 4. Introduction

### 4.1 General overview

All living organisms are conformed by one or more functional units called cells. We consider a living organism when it has the capacity to reproduce and transfer their hereditary information to the offspring during the process of cell division. Each cell is mainly composed by three different molecules with different purposes: (1) deoxyribonucleic acid (DNA), which stores the information to live and reproduce, (2) ribonucleic acid (RNA), which selects the information stored in the DNA and allows its decoding into (3) proteins, which constitute the main product of the cell with a wide range of functions. In addition, the synthesis of DNA, RNA and proteins are sequential and unidirectional, what is known as Central Dogma of Molecular Biology (Crick 1970). First, the transcription from DNA to RNA occurs, and second, the translation from RNA to proteins (**Figure 4.1**). However, the information from proteins cannot be transferred back to RNA or DNA (only in some viruses, there are exceptions such as the reverse transcription from RNA to DNA or the RNA replication) (**Figure 4.1**, dashed lines).

The building blocks of DNA are conformed by four different nucleotides, each of which is composed by a deoxyribose sugar, a phosphate group and one of four nitrogen-containing bases: cytosine (C), guanine (G), adenine (A) or thymine (T) (**Figure 4.2**, 1-3). The deoxyribose sugar is bound to the base by a  $\beta$ -glycosidic bound (**Figure 4.2**, 4). The nucleotides are joined to one another by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in a single stranded chain of DNA (**Figure 4.2**, 5). The DNA sequences are notated based on the sugar unbound ends from 5' to 3' by convention because they are polarized (e.g. GGA $\neq$ AGG). The nitrogen-containing bases of two complementary DNA strands are bound together by hydrogen bonds (H-bond) and twisted around each other to form a right-handed double helix (**Figure 4.2**, 6). According to the Canonical base pairing rules: A pairs with T (A:T) by two H-bonds and C pairs with G (C:G) with three H-bonds, what makes C:G a stronger base pair than A:T (Watson and Crick 1953). Chromosomal DNA is packaged inside the nuclei with the help of histones. These are positively-charged proteins that strongly adhere to negatively-charged DNA and form complexes called nucleosomes. Each nucleosome is composed of DNA wound 1.65 times



**Figure 4.1:** Central Dogma of Molecular Biology where the information decoding direction is shown. Solid arrows correspond to Crick's first proposal in 1970. Dashed arrows correspond to later discoveries in some viruses.

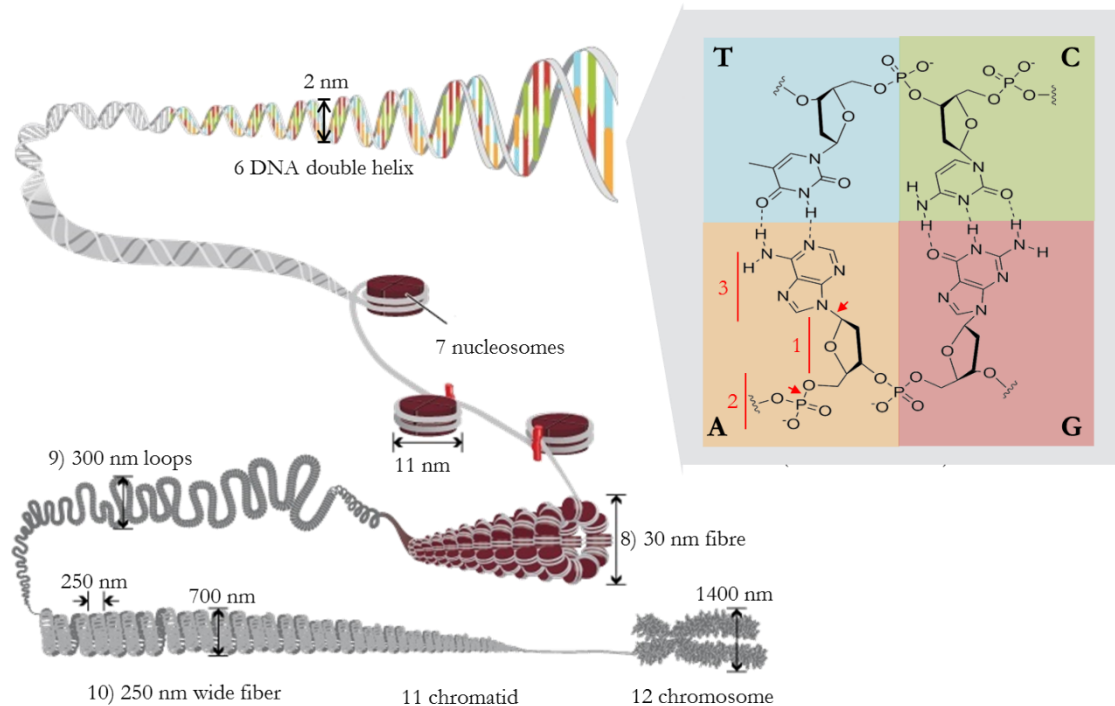
around eight histone proteins (**Figure 4.2**, 7). Nucleosomes fold up to form a 30-nanometre chromatin fibre, which forms loops averaging 300 nanometres in length (**Figure 4.2**, 8-9).

The 300 nm fibres are compressed and folded to produce a 250 nm-wide fibre, which is tightly coiled into the chromatid (700 nm width) of a chromosome (1400 nm width) (**Figure 4.2**, 10-12). The chemical structure of RNA is similar to the DNA but with some singularities: the sugar of RNA nucleotides is a ribose instead of deoxyribose, what confers to RNA molecules less stability and more predisposition to be degraded. The four nitrogen-containing bases of RNA are A, G, C and Uracil (U) instead of T. The canonical base pairing rules for RNA include A:U, C:G and U:G (note that T:G is not stable in DNA).

#### 4.1.1 The genome

The entire complement of DNA molecules in a cell is called the genome. Only a small portion of the genome is known to contain functional genes. The standard definition of a gene consists of a DNA segment that codes for a protein or for a functional RNA molecule. Genes can be classified in four different types (Cavalier-Smith 1985; Watson 1987; Lewin 1994; Li 1997):

**Protein-coding genes:** also known as structural or productive genes, are those genes that end up producing a functional protein. The transcription product of a gene is called messenger RNA (mRNA) or transcript, and the region of the gene where it comes from is defined as messenger DNA (mDNA) (**Figure 4.3**). The enzymes in charge of transcription are the RNA-polymerases (RNAPol), concretely RNAPol II in Eukarya. The first nucleotide that is transcribed is the transcription initiation site and is designated with the number 1

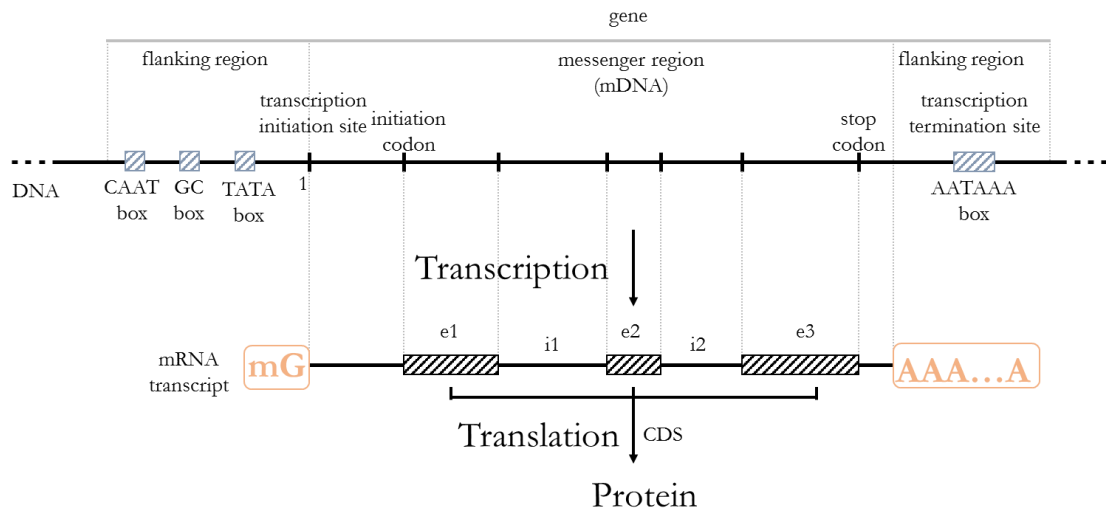


**Figure 4.2:** chromosomal DNA packaging. The deoxyribose sugar (1) is bound to a phosphate group (2) and to a base (3) by an ester bond and a  $\beta$ -glycosidic bond respectively (4,5 red arrows). Two complementary DNA strands are bound together by hydrogen bonds (H-bond) and twisted around each other to form a right-handed double helix (6). DNA is packaged inside the nuclei with the help of histones (7). Nucleosomes fold up to form a 30-nanometre chromatin fibre (8), which forms loops averaging 300 nanometres in length (9). The 300 nm fibres are compressed and folded to produce a 250 nm-wide fibre (10), which is tightly coiled into the chromatid (11) of a chromosome (12). Modified from Pierce, Benjamin. *Genetics: A Conceptual Approach*, 2nd ed.

(**Figure 4.3**). From here, upstream nucleotides are numbered negatively and not transcribed, and downstream nucleotides are numbered positively and transcribed until the transcription termination site. The region of mDNA that is finally translated into a protein is called coding DNA sequence (CDS) (**Figure 4.3**, black dashed boxes). The flanking regions neighbouring the mDNA are not transcribed (untranscribed regions), but they are considered part of the gene because they enclose regions that regulate the process of transcription. The 5' flanking region contains signals that promote the transcription defined as promoters and the 3' flanking region contains signals for the termination of the transcription process. Some regulatory elements can be found at considerable distances from the mDNA, which makes difficult to delineate with precision the points at which a gene begins and ends. The initial binding of RNAPol II is controlled by the promoters CAAT and CG boxes in Eukarya (**Figure 4.3**, blue dashed boxes), whereas in Bacteria is controlled by the -10 and -35 sites, so called because they are respectively placed at positions -10 and -35. Additionally, in



## Introduction



**Figure 4.3:** Typical gene distribution in Eukarya. In DNA, flanking regions are untranscribed but contain conserved sequences that controls the process of transcription (blue dashed boxes). Once the gene is transcribed, the mRNA maturation includes 5'-capping, 3'-polyadenylation (orange squares) and the introns (i1, i2) removal. The Coding Sequence (CDS) is finally translated into a protein.

Eukarya, most of genes present the TATA box promoter that controls the choice of transcription initiation site. In Bacteria, some mRNAs can encode for more than one protein what is referred to as operon or polycistronic mRNA. Operons are very rare in Eukarya. In Eukarya, mRNA suffers a series of modifications before it can be translated, what is known as mRNA maturation. mRNA maturation is composed by three steps that occur sequentially. First, a methylated G (mG) is linked at the 5'-end of the transcript while it is still being transcribed. This process is called 5' capping and prevents the mRNA from being degraded at 3' end (**Figure 4.3**, orange squares). Next, when the transcription is close to finish, the AATAAA box promotes the cleavage of the transcript at the transcription termination site which releases the pre-mRNA molecule (**Figure 4.3**, blue boxes). Immediately about 200 As are added at the 3' of the pre-mRNA by the enzyme polyA polymerase (PAP) (**Figure 4.3**, red squares). This process is known as polyadenylation (polyA) of mRNA and is important for the nuclear export, translation, and stability of the mRNA. Finally, the introns, which are interleaved with exons, are spliced out. This is a sophisticated process known as splicing where in most cases a complex of proteins and RNA molecules called spliceosome take part. The additional steps involved in eukaryotic mRNA maturation create a molecule with a much longer half-life than the bacterial mRNA. Indeed, there is not mRNA maturation in Bacteria, which means the absence of 5' capping, polyA or introns in bacterial mRNAs.

**RNA specifying genes:** these genes are only transcribed but not translated because the RNA transcript is the functional molecule. Ribosomal RNAs (rRNAs) and transfer RNAs

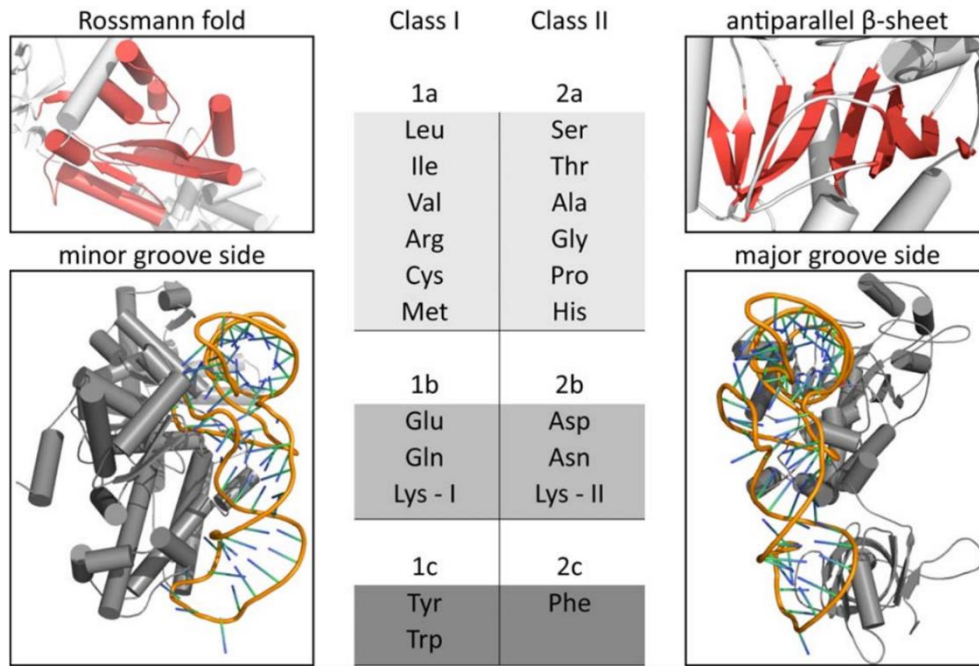
(tRNAs) are the most abundant RNA genes, both are essential for the process of translation and, in a general way, there is a correlation between the RNA gene quantification and the genome size of the organisms (Prokopowich, Gregory et al. 2003). RNA genes also undergo their own process of maturation before being functional. tRNA transcripts are cleaved at 5' and 3' ends (tRNA trimming), followed up by the addition of the invariant CCA sequence at 3' end of the tRNA (CCA addition) (Li and Deutscher 1996; Tomita and Weiner 2001; Betat and Morl 2015; Wende, Bonin et al. 2015). Moreover, a few eukaryotic tRNA genes possess short introns that have to be spliced out. Apart from maturation, RNA molecules, especially those from tRNAs, undergo extensive nucleotide modifications to become fully functional. We will extend this content in section 4.2.4. RNA molecules also have several alternative functions such as: (1) small nuclear RNAs (snRNAs), functioning in a variety of nuclear processes, including splicing of pre-mRNA. (2) Small nucleolar RNAs (snoRNAs), used to process and chemically modify rRNAs. (3) MicroRNAs (miRNAs), which regulate gene expression typically by blocking translation of selective mRNAs. (4) Small interfering RNAs (siRNAs), which turn off gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures.

**Untranscribed genes:** these genes consist of regions of the genome that regulate different processes such as: replicator genes, which define the initiation and termination of DNA replication; telomeric sequences, which are short and repetitive sequences at the end of eukaryotic chromosome acting as a protective cap against the exonucleolytic degradation; recombination genes, where the recombination enzymes bind during the meiosis process; segregator genes, which are sites for the attachment of the spindle machinery such as the centromeres; and constructional sites, which determine the chromosomal structure as supercoils or fragile sites.

**Pseudogenes:** these are genes that resemble a functional gene but with defects that make them non-functional. Pseudogenes are ubiquitous along the genome, are notated with prefix  $\Psi$ - and the name of the gene that resemble. Most of pseudogenes are not transcribed. If transcribed, most of them are not translated and only a little portion is transcribed and translated.

#### 4.1.2 The process of translation

During the process of translation, a molecule of mRNA is decoded codon by codon by means of tRNAs into a sequence of amino acids that will constitute a protein. A codon is a



**Figure 4.4:** aaRSs are classified depending on their catalytic site folding type. Class I has a Rossmann fold and recognize the tRNAs by their minor groove, whereas class II has an antiparallel  $\beta$ -sheet and recognize the tRNAs by their major groove. Classes I and II are structurally subdivided into subclasses a, b and c. The chemical nature or steric shape of the amino acid side chain is roughly paralleled across each subclass.

sequence of 3 consecutive nucleotides. tRNAs act as universal adaptors of the genetic code, able to act as bridges that convert the genomic information into functional proteins. The system is decoded in two steps. In the first step, tRNAs are charged with their cognate amino acids to form aminoacyl-tRNAs (aa-tRNA). This process is referred to as aminoacylation of tRNA, and is carried out by the aminoacyl tRNA-synthetases (aaRSs). In the second step, adjacent non-overlapping triplets of nucleotides of the mRNA (codons) are recognized by complementary sequences of 3 nucleotides (anticodon) of the aa-tRNA. The recognition between codon and anticodon proceeds through H-bonds in the ribosomal complex and has been extensively reviewed (Kapp and Lorsch 2004; Dale and Uhlenbeck 2005; Jackson, Hellen et al. 2010).

**aaRSs** are present in all the living organisms, these enzymes bind to the 3' end of their cognate tRNAs and catalyse the formation of the covalent bond between the tRNA and its cognate amino acid. Usually, organisms present 20 different aaRSs, one for each amino acid, and the catalytic core is preserved in all of them (Nagel and Doolittle 1991). aaRSs are classified according to their three-dimensional structure in classes I, that have a conserved

catalytic domain based on a Rossman fold motif, and class II, that contain a conserved antiparallel  $\beta$ -sheet motif (**Figure 4.4**). Each class contains 10-11 aaRSs enzymes and is subdivided in three subclasses a, b and c, closely related in sequence (**Figure 4.4**) (Eriani, Delarue et al. 1990; Cusack 1997). Class I enzymes bind to the minor groove side of the acceptor stem of the tRNA whereas class II enzymes bind to the major groove side of the stem (**Figure 4.4**) (Fraser and Rich 1975; Rould, Perona et al. 1989; Cramer, Englisch et al. 1991; Ruff, Krishnaswamy et al. 1991; Cavarelli, Eriani et al. 1994; Arnez and Moras 1997). The amino acids from each subclass harbour certain relationships, for instance, amino acids from subclass Ic and IIc are exclusively aromatic, or amino acids Glu and Gln from subclass Ib and Asp and Asn from subclass IIb have a very similar structure. When a tRNA is missacylated with a wrong amino acid, aaRSs, along with the tRNA, present editing functions acting as a ribonucleoprotein that translocate the missactivated amino acid to the editing domain where is cleared by hydrolysis (Baldwin and Berg 1966; Schimmel and Schmidt 1995). Editing is essential for cell viability and editing domains are universally conserved even in the most deeply rooted organisms near the base of the tree of life. As a result of binding to opposite grooves of the acceptor stem, molecular modelling studies have shown that 2 aaRSs of the same subclass but opposite classes (Ia-IIa, Ib-IIb and Ic-IIc) can simultaneously fit on the same tRNA molecule and that other pairings, apart from these, are sterically forbidden (Ribas de Pouplana and Schimmel 2001). An example of these complexes is found in the methanogenic archaea *Methanosarcina barkeri* where the two LysRSs, from classes I and II bind simultaneously to the non-canonical tRNA pyrrolysine (Pyl) (Ibba, Morgan et al. 1997; Polycarpo, Ambrogelly et al. 2003). These relationships suggest the possibility that in an early environment, aaRSs played the role of covering and protecting the acceptor stem, similar to chaperons, from the assaults of high temperatures, nucleases and harsh chemicals.

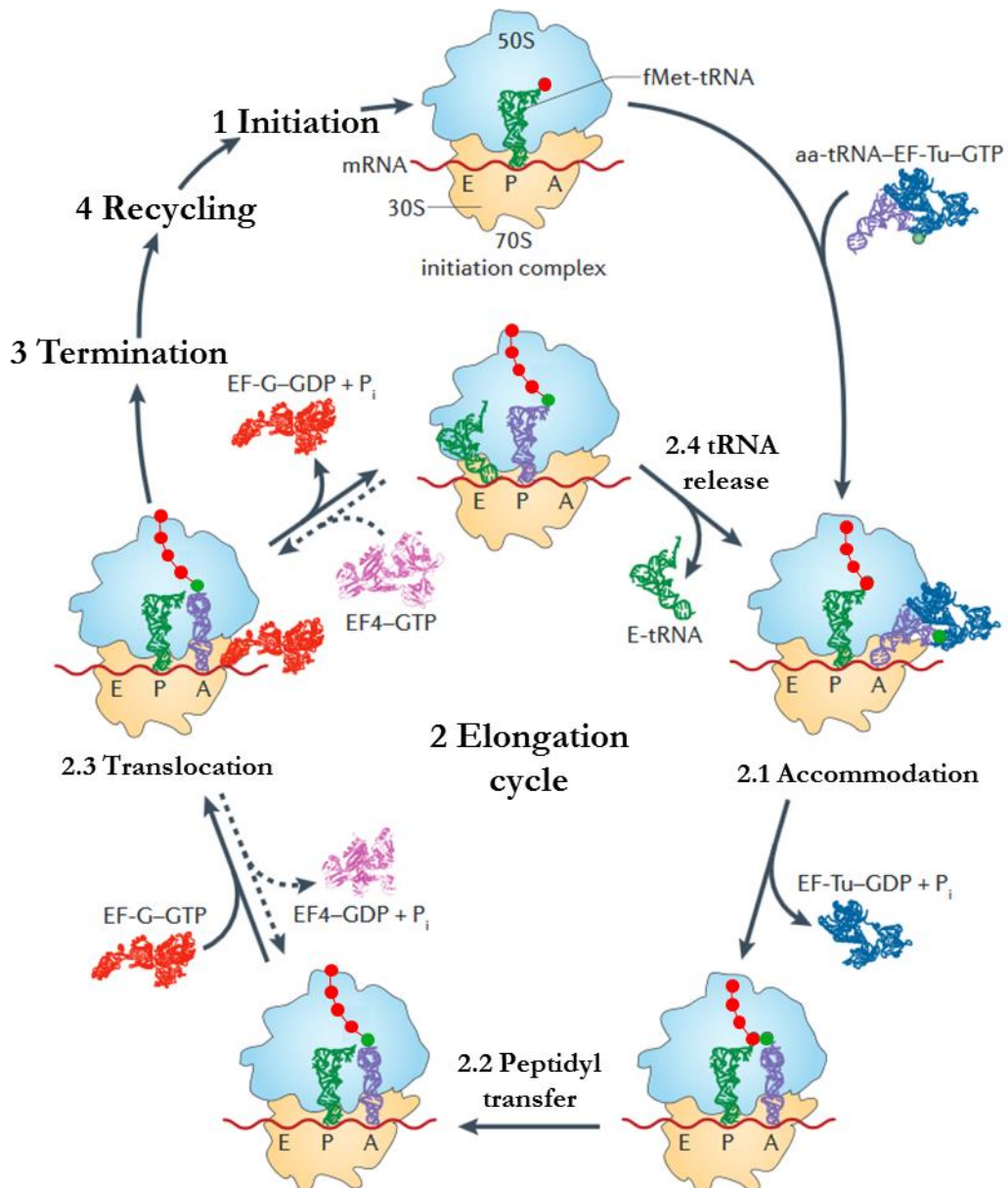
**The ribosome** is a complex molecular machine, found in all living organisms, that serves as the place for mRNA decoding and the appropriate protein synthesis. The ribosome structure, consists of two ribonucleoprotein subunits composed about 70% of rRNAs and 30% of ribosomal proteins. In fact, the ribosome is a ribozyme because rRNA is responsible for the catalytic activity of the peptide bond formation (Moore, Atchison et al. 1975; Barta, Steiner et al. 1984; Moazed and Noller 1989). In addition, crystallographic work has shown that there are no ribosomal proteins close to the reaction site for polypeptide synthesis suggesting that the protein components of ribosomes do not directly participate in the catalysis (Nissen, Hansen et al. 2000). One of the subunits (large subunit) is about twice the size of the other (small subunit). Thus, prokaryotic ribosomes sediment at 70 Svedberg units (S) (50S for the

## Introduction

large subunit and 30S for small one) whereas the cytoplasmic ribosomes from Eukarya sediment at 80S (60S for the large subunit and 40S for small one). Mitochondria and chloroplast organelles also have their own ribosomes that sediment at 70S and resemble to the bacterial ones reflecting the endosymbiotic theory that suggests that these organelles are descendants of bacteria (Benne and Sloof 1987; Alberts 1998).

The codon-anticodon recognition takes place in the small subunit, whereas the acceptor stem of the tRNA binds to the large subunit of the ribosome, which promotes the peptide bond formation through the peptide-transfer centre. The ribosome has 3 tRNA-mRNA binding sites (A, P and E), which correspond to 3 consecutive codons of the mRNA (**Figure 4.5**). The A-site ensures that the correct aa-tRNA is selected from the pool of charged tRNAs based on the codon of the mRNA, the P-site is where the peptidyl-tRNA is bound, and the E-site is where the deacylated tRNA passes and is finally discharged from the ribosome.

There are four stages that occur consecutively during the ribosomal protein translation (**Figure 4.5**): (1) In initiation, a special tRNA for initiation charged with methionine (Met-tRNA<sub>i</sub>) forms a complex with initiation factors (which differs from prokaryotes to eukaryotes) and uses guanosine triphosphate (GTP) as a source of energy. This complex binds to the P-site, which starts to scan down the mRNA from 5' to 3' until an AUG initiation codon is found. (2) In elongation cycle, the appropriate aa-tRNA is selected for the A-site mRNA codon which forms a complex with elongation factors and GTP. The complex binds to the A site of the small subunit, hydrolyses GTP, releases GDP and elongation factors, and redirects the acceptor stem of the aa-tRNA to bind the A site of the large subunit. This process is referred to as accommodation and at this stage the deacylated tRNA present in the E site leaves the ribosome (**Figure 4.5**, 2.1 and 2.4). Thereafter, the nascent peptide that is esterified to the 3'-terminal ribose of the tRNA in the P-site is transferred to the amino group of the aa-tRNA bound to the A-site, which elongates the nascent peptide by one amino acid. The tRNA in the P-site is left deacylated and the tRNA in the A-site is charged with the elongated peptide (**Figure 4.5**, 2.2). Translocation is the final step of the cycle, where the ribosome advances along its mRNA by one codon in 3' direction, leaving the deacylated tRNA to the E site and the peptidyl-tRNA to the P-site and returning the ribosome to the initial state but having advanced one codon and added one amino acid to the nascent peptide (**Figure 4.5**, 2.3). The protein elongation cycle is repeated until a stop codon is encountered. (3) Termination occurs when one of the three stop codons (UAA, UAG and UGA) arrives to the A-site. These codons are not recognized by any tRNA but by releasing factors that



**Figure 4.5:** The four ribosomal stages (Initiation, Elongation, Termination and Recycling) accomplished during the process of translation. (1) the initiator tRNA (formylmethionine tRNA (fMet-tRNA)) is bound to the P-site of the ribosome and interacts with the start codon AUG in mRNA. (2) Elongation factor Tu (EF-Tu), and guanosine 3-phosphate (GTP) bind the aa-tRNA forming a ternary complex (aa-tRNA-EF-Tu-GTP) that binds the A-site of the ribosome once a successful decoding between the tRNA and the codon placed at the A-site of the mRNA. (2.1) GTP is hydrolyzed to guanosine 2-phosphate (GDP). An inorganic phosphate (P<sub>i</sub>) is released along with EF-Tu-GDP. aa-tRNA swings into the A-site. (2.2) The nascent peptide (red chain) is transferred to the from the peptidyl-tRNA in the P-site to the aa-tRNA in the A-site, extending the peptide chain by one amino acid. (2.3) EF-G-GTP promotes the translocation of the complex by one codon. The tRNAs in the P and A sites are now in E and P sites respectively. EF4-GTP avoid the translocations process mobilizing stalled ribosomes (dashed arrows). (2.4) The uncharged tRNA placed in the E-site is released from the complex. (3) The elongation process is cyclically repeated until a stop codon enters the A-site. Releasing factors promotes the hydrolysis of the ester bound in the peptidyl-tRNA. (4) Ribosome recycling factors promotes the ribosome disassembly, after which the initiation complex is formed again. Modified from (Yamamoto, Qin et al. 2014).

## Introduction

trigger the hydrolysis of the ester bond in the peptidyl-tRNA, releasing the newly synthesized protein from the ribosome. (4) Ribosome recycling is the last step and is essential for cell viability. Despite the release of the polypeptide in termination step, ribosomes remain bound to the mRNA and tRNA. Ribosome recycling factors and elongation factors promote the ribosomes to be ultimately released from the mRNA and split into subunits that become free to bind new mRNA and start the process again (Hirokawa, Demeshkina et al. 2006).

### 4.1.3 The genetic code

The genetic code is a common language for all organisms to translate mRNA molecules, which are sequentially read by tRNAs, to amino acid sequences of proteins. A codon is a sequence of 3 consecutive nucleotides; since there are 4 different nucleotides, the possible codon combinations are  $4^3 = 64$ . 61 out of 64 codons are classified as sense codons because they code for amino acids. The remaining 3 (usually UAA, UAG and UGA) are called stop codons or non-sense codons because they indicate the end of translation. Most organisms use the same genetic code, referred to as standard or universal genetic code (**Table 4.1**). Nevertheless, some organisms or organelles, introduced small variations to the standard genetic code. For instance, some bacterial organisms from genus *Mycoplasma* use the codon UGA to code the amino acid Trp, being the stop codons reduced to UAA and UAG. Another example is in mitochondrial translation in vertebrates with four variations compared with the standard code are found (Barrell, Bankier et al. 1979). The number of different amino acids and the number of different codons is 20 and 64 respectively, and is conserved across all domains of life with few exceptions. These differences between the number of amino acids and codons cause a degeneration of the genetic code in the sense that one amino acid can be decoded by more than one codon, in fact, 18 out of 20 amino acids are degenerated (**Table 4.1**). The family of codons that code for the same amino acid are called synonymous codons and they often share the first and the second nucleotides. Conversely, the genetic code behaves unambiguously because each codon only codes for one amino acid although there are few exceptions to this rule such as selenocysteine that it is encoded by a special using the UGA codon, which is normally a stop codon (Baranov, Gesteland et al. 2002; Donovan and Copeland 2010).

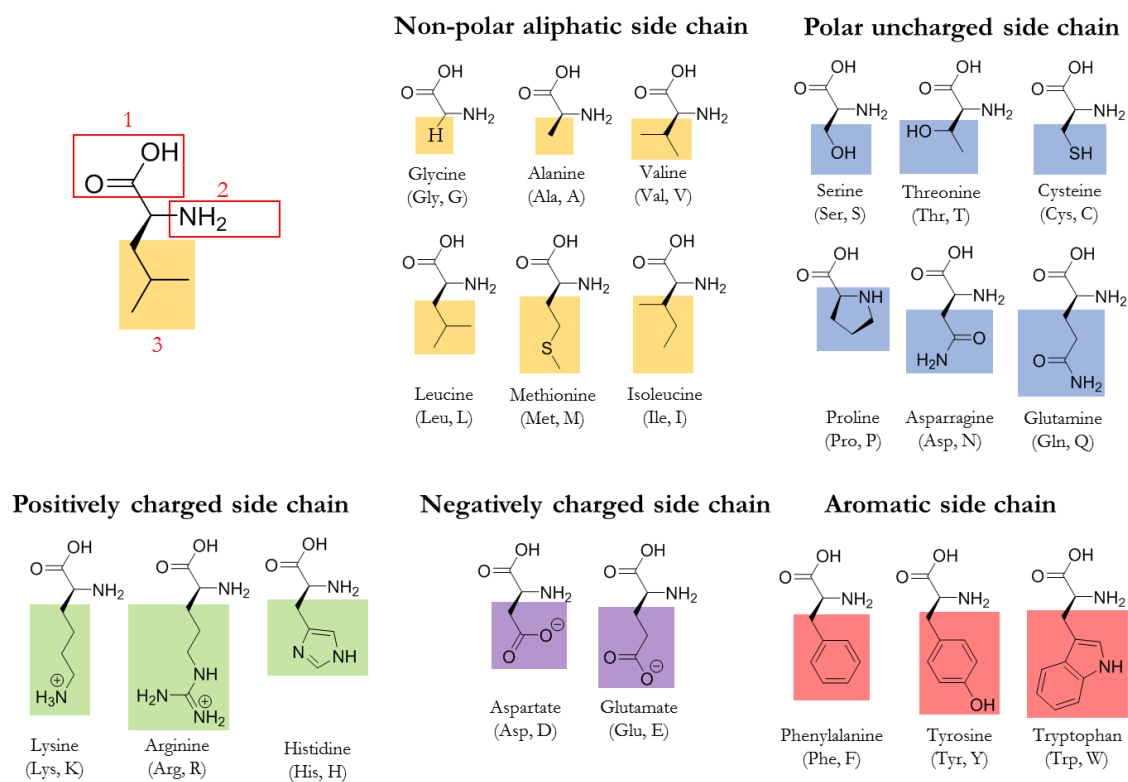
**Table 4.1:** Universal genetic code.

		Second Codon Letter								
		U		C		A		G		
First Codon Letter	U	F	Phe	S	Ser	Y	Tyr	C	Cys	U
		F	Phe	S	Ser	Y	Tyr	C	Cys	C
		L	Leu	S	Ser	Stop		Stop		A
		L	Leu	S	Ser	Stop		W	Trp	G
	C	L	Leu	P	Pro	H	His	R	Arg	U
		L	Leu	P	Pro	H	His	R	Arg	C
		L	Leu	P	Pro	Q	Gln	R	Arg	A
		L	Leu	P	Pro	Q	Gln	R	Arg	G
	A	I	Ile	T	Thr	N	Asn	S	Ser	U
		I	Ile	T	Thr	N	Asn	S	Ser	C
		I	Ile	T	Thr	K	Lys	R	Arg	A
		M	Met	T	Thr	K	Lys	R	Arg	G
	G	V	Val	A	Ala	D	Asp	G	Gly	U
		V	Val	A	Ala	D	Asp	G	Gly	C
		V	Val	A	Ala	E	Glu	G	Gly	A
		V	Val	A	Ala	E	Glu	G	Gly	G

#### 4.1.4 Proteins

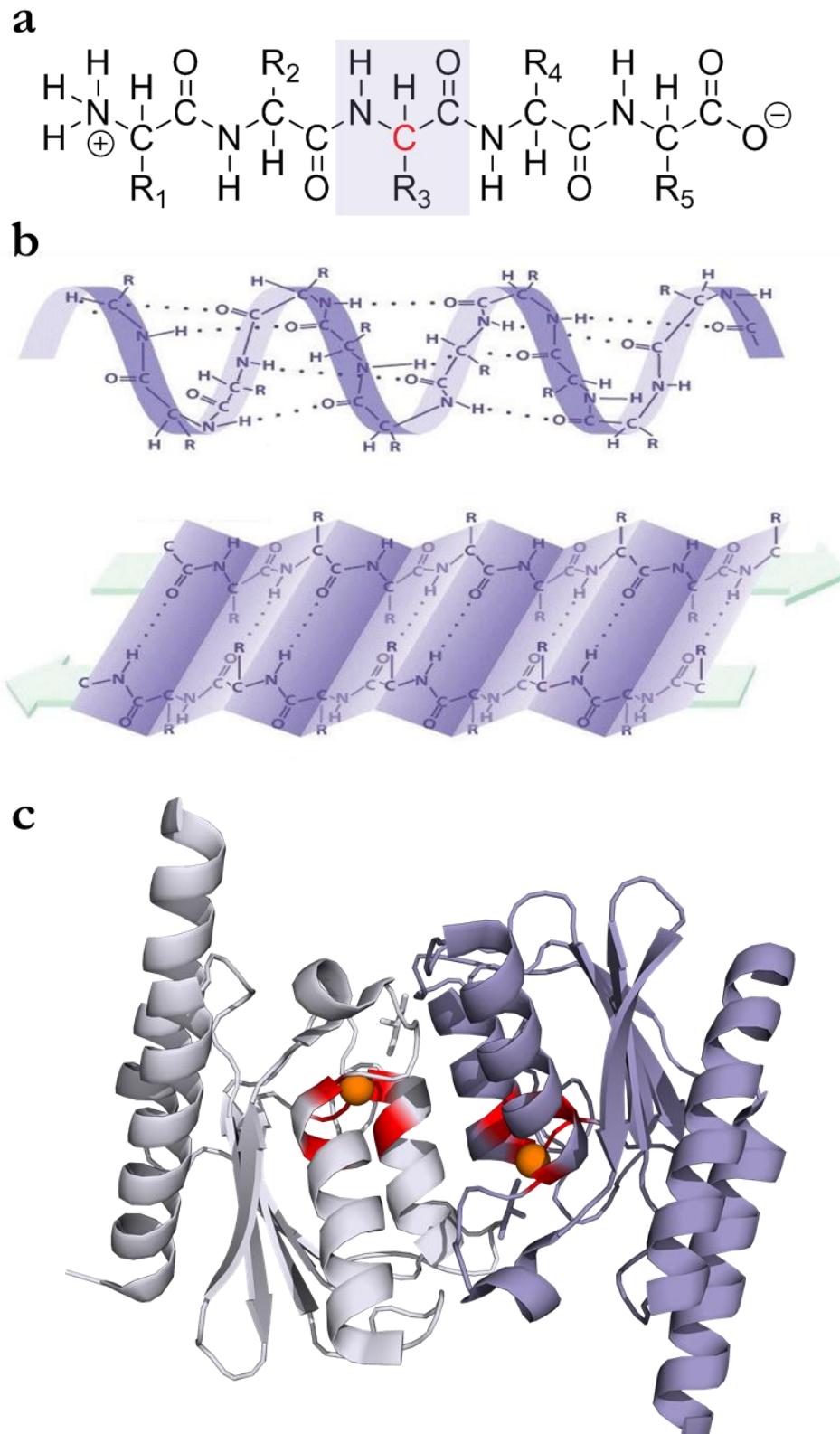
There are 20 canonical amino acids that conform the elementary structure units of proteins. The structure of an amino acid is defined by a central chiral Carbon (except for Gly) that is bound to: a hydrogen, an amino group (N-terminal), a carboxylic group (C-terminal) and a variable R group or side chain that defines each amino acid (**Figure 4.6**, 1-3). Amino acids are classified according to whether their R groups are polar, hydrophobic, positively or negatively charged, aromatic or aliphatic (**Figure 4.6**, different coloured R groups). A chain of amino acids (also called residues) covalently linked by peptide bonds, is called a polypeptide. Each polypeptide is polarized (e.g. Ala-Gly  $\neq$  Gly-Ala) and the convention for its notation is from N-terminal to C-terminal residues. The length of a polypeptide is measured by the number of residues and varies from a few amino acids to some thousands. A macromolecule composed of one or more polypeptides, also called subunits, is defined as a protein. Four levels of structural organization are usually mentioned when dealing with proteins. The primary structure is the linear arrangement of amino acids along the polypeptides, also known as protein sequence (**Figure 4.7a**). The secondary structure represents periodical structures of residues that are consecutive in the protein sequence. The most common secondary structures are  $\alpha$ -helix and  $\beta$ -sheet.  $\alpha$ -helix is a roadlike entity





**Figure 4.6:** Chemical structures and classification of the twenty-different natural amino acids. Amino group (1), carboxyl group (2) and the R group (3) is depicted for amino acid leucine. Amino acids are classified according to different properties of their R groups which are differently coloured. Three-letter and one-letter abbreviations are indicated for each amino acid.

stabilized by H-bonds each four-interval residue.  $\beta$ -sheet is a set of residue strands connected laterally in parallel or antiparallel form and stabilized by H-bonds from adjacent strands (**Figure 4.7b**). Protein regions that do not have any secondary structure remain as random coils. The tertiary structure represents the spatial arrangement of secondary structure elements. The forces that drive these arrangements are from different nature, such as H-bonds, hydrophobic interactions, salt bridges between positively and negatively charged residues, as well as covalent disulphide bonds between pairs of cysteines. Proteins embedded within an aqueous environment tend to adopt a globular shape where hydrophilic residues tend to be outside while the hydrophobic ones tend to be buried inside. The quaternary structure refers to the spatial arrangement of each subunit of the protein and the nature of its contacts. Frequently, proteins require of prosthetic groups or cofactors, which are small molecules of non-protein nature that allow a correct protein folding and use to be involved in the formation of the active site. Examples of cofactors are vitamins or inorganic metal ions such as Fe, Mn, Mg, Co, Cu or Zn, define the three-dimensional structure of a protein, crystallographic techniques are often used, however they are not usually easy to obtain,



**Figure 4.7:** The 4 levels of protein structure. (a) Primary structure. The third amino acids of the sequence is squared in blue and its chiral carbon is red. (b) Secondary structures  $\alpha$ -helix (above) and  $\beta$ -sheet (below). Modified from (Gupta 2017). (c) Tertiary and quaternary structure. Crystal structure of *Staphylococcus aureus* TadaA (Losey, Ruthenburg et al. 2006). The residues that confirms the catalytic centre are depicted in red. The  $\text{Zn}^{+2}$  cofactors are represented as orange balls.

## Introduction

especially in regions that remain as a random coil. The structure of a protein is also predicted by computational methods but with less accuracy.

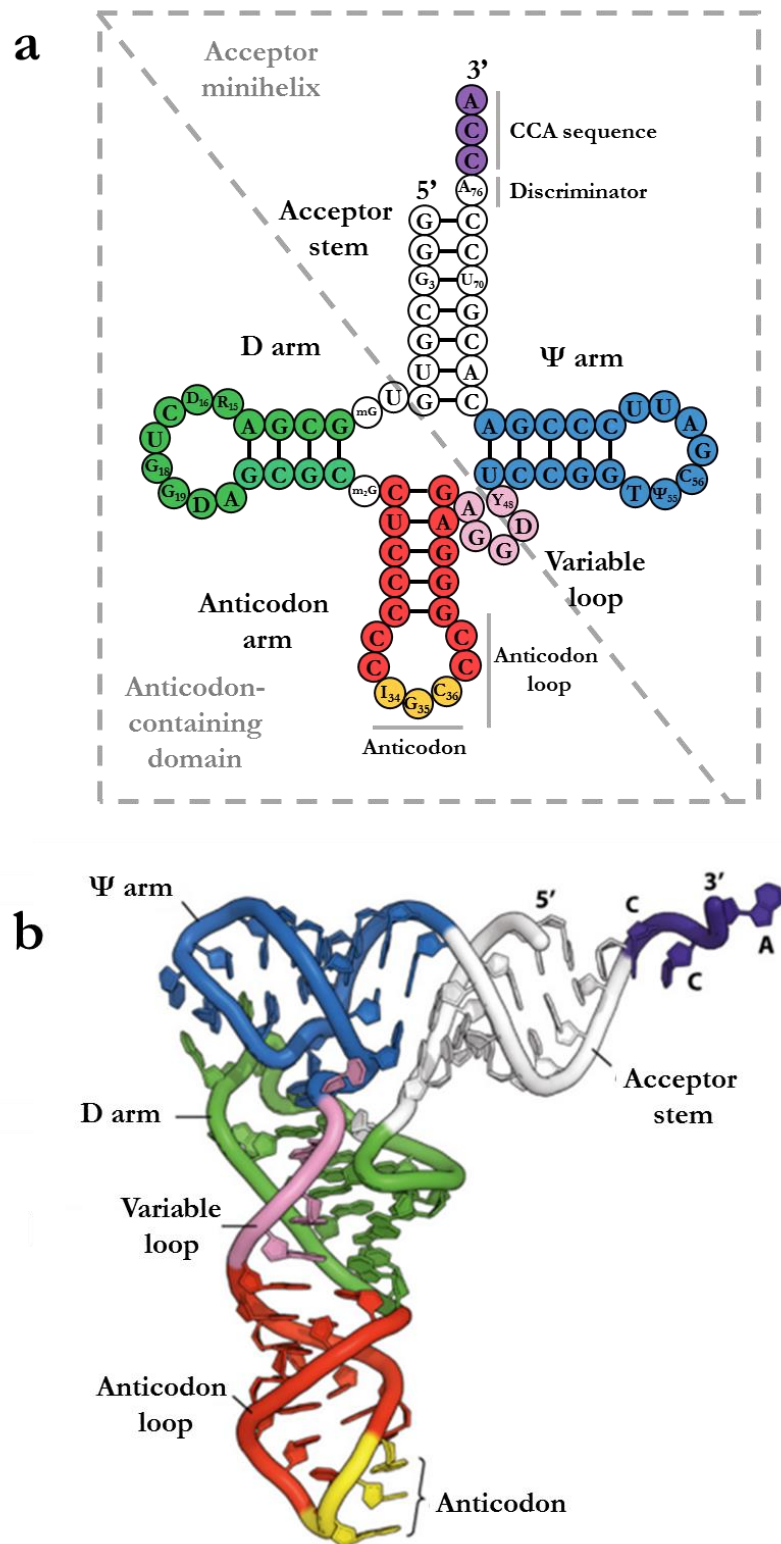
To give an example, TadA from *Staphylococcus aureus* is a bacterial protein formed by 2 identical polypeptides (homodimer) with a length of 156 residues each. The secondary structure of TadA is formed by 5  $\alpha$ -helices and 1  $\beta$ -sheet of 5 parallel strands and diverse random coils. The secondary, tertiary and quaternary structures of TadA are depicted in **Figure 4.7c**. The active site of TadA is enclosed in a pocket formed by two helices of each subunit (highlighted in red) and needs the presence of the prosthetic group  $Zn^{+2}$  (orange balls) (Losey, Ruthenburg et al. 2006).

Most of the cellular processes are carried out by proteins, being the chief actors to perform the information encoded in genes. The proteome makes up half of the dry weight of the cell. The most common function for proteins is to act as enzymes, which accelerate the chemical reactions by decreasing the activation energy barrier, and are essential for cell surviving and homeostasis. Among the multiple functions of the proteins, cell signalling is an important one where the proteins transmit a signal to a receptor to promote a process, for example, antibodies are proteins from the immune system that bind to foreign substances and promote their elimination. Proteins are also crucial for structural functions such as actin and tubulin that make up the cytoskeleton, or myosin and kinesin that allow cellular motility.

## 4.2 Transfer RNA (tRNA)

### 4.2.1 tRNA structure

tRNAs are relatively short molecules (75-95 bp) with a strongly conserved ‘cloverleaf’ secondary structure (Holley 1965; Sprinzl, Horn et al. 1998) that encloses three different arms (D,  $\Psi$  and anticodon arms) and one acceptor stem (**Figure 4.8a**). The D arm owes its name because of a dihydrouridine (D) at position 16, and the  $\Psi$  arm owes its name to a conserved pseudouridine ( $\Psi$ ) at position 55 (**Figure 4.8a**) (Bjork, Durand et al. 1999). Each arm has a loop of non-bounded nucleotides at the external part, and some tRNAs have another extra loop between the T-arm and the anticodon arm called variable loop (**Figure 4.8a**). The three outermost nucleotides in the anticodon loop use to be at positions 34 to 36 and represent the anticodon. The acceptor stem consists of 7 to 9 nucleotides of double strand, 1 unpaired nucleotide (discriminator) and finally the conserved sequence CCA that is post-transcriptionally added at the 3' end (**Figure 4.8a**). As shown in section 4.1.2, the cognate amino acid is correctly attached to the A76 of the CCA by aaRSs enzymes (Rich and RajBhandary 1976; Sprinzl, Horn et al. 1998). The tRNA notation is given by its anticodon



**Figure 4.8:** (a) tRNA secondary cloverleaf structure. The tRNA anticodon (yellow) pairs with its cognate mRNA codon (brown). The anticodon arm (red), D-arm (green), T-arm (i.e. Ψ arm) (blue), variable loop (pink), anticodon stem (white) and the CCA (violet) are depicted with the most common tRNA modifications (see legend) (for more information see section 4.2.4). (b) tRNA 3D L-shaped structure coloured with the same parameters in (a).

## Introduction

and the amino acid that is charged. For example, a tRNA charged with amino acid Ala and anticodon GCA is noted as tRNA<sup>Ala</sup><sub>GCA</sub> or tRNA<sub>Ala</sub> (GCA). tRNAs have a conserved L-shaped three-dimensional structure by reason of conserved interactions between G<sub>18</sub>G<sub>19</sub> with Ψ<sub>55</sub>C<sub>56</sub> and Y<sub>48</sub> with R<sub>15</sub> (Levitt interaction) (Levitt 1969; Kim, Sussman et al. 1974) where the anticodon and the acceptor stem lie at opposite ends of the molecule (**Figure 4.8b**).

A molecule of tRNA evolved from 2 primitive domains: the acceptor minihelix, that corresponds to the acceptor stem and the Ψ-arm, and the anticodon-containing domain, that corresponds to the D-arm and anticodon arm (**Figure 4.8a**) (Kim, Suddath et al. 1974; Robertus, Ladner et al. 1974; Francklyn and Schimmel 1989). While the acceptor minihelix binds to 50S ribosomal subunit, the anticodon minihelix binds to the 30S ribosomal subunit (Atkins, Gesteland et al. 2011). G<sub>3</sub>U<sub>70</sub> is a conserved base pair found in the acceptor stem that is shown to be specific for Ala aminoacylation in tRNA<sup>Ala</sup><sub>GCN</sub> in *Escherichia coli*. Interestingly, if this base pairing is transferred to another tRNA different from Ala, the aminoacylation turns for Ala (Hou and Schimmel 1988; McClain and Foss 1988; Hou and Schimmel 1989). These results suggest that the acceptor minihelix appeared first with a primordial secondary genetic code enclosed into its nucleotides that determines its aminoacylation, and that the anticodon-containing domain appeared before. In fact, it has been shown that the anticodon sequences (genetic code) are related with the determinants for minihelix aminoacylation (secondary genetic code) (Rodin, Rodin et al. 1996; Ribas de Pouplana and Schimmel 2001).

### 4.2.2 tRNA gene copy number and codon usage bias

tRNA genes are classified according to their anticodon sequence. There are 61 possible tRNA genes, as many as the number of sense codons, but all the species use less than 61, and more than 22 different tRNAs to translate their proteome (Yokobori, Kitamura et al. 2013). Therefore, some tRNA molecules recognize more than one codon (see section 4.2.3), and some amino acids are charged to more than one different tRNA. The family of tRNAs that code for the same amino acid are termed isoacceptors. Most organisms present among 40-50 different tRNA genes. For instance, there are 40 different tRNAs in *Escherichia coli*, 48 in *Caenorhabditis elegans* or 51 in *Homo sapiens* (Lowe and Eddy 1997). Some archaeal or bacterial species with a reduced genome, and organelles such as mitochondria, also reduced its number of tRNA-encoding genes (Dufresne, Garczarek et al. 2005). For example, the bacterial *Mycoplasma mobile* has only 27 different tRNA genes. Human mitochondrial genome,

**Table 4.2:** num-tRNA (1<sup>st</sup> column) and CU (2<sup>nd</sup> column) comparison for Homo sapiens hg19. Numbers are coloured low-medium-high in red-yellow-green. The amino acid for each tRNA is depicted in the 3<sup>rd</sup> column. Ø means stop codon. Codons are uppercased and anticodons are lowercased.

		2nd Codon - 35th tRNA													
		Ua			Cg			Au			Gc				
1 <sup>st</sup> Codon - 36 <sup>th</sup> tRNA	Ua	num-tRNA	CU		num-tRNA	CU		num-tRNA	CU		num-tRNA	CU		3 <sup>rd</sup> Codon - 34 <sup>th</sup> tRNA	
		0	1.8	F	11	1.5	S	1	1.2	Y	0	1.1	C	Ua	
		12	2	F	0	1.8	S	14	1.5	Y	30	1.3	C	Cg	
		7	0.8	L	5	1.2	S	2	0.1	Ø	3	1.3	Ø	Au	
	7	1.3	L	4	0.4	S	1	0.1	Ø	9	0.2	W	Gc		
	Cg	12	1.3	L	10	1.8	P	0	1.1	H	7	0.5	R	Ua	
		0	2	L	0	2	P	11	1.5	H	0	1	R	Cg	
		3	0.7	L	7	1.7	P	11	1.2	Q	6	0.6	R	Au	
		10	4	L	4	0.7	P	20	3.4	Q	4	1.1	R	Gc	
	Au	14	1.6	I	10	1.3	T	2	1.7	N	0	1.2	S	Ua	
		3	2.1	I	0	1.9	T	32	1.9	N	8	2	S	Cg	
		5	0.8	I	6	1.5	T	16	2.4	K	6	1.2	R	Au	
		20	2.2	M	6	0.6	T	17	3.2	K	5	1.2	R	Gc	
	Gc	11	1.1	V	29	1.8	A	0	2.2	D	0	1.1	G	Ua	
		0	1.5	V	0	2.8	A	19	2.5	D	15	2.2	G	Cg	
		5	0.7	V	9	1.6	A	13	2.9	E	9	1.7	G	Au	
16		2.8	V	5	0.7	A	13	4	E	7	1.7	G	Gc		

contains a minimal but complete set of 22 tRNA genes (Suzuki, Nagao et al. 2011). However, the number of tRNA-encoding genes in numerous mitochondrial genomes is insufficient for proper protein synthesis to occur and therefore, nuclear-encoded tRNAs are imported into the mitochondria (Akashi, Takenaka et al. 1998; Schneider and Marechal-Drouard 2000; Rinehart, Krett et al. 2005; Kamenski, Kolesnikova et al. 2007; Mager-Heckel, Entelis et al. 2007). The protozoans *Trypanosoma brucei* and *Leishmania tarentolae* represent the most extreme situation because their mitochondrial genomes are completely devoid of tRNA genes (Salinas, Duchene et al. 2008). The discovery of these species and organelles with a reduced tRNA content, opened the debate about how many tRNAs are needed, and which are the mechanisms to allow the translation of all the proteins. This field has been extensively reviewed in (Suzuki, Nagao et al. 2011; Yokobori, Kitamura et al. 2013).

tRNA genes are often present in multiple copies. The number of copies for each tRNA gene with the same anticodon in a genome is defined as tRNA gene copy number (num-tRNA) and varies widely from one copy to even hundreds of copies. The num-tRNA distribution is not uniform across all domains of life. For instance, A34 tRNA genes are the most abundant

## Introduction

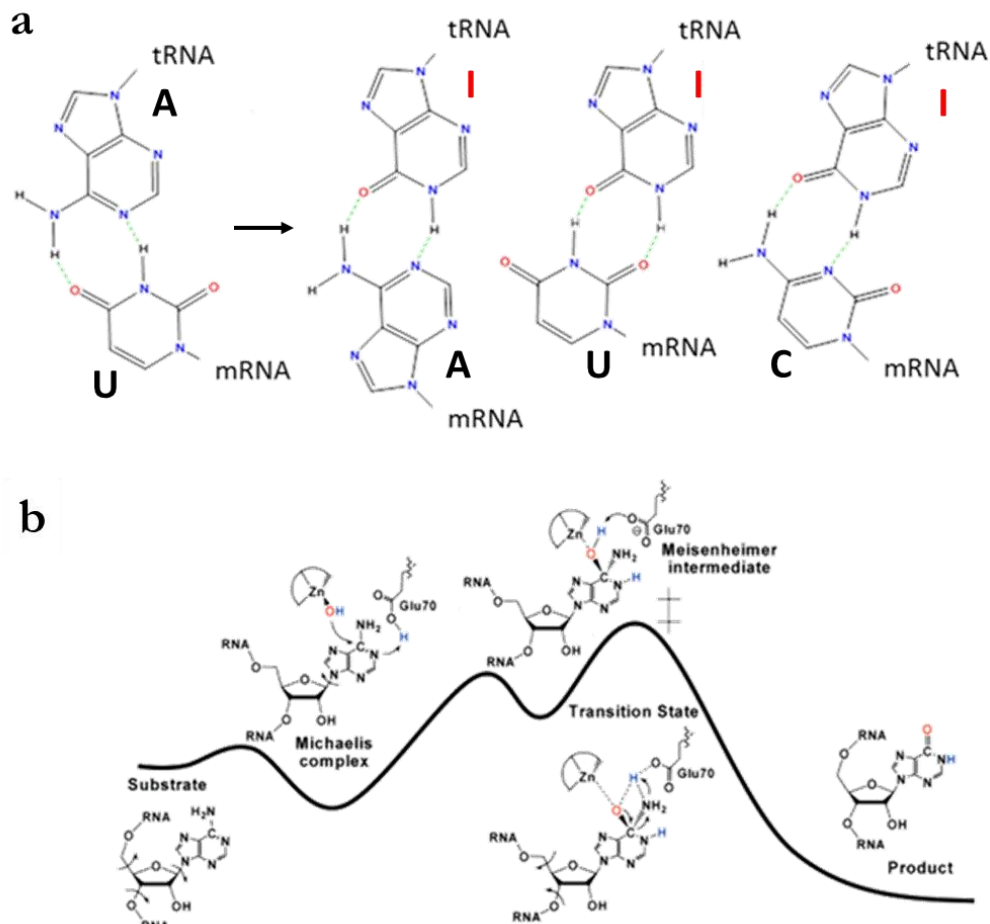
in Eukarya for amino acids T, A, P, S, L, I, V and R (TAPSLIVR), but trend to be absent in Bacteria except for tRNA<sup>Arg</sup> (ACG). On the contrary, G34 tRNA genes are absent in Eukarya but are present in Bacteria for TAPSLIVR (**Table 4.2**).

There is a correlation between the num-tRNA and the intracellular tRNA levels in low-complexity organisms such as bacteria and yeast (Ikemura 1981; Ikemura 1985; Sorensen and Pedersen 1991; Kanaya, Yamada et al. 1999; Dittmar, Mobley et al. 2004; Tuller, Carmi et al. 2010). Thus, the num-tRNA has often been used as a proxy for tRNA abundance in the cell. However, in complex organisms such as human, the situation seems to be more elaborated where the tRNA expression levels are tissue-specific and regulated by epigenetic signature and chromatin state (Dittmar, Goodenbour et al. 2006; Ernst, Kheradpour et al. 2011; Mahlab, Tuller et al. 2012).

On the other hand, no apparent correspondence has been found between codon usage of human CDSs and the abundance of iso-accepting tRNAs (**Table 4.2**) (Kanaya, Yamada et al. 2001; Urrutia and Hurst 2001; Duret 2002; Urrutia and Hurst 2003; Comeron 2004). This lack of correlation suggest that additional levels of regulation must optimize the process of translation.

Since the seventies, several studies have shown the specialization of the content of certain tRNA species for specialized cells which is consistent with the synthesis of proteins of unusual amino acid composition that usually, are generated through multiple repetitions of simple sequences. This studies include: (1) the tRNA content in rabbit reticulocytes that are specialized for the synthesis of haemoglobin, which constitutes more that 80% of total protein expression in these cells (Smith and McNamara 1971; Smith, Meltzer et al. 1974). (2) The translation of extremely codon-biased mRNAs that codes for fibroin and sericin (protein components of silk) in the salivary glands of some arthropods such as the silkworm *Bombyx mori* requires a unique and highly skewed pool of tRNAs specifically adapted to favour the translation of these tRNAs (Chevallier and Garel 1982; Li, Ye et al. 2015). (3) The synthesis of polyproline proteins is promoted by the universal elongation factor P (EF-P) (orthologous EF5 in Eukarya and Archaea). EF-P binds to the ribosome and stimulates the peptidyl-transferase avoiding the ribosome stalling in stretches of poly-proline codons (Doerfel, Wohlgemuth et al. 2013; Lassak, Wilson et al. 2016).

It have been recently shown that the activity of two tRNA modification enzymes (tRNA-dependent adenosine deaminases (ADATs) in Eukarya and tRNA-dependent uridine methyltransferases (UMs) in Bacteria), both acting in base 34 of the anticodon and both increasing the codon-pairing ability, improves the correlation between codon usage and num-



**Figure 4.9:** (a) Chemical structure of Inosine wobbling. Hydrogen bridges are represented with green dashed lines. (b) A-to-I hydrolytic deamination. The transition state is the highest energetic barrier along the reaction coordinate. The overall energy profile is relative and the corresponding structures are displayed with the arrows for bond rotation or electron flow in each step. the  $Zn^{+2}$ -chelated hydroxyl and the proton of the catalytically essential Glu are highlighted. Modified from (Luo and Schramm 2008)

tRNA. This results evidence a specific regulation for Eukarya and Bacteria used to increase the translational efficiency of their respective genomes (Novoa, Pavon-Eternod et al. 2012).

#### 4.2.3 tRNA wobbling

Francis H. C. Crick in 1966 suggested that while the standard base pairing may be used in the first two positions of the triplet between the codon and the anticodon, there may be a higher pairing permissiveness between the third base of the mRNA codon and the first position of the tRNA anticodon (wobble position or position 34) what is known as tRNA wobbling. This hypothesis is explored systematically and could explain the general nature of degeneracy of the genetic code. It is assumed that the bases can be paired (form at least two H-bonds) in many different ways, but being limited by steric effects drove by the ribosome that will ensure that all tRNA molecules are presented to the mRNA in the same way (Crick 1966). In fact, whereas the standard base pairings have the same position for the glycosidic

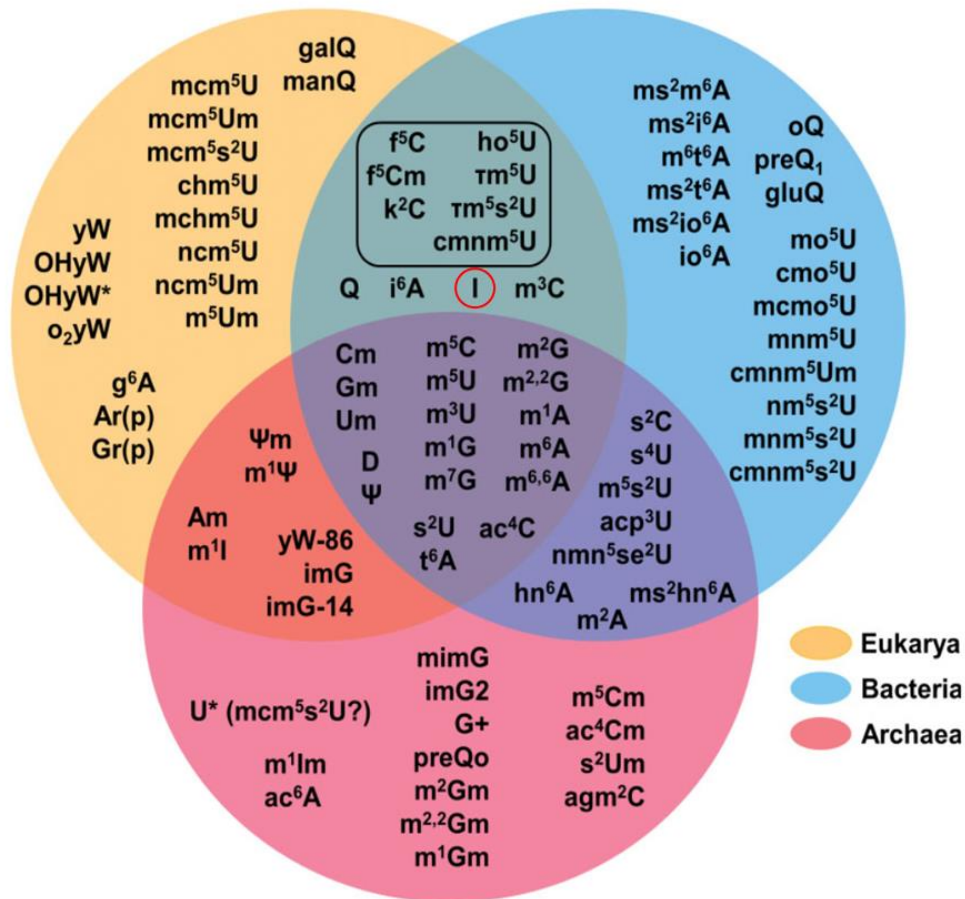


## Introduction

bond, the wobble pairings occupy alternative positions (Crick 1966). With these assumptions, the possible pairs are: (1) U34 can pair with A and wobble pair with G, (2) G34 can pair with C and wobble pair with U, and (3) I34 can pair with C and wobble pair with U and A (**Figure 4.9a**). For simplicity we say that I can wobble pair with C, U or A. Based on wobble hypothesis, 32 different tRNAs are needed at least to translate all 64 codons (Crick 1966). However, several genetic systems, such as organelles and some bacterial parasites encode fewer tRNAs than the theoretically required for the translation of all codons (see section 4.2.2). According to this, wobble theories have to be extended to explain the translation in these cases. The superwobbling suggests that U34 can wobble pair with any of the four bases (Shinozaki, Ohme et al. 1986; Pfitzinger, Weil et al. 1990), whereas the 2 out of 3 hypothesis suggests that only the two first codon bases need to be paired and the wobble base remains unpaired (Lagerkvist 1986; Shinozaki, Ohme et al. 1986; Sibling, Dirheimer et al. 1986; Delannoy, Le Ret et al. 2009). The two out of three requires a high stability of the base pairings, especially for the second one, where C:G is preferred. Both methods have the same consequence: a single tRNA can translate its cognate 4-codon family, independently of which base is placed at the third position.

### 4.2.4 tRNA modifications

Many nucleotides of the tRNA are post-transcriptionally modified to become a fully functional tRNA molecule. All nucleic acids in cells undergo chemical modifications including deamination, isomerization, glycosylation, thiolation, transglycosylation, methylations, etc. However, tRNAs undergo by far, the most numerous and chemically diverse post-transcriptional modifications, being modified approximately 17% of tRNA nucleotides, whereas in other RNA, molecules are modified about 1-2%. Decades of study have revealed more than 100 different ribonucleotide modifications that become crucial for tRNA structure, function and stability, and have a profound and generalized effect on protein synthesis (Grosjean, de Crecy-Lagard et al. 2010; Phizicky and Hopper 2010; El Yacoubi, Bailly et al. 2012; Jackman and Alfonzo 2013). Moreover, hypomodified tRNAs are targeted for degradation (Phizicky and Hopper 2010), and defects in tRNA modifications enzymes have been linked with human diseases such as cancer, type 2 diabetes, neurological disorders, and mitochondrial-linked disorders (Torres, Batlle et al. 2014).



**Figure 4.10:** tRNA modifications diversity classified according to the domains of life. Modifications enclosed in a box inside the Eukarya-Bacteria intersection are found in organelles reflecting the endosymbiotic theory that suggests that these organelles are descendants of bacteria. Inosine is circled in red. For common used symbols and abbreviations see section 3 (Abbreviations). Various combinations of the modifications listed above are indicated with combinations of multiple symbols; i.e., nm<sup>5</sup>s<sup>2</sup>U = 5-methylaminomethyl 2-thio uridine. Modified from (Jackman and Alfonzo 2013)

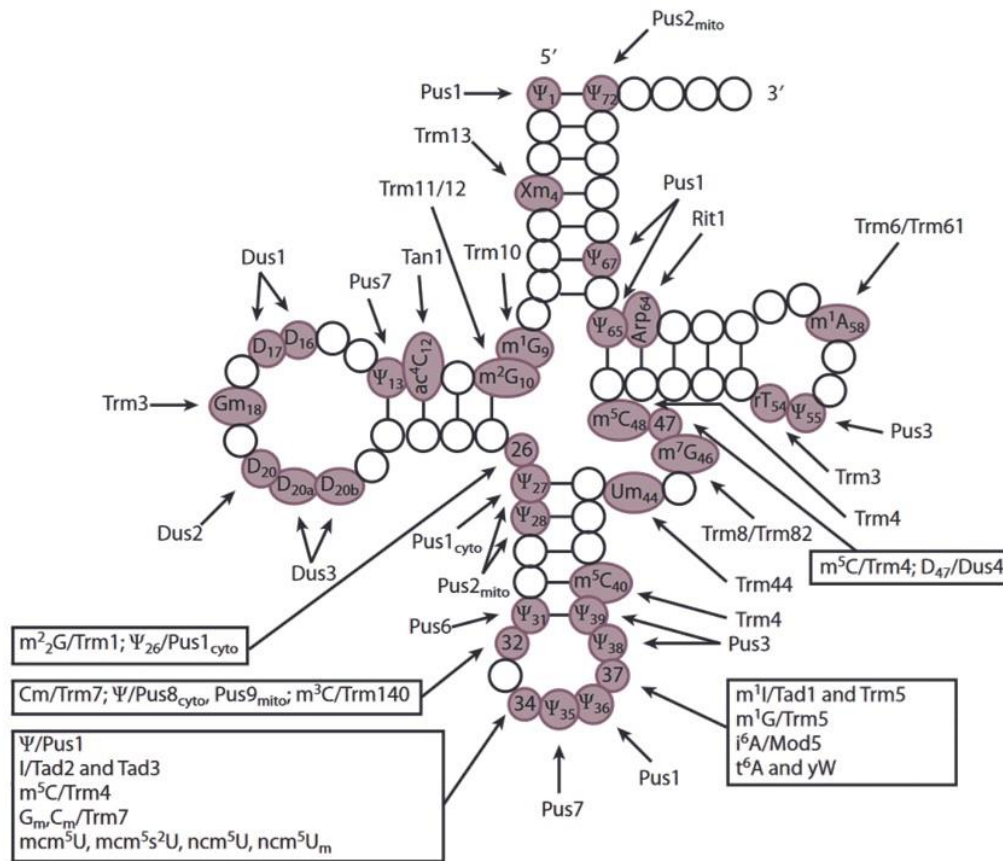
Certain tRNA modifications are present in all three domains of life, and in some cases, the same modification may appear at different positions being catalysed by different enzymes (Figure 4.10). Despite the chemical conservation of these modifications, the enzymes that synthesize them are not necessary evolutionary conserved, resulting in enzymatic convergent evolution as a way to cope with the selective pressure for a specific modification at a given nucleotide position. This core group of modified nucleotides that are conserved in the three domains, are generally characterized by relatively simple chemical structures, such as: addition of one or two methyl groups, replacement of oxygen with sulphur, isomerization of uridine (pseudouridines, Ψ), reduction of uridine (dihydrouridine, D), addition of other relatively small chemical functional groups (acetylation and threonylation), etc.

## Introduction

We can divide tRNA modifications in two major groups. Those that target the functional centres of the tRNA (i.e. anticodon loop and the acceptor stem) placed at opposite ends of the L-shaped three-dimensional structure, and those placed in the main body of the tRNA. Modifications placed in the main body of tRNAs are not usually essential for cell viability, they act in concert and in relatively subtle ways and they are involved in structural and stabilizing roles. For instance, pseudouridines are ubiquitous throughout the tRNAs and they favour the 3'-endo sugar pucker associated with the A-form RNA helices (the A-form helix geometry results in a very deep and narrow major groove and a shallow and wide minor groove) which increases the rigidity of the tRNA (Durant, Bajji et al. 2005). Alternatively, dihydrouridines are thought to promote the opposing 2'-endo sugar pucker associated with conformational flexibility in RNA (Dalluge, Hashizume et al. 1996; El Yacoubi, Bailly et al. 2012). Interestingly, tRNAs from cryophilic Archaea show a higher abundance of dihydrouridine relative to their thermophilic archaeal counterparts suggesting a tRNAs higher flexibility demand under low temperature conditions (Edmonds, Crain et al. 1991). Modifications placed in the functional centres of the tRNA have direct effects on decoding and protein synthesis and some of them are essential for cell survival. Some modifications in the acceptor stem (or close to it) are involved in the correct recognition of tRNA with aaRSs enzymes. For example, post-transcriptional addition of G at 5'-end of tRNA<sup>His</sup> is critical for charging it with His by histidinyl-tRNA synthetase (HisRS) in yeast (Rudinger, Florentz et al. 1994). Another example in yeast is 2'O-ribosyl phosphate modification at position 64 that exclusively acts on initiator tRNA<sup>iMet</sup>, discriminating it from elongator tRNA<sup>Met</sup> that is not modified (Astrom and Bystrom 1994).

Modifications in the anticodon loop are usually involved in translation accuracy and their absence could lead to problems to the ribosome to read the proper frame (frameshifting) and the consequent synthesis of aberrant proteins.

Particularly, positions 34 (wobble position) and 37 are modified in almost every tRNA and they comprise, by far, the largest chemical diversity of tRNA modifications directly contributing to the stability of tRNA-mRNA interaction during the process of decoding (Agris, Vendeix et al. 2007) (**Figure 4.11**). Modifications at position 37 maintain an open loop confirmation, sterically preventing base pairing with neighbouring nucleotides on the other side of the loop, such as the conserved U33, and also aid in the formation of the canonical U-turn in the tRNA structure, important for codon recognition and to prevent frameshifting (Ashraf, Ansari et al. 1999). For instance, tRNAs with a G at position 37 are methylated to m<sup>1</sup>G37 and the absence of this modification is associated with an increase of



**Figure 4.11:** tRNA modifications with the modification enzymes found in *Saccharomyces cerevisiae* cytoplasmic and mitochondrial tRNAs. The nucleotide position for each modification is indicated in the secondary structure (arrows). For common used symbols and abbreviations see section 3. For additional information see (Grosjean 2005). Modified from (El Yacoubi, Bailly et al. 2012).

frameshifting (Bjork, Jacobsson et al. 2001) and severe growth phenotypes (Urbonavicius, Qian et al. 2001).  $m^1G37$  is a primordial modification found in all domains of life, and in some cases it is further modified to more complex modifications such as wyosine (Noma, Kirino et al. 2006; Noma and Suzuki 2006; de Crecy-Lagard, Brochier-Armanet et al. 2010) and its derivatives such as wybutosine (Noma, Kirino et al. 2006; Noma and Suzuki 2006; de Crecy-Lagard, Brochier-Armanet et al. 2010). The tRNA modification enzymes responsible for  $m^1G37$  methylation are notated as SAM-dependent methyltransferases, and they are not evolutionary conserved between Eukarya (Trm5) and Bacteria (TrmD) (Christian and Hou 2007) despite they converged to the same enzymatic reaction. Convergent evolution is typical of many tRNA modifications and prevalent among many methyltransferases.

The wobble position (or position 34) is the tRNA nucleotide with more diverse chemical modifications, where they increase or restrict the wobbling capacities, altering the codon-anticodon recognition. U34-containing tRNAs, especially those from tRNA<sup>Gln</sup>, tRNA<sup>Glu</sup> and tRNA<sup>Lys</sup> are usually thiolated to  $s^2U$  and then hypermodified at the C5 position of the

## Introduction

pyrimidine ring with methylations, acetylations or with the addition of an entire sugar. These modifications, lead to increased codon rigidity and may result determinant for aminoacylation, translation efficiency and fidelity (Ashraf, Sochacka et al. 1999; Madore, Florentz et al. 1999; Tisne, Rigourd et al. 2000; Bjork, Huang et al. 2007; Johansson, Esberg et al. 2008). For example, the U34 modification  $xm^5s^2U34$  pairs with A and G, and  $xmo^5U34$  pairs with A, G and U regulating the wobble pairing since U34 might wobble pair with any of the four standard nucleotides by superwobbling (Lim and Curran 2001; Agris 2004). In terms of evolution, the enzymes responsible for  $s^2U34$  thiolation, evolved independently in the eukaryotic and bacterial lineage. The eukaryotic mitochondrial pathway resembles the bacterial one, supporting the endosymbiotic theory (Leidel, Pedrioli et al. 2009).

One of the best characterized modification in the wobble position is the hydrolytic deamination of A34 to inosine 34 (I34) (**Figure 4.9b**). Inosine is a G analogue that, according to Crick's wobble hypothesis, can wobble pair with the nucleotides C, U or A in the third position of the appropriate codons in the mRNA, enlarging the capacity of a single I34 tRNA molecule, to decode three different synonymous codons (**Figure 4.9a**). This modification is essential for viability and represents one of the most extreme cases of base pairing flexibility in protein synthesis, which increase the decoding efficiency and prevent frameshifts. (Schaub and Keller 2002). In Bacteria, the homodimeric enzyme tRNA adenosine deaminase A (TadA) modifies the nucleotide A34 of tRNA<sup>Arg</sup> (ACG) to I34 (Wolf, Gerber et al. 2002). TadA is an essential enzyme in *E. coli* (Wolf, Gerber et al. 2002) with a conserved catalytic center coordinated with a  $Zn^{+2}$  ion that binds to one histidine, two cysteines and one molecule of water responsible for the nucleophilic attack of the amino group mediated by an essential glutamic acid (**Figure 4.9b**) (Gerber and Keller 1999; Elias and Huang 2005; Spears, Rubio et al. 2011). In terms of evolution, a bacterial TadA gene was transferred to eukaryotes, probably during the endosymbiotic event, and through duplication and divergence, two different copies formed the heterodimeric enzyme Adenosine Deaminase Acting on tRNAs (ADAT) formed by ADAT2 and ADAT3. ADAT has been shown to be essential in humans (Torres, Pineyro et al. 2015), yeast (Gerber and Keller 1999; Tsutsumi, Sugiura et al. 2007), plants (Zhou, Karcher et al. 2014) and protists (Rubio, Pastar et al. 2007). The subunit ADAT2 holds a high identity with TadA, including the same conserved catalytic centre, whereas the subunit ADAT3 is less conserved and it has been suggested to play a role in substrate recognition (Gerber and Keller 1999; Wolf, Gerber et al. 2002). ADAT increased its substrate repertoire up to eight different A34 tRNAs that code for threonine, alanine, proline, serine, leucine, isoleucine, valine and arginine (TAPSLIVR) (Sprinzl, Horn et al.

1998; Gerber and Keller 1999). In Archaea, I34 has not been found in the anticodon of any tRNA. The crystal structure of TadA and the ADAT2-ADAT2 (pubmed 3DH1) homodimer has been resolved in several organisms, (**Figure 4.7c**) (Elias and Huang 2005; Kuratani, Ishii et al. 2005; Losey, Ruthenburg et al. 2006; Lee, Kim et al. 2007) but it still has not been resolved for the heterodimer ADAT2-ADAT3, probably because of the variability of the N-terminal end of ADAT3 (Spears, Rubio et al. 2011). In humans, ADAT has been localized in the nucleus although separately ADAT2 localizes mainly to nucleus and ADAT3 mainly to cytoplasm. These results indicate that I34 is incorporated in the nucleus at the precursor tRNA level and suggest that ADAT3 translocates to nucleus in an ADAT2-dependent manner (Torres, Pineyro et al. 2015). We have recently shown that the activity of ADAT effectively modifies the pool of tRNAs available for each codon, and aligns the correlation between the tRNA gene copy number and the codon usage bias in eukaryotes (Novoa, Pavon-Eternod et al. 2012). Although TadA and ADAT have been shown to be essential, the nature of the selection force driving the evolution from TadA to ADAT remains an open question that will be tackled in this thesis.

### **4.3 Molecular evolution and phylogeny**

#### **4.3.1 A brief history**

Molecular evolution is a discipline that comes from two separated branches of study: the molecular biology and the reconstruction of evolution history of organisms. Molecular biology includes the evolution of macromolecules, searching the mechanisms and the causes of sequence changes both at the genomic and proteomic level. The reconstruction of evolution history of organisms begins at the turn of XX century, when studies in immunochemistry showed that serological cross-reactions were stronger for more closely related organisms. This allowed to infer the phylogenetic reactions among various groups of animals where, for instance, apes were determined as the closest human relatives (Nuttall and Inchley 1904). In 1950s sequencing methods such as starch-gel electrophoresis were improved procuring a new and promising source of genomic and proteomic sequences that were more informative, easier to analyze and could be used for phylogenetic reconstruction within and between species. For instance, in 1952, the first complete sequence of a protein (insulin) was published by F. Sanger (Sanger and Thompson 1952). These data revealed that amino acid substitutions occurred non-randomly, as well as the existence of conserved regions in proteins that avoid substitutions (Sanger 1952). Yet, most of these substitutions

Original sequence	A U G Met	C A G Gln	U C A Ser
Silent mutation	A U G Met	C A <b>A</b> Gln	U C A Ser
Missense mutation	A U G Met	C <b>C</b> G Pro	U C A Ser
Nonsense mutation	A U G Met	<b>U</b> A G STOP	U C A
Frameshift mutation	A U G Met	C - - U C A	

**Figure 4.12:** Different types of mutation based on the effect they produce in amino acid translation. Punctual mutations and indel are highlighted in red. Coded amino acids are in grey.

did not have a significant effect on the biological activity what is known as neutral evolution, although a small number of them may account for large differences in the activity of two related proteins. Synonymous mutations, which cause no amino acid change, were considered a good example of neutral evolution. However synonymous codons for an amino acid may have different fitness because the tRNAs recognizing those codons (that could be different or not) may have different binding affinities or different concentrations in the cell (Richmond 1970). Nonetheless, studies in the pattern and rate of protein substitutions revealed that amino acids that are similar in physicochemical properties are interchanged more frequently than dissimilar ones. In 1962 the molecular clock hypothesis was proposed by Émile Zuckerkandl and Linus Pauling because they noticed that the rate of amino acid substitution in hemoglobin remains constant among different lineages (Zuckerkandl, Jones et al. 1960; Zuckerkandl and Pauling 1965). This hypothesis had an impact on the development of molecular evolution because the time in prehistory when two life forms diverged must be inferred. However, it also provokes a great deal of controversy because the concept of constancy is opposed to the erratic evolution of morphological and physiological events. The accumulation of genetic information prompted the development of methods of genetic distance measures, sequence alignments and tree-making.

### 4.3.2 Mutations

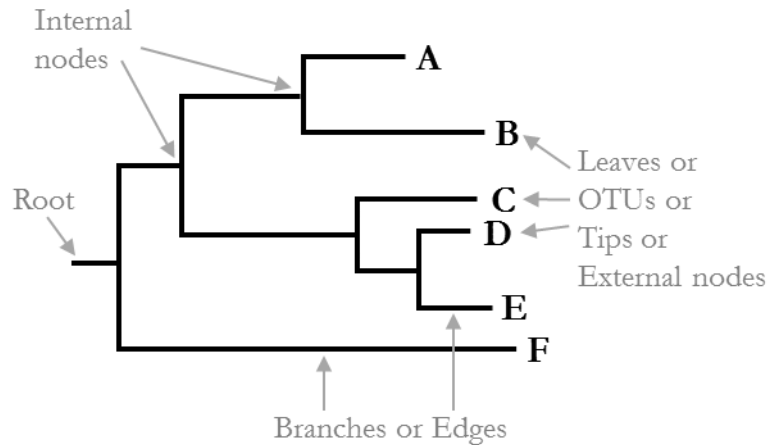
Mutations is a permanent alteration of DNA sequence caused by errors during DNA replication, DNA repairing or caused by the exposure to mutagenic chemical agents or radiation. Mutations can be classified by the type of change they produce into: (1) substitutions, where one nucleotide is replaced by another, (2) recombination, where a

sequence of consecutive nucleotides is replaced by another, (3) insertions and deletions, where a sequence of one or more nucleotides are inserted or removed respectively from the original sequence (**Figure 4.12**). Note that, when it is not known which one of the sequences is the original and which one is the mutated, it is not possible to know whether the mutation came by insertion or deletion, thus insertions and deletions are collectively called indels. Point mutations are those that only affect one nucleotide, whereas a whole region is affected we referred to as segmental mutations.

Point mutation are classified into: (1) transitions where purines (A, G) or pyrimidines (C, T), are substituted between them, and (2) transversions, where a purine is substituted by a pyrimidine or vice versa. When point mutations occurs in a coding sequence, they can also be classified by the effect on the amino acid that is translated: (1) synonymous or silent mutation generates a synonymous codon and therefore the amino acid does not change, (2) missense mutation generates a non-synonymous codon and therefore the amino acid is changed for another one, finally (3) nonsense mutation generates a stop codon that prematurely ends the translation process resulted in the production of a truncated protein. A frameshift mutation is produced by an indel in a coding-region with a length in nucleotides not multiple of three. The reading frame and consequently the encoded amino acids are altered from the mutation and beyond until a new stop codon is found, resulting in a mutated protein with abnormal length (**Figure 4.12**).

The process where a gene, or part of a gene is copied (indel) elsewhere in the genome is referred to as gene duplication. This process is a fundamental source of genetic diversity because the duplicated gene is probably redundant and free to selective contains. Therefore, it will accumulate mutations and diverge from its original ancestor. The duplicated gene might lose its functional activity eventually becoming a pseudogene or even being deleted from the genome (indel). Another possibility is that natural selection favors mutations of the duplicated gene, probably with a slightly altered functional activity compared with the original, and become part of the genome as a new gene copy, or even as a new gene if the functional activity becomes far diverged. Mutations do not occur with the same frequency, throughout the genome. There exist regions, so called hotspots of mutation, that are more prone to me mutated than others. Regions with a high concentration of consecutive nucleotides CG (CpG islands) tends to methylate the C, which in replication is substituted by T. Also, in Bacteria the dinucleotide TT and short palindrome sequences are mutation hotspots.





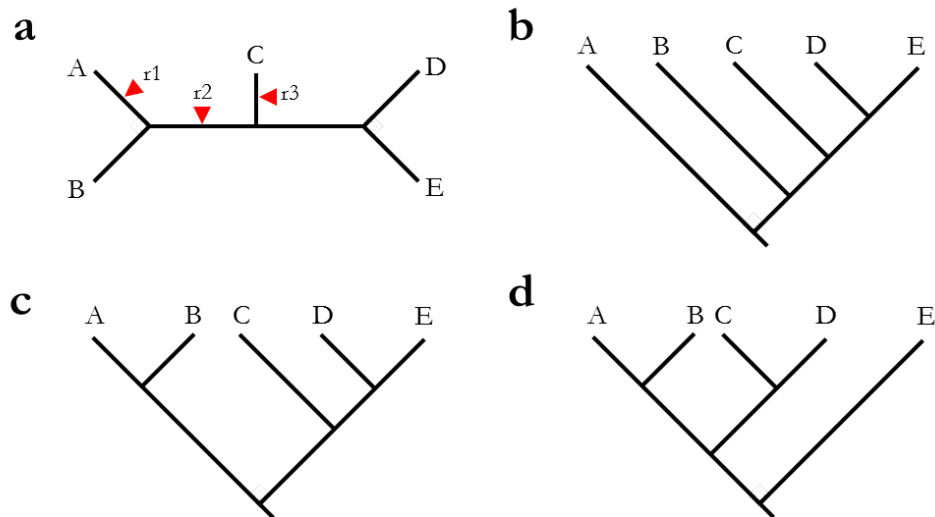
**Figure 4.13:** A simple phylogenetic tree. Leaves are connected by branches and internal nodes defining the topology of the tree. The root is defined as the ancestor of all the leaves and defines a time direction.

### 4.3.3 Molecular phylogenetics

All life on earth is related by the last universal common ancestor (LUCA). Each group of living organisms have its common ancestor and the closer they are, the more recent the ancestor is. Molecular phylogeny uses molecular data (basically DNA and protein sequences) of each organism to determine how related they are and illustrate their evolution by means of a phylogenetic tree.

**Tree terminology:** a tree is a set of external and internal nodes, that represents the organisms (or sequences) and its ancestors respectively, connected by a set of branches (or edges) whose pattern defines a specific topology (**Figure 4.13**). A tree is used to model the evolutionary history of a group of organisms or sequences and its topology represents the phylogeny or the evolution of the tree. Each external node (or leaf, or operational taxonomic unit (OTU), or tip) contain a portion of the available molecular data, represented by a sequence or a set of sequences that might come from extant or extinct organisms. The rest of the tree is inferred using phylogenetic models.

The root of the tree is defined as the ancestor of all the sequences. From the root to the leaves, there is a direction corresponding to evolutionary time or amount of changes (**Figure 4.13**). However, not all the trees have a root. Unrooted trees lack a root, and only specifies a topology between their external nodes, and hence, do not allow us to make assumptions about ancestors and descendants (**Figure 4.14a**). Unrooted trees might be rooted on one of their branches. From one unrooted tree, there are associated several different rooted trees

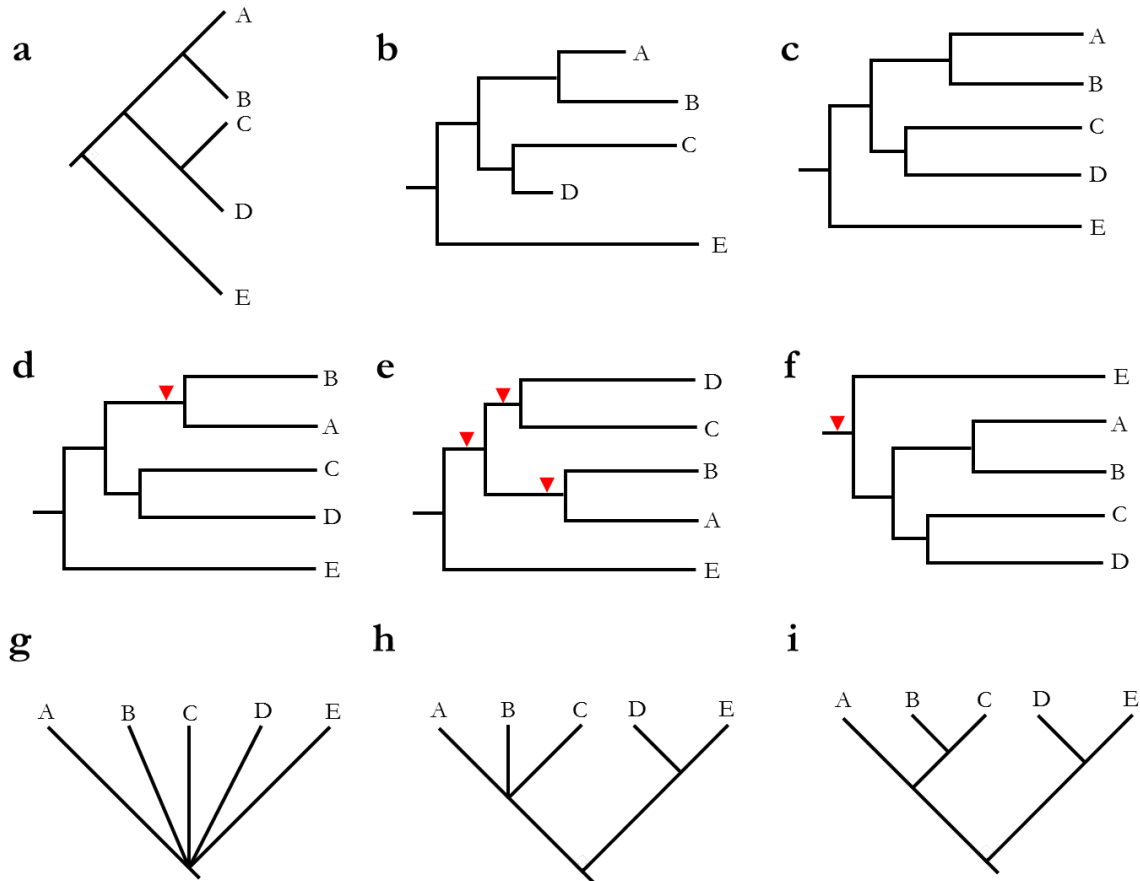


**Figure 4.14:** (a) Unrooted tree. Red arrows indicate three possible rooting sites r1-r3 that corresponds with (b-d) rooted trees respectively.

depending on where the root is placed. The number of rooted trees that come from the unrooted tree increases exponentially with the number of leaves (**Figure 4.14b**). For instance, from an unrooted tree of 4 nodes there are 5 possible unrooted trees, however for a relatively modest 20-leaves unrooted tree, the number of possible rooted trees ascend to the astronomic number of 8200 trillion.

Most of tree-making methods yields unrooted trees, that eventually might be rooted by the addition of an outgroup. An outgroup is a leaf that determines where the root is placed because it contains contrasted external information that clearly indicates that it has branched off earlier than the sequences under study. The selection of an outgroup must be appropriate. If the outgroup is evolutionally too far from the rest of the sequences, their alignment will be of poor quality, creating a non-confident tree. Otherwise, if the outgroup is too close from the rest of the sequences, the place where the root is determined must be wrong, and so the time direction.

**Types of trees:** depending on the information placed on their branches, phylogenetic trees are classified in three different classes (**Figure 4.15**): (a) cladograms, where their branches simply connect the nodes with no additional information, (b) additive trees, where branch lengths correspond to some attribute, typically the amount of evolutionary change, and (c) dendrograms or ultrametric trees, where all the leaves are equidistant from the root and so, the branch lengths represent the evolutionary time. Note that trees do not vary their

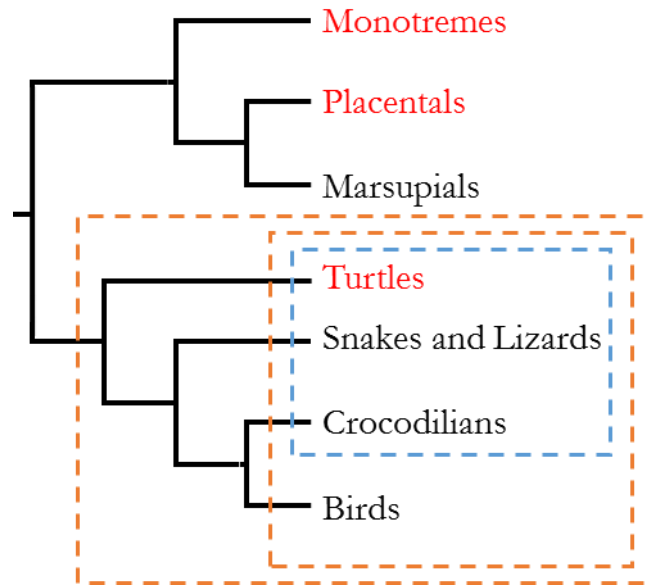


**Figure 4.15:** Different types of trees. (a) cladogram, (b) additive tree and (c) ultrametric tree where all the leaves are equidistant from the root. (d-f) Three trees with exactly the same interpretation as tree (c) because only rotations to internal nodes (red arrows) were applied. (g-i) Different levels of resolution on a phylogenetic tree. (g) Completely unresolved tree also called star tree. (h) Partially resolved tree with one polytomy. (i) Fully resolved tree without polytomies.

interpretation when rotations to any of their internal nodes are applied (**Figure 4.15d-f**). In these sense trees are like mobiles, whatever you rotate their objects, the connection between them remains unaltered.

**Tree shape:** the number of branches that connect an internal node is referred to as node's degree. If a node's degree is higher than three (one ancestor and two immediate descendants) it is defined as a polytomy. A tree is considered fully resolved when it does not have any polytomy (**Figure 4.15g-i**). Typically, polytomies are treated as uncertainty about phylogenetic relationships, the lineages probably did not diverge at once but the information provided is insufficient to resolve the actual order of divergence.

A group of leaves are considered monophyletic if there exist a common ancestor for exactly all the leaves considered without any additional leaf. Otherwise the group is defined polyphyletic (**Figure 4.16**). When in a polyphyletic group only a few leaves are missing to

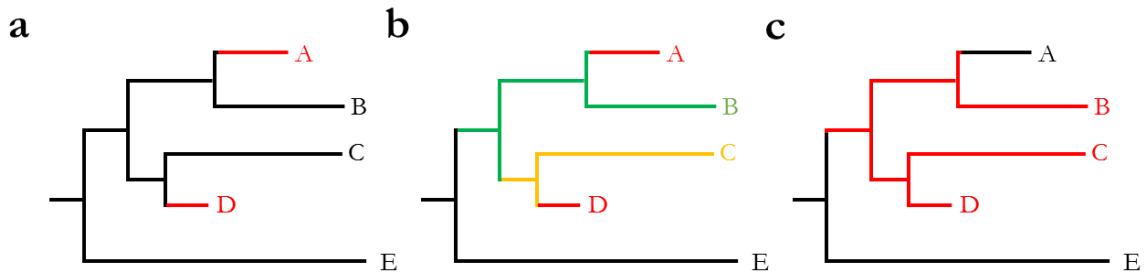


**Figure 4.16:** Monophyletic group or clade of Amniota (orange dashed box). Paraphyletic group of Reptilia (blue dashed box). Polyphyletic group (red leaves).

become monophyletic it might also be referred to as paraphyletic group. A typical example of paraphyly is found in the classical taxonomic assignment of reptiles, where birds are not included because of their anatomic differences, even though crocodiles are molecularly more related to birds than they are to the rest of the reptiles (**Figure 4.16**). A clade is defined as a monophyletic group that includes all the descendant nodes from the common ancestor. However, monophyletic groups and clades are often used exchangeably.

**Character change reconstruction:** the set of leaves of a phylogenetic tree have different intrinsic properties that could be directly related to the molecular data used to perform the tree (e.g. the distribution of a point mutation along the sequences) or could also be not directly related. In phylogenetic studies, it is interesting to reconstruct the history of the character changes based on the tree topology. We distinguish between three different types of evolution (**Figure 4.17**): (1) parallel evolution, where the same character evolved independently from the same ancestral condition in two different leaves of the tree, (2) convergent evolution, where the same character evolved independently from different ancestral conditions in two different leaves of the tree, and (3) secondary loss, where the character apparently reversed to the ancestral conditions.

**Tree-making methods** are the mathematical methods that allows the conversion of molecular data into a phylogenetic tree. There are several, and still increasing, different



**Figure 4.17:** Three different types of evolution. Black branches indicate ancestral character condition. Red branches correspond to the new character condition. Green and yellow branches correspond to an intermediate character condition. (a) Parallel evolution implies independent evolution of same new character from the same ancestral condition. (b) Convergent evolution implies independent evolution of same new character but from different ancestral conditions. (c) Secondary loss implies reversion to the ancestral condition.

number of methods depending on how the data is interpreted, because no single method performs well under all circumstances. Here we focus on maximum likelihood (ML) method. This method considers different possible trees based on the character state configurations among their sequences, associates a ML value for each one and selects the preferred tree as those with the highest ML value. The different states between the sequences are based on a probabilistic model that specify the transition probabilities from one state to another in a time interval. For instance, if the molecular dataset of the tree are protein sequences, the model might specify the probability to mutate one amino acid (or a gap) to another amino acid (or a gap) in a specific time. ML probabilistic methods were developed for genomic sequences (Cavalli-Sforza and Edwards 1967) and for protein sequences (Felsenstein 1973; Felsenstein 1981) several years ago. However, it becomes popular in the last twenty years, because its high computational uptake has been overcome (Belhassen, Domme et al. 1991).

**Bootstrapping:** In order to perform a phylogenetic tree, the molecular data have to be processed as follows: (1<sup>st</sup>) all the sequences have to be aligned together in a multiple sequence alignment (MSA). This transforms the set of sequences in a single matrix where each column corresponds to a character and each row corresponds to a leaf of the tree. (2<sup>nd</sup>) the matrix has to be sampled, that is, some of the columns but not all, are selected to perform the tree. Therefore, the topology of the tree depends ultimately on which characters have been sampled. Bootstrapping consist on resampling the characters and rebuild the tree several times (100 to 1000 times) in order to test the robustness of the topology. A consensus tree

is plotted based on the commonality among all the resampled trees. Each node of the consensus tree has associated a bootstrap value, indicating the percentage of recovery from all the samples. As a rule, bootstrap values  $\sim 70\%$  indicates a moderate support whereas BS  $> 90\%$  indicates a good support.



## 5. Objectives

In higher organisms, there is apparently no correlation between num-tRNA and codon usage bias. However, this correlation is improved if the tRNA modifications caused by ADAT are taken into account, suggesting that the emergence of ADAT in Eukarya contributed to the evolution of genomic codon composition and tRNA gene content differences (Novoa, Pavon-Eternod et al. 2012).

Based on these results, we attempt to unravel the influence of ADAT activity in human translation. (i) Are there regions in the human transcriptome (ADAT stretches) whose translation is prone to be regulated by ADAT? If so, which is the codon composition of these regions? We also extended our initial analysis in a series of representative eukaryotic and bacterial organisms spanning the whole tree of life, to address the following questions: (ii) How do ADAT stretches, tRNA gene composition and ADAT proteins evolve across species, is there any relation between them? (iii) Is there any organism that behaves differently than expected? (iv) Which are the differences between Eukarya (ADAT2-ADAT3) and Bacteria (TadA)? (v) How did ADAT evolve from TadA? (vi) Is I34 involved in the regulation of gene expression levels?





## 6. Publications

### 6.1 PhD advisor Report

Mr. Àlbert Rafels Ybern joined our group in 2013 as a Ph.D. student to work on the computational analysis of the molecular function of I34 in transfer RNAs, and the evolution of the machinery linked to the synthesis of this modified base. He was accepted to this position because of his unusual academic background that includes a B.Sc. in Mathematics and a M.Sc. in Biomedical sciences. That unusual training background made him the best candidate to tackle what was, essentially, a computational problem framed within a complicated biological context.

As was to be expected, Àlbert required an initial adaptation phase to familiarize himself with the biology associated of gene translation, including the complex field of tRNA modifications, roles of I34, and more general topics such as evolution, molecular biology, biochemistry, etc. Over the last four years he has greatly progressed in his understanding of these scientific fields, and he has managed to apply his knowledge of computational methods for mathematic analysis to the problems that constitute the body of his Ph.D. thesis.

Through this work, Àlbert has studied a virtually unexplored aspect of the role of tRNA modifications on proteins synthesis, and the evolutionary forces that drove the extant distribution of I34 in Bacteria and Eukarya. He has also been instrumental in the characterization of the phylogenetic distribution of I34, and in the identification of species whose I34 content allows to us draw conclusions about the mechanisms that rule tRNA identity and nucleotide modification expansions.

The Ph.D. thesis that Àlbert will be defending is a very original piece of scientific work, which truly explores previously unknown aspects of biological knowledge and reveals fundamental mechanisms that shaped the evolution of bacteria and eukaryotes. My laboratory is very proud of the research presented in this thesis, and of the conclusions that it reaches. The importance of the work is reflected on the publication record that it has generated. It is uncommon for a Ph.D. candidate to be able to present three first-author publications, and I believe each of the three articles to be ground-breaking in its own right.

Note that in **Publication 3**, Àlbert Rafels performed the bioinformatics analyses corresponding to all the figures except **Figure 2**, **Figure S1**, and **Table S2**. The co-author

## Publications

Adrian Torres performed the experimental procedures corresponding to **Figure 2**, **Figure S1**, and **Table S2** and did not use for any other PhD thesis.

Lluís Ribas de Pouplana

Gene Translation Laboratory

Institute for Research in Biomedicine (IRB)

## 6.2 Publication 1

**Rafels-Ybern, A.,** C. S. Attolini and L. Ribas de Pouplana (2015). “Distribution of ADAT-Dependent Codons in the Human Transcriptome.” *Int J Mol Sci* 16(8): 17303-17314.

Impact Factor 5 years (2016): 3.48



Article

## Distribution of ADAT-Dependent Codons in the Human Transcriptome

Àlbert Rafels-Ybern <sup>1</sup>, Camille Stephan-Otto Attolini <sup>1</sup> and Lluís Ribas de Pouplana <sup>1,2,\*</sup>

<sup>1</sup> Institute for Research in Biomedicine (IRB), Parc Científic de Barcelona, C/Baldiri Reixac 10, 08028 Barcelona, Spain; E-Mails: albert.rafels@irbbarcelona.org (A.R.-Y.); camille.stephan@irbbarcelona.org (C.S.-O.A.)

<sup>2</sup> Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

\* Author to whom correspondence should be addressed; E-Mail: lluis.ribas@irbbarcelona.org; Tel.: +34-93-40-34868.

Academic Editor: Michael Ibba

Received: 30 April 2015 / Accepted: 6 July 2015 / Published: 29 July 2015

---

**Abstract:** Nucleotide modifications in the anticodons of transfer RNAs (tRNA) play a central role in translation efficiency, fidelity, and regulation of translation, but, for most of these modifications, the details of their function remain unknown. The heterodimeric adenosine deaminases acting on tRNAs (ADAT2-ADAT3, or ADAT) are enzymes present in eukaryotes that convert adenine (A) to inosine (I) in the first anticodon base (position 34) by hydrolytic deamination. To explore the influence of ADAT activity on mammalian translation, we have characterized the human transcriptome and proteome in terms of frequency and distribution of ADAT-related codons. Eight different tRNAs can be modified by ADAT and, once modified, these tRNAs will recognize NNC, NNU and NNA codons, but not NNG codons. We find that transcripts coding for proteins highly enriched in these eight amino acids (*ADAT-aa*) are specifically enriched in NNC, NNU and NNA codons. We also show that the proteins most enriched in *ADAT-aa* are composed preferentially of threonine, alanine, proline, and serine (*TAPS*). We propose that the enrichment in ADAT-codons in these proteins is due to the similarities in the codons that correspond to *TAPS*.

**Keywords:** tRNA modification enzymes; ADAT2-ADAT3; codon degeneracy; tRNA gene copy number

---

## 1. Introduction

The genetic code is degenerate, as the number of amino acids coded is smaller than the number of possible codons, and multiple codons can code for the same amino acid. Typically, the multiple codons that can correspond to a single amino acid are not equally abundant in the genome. This codon bias is a signature of genomes, and can vary widely from species to species.

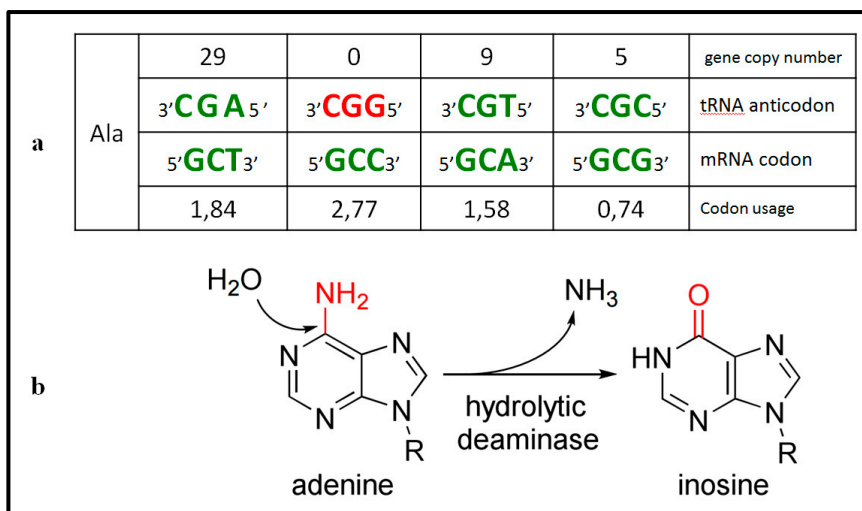
In addition, the number of tRNAs with different anticodons is always smaller than the number of codons used in any species, because a single tRNA anticodon may pair with more than one codon. This is due to the fact that a higher pairing permissiveness exists between the third position of the mRNA codons and the first position of the tRNA anticodon. This is still known as wobble or degenerate pairing [1], although recent crystallographic data has shown that the ribosome enforces a Watson–Crick geometry even at the third position of the codon-anticodon interaction [2].

We have recently shown that codon bias and tRNA gene copy number in eukaryotes were influenced by the emergence of heterodimeric adenosine deaminases acting on tRNAs (ADAT), which deaminate A34 to I34 in those tRNAs with ANN anticodons that decode threonine, alanine, proline, and serine (*TAPS*), and leucine, isoleucine, valine, and arginine (*LIVR*) [3] (Figure 1b). Here we will refer to these amino acids as *ADAT-aa*.

ADAT most likely evolved from the homodimeric bacterial adenosine deaminase TadA, which acts solely on *tRNA*<sup>Arg</sup><sub>ACG</sub> [4,5]. The emergence of ADAT was instrumental in the enrichment of genes coding for *TAPS* and *LIVR* tRNAs with ANN anticodons in eukaryotes. With the exception of *tRNA*<sup>Arg</sup><sub>ACG</sub>, these tRNAs are virtually absent in bacteria and archaea [6]. The activity of ADAT effectively modifies the pool of tRNAs available for each codon, and aligns the correlation between codon usage and tRNA gene copy number in eukaryotes [3].

Inosine 34-modified tRNAs can “wobble” pair with A, C or U, and this solves the apparent riddle offered by the abundance of C-ended codons coding for *TAPS* and *LIVR* and the complete absence of the corresponding tRNAs with GNN anticodons in eukaryotes (Figure 1a) [7]. ADAT was first shown to be an essential enzyme in yeast, and has been later characterized in *Trypanosoma*, and *Arabidopsis* [5,8–10]. We have recently shown that, in *Homo sapiens*, the modification by ADAT of the eight cytoplasmic tRNAs, *tRNA*<sup>Ala</sup><sub>IGC</sub>, *tRNA*<sup>Pro</sup><sub>IGG</sub>, *tRNA*<sup>Thr</sup><sub>IGU</sub>, *tRNA*<sup>Val</sup><sub>IAC</sub>, *tRNA*<sup>Ser</sup><sub>IGA</sub>, *tRNA*<sup>Arg</sup><sub>ICG</sub>, *tRNA*<sup>Leu</sup><sub>IAG</sub> and *tRNA*<sup>Ile</sup><sub>IAU</sub>, takes place predominantly in the nucleus, during the maturation process of these molecules [11].

Generally speaking, the study of the influence of anticodon modifications on the translation of specific codons has recently led to the realization that tRNA populations can act as a new layer of gene translation regulation through the modulation of their anticodon modification status, or through changes in the expression levels of different tRNA genes [12–17]. In the case of ADAT, and despite its importance in the evolution of eukaryotic genomes, little is known about its potential role in translation regulation [12]. Here we present the first analysis of the distribution of ADAT-related codons in the human transcriptome, and computationally characterize the proteins most rich in *ADAT-aa*.



**Figure 1.** (a) Codon–anticodon relation for ADAT-related alanine amino acid. tRNA copy number and codon usage is shown for each pair, note that  $tRNA_{GCG}^{Ala}$  do not exist in human (red nucleotides); (b) Hydrolytic deamination: Adenine is converted into an inosine throughout a hydrolytic deamination reaction.

To try to identify those polypeptides whose translation is more likely to be influenced by ADAT activity, we have first classified the human proteome according to the abundance of *ADAT-aa* in each protein. We have used two different methods to determine the distribution of *TAPS*-, and *LIVR*-coding triplets in the human genome: a *half-gene* analysis, and a *running-window* approach. Those transcripts with a significantly increased proportion of these triplets have been analyzed for their composition in ADAT-preferred codons, to test if these are enriched with respect to G-ended codons in these proteins.

We show that, in general, ADAT codons (those that can be recognized by tRNAs modified by ADAT) are generally preferred to G-ended codons (not recognizable by ADAT) in the human genome. Moreover, this preference increases in proteins enriched in *ADAT-aa*. Interestingly, although we included both *TAPS* and *LIVR* in the search for proteins highly enriched in ADAT-related codons, we find that the most biased human protein sequences in this regard are only enriched in *TAPS*.

Coherently, in these sequences only the triplets for *TAPS* are enriched for ADAT-dependent codons, indicating that the activity of ADAT may be important for the translation of gene regions coding for long stretches of *ADAT-aa*. We argue that this enrichment may be explained by the fact that codons for *TAPS* occupy a close position in the genetic code, where all of them share the same second base.

More importantly, our results hint at the possibility that the emergence of ADAT allowed eukaryotic cells to produce highly repetitive protein sequences that bacterial or archaeal ribosomes may be unable to translate due to the absence of I34-containing tRNAs in these organisms.

## 2. Materials and Methods

### 2.1. Definitions

*ADAT-aa* are defined as those amino acids that are charged to tRNAs that can be modified by ADAT (Thr, Ala, Pro, Thr, Ser (*TAPS*), and Leu, Ile, Val, and Arg (*LIVR*)). *C* is defined as the set of all the 64 codons. *A* is defined as the total set of codons that code for any of the *ADAT-aa*, and corresponds to the



37 codons present in Figure S6.  $D$  is defined as the subset of 24 codons of  $A$  that are recognized by tRNAs modified by *ADAT* at position 34 (Figure S6 codons that can “wobble” pair with I34 anticodons). For a given region  $t$  of a coding sequence, the amount of codons in  $t$  that belongs to  $C$ ,  $A$  or  $D$  is defined as  $c(t)$ ,  $a(t)$  or  $d(t)$ , respectively. *ADAT stretch* are those regions that have a high value for  $a(t)/c(t)$  compared with the rest of the transcriptome. We will define the stretch in more detail in the next section.

## 2.2. Human Transcriptome Retrieval

We have analyzed 28,870 human Coding Sequences (CDSs) that conform the human transcriptome. All these sequences have been downloaded from the Consensus CDSs (CCDS) project [18]. Only CDSs with start codon, stop codon and a number of nucleotides multiple of 3 were used for our analysis. Only 48 CCDSs were eliminated.

## 2.3. Identification of Stretches by the Halves-Genes Method

To carry on this analysis we developed the *Halves-genes method*. Each CDSs of the human transcriptome is recursively divided into halves and for each region  $a(t)$  is calculated (Figure S1). We divided each CDS until sections of ~15 codons were reached. When the region to be divided had an odd number of codons, the first half was assigned one codon more than the second (Figure S1). Note that this method has the disadvantage that each CDS is represented several times but with different lengths. However, as all the CDSs are equally treated, there is no bias in the final data.

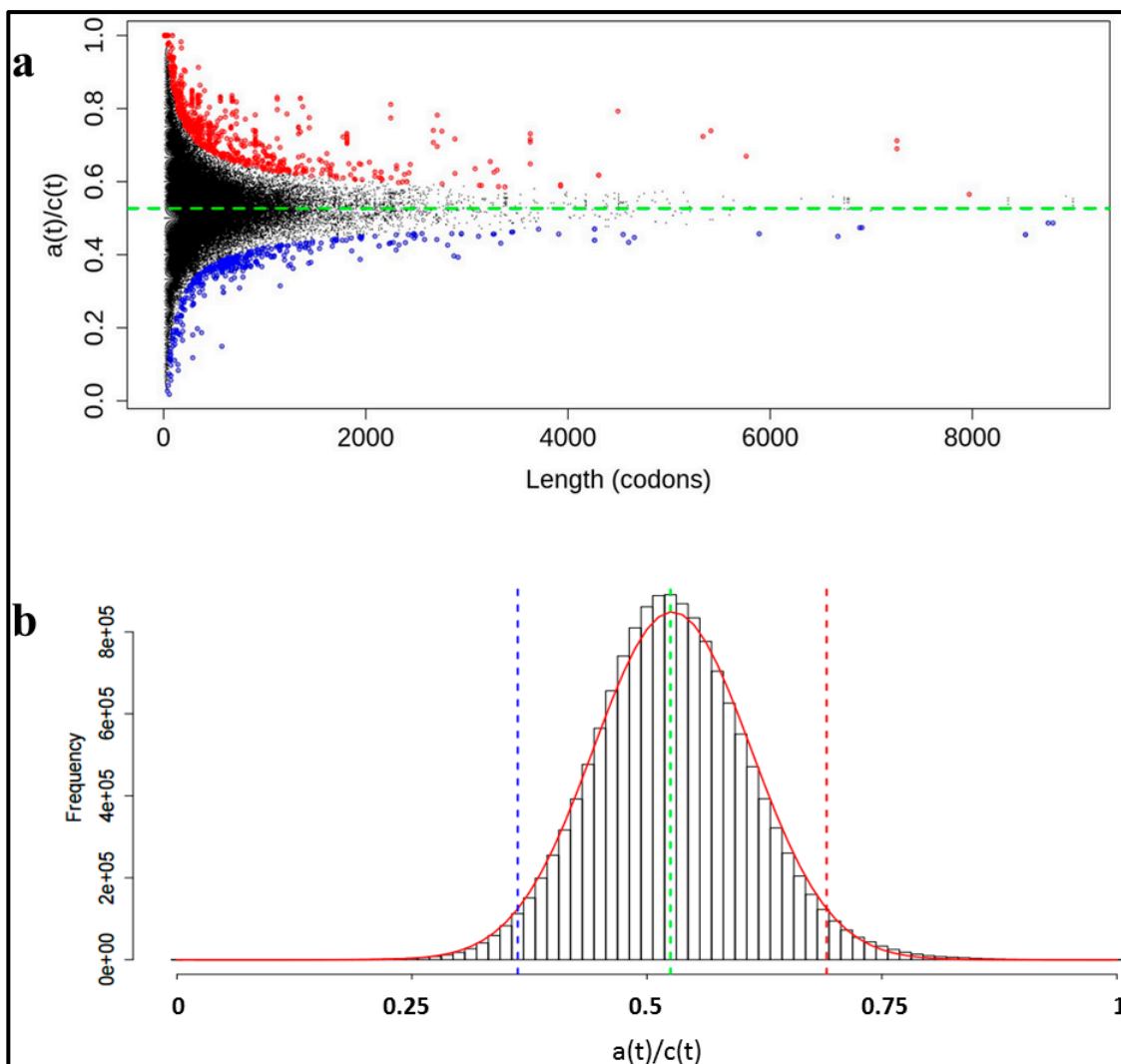
The variability in Figure S7 was measured by Interquartile Range (IQR). IQR is equal to the difference between the 1st and the 3rd quartiles ( $IQR = Q_3 - Q_1$ ). The density plot in Figure S5 was computed using the *smoothScatter* function in *graphics* package for R. Multiple linear regression in Figure S5 was computed using the *segmented* package for R with seeds 0.3 and 0.7 [19] and the slopes were obtained with the function *slope*.

## 2.4. Identification of Stretches by the Running Windows Method

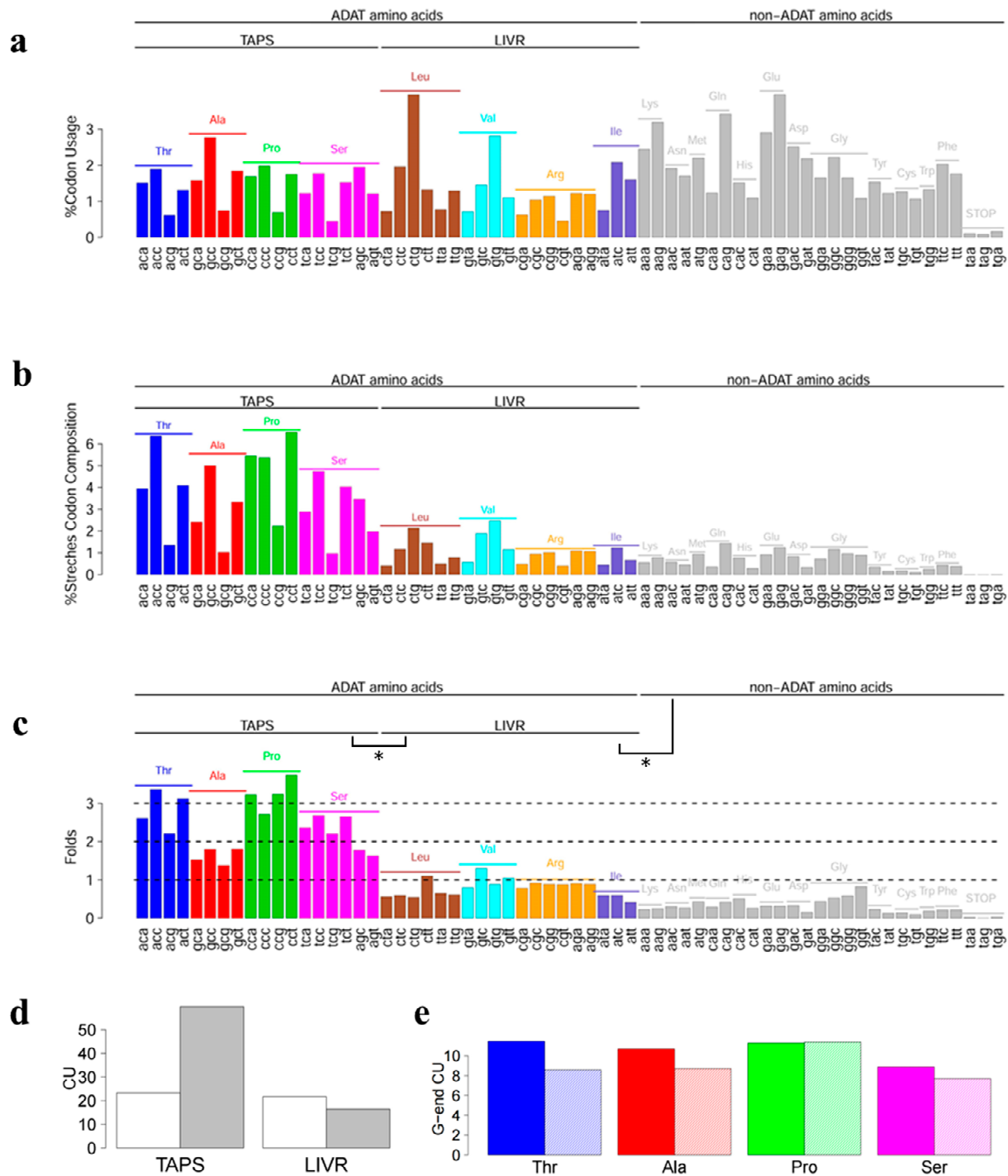
The False Discovery Rate (*FDR*) is the ratio between the expected values and the obtained values (red line, and histogram bars respectively in Figure 2b). Wilcoxon test (Figure 3) was computed using *stats* package for R.

To study in more detail the presence of *stretches* of  $A$  codons in the human transcriptome, we applied the *Running Windows method* (Figure S4) based on software that applies similar methodologies [20–22]. For each CDS of the human transcriptome a window (fragment of the sequence with a fixed size) slides codon by codon from the beginning to the end of the sequence (Figure S4a). For each window,  $a(t)$  is calculated and represented with respect to its location (Figure S4b). To fix the window size we took advantage of the previous method (*Halves-genes method*) to find a region length as small as possible but with a low variability (Figure S7). We fixed a window length of 80 codons because it has a low variability (IQR ~40%) and because this length approximates the average size of single protein domains [23]. We arbitrarily limited future analyses to windows enriched in  $A$  codons with  $FDR < 0.2$ , *i.e.*, those windows with  $a(t)$  comprised in the interval 67–80 (Figure S3). We define an *ADAT stretch* as those regions corresponding to a window, or a set of consecutive windows with this enrichment (Figure S4). Two (or

more) windows are considered consecutive if the intersection between them in the cognate CDS is not void. This corresponds to 560 sequences containing stretches ( $\geq 80$  codons) of *A codons* in 242 different human genes.



**Figure 2.** (a) Distribution of the human transcriptome using the *Halves method* to study the enrichment in *A codons*. The dashed green line shows the mean of *A codons* in the human transcriptome (0.527). Black dots correspond to all the regions obtained with this method (see Section 2.3). Red circles correspond to the enriched regions (2666). Blue circles correspond to the unenriched regions (412). Both regions are calculated supposing a binomial distribution with  $p$ -value  $< 10^{-11}$ ; (b) Running Windows Method distribution.  $a(t)$  enrichment for all the windows. Red line is the normal distribution following this histogram. Blue and red dashed lines show where the tails of the distribution represents a 5%. Green dashed line shows the mean.



**Figure 3.** (a) Codon usage of the human transcriptome [24]; (b) Codon usage for *ADAT stretches*. A codons are colored with a different color for each ADAT amino acid. (a,b) Each percentage is measured with respect to the total of codons, thus all the bars sums 100%; (c) Fold increase of *ADAT stretches* (b) normalized by human codon usage (a). There are significant differences (see asterisks \*) between sets of *TAPS* and *LIVR* ( $p$ -value =  $10^{-7}$ ), and sets of *LIVR* and the *non-ADAT amino acids* ( $p$ -value =  $5 \times 10^{-8}$ ). One-tail Wilcoxon test was applied in both cases with confidence level 0.95; (d) Codon usage for *TAPS* or *LIVR* codons in the human transcriptome (white) or in the *ADAT stretch* regions (grey); (e) Codon usage for G-ended codons for *TAPS* in the human transcriptome (colored) or in the *ADAT stretches* (dashed colored).

### 3. Results and Discussion

#### 3.1. Identification of Human Proteins Highly Enriched in ADAT-aa

Using the two strategies described in Section 2, we identify those transcript sequences more enriched in *A codons*, which correspond to protein regions that are highly enriched in *ADAT-aa* compared to the rest of the proteome. Figure 2a shows the distribution of  $a(t)/c(t)$  of all the fragments analyzed, as a function of their sequence length using the *Halves-gene method* (Figure S1). The distribution is centered on the mean of *A codons* for the human transcriptome (0.527, dashed green line). The variability in the y-axis decreases as the length of the fragments increases. We detect a number of sequences (red and blue points) that significantly deviate from an expected random distribution. This behavior is expected because of the non-random nature of the genome.

Supposing that all the samples follow a binomial distribution, we identified outliers with a  $p$ -value  $< 10^{-11}$ . There are 2666 samples enriched in *A codons*, and 412 depleted of *A codons*. Therefore the distribution of the outliers is not well-balanced (Figure S2), with a tendency of the CDSs to create regions highly enriched in *A codons*.

#### 3.2. Stretches of A codons Are Composed Preferentially by Triplets Coding for TAPS

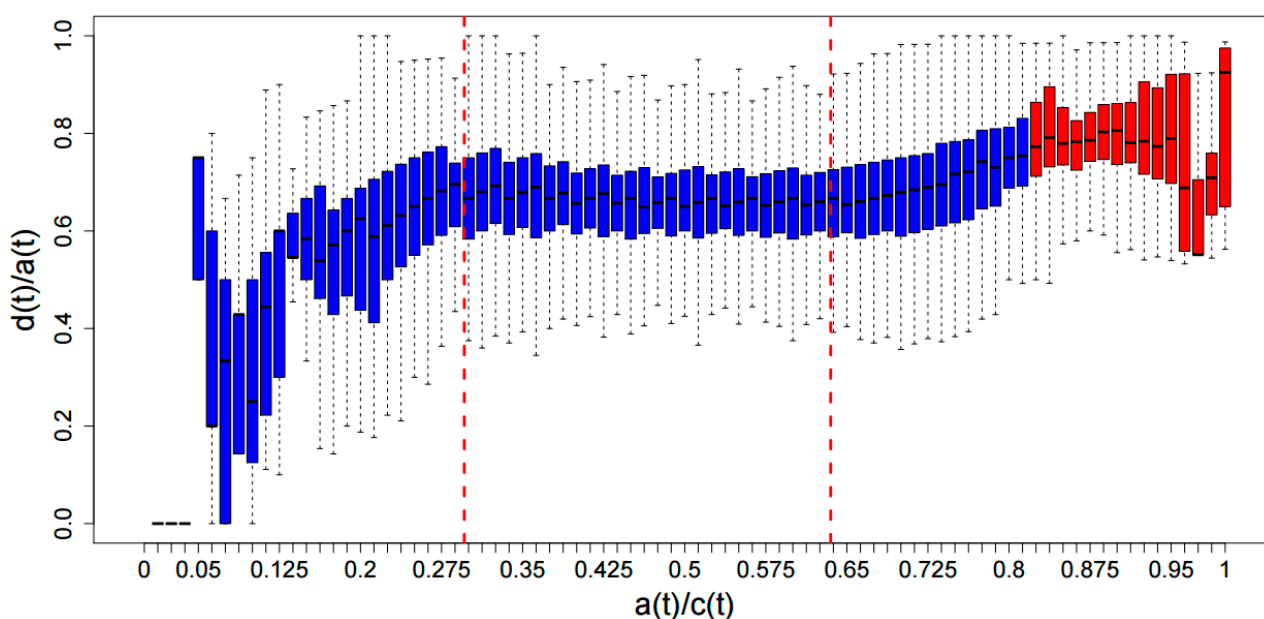
The *Running Windows method* gives a different distribution compared to the *Halves-gene method* due to the different nature of the data (Figure S4, Section 2). Figure 2b shows the distribution of all the windows for the whole transcriptome, based in the abundance of *A codons*. The most frequent value is at  $a(t)/c(t) = 0.525$  (green line). Note that in this method  $c(t)$  have a constant value of 80 codons. The mean for all the windows is at  $a(t)/c(t) = 0.526$  and the codon usage for *A codons* is 0.527. If normality of the data is assumed (red line), 95% of windows are [24] comprised between the region [0.362, 0.7] (region between blue and red lines). The windows outside of this central region are either depleted or enriched in *A codons*. The number of enriched windows is  $5.2 \times 10^5$  while the number of unenriched windows is  $2.7 \times 10^5$ . Note that, when comparing the two tails, *FDR* for the enriched tail is always lower than the *FDR* for the unenriched tail (Figure S3), thus the distribution is not symmetric and indicates again a preference in the human transcriptome for sequences enriched in *A codons*.

Figure 3b shows the composition of individual codons for all the *ADAT stretches*, and Figure 3a shows the codon usage for the whole human transcriptome. Surprisingly, not all the *A codons* are equally enriched in *ADAT stretches*. Codons for *TAPS* strongly predominate over codons for *LIVR* ( $p$ -value =  $10^{-7}$ ) (Figure 3d; Table S1), indicating that the enrichment of ADAT-dependent codons is not uniform, and that the concentration of *TAPS* can reach much higher values in proteins than the concentration of *LIVR*.

Figure 3c shows the fold change for codon composition comparing the stretches with the rest of the genome. Each residue in the *TAPS* group reaches enrichments of more than two-fold with respect to the mean, with threonine and proline reaching almost three-fold increases. Strikingly the concentration of *LIVR* codons decreases in the *ADAT stretches* (Figure 3d; Table S2). Finally, G-ended codons for *TAPS* are significantly decreased in the *ADAT stretches*, with the sole exception of Pro, which remains stable (Figure 3e; Table S3).

### 3.3. ADAT Stretches Are Composed Preferentially of D codons

We asked whether *ADAT stretches* would be significantly enriched in *D codons*, and depleted of G-ended codons. To carry out this analysis we calculated the relative concentration of *D codons*, that is,  $d(t)/a(t)$  for all the samples. Figure 4 shows a boxplot graph of  $d(t)/a(t)$  for all the windows (blue bars) and all the stretches (red bars) belonging to the respective interval in  $a(t)/c(t)$ . Those boxplots that correspond to an stretch are plotted in red located at the region  $a(t) > 67$ . Figure S5 shows a density plot for all the samples where the points correspond to the means in the boxplot intervals graph, and the black lines correspond to a multiple linear regression based on the mean values. Two breakpoints can be seen when  $a(t)/c(t)$  is 0.296 and 0.635 (Figure 4, dashed red lines). The behavior of the data is well differentiated and can be divided into three regions. The region  $a(t)/c(t)$  in  $[0, 0.296]$  comprises only a 0.27% of the of windows and therefore the linear regression is not considered. The region  $a(t)/c(t)$  in  $[0.296, 0.635]$  comprises 89.48% of windows and the linear regression is essentially flat (slope  $-0.07 \pm 0.03$ ), indicating a non-dependence between  $a(t)/c(t)$  and  $d(t)/a(t)$ . Finally,  $a(t)/c(t)$  in  $[0.635, 1]$  contains 10.25% of windows with a slope of  $0.52 \pm 0.03$ , showing that there is a clear dependence between  $a(t)/c(t)$  and  $d(t)/a(t)$  in this region that contains both the *ADAT stretches* (red boxes) for the highest values and non-stretched regions (blue boxes).



**Figure 4.** Graph of boxplots for concentration of *A codons*,  $a(t)/c(t)$ , versus the relative concentration of *D codons*,  $d(t)/a(t)$ . Blue boxes correspond to current windows, whereas the red ones correspond to the *ADAT stretches*. Multiple linear regression based on the mean values found two breakpoints at  $a(t)/c(t)$  0.296 and 0.635 (dashed red lines) (See Figure S5 for more details).

## 4. Conclusions

The biological significance of inosine at position 34 of anticodons is generally linked to the pairing ability of this nucleotide, which allows harboring anticodons to recognize codons with C, U and A (Figure 5b) [4]. This is in contrast to adenosine, which is supposed to favor a Watson–Crick interaction

with uridine, although several reports demonstrate that adenosine is capable of pairing with any base in the third position of the codon (Figure 5b) [4,25]. Thus, it remains unclear whether inosine is widely used in eukaryotes to restrict or expand the pairing capacity of A34 containing codons [4,26–28].

It is clear that inosine is important to balance codon usage and tRNA gene copy number in eukaryotes, and that highly translated genes in these species tend to be enriched in ADAT-dependent codons [3]. In the light of the growing realization of the regulatory role that modification enzymes, and fluctuations in tRNA populations, play in the regulation of specific gene programs, it is important to determine if inosine is also involved in the regulation of gene expression levels.

To start addressing this question we have begun to characterize how ADAT-dependent codons are distributed in eukaryotic proteomes. This initial analysis is required to try to identify sections of the transcriptome potentially more dependent on the ADAT activity levels. Our approach has been, first, to screen the complete human transcriptome and classify its genes on the basis of the abundance of *ADAT-aa*. Using this initial curation, we have identified those transcripts whose proportion of codons for such *ADAT-aa* is significantly enriched, and we have used this subset of sequences to determine the relative enrichment of each *ADAT-aa*, and the variation in D and G-ended codons in this group of sequences with respect to the whole transcriptome.

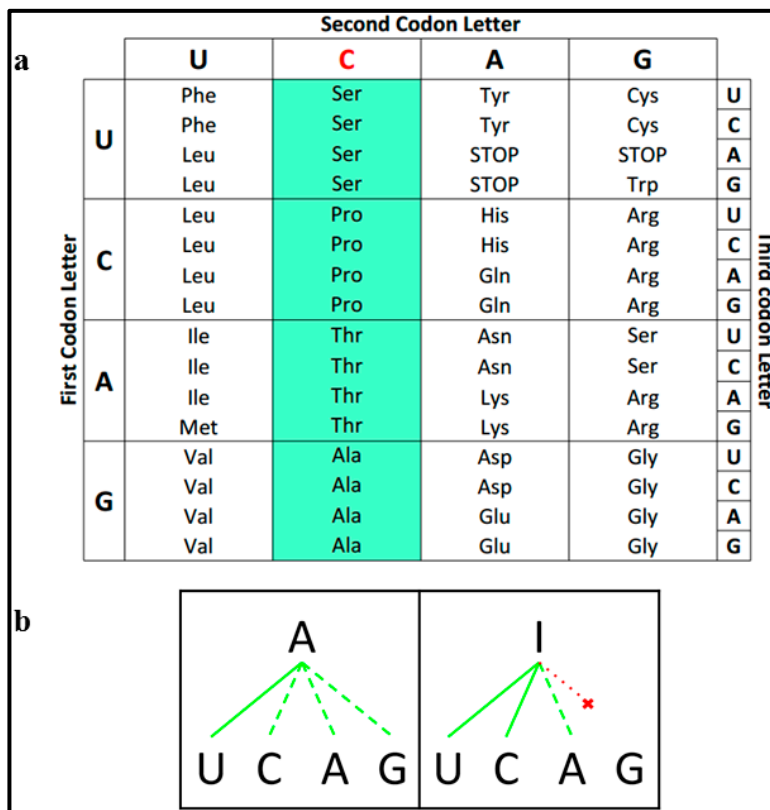
Our results show, first, that the human transcriptome is biased towards proteins enriched in *ADAT-aa*. Interestingly, the majority of these proteins are specifically enriched in *TAPS*, but not in *LIVR*. Physicochemical parameters specific to these residues may explain why *TAPS* can reach higher relative frequencies than *LIVR* in human proteins. At the same time, functional features of proteins rich in *TAPS* must have driven the selection of these extremely biased protein sequences.

Our data also shows that the more enriched in *ADAT-aa* a region is, the higher its tendency to use D codons instead of G-ended codons. Therefore, transcripts coding for stretches of *ADAT-aa* are composed preferentially by D codons, and this composition increases with the length and quality of the stretch. This observation confirms the notion that ADAT-modified tRNAs are preferred in eukaryotic translation, and suggests that the selective force behind this selection is increased when the frequency of *ADAT-aa* rises.

The reason why stretches of TAPS amino acids are enriched in ADAT-dependent codons remains to be determined. In this regards, it is interesting to notice that the four amino acids that are enriched in the stretches (threonine, alanine, proline, and serine) are all coded by four-box codon sets that share the same second base (C) (serine is also coded by two additional codons) (Figure 5a). Thus, the selectivity between these four codon sets depends only upon recognition of the first codon base. Under these circumstances, it is possible that the proposed higher selectivity of I over A may be preferred to minimize the possibility of decoding errors, particularly in highly repetitive transcript regions such as the stretches identified in our analysis.

An important corollary of our analysis is the potential role of anticodon modifications in allowing ribosomal protein synthesis machineries to access new protein sequence spaces. Several examples of codon and amino acid compositions are known that impair ribosomal functional and, in some cases, require additional cofactors to allow the ribosome to progress through these regions [29–32]. It is similarly conceivable that certain highly repetitive transcript sequences may be inaccessible to ribosome processing unless new functional improvements that increase efficiency or selectivity can be found. The selection of modified bases, such as inosine, that possibly allow species to synthesize proteins previously

unavailable may be a major driving force in speciation. A detailed evolutionary analysis of ADAT function in the eukaryotic lineage will contribute to test this hypothesis.



**Figure 5.** (a) Codon usage table. Highlighted in green are all the codons that translate for *TAPS* that share a common C in their second position (red); (b) Schematic representation of base pairing between Adenine (or Inosine) and the rest of the classical bases. Continuous green lines indicate preferred pairings and dashed green lines indicate poor pairings. Dashed red line indicates no pairing.

### Supplementary Materials

Supplementary materials can be found at <http://www.mdpi.com/1422-0067/16/08/17303/s1>.

### Acknowledgments

We thank the members of the Ribas Lab as well as Bioinformatics-Biostatistics unit of IRB for helpful discussions. We would thank Salva Guardiola for help in the design of Figure 1. Àlbert Rafels-Ybern was supported by an FPI Spanish grant 2012. This work was supported by grant BIO2012-32200 from the Spanish Ministry of Economy and Innovation to Lluís Ribas de Pouplana.

### Author Contributions

Àlbert Rafels-Ybern performed the experiments and wrote the manuscript. Camille Stephan-Otto Attolini supervised the mathematical and statistical analysis of the data. Lluís Ribas de Pouplana directed the research and wrote the manuscript.

## Conflicts of Interest

The authors declare no conflict of interest.

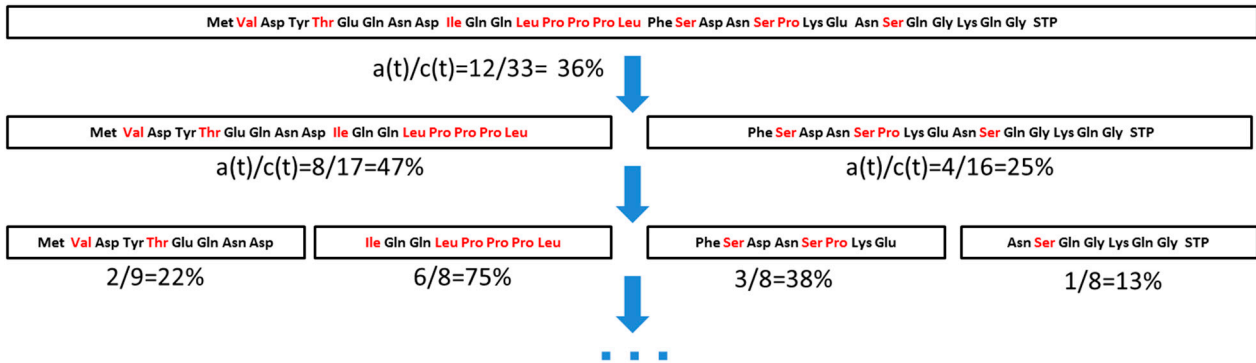
## References

1. Crick, F.H. On protein synthesis. *Symp. Soc. Exp. Biol.* **1958**, *12*, 138–163.
2. Demeshkina, N.; Jenner, L.; Westhof, E.; Yusupov, M.; Yusupova, G. New structural insights into the decoding mechanism: Translation infidelity via a G.U pair with Watson-Crick geometry. *FEBS Lett.* **2013**, *587*, 1848–1857.
3. Novoa, E.M.; Pavon-Eternod, M.; Pan, T.; Ribas de Pouplana, L. A role for tRNA modifications in genome structure and codon usage. *Cell* **2012**, *149*, 202–213.
4. Auxilien, S.; Crain, P.F.; Trewyn, R.W.; Grosjean, H. Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of inosine 34 in the anticodon of transfer RNA. *J. Mol. Biol.* **1996**, *262*, 437–458.
5. Gerber, A.P.; Keller, W. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science* **1999**, *286*, 1146–1149.
6. Marck, C.; Grosjean, H. tRNomics: Analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA* **2002**, *8*, 1189–1232.
7. Crick, F.H. Codon—Anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* **1966**, *19*, 548–555.
8. Rubio, M.A.; Pastar, I.; Gaston, K.W.; Ragone, F.L.; Janzen, C.J.; Cross, G.A.; Papavasiliou, F.N.; Alfonzo, J.D. An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7821–7826.
9. Spears, J.L.; Rubio, M.A.; Gaston, K.W.; Wywiał, E.; Strikoudis, A.; Bujnicki, J.M.; Papavasiliou, F.N.; Alfonzo, J.D. A single zinc ion is sufficient for an active *Trypanosoma brucei* tRNA editing deaminase. *J. Biol. Chem.* **2011**, *286*, 20366–20374.
10. Zhou, W.; Karcher, D.; Bock, R. Identification of enzymes for adenosine-to-inosine editing and discovery of cytidine-to-uridine editing in nucleus-encoded transfer RNAs of *Arabidopsis*. *Plant Physiol.* **2014**, *166*, 1985–1997.
11. Torres, A.; Piñeyro, D.; Rodríguez-Escribà, M.; Camacho, N.; Reina, O.; Saint-Leger, A.; Filonava, L.; Batlle, E.; de Pouplana, L.R. Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Res.* **2015**, *43*, doi:10.1093/nar/gkv277.
12. Novoa, E.M.; Ribas de Pouplana, L. Speeding with control: Codon usage, tRNAs and ribosomes. *Trends Genet.* **2012**, *28*, 574–581.
13. Yona, A.H.; Bloom-Ackermann, Z.; Frumkin, I.; Hanson-Smith, V.; Charpak-Amikam, Y.; Feng, Q.; Boeke, J.D.; Dahan, O.; Pilpel, Y. tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* **2013**, *2*, e01339, doi:10.7554/eLife.01339.
14. Bauer, F.; Hermand, D. A coordinated codon-dependent regulation of translation by Elongator. *Cell Cycle* **2012**, *11*, 4524–4529.
15. Bauer, F.; Matsuyama, A.; Candiracci, J.; Dieu, M.; Scheliga, J.; Wolf, D.A.; Yoshida, M.; Hermand, D. Translational control of cell division by Elongator. *Cell Rep.* **2012**, *1*, 424–433.

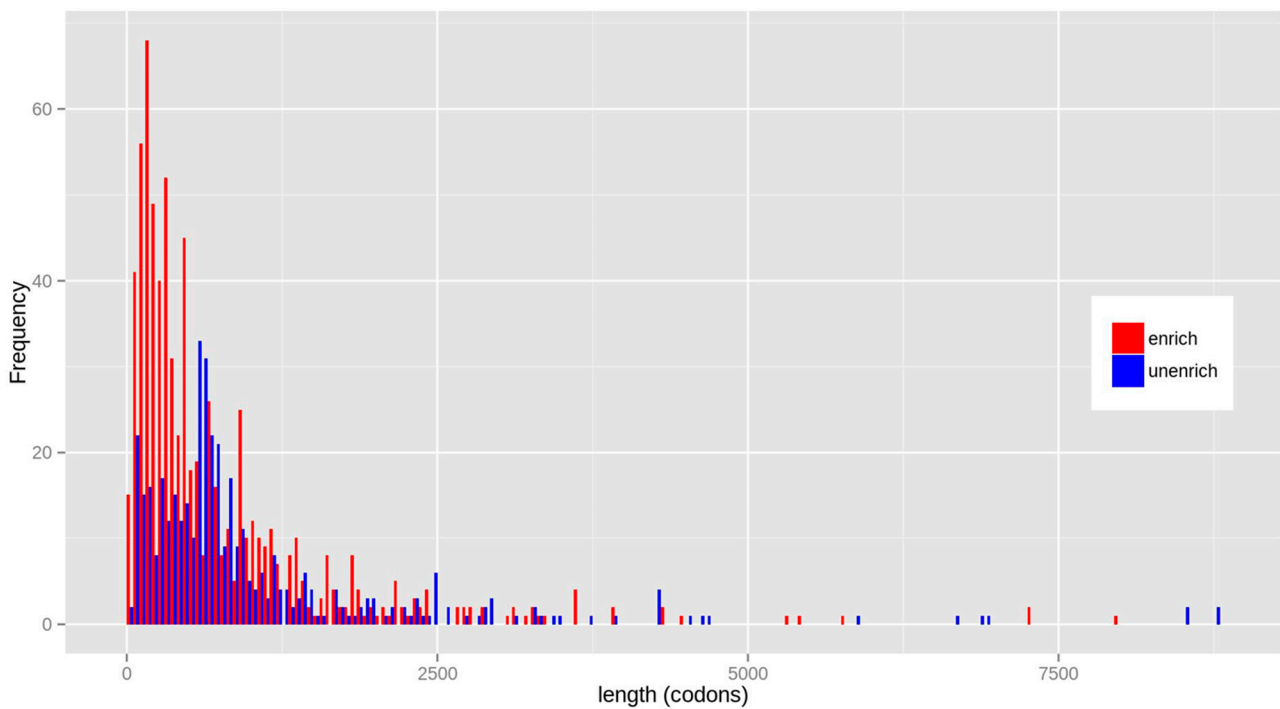


16. Fernandez-Vazquez, J.; Vargas-Perez, I.; Sanso, M.; Buhne, K.; Carmona, M.; Paulo, E.; Hermand, D.; Rodríguez-Gabriel, M.; Ayté, J.; Leidel, S.; *et al.* Modification of tRNA(Lys) UUU by elongator is essential for efficient translation of stress mRNAs. *PLoS Genet.* **2013**, *9*, e1003647, doi:10.1371/journal.pgen.1003647.
17. Huang, B.; Johansson, M.J.; Bystrom, A.S. An early step in wobble uridine tRNA modification requires the Elongator complex. *RNA* **2005**, *11*, 424–436.
18. CCDS Database. Available online: <http://ncbi.nlm.nih.gov/CCDS/> (accessed on 13 July 2015).
19. Muggeo, V.M.R. Segmented: An R package to fit regression models with broken-line relationships. *R News* **2008**, *8*, 20–25.
20. Librado, P.; Rozas, J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **2009**, *25*, 1451–1452.
21. McDonald, J.H. Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol. Biol. Evol.* **1996**, *13*, 253–260.
22. Hutter, S.; Vilella, A.J.; Rozas, J. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinform.* **2006**, *7*, 409, doi:10.1186/1471-2105-7-409.
23. Sander, C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **1991**, *9*, 56–68.
24. Genomic tRNA Database. Available online: <http://gtrnadb.ucsc.edu> (accessed on 13 July 2015).
25. Alkatib, S.; Scharff, L.B.; Rogalski, M.; Fleischmann, T.T.; Matthes, A.; Seeger, S.; Schöttler, M.A.; Ruf, S.; Bock, R. The contributions of wobbling and superwobbling to the reading of the genetic code. *PLoS Genet.* **2012**, *8*, e1003076, doi:10.1371/journal.pgen.1003076.
26. Haumont, E.; Fournier, M.; de Henau, S.; Grosjean, H. Enzymatic conversion of adenosine to inosine in the wobble position of yeast tRNA<sup>Asp</sup>: The dependence on the anticodon sequence. *Nucleic Acids Res.* **1984**, *12*, 2705–2715.
27. Leonard, G.A.; Booth, E.D.; Hunter, W.N.; Brown, T. The conformational variability of an adenosine inosine base-pair in a synthetic DNA dodecamer. *Nucleic Acids Res.* **1992**, *20*, 4753–4759.
28. Torres, A.G.; Pineyro, D.; Filonava, L.; Stracker, T.H.; Batlle, E.; Ribas de Pouplana, L. A-to-I editing on tRNAs: Biochemical, biological and evolutionary implications. *FEBS Lett.* **2014**, *588*, 4279–4286.
29. Pavlov, M.Y.; Watts, R.E.; Tan, Z.; Cornish, V.W.; Ehrenberg, M.; Forster, A.C. Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 50–54.
30. Guimaraes, J.C.; Rocha, M.; Arkin, A.P. Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. *Nucleic Acids Res.* **2014**, *42*, 4791–4719.
31. Ude, S.; Lassak, J.; Starosta, A.L.; Kraxenberger, T.; Wilson, D.N.; Jung, K. Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. *Science* **2013**, *339*, 82–85.
32. Doerfel, L.K.; Wohlgemuth, I.; Kothe, C.; Peske, F.; Urlaub, H.; Rodnina, M.V. EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. *Science* **2013**, *339*, 85–88.

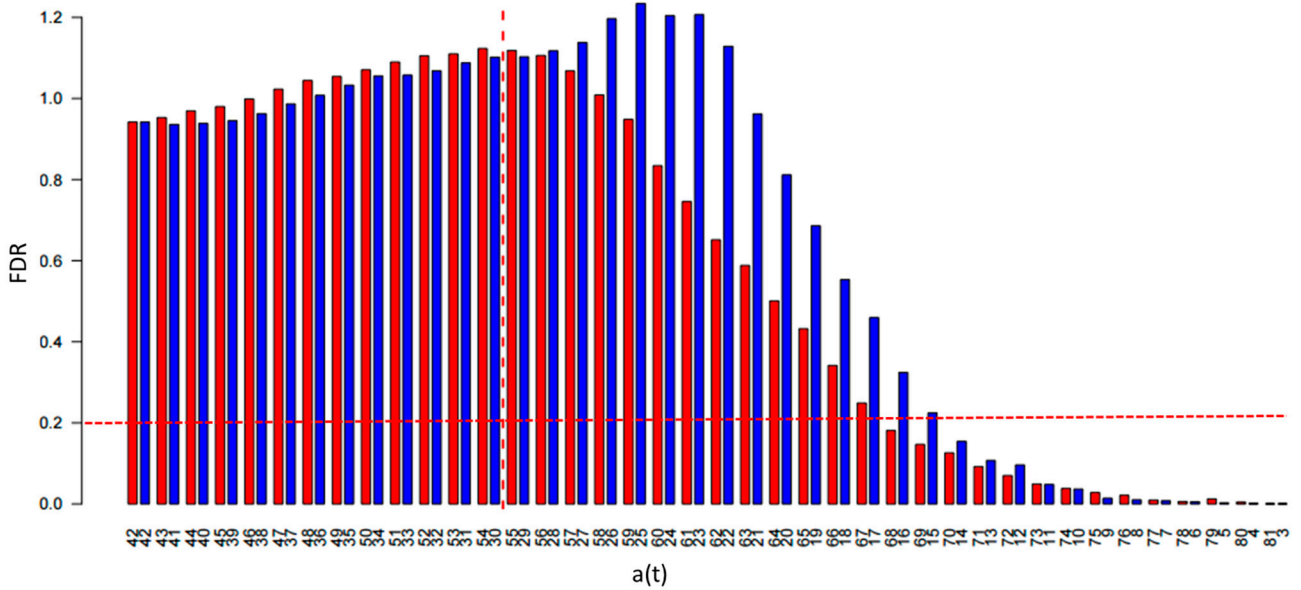
# Supplementary Information



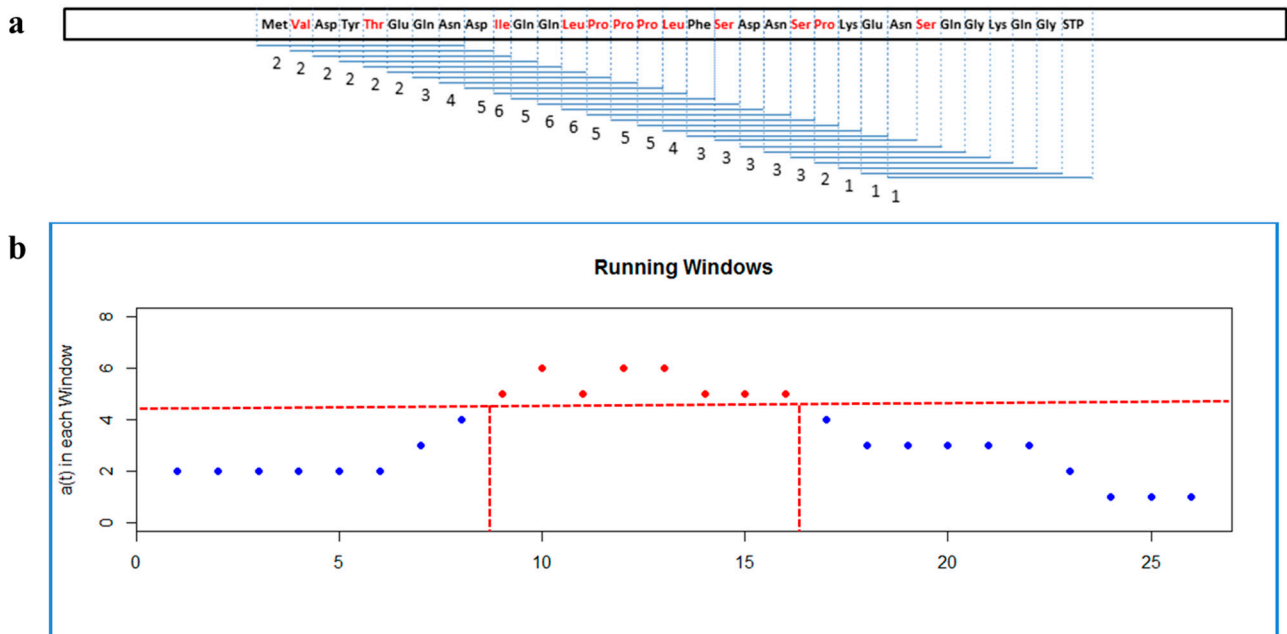
**Figure S1.** *Halves-gene* method layout. Each sequence is divided into halves recursively (blue arrows) until the regions are small enough. For each region  $a(t)/c(t)$  is calculated. Red amino acids correspond to *A* codons.



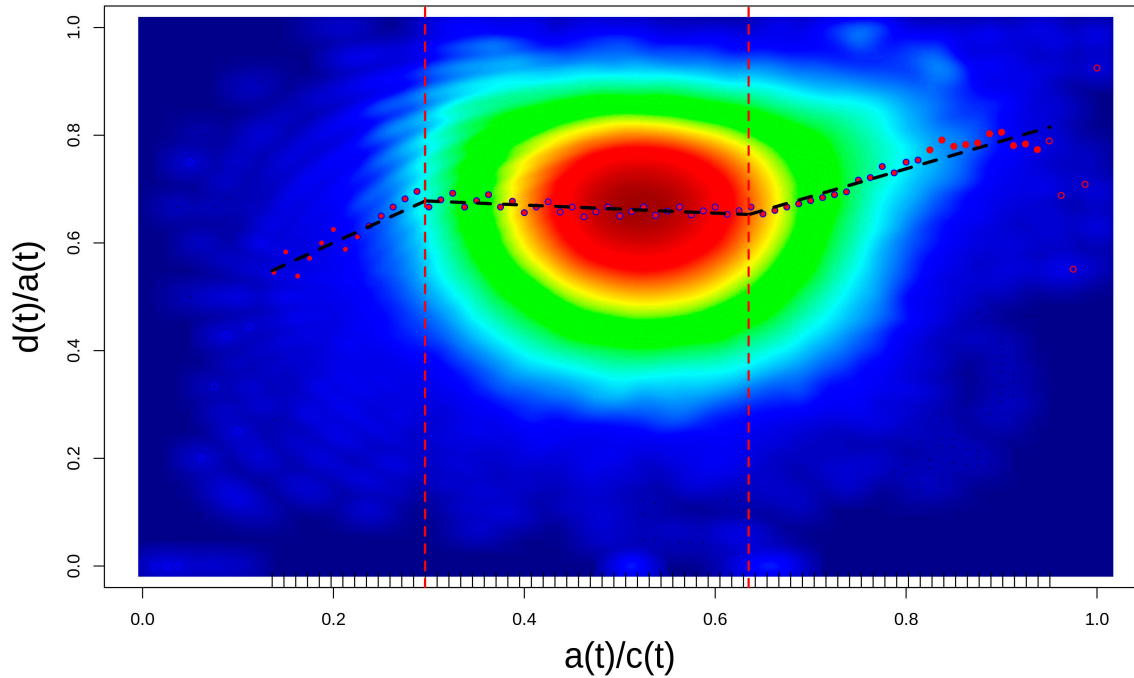
**Figure S2.** Histogram comparing the enriched (red) and the unenriched (blue) outliers from Figure 2a.



**Figure S3.** Histogram comparing the False Discovery Rate (FDR) symmetry of Figure 2b. Tails of 5% (from the horizontal dashed red line to the right) for enriched windows (red bars) and unenriched windows (blue bars). *ADAT stretches* are considered for those windows with  $FDR < 0.2$  (vertical dashed red line), or equivalently  $a(t) > 67$ .



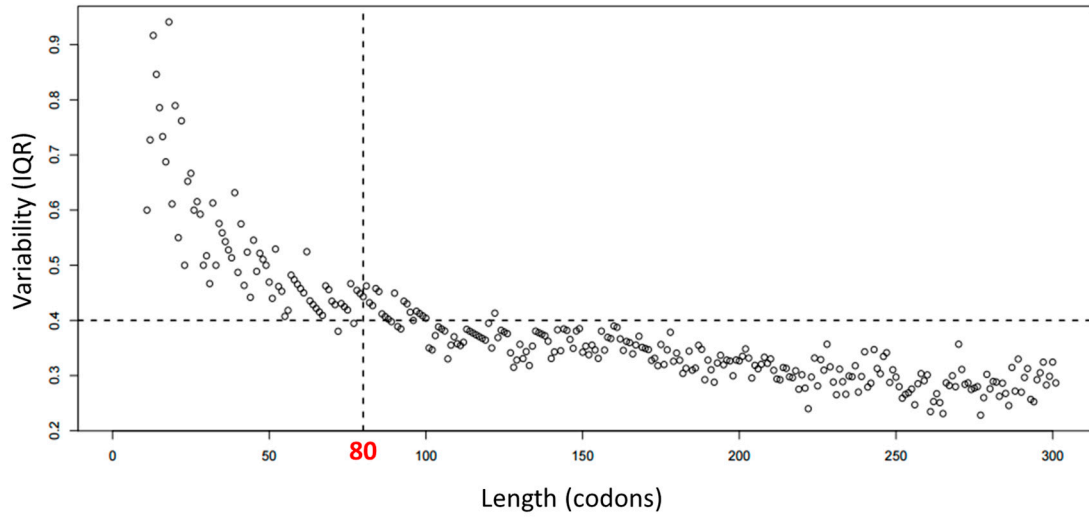
**Figure S4.** *Running-windows* layout (a) Scheme of action for this method. For each sequence the window slides codon by codon from the beginning to the end (dashed blue lines) and  $a(t)$  is calculated; (b) Distribution of  $a(t)$  for each window, the enriched windows correspond to the highest  $a(t)$  values. In this example, there is one single stretch (between horizontal dashed red lines) that consists of eight consecutive enriched windows (red points). Red amino acids correspond to *A codons*.



**Figure S5.** Density plot between  $a(t)/c(t)$  and  $d(t)/a(t)$ . The dots correspond to the mean values in Figure 4 and the dashed black lines correspond to a multiple linear regression based on these values. There are two breakpoints at  $a(t)/c(t)$  equal to 0.296 and 0.635 (dashed red lines).

Ala	29	0	9	5	Ser	11	0	4	5	0	8
	3'CGI5' 5'GCT3'	3'CGG5' 5'GCC3'	3'CGT5' 5'GCA3'	3'CGC5' 5'GCG3'		3'AGI5' 5'TCT3'	3'AGG5' 5'TCC3'	3'AGT5' 5'TCA3'	3'AGC5' 5'TCG3'	3'TCA5' 5'AGT3'	3'TCG5' 5'AGC3'
	1,84	2,77	1,58	0,74		1,52	1,77	0,44	1,22	1,21	1,95
Pro	10	0	4	7	Arg	7	0	4	6	5	6
	3'GGI5' 5'CCT3'	3'GGG5' 5'CCC3'	3'GGT5' 5'CCA3'	3'GGC5' 5'CCG3'		3'GCI5' 5'CGT3'	3'GCG5' 5'CGC3'	3'GCT5' 5'CGA3'	3'GCC5' 5'CGG3'	3'TCC5' 5'AGG3'	3'TCT5' 5'AGA3'
	1,75	1,98	0,69	1,69		0,45	1,04	1,14	0,62	1,2	1,22
Thr	10	0	6	6	Leu	7	0	4	6	5	6
	3'TGI5' 5'ACT3'	3'TGG5' 5'ACC3'	3'TGT5' 5'ACA3'	3'TGC5' 5'ACG3'		3'GAI5' 5'CTT3'	3'GAG5' 5'CTC3'	3'GAT5' 5'CTA3'	3'GAC5' 5'CTG3'	3'AAC5' 5'TTG3'	3'AAT5' 5'TTA3'
	1,31	1,89	0,61	1,51		1,32	1,96	3,96	0,72	1,29	0,77
Val	11	0	16	5	Ala	14	3	5	tRNA copy number 3'anticodon5' 5'Codon3' codon usage		
	3'CAI5' 5'GTT3'	3'CAG5' 5'GTC3'	3'CAT5' 5'GTA3'	3'CAC5' 5'GTG3'		3'TAI5' 5'ATT3'	3'ATG5' 5'ATC3'	3'CGT5' 5'ATA3'			
	1,1	1,45	2,81	0,71		1,6	2,08	0,75			

**Figure S6.** Codon-anticodon relationships for *ADAT-aa*. tRNA copy number and codon usage is shown for each pair, as indicated in bottom-right legend. Inosine 34 and those nucleotides recognized throughout “wobble” pairing are depicted in pink. Green lines shows which codon-anticodon pairings takes place. Anticodons depicted in red do not exist in human genome.



**Figure S7.** Variability corresponding to Figure 2a calculated by Interquartile Range (IQR).

**Table S1.** Data corresponding to Figure 3a–c. *aa*: amino acids, *cod*: codons, *CU*: codon usage, *Str*: ADAT stretches dataset, *All*: Human transcriptome dataset, *fold*: rate Str/All. Each percentage is measured with respect to the total of codons, thus Str and All columns add up to 100%.

aa	cod	CU		fold	aa	cod	CU		fold	aa	cod	CU		fold
		Str	All				Str	All				Str	All	
Thr	aca	3.94	1.51	2.61	Leu	cta	0.4	0.72	0.56	Lys	aaa	0.55	2.44	0.23
	acc	6.36	1.89	3.37		ctc	1.17	1.96	0.60		aag	0.79	3.19	0.25
	acg	1.35	0.61	2.21		ctg	2.14	3.96	0.54	Asn	aac	0.57	1.91	0.30
	act	4.09	1.31	3.12		ctt	1.45	1.32	1.10		aat	0.45	1.7	0.26
Ala	gca	2.4	1.58	1.52	tta	0.5	0.77	0.65	Met	atg	0.94	2.2	0.43	
	gcc	4.99	2.77	1.80	ttg	0.79	1.29	0.61	Gln	caa	0.35	1.23	0.28	
	gcg	1.02	0.74	1.38	Val	gta	0.57	0.71		0.80	cag	1.43	3.42	0.42
	gct	3.32	1.84	1.80		gtc	1.89	1.45	1.30	His	cac	0.77	1.51	0.51
Pro	cca	5.45	1.69	3.22	gtg	2.48	2.81	0.88	cat		0.28	1.09	0.26	
	ccc	5.37	1.98	2.71	gtt	1.15	1.1	1.05	Glu	gaa	0.92	2.9	0.32	
	ccg	2.23	0.69	3.23	Arg	cga	0.48	0.62		0.77	gag	1.25	3.96	0.32
	cct	6.54	1.75	3.74		cgc	0.95	1.04	0.91	Asp	gac	0.83	2.51	0.33
Ser	tca	2.88	1.22	2.36	cgg	1.01	1.14	0.89	gat		0.33	2.18	0.15	
	tcc	4.75	1.77	2.68	cgt	0.39	0.45	0.87	Gly	gga	0.71	1.65	0.43	
	tcg	0.97	0.44	2.20	aga	1.1	1.22	0.90		ggc	1.16	2.22	0.52	
	tct	4.03	1.52	2.65	agg	1.06	1.2	0.88		ggg	0.96	1.65	0.58	
agc	3.46	1.95	1.77	Ile	ata	0.44	0.75	0.59		ggt	0.89	1.08	0.82	
agt	1.97	1.21	1.63		atc	1.24	2.08	0.60	Tyr	tac	0.35	1.53	0.23	
				att	0.67	1.6	0.42	tat		0.15	1.22	0.12		
								Cys	tgc	0.17	1.26	0.13		
									tgt	0.1	1.06	0.09		
								Trp	tgg	0.24	1.32	0.18		
								Phe	ttc	0.43	2.03	0.21		
									ttt	0.38	1.76	0.22		
								STOP	taa	0	0.1	0.00		
									tag	0	0.08	0.00		
									tga	0	0.16	0.00		

**Table S2.** Data corresponding to Figure 3d. Same notation as in Table S1. *4box* means that for the amino acids Ser, Leu and Arg, only the 4box XXN codons were taken into account.

	CU	
	Str	All
<b>TAPS (4box)</b>	59.69	23.31
<b>LIVR (4box)</b>	16.43	21.71

**Table S3.** Data corresponding to Figure 3e. Same notation as in Table S2. *G*: G-ended codon for the corresponding amino acid. %*G*: ratio *G*/*4box*.

aa	cod	CU		aa	cod	CU	
		Str	All			Str	All
Thr	4box	15.74	5.32	Leu	4box	5.16	7.96
	G	1.35	0.61		G	2.14	3.96
	%G	8.58	11.47		%G	41.47	49.75
Ala	4box	11.73	6.93	Val	4box	6.09	6.07
	G	1.02	0.74		G	2.48	2.81
	%G	8.70	10.68		%G	40.72	46.29
Pro	4box	19.59	6.11	Arg	4box	2.83	3.25
	G	2.23	0.69		G	1.01	1.14
	%G	11.38	11.29		%G	35.69	35.08
Ser	4box	12.63	4.95	Ile	4box	2.35	4.43
	G	0.97	0.44		G	-	-
	%G	7.68	8.89		%G	-	-

### 6.3 Publication 2

**Rafels-Ybern, A.,** A. G. Torres, X. Grau-Bove, I. Ruiz-Trillo and L. Ribas de Pouplana (2017). “Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla.” *RNA Biol*: 1-8.

Impact factor (2016): 3.9





RESEARCH PAPER



# Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla

Àlbert Rafels-Ybern<sup>a</sup>, Adrian Gabriel Torres<sup>a</sup>, Xavier Grau-Bove<sup>b,c</sup>, Iñaki Ruiz-Trillo<sup>b,c,d</sup>, and Lluís Ribas de Pouplana<sup>a,d</sup>

<sup>a</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, Barcelona, Catalonia, Spain; <sup>b</sup>Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Catalonia, Spain; <sup>c</sup>Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Catalonia, Spain; <sup>d</sup>ICREA, Pg. Lluís Companys 23, Barcelona, Catalonia, Spain

## ABSTRACT

The modification of adenosine to inosine at position 34 of tRNA anticodons has a profound impact upon codon-anticodon recognition. In bacteria, I34 is thought to exist only in tRNA<sup>Arg</sup>, while in eukaryotes the modification is present in eight different tRNAs. In eukaryotes, the widespread use of I34 strongly influenced the evolution of genomes in terms of tRNA gene abundance and codon usage. In humans, codon usage indicates that I34 modified tRNAs are preferred for the translation of highly repetitive coding sequences, suggesting that I34 is an important modification for the synthesis of proteins of highly skewed amino acid composition. Here we extend the analysis of distribution of codons that are recognized by I34 containing tRNAs to all phyla known to use this modification. We find that the preference for codons recognized by such tRNAs in genes with highly biased codon compositions is universal among eukaryotes, and we report that, unexpectedly, some bacterial phyla show a similar preference. We demonstrate that the genomes of these bacterial species contain previously undescribed tRNA genes that are potential substrates for deamination at position 34.

## ARTICLE HISTORY

Received 6 June 2017  
Revised 12 July 2017  
Accepted 17 July 2017

## KEYWORDS

Translation; Evolution;  
Speciation; tRNA;  
Transcriptome; mRNA; ADAT;  
TadA; CDS

## Introduction


Transfer RNAs (tRNA) are the universal adaptors of the genetic code,<sup>1</sup> linking codons to their cognate amino acids during protein synthesis. While the structure of the genetic code is mostly conserved across all domains of life, the genomic composition of tRNA genes varies widely between species.<sup>2</sup> Some archaeal species have only 20–25 different tRNAs, while some Eukaryotes can present up to 40–45 different tRNAs in their genomes.<sup>3,4</sup> Importantly, the number of different tRNAs is always lower than the number of codons used and, therefore, some tRNAs need to recognize more than one codon. This is achieved by a higher pairing permissiveness between the third position of the mRNA codons and the first position of the tRNA anticodon in what is known as “wobble” or degenerate pairing.<sup>5</sup> Due to codon degeneracy those amino acids coded by at least four codons (threonine, alanine, proline, serine, glycine, leucine, valine and arginine) can, in principle, be specifically incorporated with just two specific pairings (Fig. 1a).<sup>6</sup>


Posttranscriptional tRNA modifications are essential for tRNA function. While certain tRNA modifications are present in all living organisms, others are kingdom specific.<sup>4</sup> In general, modifications in the acceptor stem of tRNAs are important for amino acid charging, modifications in the main body of the tRNA can affect tRNA structure and stability, and modifications in the anticodon loop of tRNAs modulate codon recognition.<sup>7–9</sup> Position 34 in the anticodon pairs with the third base of codons,

and harbors the widest variety of known modifications.<sup>10,11</sup> Modifications at position 34 fine-tune wobble pairing and extend or restrict the number of codons that anticodons recognize. We have shown that an improved correlation between codon usage and tRNA gene copy number in Bacteria and Eukaryotes can be observed when the effects that modified bases at position 34 have upon codon recognition are taken into account.<sup>6</sup>

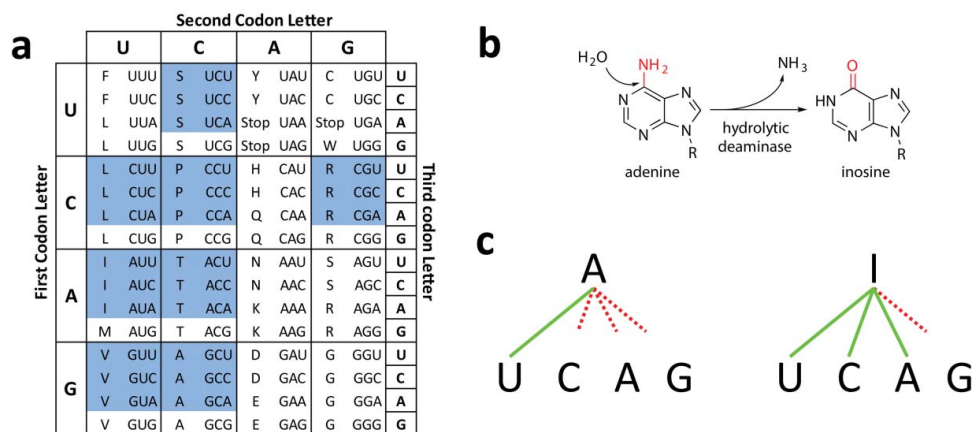
In bacterial genomes, tRNA genes complementary to codons for four-box and six-box amino acids predominantly have guanine at position 34 (G34) over adenosine at this position (A34), with the exception of genes for tRNA<sup>Arg</sup>. In eukaryotic genomes, on the other hand, tRNA genes for the same codons almost exclusively use A34.<sup>6</sup> These differences in decoding strategies between Bacteria and Eukarya are explained by the presence or absence of a modified base (inosine) at position 34 (I34) in these tRNAs. I34 is absent in archaeal tRNAs, occurs in bacterial tRNA<sup>Arg</sup> (ACG), and in eukaryotic tRNA<sup>Ala</sup> (AGC), tRNA<sup>Pro</sup> (AGG), tRNA<sup>Thr</sup> (AGT), tRNA<sup>Val</sup> (AAC), tRNA<sup>Ser</sup> (AGA), tRNA<sup>Arg</sup> (ACG), tRNA<sup>Leu</sup> (AAG) and tRNA<sup>Ile</sup> (AAT).<sup>12,13</sup>

The conversion by hydrolytic deamination of adenosine (A) to inosine (I) at position 34 of tRNAs is catalyzed by adenosine deaminases (Fig. 1b).<sup>12–18</sup> I34-modified tRNAs can wobble pair with codons ending in A3, C3 or U3, but not G3, expanding to 3 the number of codons that genetically encoded A34 tRNAs can recognize (Fig. 1c).<sup>5,13</sup> However, the selective pressure that drove

**CONTACT** Lluís Ribas de Pouplana ✉ [lluis.ribas@irbbarcelona.org](mailto:lluis.ribas@irbbarcelona.org)  Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Carrer Baldiri Reixac, 08028 Barcelona, Catalonia, Spain.

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2017 Taylor & Francis Group, LLC



**Figure 1.** Impact of I34 in translation. (a) Standard genetic code. Blue cells correspond to ADAT-sensitive codons. (b) Inosine modification by hydrolytic deamination. (c) Schematic representation of possible base pairings between Adenine and Inosine. Continuous green lines indicate preferred pairings and dashed red line indicates poor or no pairing.

the increase from one I34-containing tRNA in bacteria to eight modified tRNAs in eukaryotic translation is still unclear.<sup>13,19</sup> We have recently shown that tRNA<sup>Gly</sup> escaped this process because its anticodon structure is incompatible with the presence of an adenosine at position 34.<sup>20</sup>

In Bacteria, an essential homodimer called tRNA adenosine deaminase A (TadA) modifies tRNA<sup>Arg</sup> (ACG) to tRNA<sup>Arg</sup> (ICG).<sup>21</sup> tRNA<sup>Arg</sup> (ACG) is, to date, the only A34-containing tRNA known to be deaminated to I34 in Bacteria, and this explains why bacterial arginine codons are preferentially translated by tRNAs initially transcribed with A34. A bacterial *tadA* gene was transferred to eukaryotes (possibly during the mitochondrial endosymbiotic event) where, through duplication and divergence, evolved into the heterodimeric enzyme adenosine deaminase acting on tRNA (ADAT, formed by the subunits ADAT2 and ADAT3). This acquisition increased the substrate repertoire of ADAT to eight different tRNAs,<sup>13,22</sup> which are preferentially used by eukaryotes to translate the codons for threonine, alanine, proline, serine, leucine, isoleucine, valine and arginine (hereinafter “ADAT amino acids”).

We have shown that the impact of ADAT activity in the translation of the human genome is particularly relevant to the synthesis of proteins highly enriched in ADAT amino acids.<sup>23</sup> We found that human ORFs enriched in ADAT amino acids are also significantly enriched in codons that are translated by I34-containing tRNAs (hereinafter “ADAT-sensitive codons”) (23). Interestingly, although eight amino acids are translated by I34-containing tRNAs, only four amino acids (T, A, P, and S; hereinafter “TAPS”) are found to accumulate to high levels (over 84% of positions) in human proteins.<sup>23</sup>

Here we have extended our initial analysis of the human proteome to 64 eukaryotic and 980 bacterial species spanning the whole tree of life. Again we find that, for both Eukarya and Bacteria, proteins enriched in ADAT amino acids (>84% threshold) show an enrichment limited to TAPS. When all eukaryotic sequences are considered, a bias toward the use of ADAT-sensitive codons in genes coding for proteins enriched in ADAT amino acids is apparent. Similarly, a positive correlation exists between length of the stretches enriched in ADAT amino acids and enrichment in ADAT-sensitive codons. Bacterial sequences as a whole do not display such correlations.

We have continued this analysis organizing the organisms into different groups (Table S1), to check whether our findings within eukaryotes and bacteria are applicable to all the groups. Surprisingly, this additional analysis revealed that neither eukaryotes nor bacteria behave uniformly with regards to the frequency of ADAT-sensitive codons in their genomes. Instead, our data indicates that the evolution of I34 as an adaptation for the synthesis of TAPS-enriched proteins is different between bacterial and eukaryotic phyla. In support for this idea we present new evidence that bacterial species showing an unexpected enrichment in ADAT-sensitive codons contain in their genomes tRNA genes with A34 that are cognate for amino acids other than arginine and could possibly also be deaminated by TadA.

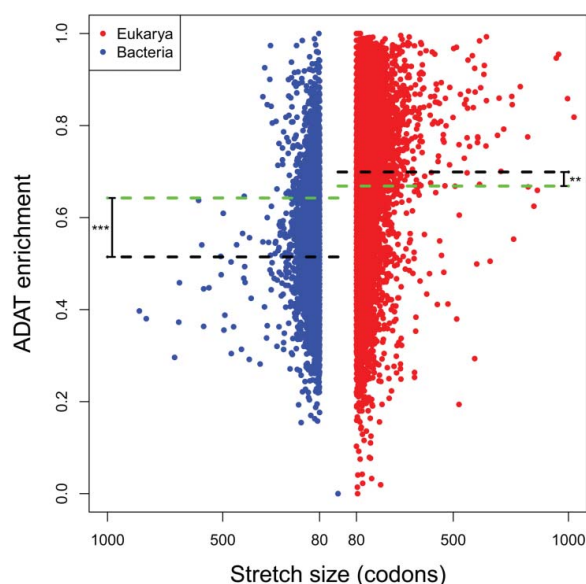
We postulate that this situation is an adaptation of these species that allows them to translate a larger number of transcripts enriched in ADAT-sensitive codons. We propose that I34 was selected in eukaryotes because of its contribution to the potential diversity of the eukaryotic proteome.<sup>24,25</sup> It is possible that some bacterial phyla have undergone a similar evolution of their translation apparatus.

## Results

### *Stretches of ADAT amino acids are more abundant, and more enriched in ADAT-sensitive codons, in eukarya than in bacteria*

Using previously described methods,<sup>23</sup> we have analyzed 1044 genomes in search of transcript coding sequences (CDSs) coding for protein stretches enriched in ADAT amino acids (see Materials and Methods). Figure 2 shows the number of such stretches obtained for our set of 64 eukaryotic and 980 bacterial genomes (Table 1). As previously reported for the human transcriptome<sup>23</sup>, the observed stretches both in Eukarya and Bacteria are mainly composed of TAPS amino acids (p-value = 6.3e-13) (Fig. 3), indicating that the enrichment in TAPS within stretches probably is a consequence of the physicochemical characteristics of these amino acids and not of a phylogenetic parameter.

We find that the total number of stretches rich in ADAT amino acids in Eukarya is higher (11769) than in Bacteria



**Figure 2.** Characterization of proteins enriched in ADAT amino acids in Bacteria (blue dots) and Eukarya (red dots). Each dot represents a stretch of amino acids of length >80, with a composition of ADAT amino acids higher than 83%. Black dashed lines denote the mean enrichment in ADAT sensitive codons in stretches of ADAT amino acids. The green dashed lines denote the mean enrichment in ADAT-sensitive codons for all CDSs analyzed. There are significant differences between black and green dashed lines (p-value = 0.002 for Eukarya, p-value < 2.2e-16 for Bacteria). Stretches higher than 1000 codons were not considered.

(7747) (Table 1) (a 4-fold increase when corrected for the total number of CDSs analyzed). Similarly, the mean stretch length is also significantly longer in eukaryotic proteomes (Table 1). The number of eukaryotic stretches is higher than the number of bacterial stretches at any length interval (Fig. 4), and this difference reaches a maximum of eight fold at the interval 227–236 codons (Fig. 4, black line). However, this differences in stretch length are likely not due to general differences in gene length between eukaryotes and bacteria, as no correlation was apparent between the length of ADAT stretches and the length of their corresponding CDSs for any species (data not shown).

When considering all coding sequences, eukaryotic genes display a significantly higher frequency of ADAT-sensitive codons with respect to bacterial genes (Table 1, p-value = 0.016). This difference is further increased when only genes coding for proteins enriched in ADAT amino acids are considered (Table 1, p-value < 2.2e-16). In fact, in Eukaryotes, a significant increase of ADAT-sensitive codons is seen in genes coding for proteins rich in ADAT amino acids (p-value < 2.2e-

**Table 1.** Statistics from the analysis in Figure 1.

	Bacteria	Eukarya	p-val
Num. organisms	980	64	n.a.
with stretches	531 (54%)	64 (100%)	n.a.
Num. Stretches	7047	11769	n.a.
Mean stretch size	107 ± 42	118 ± 58	< 2.2e-16
Mean stretch enrich.	0.546 ± 0.12	0.699 ± 0.15	< 2.2e-16
Mean CDSs enrich.	0.643 ± 0.08	0.668 ± 0.06	0.0159
Num. CDSs	3.39E+06	1.36E+06	n.a.
Stretch/CDS	2.08E-03	8.64E-03	n.a.

"enrich" = enrichment.

16), while the opposite is true for Bacteria (p-value = 0.002) (Fig. 2), suggesting a domain-specific bias in codon composition within ADAT stretches. Importantly, there is no correlation between the number of CDSs (or the GC content) and the number of stretches present in any given organism, either when comparing Eukaryotes versus Bacteria or when comparing individual phyla (see below), meaning that our observations are not biased by the size of the genomes, nor by the CG content of the genomes used in our analyses (Figure S2, S7 and S8).

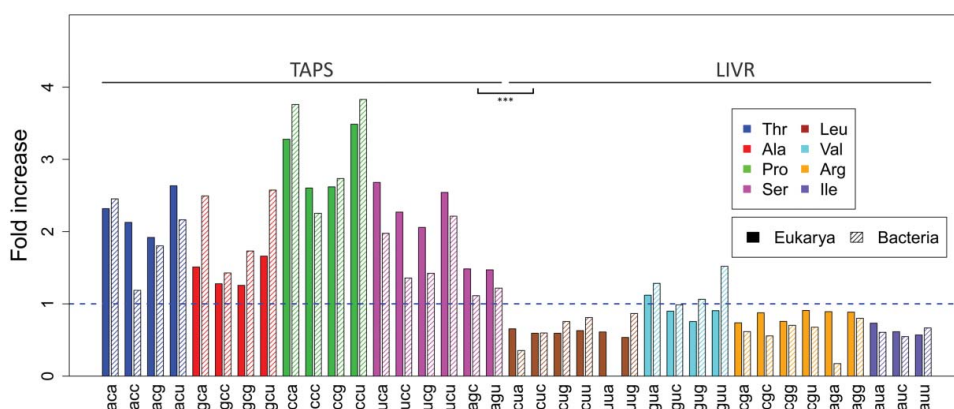
We next compared the median enrichment in ADAT-sensitive codons in stretches of different sizes in Bacteria and Eukarya (Fig. 5). In general, no correlation was apparent between the length of ADAT stretches and the length of their corresponding CDSs (data not shown). Enrichment in ADAT-sensitive codons in Eukarya was higher than in Bacteria for all the intervals. Additionally, we found a strong positive correlation of  $\Delta_{1\%} = 19$  (See Materials and Methods for  $\Delta_{1\%}$  definition) between the stretch size and enrichment in ADAT-sensitive codons in Eukarya, while in Bacteria this correlation was not present (Fig. 5).

### The enrichment in ADAT-sensitive codons in stretches behaves differently across eukaryotic and bacterial phyla

To further study ADAT stretches in Eukarya, we divided the data into the four eukaryotic groups: Metazoa, Fungi, Plantae and the rest of Eukarya (hereinafter "Protists") (Table S1). Figure 6 shows a histogram analysis of enrichment in ADAT-sensitive codons for these kingdoms according to stretch size. The highest overall enrichment in ADAT-sensitive codons within stretches is observed for Metazoa, followed by Fungi. Note that Metazoa have the longest stretches and is the only kingdom with ADAT stretches higher than 356 codons. Significant differences in enrichment in ADAT-sensitive codons can be seen for stretch sizes higher than 190 codons, where Metazoan sequences are clearly more ADAT enriched compared to Fungi (Figure S9). Both Metazoa and Plantae show a positive correlation between enrichment in ADAT-sensitive codons and stretch length ( $\Delta_{1\%} = 33$  and  $\Delta_{1\%} = 25$  respectively); while Fungi ( $R^2 = 0.27$ ) and Protists ( $R^2 = 0.01$ ) present no correlation (Fig. 6).

We performed a similar analysis by dividing the 980 Bacteria species into 20 phyla (Table S1). Figure 7 shows a histogram analysis of enrichment in ADAT-sensitive codons for these phyla according to their stretch size. Unexpectedly, cyanobacteria (cyan) and Firmicutes (yellow) show a high degree of enrichment in ADAT-sensitive codons throughout all the intervals, consistently higher than 0.65 (Fig. 7). Moreover, Firmicutes present a strong positive correlation of  $\Delta_{1\%} = 14$ , indicating that longer stretches are coded by sequences richer in ADAT-sensitive codons (Fig. 7). The rest of bacterial phyla (Fig. 7, black) have a depletion in ADAT-sensitive codons with most values below the mean when considering all the CDSs (Table 1,  $0.643 \pm 0.08$ ) for all intervals.

Firmicutes and Cyanobacteria are the two bacterial phyla with the highest overall enrichment in ADAT-sensitive codons compared with the rest of bacterial phyla (Fig. 8), and are comparable to Eukarya in terms of enrichment in ADAT-sensitive codons within stretches, with values not statistically different from those found for Protists (Fig. 8). Alphaproteobacteria (868 stretches),



**Figure 3.** Fold increase observed for each amino acid in the stretches identified in Fig. 2 for Eukarya (uniform color) and Bacteria (shading lines). Synonymous codons for ADAT amino acids are colored (legend).

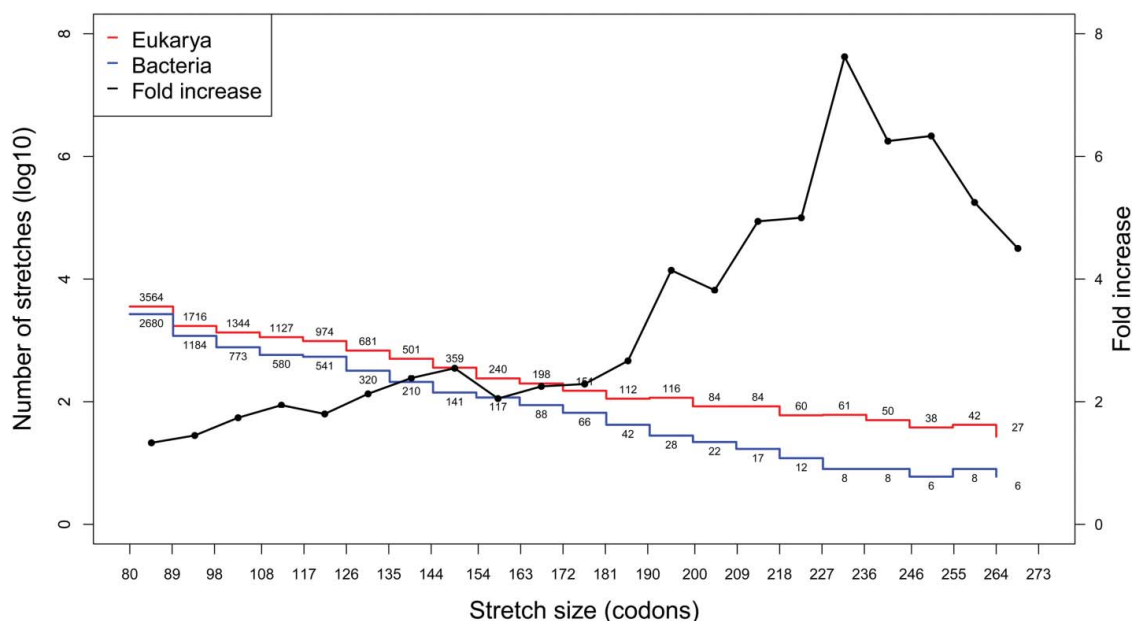
Actinobacteria (3360 stretches) and Betaproteobacteria (1115 stretches) have the lowest median enrichment in ADAT-sensitive codons despite being the bacterial phyla with most stretches, representing 57% of all the ADAT stretches (Fig. 8).

Although the discovery of proteins highly enriched in ADAT amino acids in bacteria was not unexpected given our initial analysis (Fig. 2), it was surprising to discover that the coding sequences for these stretches in Cyanobacteria and Firmicutes are also enriched in ADAT-sensitive codons and resemble Eukaryotes in this regard. Given that this observation suggests that I34 may also be playing a role in the translation of genes from these bacterial species, we searched their genomes for tRNA genes cognate for ADAT amino acids and with A34 containing anticodons. This analysis revealed that several species of Firmicutes contain this type of tRNA genes (Table S2). Possibly, these tRNAs are deaminated by TadA and influence the translation of ADAT-sensitive codons in these species.

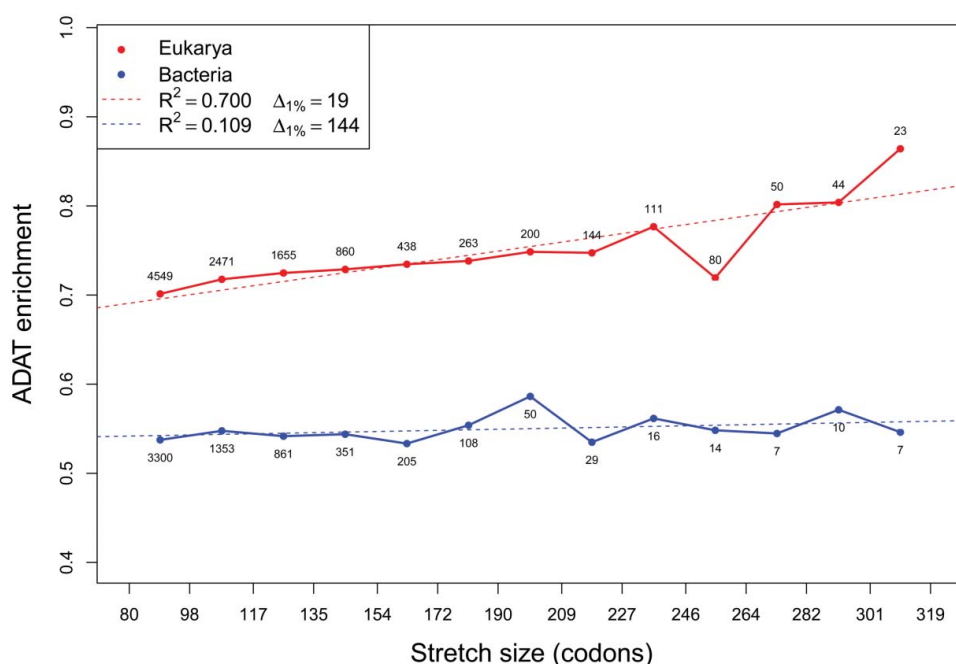
## Materials and methods

### Definitions

ADAT amino acids are defined as those amino acids that are charged to tRNAs that can be modified by ADAT in most Eukarya: Thr, Ala, Pro, Thr, Ser (TAPS), and Leu, Ile, Val, and Arg. ADAT codons are defined as the set of 37 codons that code for any of the ADAT amino acids (Fig. 1a). ADAT-sensitive codons are defined as the set of 24 codons that are recognized by I34 tRNAs (Fig. 1a, blue codons). ADAT stretches are defined as those regions in the CDSs that have an enrichment in ADAT amino acids and are found using the “Running Windows” Method.<sup>26</sup> For each ADAT stretch we define ADAT enrichment as the fraction ADAT-sensitive codons/ADAT codons. Fold increase is calculated by normalizing ADAT stretch codon composition to total CDSs codon usage, for each organism.  $\Delta_{1\%}$  represents the number of codons needed to increase the ADAT enrichment by 1% based on the slope of the linear model.



**Figure 4.** Histogram of number of ADAT stretches (in log<sub>10</sub>) for Eukarya (red) and Bacteria (blue) according to their length. ADAT stretches were calculated as in<sup>23</sup> (See Materials and Methods) and their lengths (in codons) were clustered in equidistant intervals. Black line shows the fold increase in number of stretches of Eukarya compared with Bacteria (right axis). Numbers close to red and blue lines corresponds to the number of samples for each interval. Intervals with less than five samples were omitted.

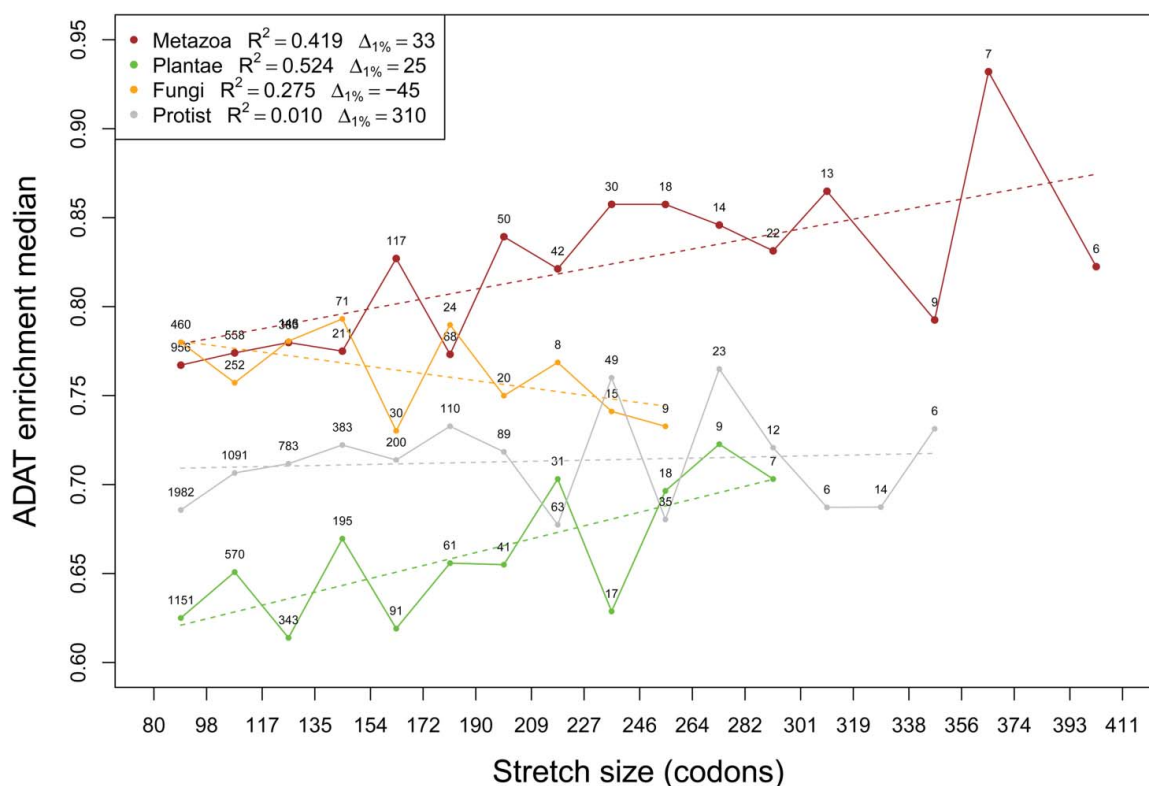


**Figure 5.** Median values of ADAT-sensitive codon frequency by length of ADAT amino acid stretches plotted for Eukarya (red) and bacteria (blue). Numbers close to the dots correspond to the number of samples for each interval. Linear regression was calculated for Eukarya (dashed red line and legend) and Bacteria (dashed blue line and legend). Intervals with less than five samples were omitted.

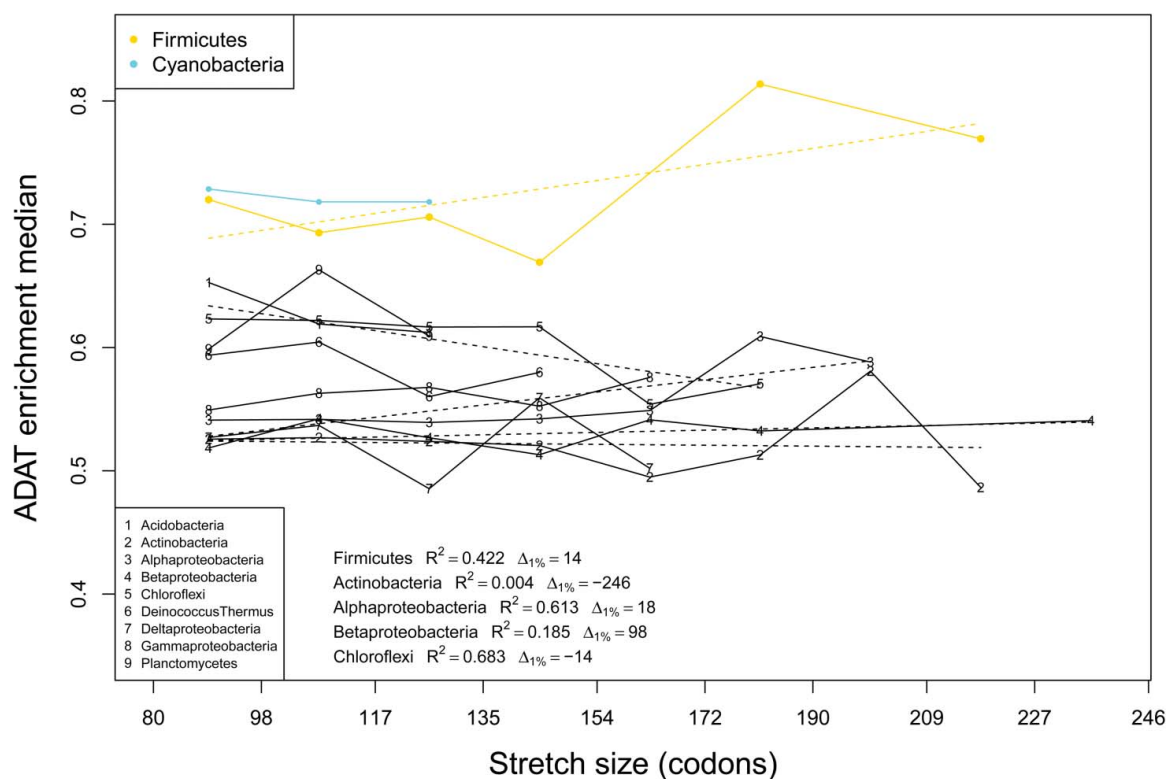
### Eukaryotic and bacterial transcriptome retrieval

We have analyzed the transcriptomes of 64 eukaryotic species and 980 bacterial species. Eukaryotic species were downloaded from Ensembl and bacterial sequences were

downloaded from NCBI. We analyzed a total of 1.13 million eukaryotic CDSs and 3.58 million bacterial CDSs. Only CDSs with a start codon, a stop codon and a number of nucleotides multiple of 3 were used for our analysis. 26% (27% in Bacteria and 19% in Eukarya) of the initial data



**Figure 6.** Median values of ADAT-sensitive codon enrichment by length of ADAT amino acid stretches plotted for different eukaryotic kingdoms (see legend). Linear regression was calculated for each kingdom (colored dashed lines and legend). Numbers close to the dots correspond to the number of samples for each interval. Intervals with less than five samples were omitted.



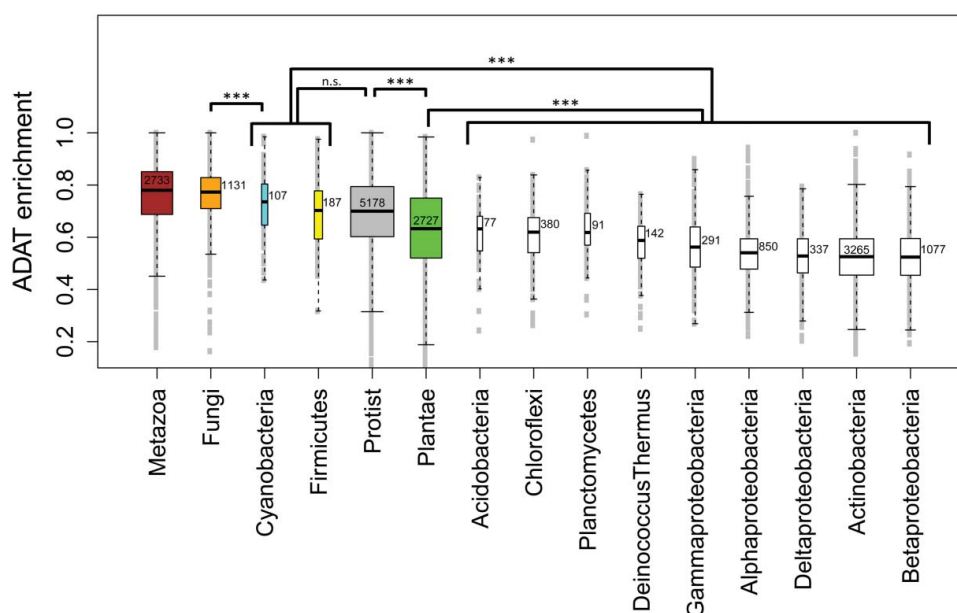
**Figure 7.** Median values of ADAT-sensitive codon enrichment by length of ADAT amino acid stretches plotted for different bacterial phyla (see legend). Phyla with less than 50 stretches were not considered. Linear regression was calculated for each phylum with more than 5 dots (dashed lines and legend). Intervals with less than five samples were omitted.

were discarded because it did not fulfill these three requisites.

#### Identification of stretches by the “running windows” method

To study in more detail the presence of stretches of ADAT codons, we applied the “Running Windows” method based on

software that applies similar methodologies.<sup>26–28</sup> For each CDS of the human transcriptome a window (fragment of the sequence with a fixed length in codons) slides codon by codon from the beginning to the end of the sequence. For each position, the percentage of ADAT codons is calculated and represented with respect to its location.<sup>23</sup> We fix the window size to 80 codons, and the threshold to be considered an enriched



**Figure 8.** Boxplot of the enrichment in ADAT-sensitive codons for four different eukaryotic kingdoms and eleven different bacterial phyla. Colored boxes correspond to eukaryotic kingdoms except two bacterial phyla, Cyanobacteria (cyan) and Firmicutes (yellow). Numbers close to the median correspond to the number of stretches for each interval. Phyla with less than 50 stretches were not considered.

window in ADAT codons was set at 83% based on our previous analysis of the human proteome<sup>23</sup> in order to have comparable datasets. We define an ADAT stretch as those regions corresponding to a window, or a set of consecutive windows, enriched in ADAT codons. Two (or more) windows are considered consecutive if the intersection between them in the cognate CDS is not void.

### Statistical analyses

In order to perform robust statistics we discarded intervals of stretch length represented by less than 5 ADAT stretches in either Bacteria or Eukarya in Figs. 3–6 (full data can be observed in Figures S3–S6 respectively). In Figs. 7 and 8 we only used for the analyses bacterial phyla with more than 50 stretches in total. In Fig. 7, linear regression was made for phyla with at least 5 points of approximation. Linear regression in Figs. 4–6 were fitted using the *lm()* function in R software. Significant differences in all the analysis were obtained using two-samples Wilcoxon test with *wilcox.test()* function in R software.<sup>29</sup> In Figure S1 we have considered as outliers those sequences longer than 1000 codons since they represent only 0.13% of the data (24 stretches out of 18840 stretches in total).

### Identification of tRNA genes

tRNA genes were predicted with *tRNAscan-SE* software.<sup>30,31</sup> We used the version 1.3.1 with the options -B for bacterial genomes and -G for eukaryotic genomes. Pseudogenes and undetermined tRNAs from standard output were discarded from the analysis.

### Discussion

I34 tRNAs modified by ADAT are essential for cell survival.<sup>12,16,18,21,32</sup> They regulate the process of translation by increasing tRNA pairing ability to synonymous codons ended in C, U or A, and avoiding those ended in G. It is clear that inosine is important to balance codon usage and tRNA gene copy number in Eukaryotes, and that highly translated genes in these species tend to be enriched in ADAT-sensitive codons.<sup>19</sup> Consistent with these observations, an initial analysis of the human genome<sup>23</sup> showed: (1) that CDSs coding for proteins rich in ADAT amino acids are overrepresented in the human genome, (2) that ADAT stretches in the human proteome are specifically enriched in TAPS but not in LIVR, and (3) that the more enriched in ADAT amino acids an ORF is, the higher its tendency to use ADAT-sensitive codons instead of G-ended codons.

To explore the impact of I34 in evolution we extended our initial analysis to proteomes from a number of Eukarya and Bacteria. This analysis was designed to identify sections of proteins potentially more dependent on I34 for translation. Our approach was to screen the complete CDSs of different organisms and classify their genes on the basis of ADAT amino acid abundance. Using this initial curation, we identified those transcripts whose proportion of codons for ADAT amino acids is significantly enriched, and we used this subset of sequences to

determine their relative enrichment in ADAT amino acids and ADAT-sensitive codons.

First, we found that the majority of eukaryotic and bacterial proteins rich in ADAT amino acids are specifically enriched in TAPS (Fig. 3). Physicochemical parameters probably explain why, among all ADAT amino acids, only TAPS can reach higher relative frequencies. Functional features of proteins rich in TAPS must have driven the selection of these extremely biased protein sequences.

Our data shows that in Eukarya ADAT stretches are longer and more frequent than in Bacteria (Figs. 2, 3 and Table 1). Eukaryotic genes coding for stretches of ADAT amino acids are generally biased towards ADAT-sensitive codons, with a positive correlation with respect to their length. Conversely, in Bacteria, the presence of stretches rich in ADAT amino acids is 4-fold lower, and the corresponding genes are generally depleted of ADAT-sensitive codons (Fig. 5 and Table 1). These results likely reflect the differences in I34 dependence between Eukaryotes and Bacteria.

Among Eukarya, only animals and plants have a positive correlation between lengths of stretches rich in ADAT amino acids with enrichment in ADAT-sensitive codons. Fungi and Protists show no such correlation. Moreover, Metazoa contain the longest ADAT stretches and the highest enrichment in ADAT-sensitive codons among Eukarya (Fig. 6). These results indicate a preference for ADAT-sensitive codons in longer stretches of ADAT amino acids in kingdoms with embryonic development where multicellularity is prevalent. In this regard, a connection between variation in codon usage and the establishment of multicellularity was already proposed by Ikemura.<sup>33</sup>

Surprisingly we have discovered that in the bacterial phyla of Firmicutes and Cyanobacteria genes for proteins rich in ADAT amino acids are also enriched in ADAT-sensitive codons (Fig. 7). This is initially surprising because, in bacteria, only tRNA<sup>Arg</sup> (ACG) is believed to be deaminated by TadA, and arginine is not significantly enriched in the proteins that we have identified. This apparent contradiction may be resolved by our discovery that the genomes of several Firmicutes code for previously unnoticed tRNAs with A34 for amino acids Threonine, Proline, Serine, Leucine and Isoleucine (Table S2). These tRNAs could potentially be deaminated to I34. Thus, it is possible that the enrichment in ADAT-sensitive codons in these species is due, at least partially, to their use of additional I34-modified tRNAs. Cyanobacteria genomes do not contain this type of tRNAs, but their stretch length is shorter than in Firmicutes, and the frequency of ADAT-sensitive codons does not increase with stretch length (Fig. 7).

Our results suggest that a dependence on ADAT activity for the synthesis of TAPS-enriched proteins is an eukaryote-wide phenomenon that possibly extends to some bacterial species. It has previously been reported that some bacterial groups have secondarily lost A34-containing tRNAs and the ability to deaminate A34 to I34.<sup>34,35</sup> Here, we show that other bacterial groups display the opposite behavior, and possibly have gained the ability to modify the A34 to I34 in tRNAs other than tRNA<sup>Arg</sup> (AGC). We would like to offer the hypothesis that the enrichment in ADAT-sensitive codons is an adaptation of the translational apparatus to improve the synthesis of proteins



extremely rich in threonine, alanine, proline, and serine. As this type of polypeptides is particularly prevalent in extracellular matrices it is possible that ADAT, I34, and ADAT-sensitive codons were important in the evolution of multicellularity.<sup>24</sup>

## Conflicts of interest

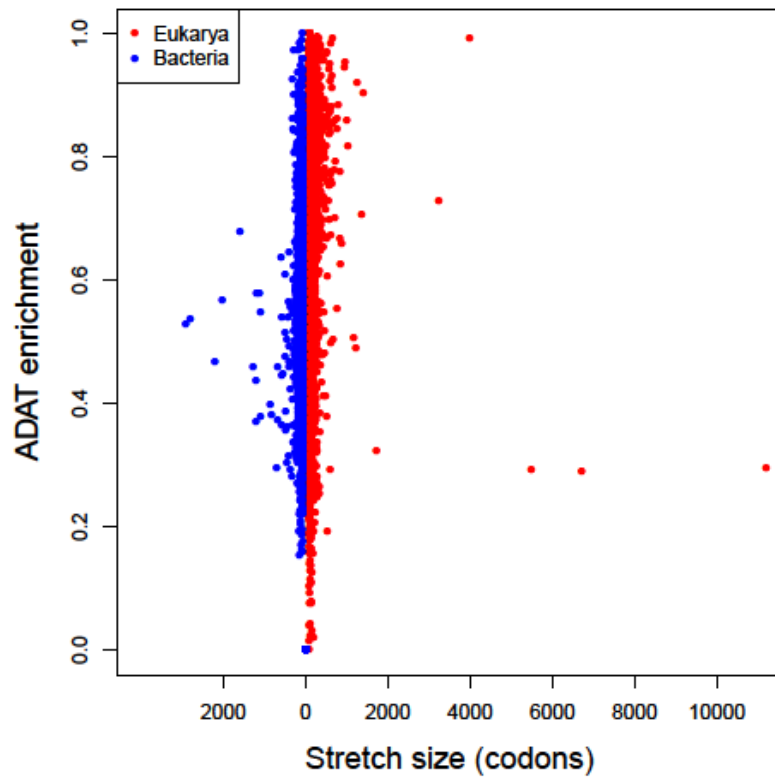
The authors declare no conflicts of interest.

## Acknowledgements

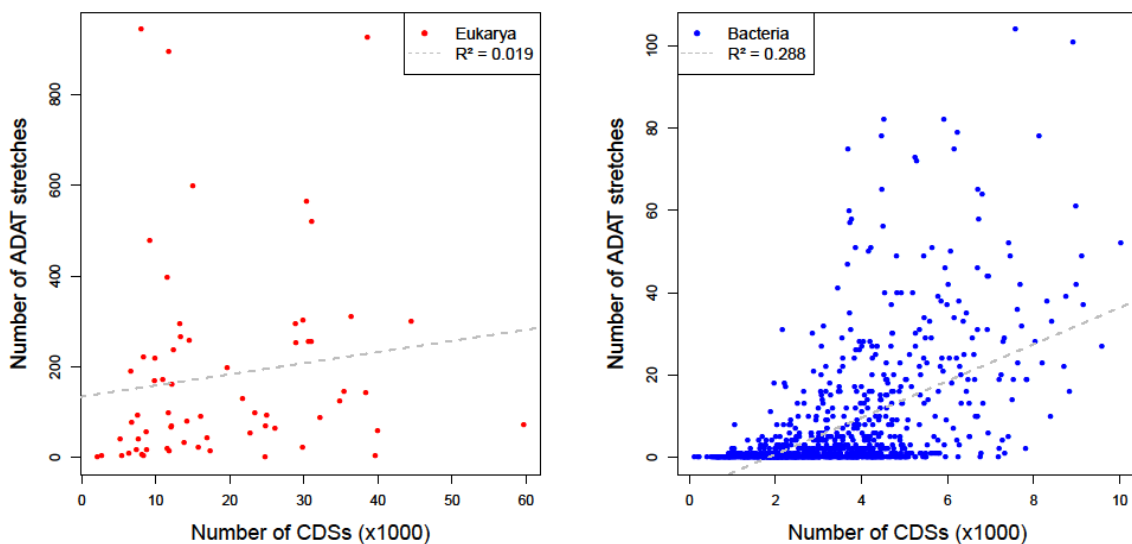
This work was supported by the Spanish Ministry of Economy and Competitiveness under Grant [BES2013-064551] to AR-Y; and [BIO2015-64572] to LRdP. IR-T acknowledges financial support from an ERC Consolidator (ERC-2012-Co -616960) grant and a grant (BFU2014-57779-P) from Ministerio de Economía y Competitividad (MINECO), the latest with help from the Fondo Europeo de Desarrollo regional (FEDER).

## References

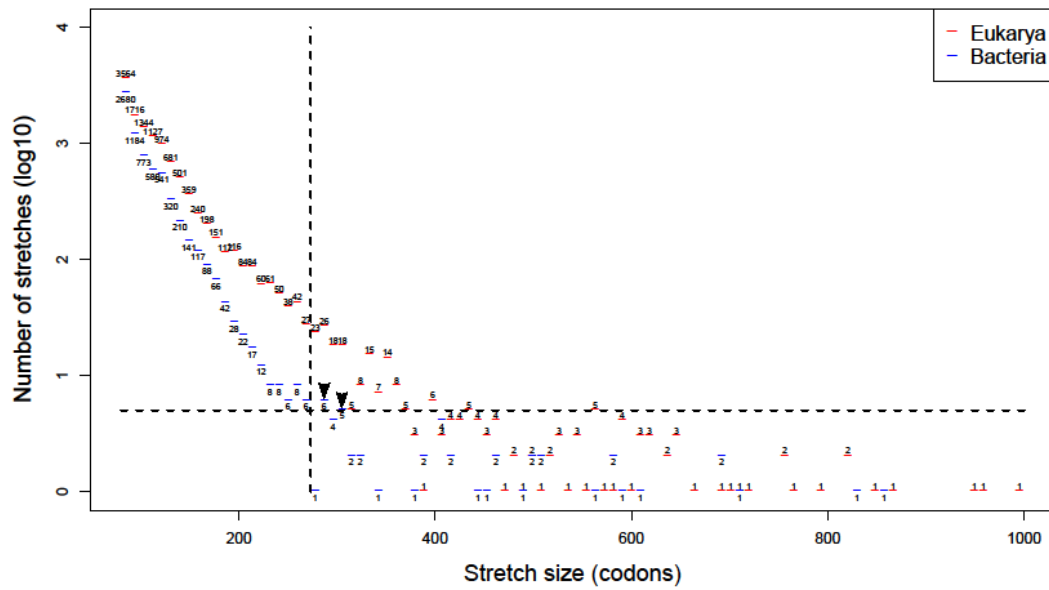
- Crick FH. On protein synthesis. *Symp Soc Exp Biol.* 1958;12:138–63.
- Marck C, Grosjean H. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA.* 2002;8(10):1189–232.
- Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* 2009;37(Database issue):D93–7. doi:10.1093/nar/gkn787.
- Grosjean H, de Crecy-Lagard V, Marck C. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett.* 2010;584(2):252–64. doi:10.1016/j.febslet.2009.11.052.
- Crick FH. Codon–anticodon pairing: the wobble hypothesis. *J Mol Biol.* 1966;19(2):548–55. doi:10.1016/S0022-2836(66)80022-0.
- Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L. A role for tRNA modifications in genome structure and codon usage. *Cell.* 2012;149(1):202–13. doi:10.1016/j.cell.2012.01.050.
- El Yacoubi B, Bailly M, de Crecy-Lagard V. Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annu Rev Genet.* 2012;46:69–95. doi:10.1146/annurev-genet-110711-155641.
- Phizicky EM, Hopper AK. tRNA biology charges to the front. *Genes Dev.* 2010;24(17):1832–60. doi:10.1101/gad.1956510.
- Towns WL, Begley TJ. Transfer RNA methyltransferases and their corresponding modifications in budding yeast and humans: activities, predictions, and potential roles in human health. *DNA Cell Biol.* 2012;31(4):434–54. doi:10.1089%2Fdnadna.2011.1437.
- Torres AG, Batlle E, Ribas de Pouplana L. Role of tRNA modifications in human diseases. *Trends Mol Med.* 2014;20(6):306–14. doi:10.1016/j.molmed.2014.01.008.
- Jackman JE, Alfonzo JD. Transfer RNA modifications: nature's combinatorial chemistry playground. *Wiley Interdiscip Rev RNA.* 2013;4(1):35–48. doi:10.1002%2Fwrna.1144.
- Gerber AP, Keller W. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Sci.* 1999;286(5442):1146–9. doi:10.1126/science.286.5442.1146.
- Torres AG, Pineyro D, Filonava L, Stracker TH, Batlle E, Ribas de Pouplana L. A-to-I editing on tRNAs: biochemical, biological and evolutionary implications. *FEBS Lett.* 2014;588(23):4279–86. doi:10.1016/j.febslet.2014.09.025.
- Auxilien S, Crain PF, Trewyn RW, Grosjean H. Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of inosine 34 in the anticodon of transfer RNA. *J Mol Biol.* 1996;262(4):437–58. doi:10.1006/jmbi.1996.0527.
- Zhou W, Karcher D, Bock R. Identification of enzymes for adenosine-to-inosine editing and discovery of cytidine-to-uridine editing in nucleus-encoded transfer RNAs of Arabidopsis. *Plant Physiol.* 2014;166(4):1985–97. doi:10.1104/pp.114.250498.
- Rubio MA, Pastar I, Gaston KW, Ragone FL, Janzen CJ, Cross GA, Papavasiliou FN, Alfonzo JD. An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. *Proc Natl Acad Sci U S A.* 2007;104(19):7821–6. doi:10.1073/pnas.0702394104.
- Elias Y, Huang RH. Biochemical and structural studies of A-to-I editing by tRNA:A34 deaminases at the wobble position of transfer RNA. *Biochemistry.* 2005;44(36):12057–65. doi:10.1021/bi050499f.
- Torres AG, Pineyro D, Rodriguez-Escriba M, Camacho N, Reina O, Saint-Leger A, Filonava L, Batlle E, Ribas de Pouplana L. Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Res.* 2015;43(10):5145–57. doi:10.1093/nar/gkv277.
- Novoa EM, Ribas de Pouplana L. Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet.* 2012;28(11):574–81. doi:10.1016/j.tig.2012.07.006.
- Saint-Leger A, Bello C, Dans PD, Torres AG, Novoa EM, Camacho N, Orozco M, A Kondrashov F, de Pouplana L. Saturation of recognition elements blocks evolution of new tRNA identities. *Sci Adv.* 2016;2(4):e1501860. doi:10.1126/sciadv.1501860.
- Wolf J, Gerber AP, Keller W. tadA, an essential tRNA-specific adenosine deaminase from *Escherichia coli*. *EMBO J.* 2002;21(14):3841–51. doi:10.1093%2Femboj%2F21.14.3841.
- Iyer LM, Zhang D, Rogozin IB, Aravind L. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Res.* 2011;39(22):9473–97. doi:10.1093%2Fnar%2F39.22.9473.
- Rafels-Ybern A, Attolini CS, Ribas de Pouplana L. Distribution of ADAT-Dependent Codons in the Human Transcriptome. *Int J Mol Sci.* 2015;16(8):17303–14. doi:10.3390/ijms160817303.
- Ribas de Pouplana L, Torres AG, Rafels-Ybern A. What Froze the Genetic Code? *Life.* 2017;7(2):14. doi:10.3390/life7020014.
- Schaub M, Keller W. RNA editing by adenosine deaminases generates RNA and protein diversity. *Biochimie.* 2002;84(8):791–803. doi:10.1016/S0300-9084(02)01446-3.
- Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 2009;25(11):1451–2. doi:10.1093/bioinformatics/btp187.
- Hutter S, Vilella AJ, Rozas J. Genome-wide DNA polymorphism analyses using Variscan. *BMC bioinformatics.* 2006;7:409. doi:10.1186/1471-2105-7-409.
- McDonald JH. Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol Biol Evol.* 1996;13(1):253–60.
- The R project for statistical computing. <https://www.r-project.org/>.
- Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* 2013;29(22):2933–5. doi:10.1093/bioinformatics/btt509.
- Fichant GA, Burks C. Identifying potential tRNA genes in genomic DNA sequences. *J Mol Biol.* 1991;220(3):659–71. doi:10.1016/0022-2836(91)90108-I.
- Tsutsumi S, Sugiura R, Ma Y, Tokuoka H, Ohta K, Ohte R, Noma A, Suzuki T, Kuno T. 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.* 2007;282(46):33459–65. doi:10.1074/jbc.M706869200.
- Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 1985;2(1):13–34.
- Yokobori S, Kitamura A, Grosjean H, Bessho Y. Life without tRNAArg-adenosine deaminase TadA: evolutionary consequences of decoding the four CGN codons as arginine in Mycoplasmas and other Mollicutes. *Nucleic Acids Res.* 2013;41(13):6531–43. doi:10.1093%2Fnar%2F41.13.6531.
- de Crecy-Lagard V, Marck C, Brochier-Armanet C, Grosjean H. Comparative RNomics and modomics in Mollicutes: prediction of gene function and evolutionary implications. *IUBMB life.* 2007;59(10):634–58. doi:10.1080/15216540701604632.



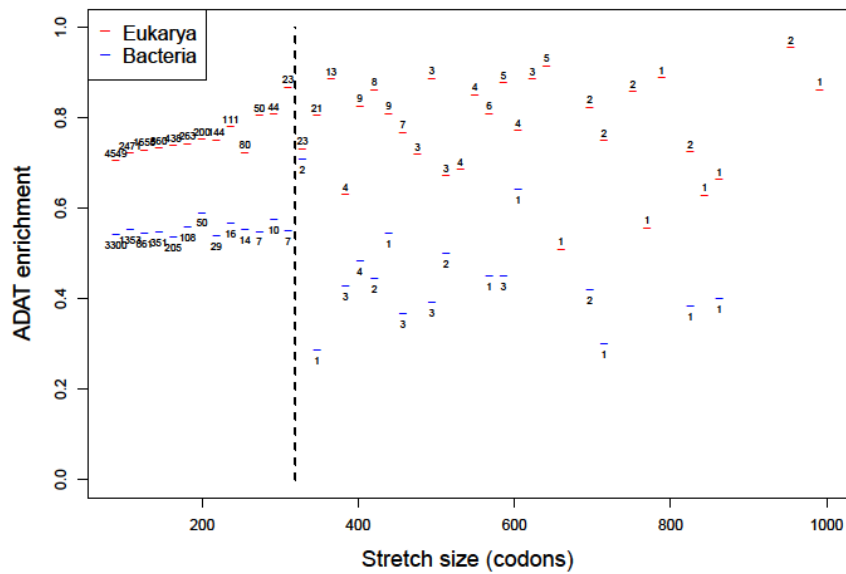
**Figure S1:** Complete dataset obtained from the identification of proteins enriched in ADAT amino acids in Eukaryotes (red dots) and Bacteria (blue dots). We have considered as outliers those sequences longer than 1000 codons since they represent only 0.13% of the data (24 stretches out of 18840 stretches in total)



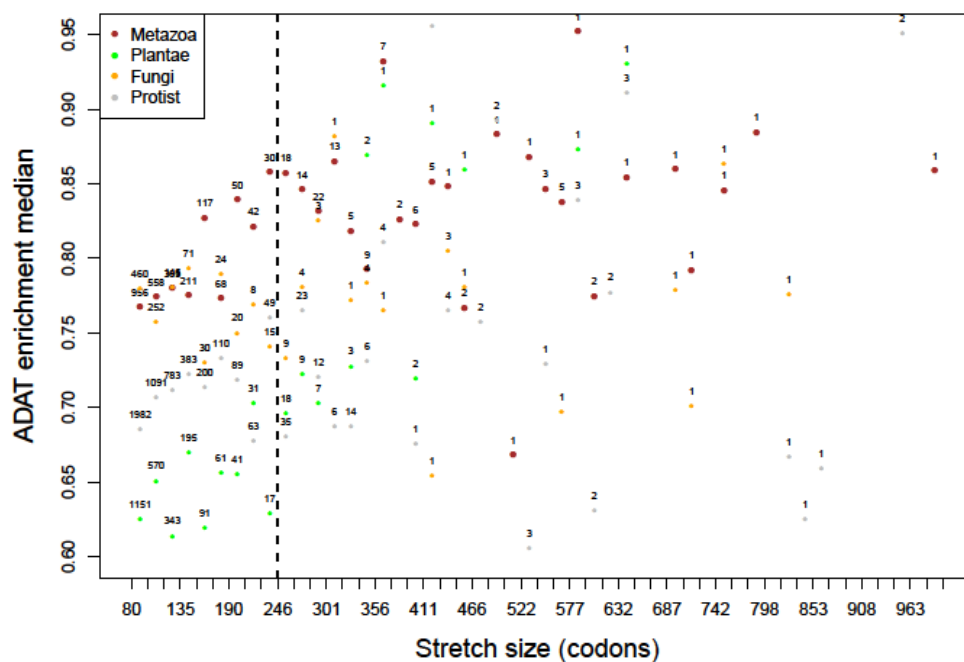
**Figure S2:** Correlation between number of CDSs analysed and sequence stretches enriched in ADAT amino acids in Eukaryotes (left graph) and Bacteria (right graph).



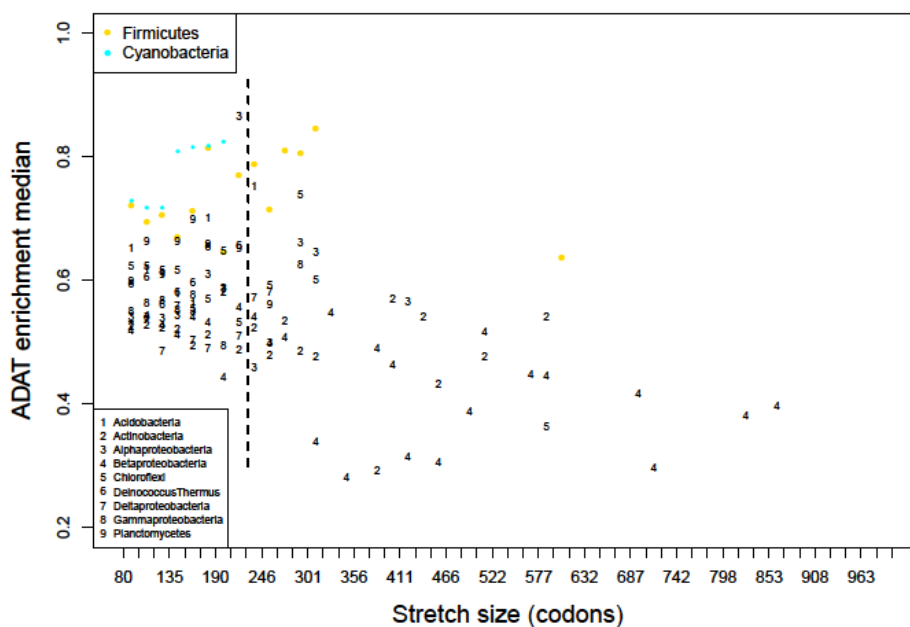
**Figure S3:** Histogram of number of ADAT stretches (in log10) for Eukarya (red) and Bacteria (blue) according to their length. Numbers close to red and blue lines corresponds to the number of samples for each interval. Intervals with less than 5 samples were omitted in Figure 4 (dashed lines). For clarity, data from intervals [282,291] and [301,310] (black triangle) were also omitted in Figure 4.



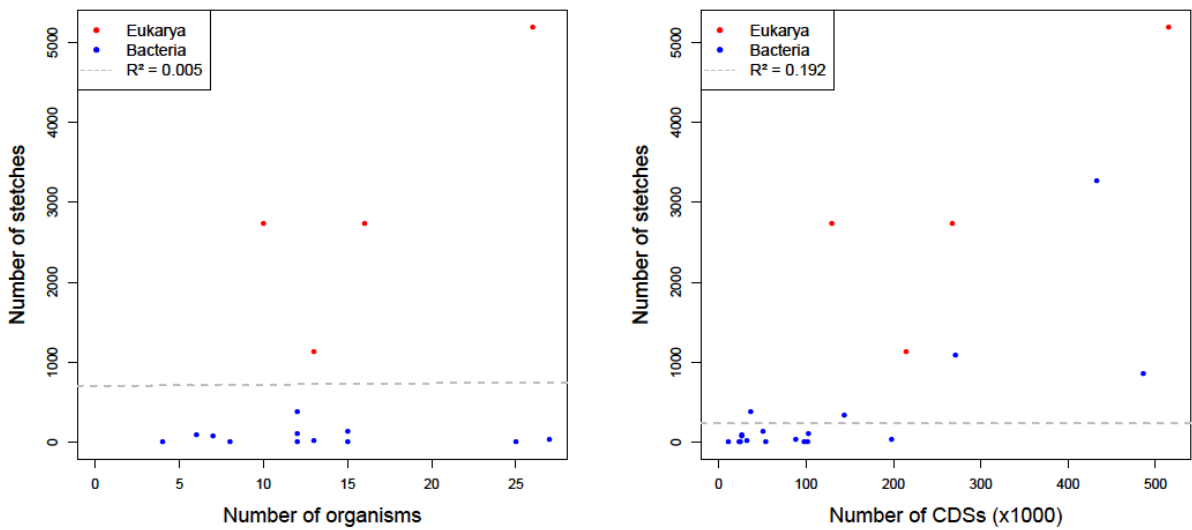
**Figure S4:** Median values of ADAT-sensitive codon frequency by length of ADAT amino acid stretches plotted for Eukarya (red) and bacteria (blue). Numbers close to the lines correspond to the number of samples for each interval. Intervals with less than 5 samples were omitted in Figure 5 (dashed line).



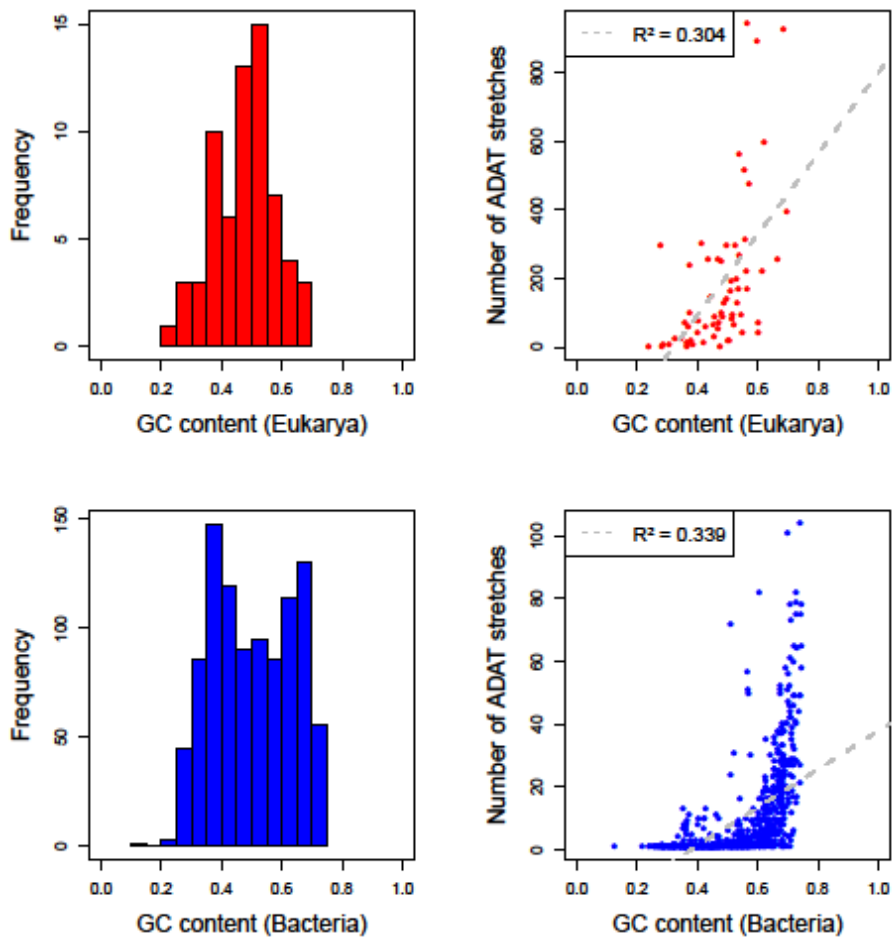
**Figure S5:** ADAT enrichment histogram with median plotted for different eukaryotic kingdoms (see legend) according to their size. Numbers close to the dots corresponds to the number of samples for each interval. Intervals with less than 5 samples were omitted in Figure 6 (dashed line).



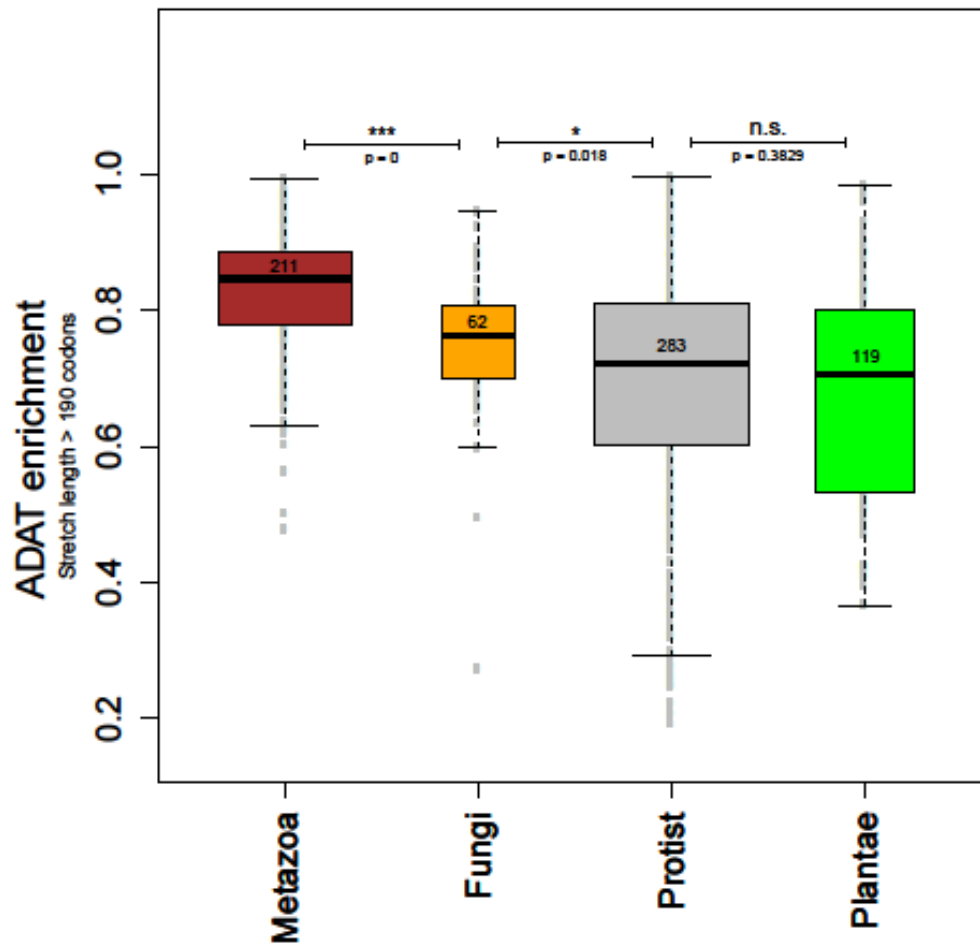
**Figure S6:** Median values of ADAT-sensitive codon enrichment by length of ADAT amino acid stretches plotted for different bacterial phyla (see legend). Phyla with less than 50 stretches are not considered. Intervals with less than 5 samples were omitted in Figure 7 (dashed line).



**Figure S7:** Correlation between number of CDSs analyzed (right graph) or number of different organisms (left graph) and the total number of ADAT stretches in four eukaryotic kingdoms (red) and twenty bacteria phyla (blue).



**Figure S8:** GC content analysis for eukaryotic (1<sup>st</sup> row) and bacterial (2<sup>nd</sup> row) organisms. Distribution of GC values (1<sup>st</sup> column) and correlation between GC content and number of ADAT stretches (2<sup>nd</sup> row) with linear regressions (dashed lines and  $R^2$  values) are shown.



**Figure S9:** Boxplot of the enrichment in ADAT-sensitive codons for four different eukaryotic kingdoms for stretches higher than 190 codons. Numbers close to the median correspond to the number of stretches for each interval.

**Table S1:** Additional information for 4 different Eukaryotic kingdoms and 20 different Bacterial phyla.

	<b>Nstr</b>	<b>Norg</b>	<b>Nstr /Norg</b>	<b>Ncds x1000</b>
<b>Kingdom</b>				
Protist	5185	26	199	515
Metazoa	2735	16	171	268
Plantae	2729	10	273	130
Fungi	1132	13	87	215
<b>Phylum</b>				
Actinobacteria	3265	123	27	432
Betaproteobacteria	1082	77	14	271
Alphaproteobacteria	850	128	7	486
Chloroflexi	380	12	32	36
Deltaproteobacteria	337	38	9	143
Gammaproteobacteria	291	173	2	578
Firmicutes	194	185	1	786
DeinococcusThermus	142	15	9	50
Cyanobacteria	107	12	9	103
Planctomycetes	91	6	15	26
Acidobacteria	77	7	11	26
Spirochaetes	31	27	1	88
Bacteroidetes	28	63	0	198
ChlamydiaeVerrucomicrobia	23	13	2	32
Chlorobi	10	12	1	24
Fusobacteria	5	4	1	11
Epsilonproteobacteria	4	25	0	98
Tenericutes	1	30	0	102
Aquificae	0	8	0	23
Thermotogae	0	15	0	54

Nstr: number of stretches. Columns means: Norg, number of organisms. Ncds, number of CDSs.

**Table S2:** List of tRNA genes cognate for ADAT amino acids with A34 found in Firmicutes.

Name	Ala	Pro	Thr	Val	Ser	Leu	Arg	Ile	Name	Ala	Pro	Thr	Val	Ser	Leu	Arg	Ile
	AGC	AGG	AGT	AAC	AGA	AAG	ACG	AAT		AGC	AGG	AGT	AAC	AGA	AAG	ACG	AAT
Anaerococcus	0	0	0	0	0	1	1	0	Leuconostoc	0	0	1	0	0	1	2	0
Bacillus	0	1	0	0	0	0	1	0	Leuconostoc	0	0	1	0	0	1	2	0
Butyrivibrio	0	1	0	0	0	1	1	0	Leuconostoc	0	0	1	0	0	1	2	0
Clostridium	0	0	0	0	0	0	1	1	Leuconostoc	0	0	1	0	0	1	2	0
Clostridium	0	0	0	0	0	2	2	0	Leuconostoc	0	0	1	0	0	1	2	0
Clostridium	0	0	0	0	0	1	1	0	Oenococcus	0	0	1	0	1	1	1	0
Clostridium	0	0	0	0	0	2	2	0	Pediococcus	0	0	0	0	0	1	2	0
Desulfosporosinus	0	0	0	0	1	0	2	0	Pediococcus	0	0	0	0	0	1	2	0
Eubacterium	0	0	0	0	0	1	1	0	Roseburia	0	0	0	0	0	1	1	0
Eubacterium	0	0	0	0	0	2	1	0	Streptococcus	0	0	0	0	0	1	3	0
Lactobacillus	0	0	1	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	3	0
Lactobacillus	0	0	1	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	2	0
Lactobacillus	0	0	1	0	0	1	2	0	Streptococcus	0	0	0	0	0	2	4	0
Lactobacillus	0	0	0	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	2	0
Lactobacillus	0	0	1	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	2	0
Lactobacillus	0	0	2	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	2	0
Lactobacillus	0	0	1	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	1	0
Lactobacillus	0	0	1	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	1	0
Lactobacillus	0	0	1	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	4	0
Lactobacillus	0	0	0	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	2	0
Lactobacillus	0	0	0	0	0	2	2	0	Streptococcus	0	0	0	0	0	1	3	0
Lactobacillus	0	0	0	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	2	0
Lactobacillus	0	0	0	0	0	2	2	0	Streptococcus	0	0	0	0	0	1	3	0
Lactobacillus	0	0	1	0	0	1	3	0	Streptococcus	0	0	0	0	0	1	2	0
Lactobacillus	0	0	1	0	0	1	2	0	Streptococcus	0	0	0	0	0	1	2	0
Lactococcus	0	0	0	0	0	2	2	0	Streptococcus	0	0	0	0	0	1	3	0
Lactococcus	0	0	0	0	0	2	2	0	Weissella	0	0	1	0	0	1	2	0
Lactococcus	0	0	0	0	0	2	2	0									

Numbers in each column show the number of genes predicted in each corresponding genome.





### 6.4 Publication 3

**Rafels-Ybern, A.**, A. G. Torres, N. Camacho, A. Herencia-Ropero, H. Roura, T. Wulff, M. Raboteg, J.C. Gutierrez, A. Bordons and L. Ribas de Pouplana (**to be submitted**).  
“Distribution of Inosine at the wobble position of transfer RNAs, and the role of this modification in the evolution of bacterial and eukaryotic proteomes.”



# Distribution of Inosine at the wobble position of transfer RNAs, and the role of this modification in the evolution of bacterial and eukaryotic proteomes

(to be submitted)

Àlbert Rafels-Ybern<sup>1§</sup>, Adrian Gabriel Torres<sup>1§</sup>, Noelia Camacho<sup>1</sup>, Andrea Herencia-Ropero<sup>1</sup>, Helena Roura Frigolé<sup>1</sup>, Thomas Wulff<sup>1</sup>, Marina Raboteg<sup>1</sup>, Juan Carlos Gutierrez<sup>2</sup>, Albert Bordons<sup>3</sup>, Lluís Ribas de Pouplana<sup>1,4</sup>

§ These authors equally contributed.

<sup>1</sup> Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Catalonia, Spain.

<sup>2</sup> Departamento de Microbiología-III. Facultad de Biología. Universidad Complutense (UCM). 28040 Madrid, Spain.

<sup>3</sup> Departament de Bioquímica i Biotecnologia. Universitat Rovira i Virgili. 43007 Tarragona, Catalonia, Spain.

<sup>4</sup> ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Catalonia, Spain.

Corresponding author: lluis.ribas@irbbarcelona.org (L. Ribas de Pouplana)

## Abstract

The modification of adenosine to inosine at the first position of transfer RNA (tRNA) anticodons (I34) is widespread among bacteria and eukaryotes. In bacteria, the modification is mostly found in tRNA<sup>Arg</sup>, and is catalysed by TadA, a homodimeric adenosine deaminase. In eukaryotes, a heterodimeric enzyme (ADAT) evolved from TadA, introduces I34 in up to eight different tRNAs. This substrate expansion, significantly influenced the evolution of eukaryotic genomes, both in terms of codon usage and tRNA gene composition. However, the selective advantages driving this process, remain unclear. Here we have studied the evolution of I34, TadA, ADAT, and their relevant codons in a large set of bacterial and eukaryotic species. Our analysis indicates that I34 and TadA were originally absent in ancestral bacterial phyla and became secondarily lost in other prokaryotes. We also show that a functional expansion of I34 to tRNAs other than tRNA<sup>Arg</sup> has occurred within bacteria, and that this process likely starts with the emergence of unmodified A34-containing tRNAs. In eukaryotes, we report on a large variability in the use of I34 in protists, while the set of ADAT substrates becomes fixed in fungi, plants and animals. Finally, we present data to support that the expansion in I34-containing tRNA was driven by their ability to facilitate the translation of proteins rich in amino acids cognate to I34-modified tRNAs.

**Keywords:** Translation; Evolution; Speciation; tRNA; Transcriptome; mRNA; ADAT; TadA; CDS.

## Introduction

Transfer RNAs (tRNAs) are heavily modified nucleic acids that translate nucleotide triplets to amino acids during ribosomal protein synthesis. The chemical modification of tRNA bases takes place during the process of maturation of tRNA gene transcripts. Over a hundred different modifications have been described across the phylogenetic tree, the vast majority of which are not universal, and only a limited number are essential (Marck and Grosjean 2002). Base modifications in tRNAs play structural and functional roles that include the stabilization of the tRNA structure, molecular recognition by tRNA-related enzymes and, when located in the anticodon stem-loop, the modulation of codon-anticodon interactions (Phizicky and Hopper 2010; El Yacoubi, Bailly et al. 2012; Towns and Begley 2012; Jackman and Alfonzo 2013). Modifications in tRNAs are important modulators of tRNA function, and are linked to a growing list of human diseases (Torres,

Battle et al. 2014). However, the details of their evolutionary origins and of the selective pressures that drove their emergence remain mostly unknown.

Inosine is the product of adenosine deamination, and is abundantly found in RNAs (Maas and Rich 2000; Nishikura 2010; Wulff and Nishikura 2010). To date, inosine has been described as a soluble nucleotide in the cytoplasm, and it is also present in messenger RNAs (mRNAs), small non-coding RNAs, and tRNAs. In tRNAs inosine is found at position 34, and further modified to methyl inosine at positions 37 and 54 (Grosjean, Auxilien et al. 1996; El Yacoubi, Bailly et al. 2012). The modification of adenosine to inosine at position 34 in tRNAs (the first position of the anticodon) is catalysed by a family of enzymes that evolved from cytidine deaminases (Gerber and Keller 2001). In bacteria, I34 modifications are catalysed by the homodimeric enzyme tRNA Adenosine Deaminase A (TadA) and were previously believed to be exclusive for tRNA<sup>Arg</sup> (Wolf, Gerber et al. 2002).

In eukaryotes the same modification is found in tRNAs coding for the amino acids Thr, Ala, Pro, Ser, Leu, Ile, Val, and Arg (TAPSLIVR), and is catalysed by the heterodimeric enzyme Adenosine Deaminase acting on tRNA (ADAT) composed of two related subunits ADAT2 and ADAT3 (Gerber and Keller 1999). I34 is absent in archaeal tRNAs (Marck and Grosjean 2002). Sequence analyses, phylogenetic studies, and structural comparisons show that ADAT2 and ADAT3 evolved through duplication and divergence from the bacterial TadA gene, possibly in an ancestral eukaryotic genome. This evolution was likely followed by the increase in the number of tRNAs modified to I34 (Auxilien, Crain et al. 1996; Gerber and Keller 1999; Elias and Huang 2005; Rubio, Pastar et al. 2007; Torres, Pineyro et al. 2014; Zhou, Karcher et al. 2014; Torres, Pineyro et al. 2015).

The emergence of ADAT and the expansion of I34 in the eukaryotic tRNAome was accompanied by a dramatic modification of the tRNA gene population in nucleated cells, which rapidly became enriched in tRNA genes coding for ADAT substrates. Concomitantly, highly expressed eukaryotic genes changed their codon composition to adapt to variations in the tRNAome. Thus, the emergence of ADAT and the expansion of I34 in eukaryotes was an important parameter in the evolution of the eukaryotic genome (Novoa, Pavon-Eternod et al. 2012). In humans, for example, I34-modified tRNAs are absolutely required to translate cognate codons ending in cytidine because genes coding for isoacceptor tRNAs with G34 are absent in our genome (Novoa, Pavon-Eternod et al. 2012).

Although TadA and ADAT have been shown to be essential in *E. coli*, *T. brucei*, *S. cerevisiae*, *S. pombe*, *A. thaliana*, *F. graminearum*, and *H. sapiens*, (Gerber and Keller

1999; Wolf, Gerber et al. 2002; Rubio, Pastar et al. 2007; Tsutsumi, Sugiura et al. 2007; Zhou, Karcher et al. 2014; Torres, Pineyro et al. 2015; Liu, Wang et al. 2016) the nature of the selection force, driving the evolution from TadA to ADAT, remains an open question. I34 changes the pairing ability of anticodons, allowing them to recognize mRNA triplets ended in uridine (U), cytidine (C), and adenosine (A) (Crick 1966; Gerber and Keller 1999; Elias and Huang 2005; Torres, Pineyro et al. 2014) and ADAT has been shown to play additional specialized functions in *T. brucei* (Gaston, Rubio et al. 2007). However, it seems reasonable to assume that the main selection parameter driving the evolution from TadA to ADAT is linked to the universal role of I34 in translation elongation.

We have shown that genes coding for TAPSLIVR-rich proteins are more abundant in eukaryotic genomes than in bacterial ones, and only the former display a preference for codons cognate to I34-containing tRNAs (Rafels-Ybern, Attolini et al. 2015). Thus, a potential selective advantage of ADAT in eukaryotes is the facilitation of translation of genetic sequences coding for TAPSLIVR-rich proteins regions (stretches).

To understand the functional parameters that drove the evolution of TadA and ADAT we have performed a comprehensive analysis of the evolution of the I34-associated molecular machinery. This includes the characterization of the phylogenetic distribution of A34-containing tRNAs (A34-tRNAs), the phylogenetic analyses of TadA and ADAT, a genomic analysis of TAPSLIVR-codon usage in prokaryotes and eukaryotes, and the experimental characterization of the substrate specificities of TadA and ADAT in relevant species.

First, we characterized the distribution of genes coding for A34-tRNAs across all major bacterial and eukaryotic groups. This analysis provided a first approximation to the distribution of I34 in the phylogenetic tree. We then developed a pipeline to identify genes coding for *bona fide* TadA or ADAT. We applied this pipeline to determine the phylogenetic distribution of TadA and ADAT genes in bacteria and eukaryotes, respectively. Our findings indicate that several ancestral bacterial groups lack both TadA and A34-tRNAs, suggesting that these species never developed the machinery to generate I34-modified tRNAs. On the other hand, limited sets of bacterial species have either lost the system secondarily, or expanded it to additional tRNA substrates.

We experimentally determined that the genome of the firmicute *Oenococcus oeni* has an expanded repertoire of A34-tRNAs that includes tRNA<sup>Arg</sup>, tRNA<sup>Leu</sup>, tRNA<sup>Ser</sup> and tRNA<sup>Thr</sup>, but only tRNA<sup>Arg</sup> and tRNA<sup>Leu</sup> are modified to I34 under standard culture conditions. Thus,

I34 tRNA expansion is not exclusive to eukaryotes, and it likely requires the emergence of unmodified A34-tRNAs.

We also found that, in eukaryotes, a division can be established between protists, which display large variability in their genomic contents of A34-tRNA genes, and fungi, plants and animals, which remain uniform in the same regard. We report that, despite the variability seen in protists, the tRNAome of the protozoan *Tetrahymena thermophila* contains the same set of I34-tRNAs than metazoans, showing that a fully functional ADAT evolved early in eukaryotic evolution. This suggests that this process is basal to all extant eukaryotes and was quickly fixed in their genomes.

Finally, we characterized the proteome composition of bacterial and eukaryotic groups in terms of TAPSLIVR-rich protein abundance and codon composition of the respective genes. We found that the abundance of TAPSLIVR-rich proteins gradually increases from bacteria to protists, and to multicellular eukaryotes. This increase is mirrored by a growing preference for codons that depend on I34-modified tRNAs.

In summary, we have attempted to dissect the evolutionary transition from TadA to ADAT and the changes linked to this evolution in terms of tRNA gene composition, I34-modified tRNAome, codon usage, and proteome composition. These analyses point at the role played by I34-tRNAs in the synthesis of TAPSLIVR-rich proteins as a potential selection force driving this evolution.

## Results

### Analysis of the phylogenetic distribution of A34-tRNAs

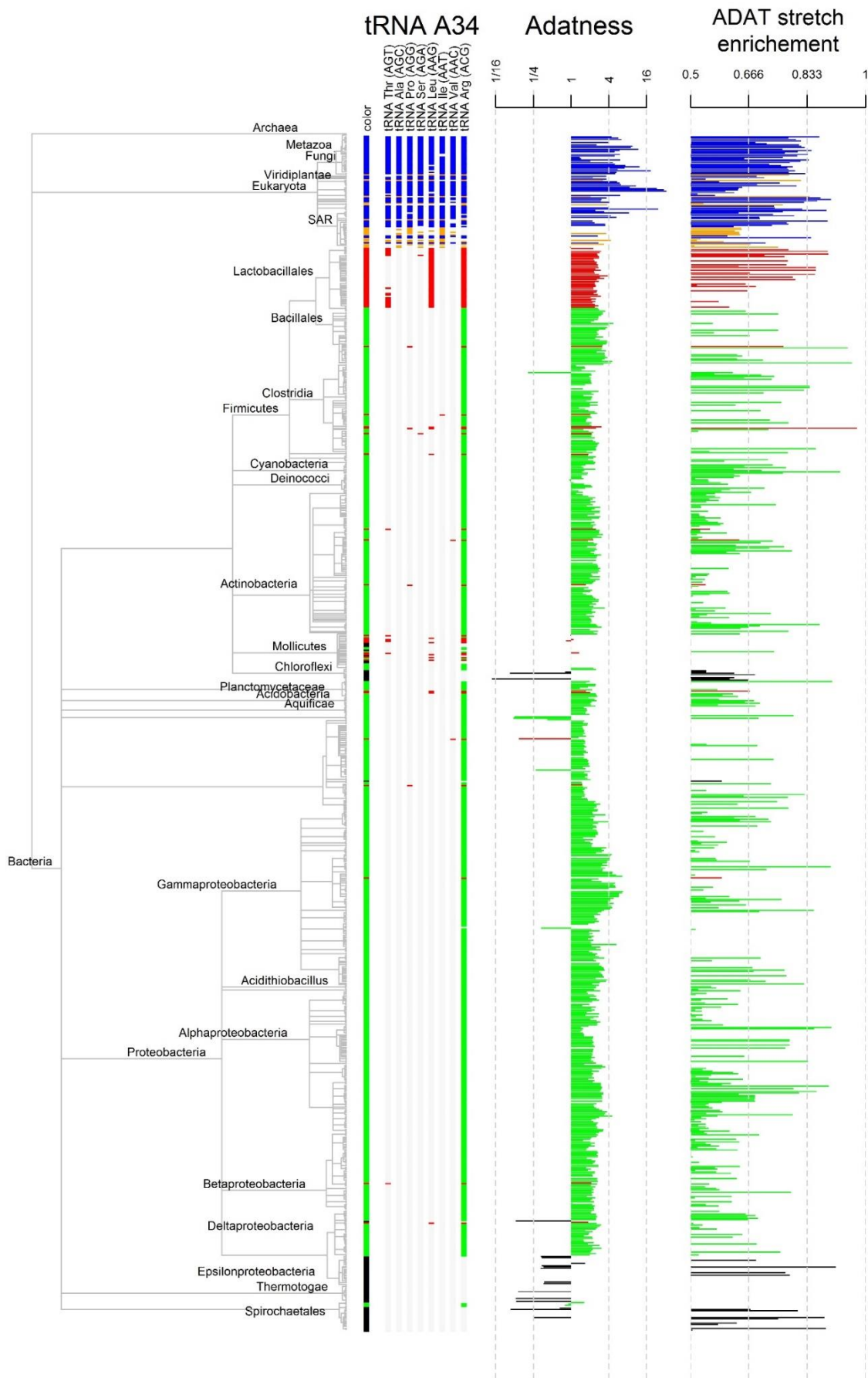
As an initial step towards the determination of the distribution of I34-tRNAs, we selected 956 bacterial and 150 eukaryotic species with completely sequenced genomes to represent all major phylogenetic phyla of both Kingdoms (see Materials and Methods). Using tRNAscan-SE software, (Lowe and Eddy 1997) we systematically searched each genome to identify tRNA genes cognate to TAPSLIVR with A at the wobble position (A34-tRNAs). Those A34-tRNAs are potential substrates for TadA or ADAT enzymes. We define A34-tRNA diversity as the number of different tRNA isoacceptors present in a genome. For example, the *H. sapiens* genome contains A34-tRNAs cognate for the amino acids T, A, P, S, L, I, V, and R (TAPSLIVR) so its A34-tRNA diversity is 8. We then used a consensus phylogenetic tree (Letunic 2015) to map the distribution of A34-tRNA diversity scores (**Figure 1**).



In the vast majority of bacterial phyla, as expected, A34 was only found in tRNA<sup>Arg</sup>, consistent with the notion that TadA only deaminates this tRNA (A34-tRNA diversity =1). Several species belonging to the orders of Tenericutes,  $\epsilon$ -proteobacteria, Chloroflexi, Spirochaetes, and Thermotogae were found to contain no detectable genes coding for A34-tRNAs. This situation was previously reported in the monophyletic class of Mollicutes that share a common ancestor with Firmicutes but have adopted parasitic lifestyle (Weisburg, Tully et al. 1989). Mollicutes presents a massive secondary genomic reduction that drove the disappearance of tRNA<sup>Arg</sup><sub>ACG</sub> and TadA (Grosjean, Breton et al. 2014). However, we discovered a widespread lack of A34-tRNAs in bacterial groups that do not display genome reduction and are considered to be ancestral in the bacterial phylogenetic tree (**Figure 1, Table 1**), opening the possibility that ancestral bacterial clades initially lacked genes for A34-tRNAs. The fact that most of the species analyzed in the Chloroflexi, Spirochaetes, and Thermotogae groups lack A34-tRNAs, whereas most Tenericutes contain tRNA<sup>Arg</sup><sub>ACG</sub> genes, supports the possibility that A34-tRNAs were absent in ancestral bacteria, and were lost secondarily in Tenericutes (**Figure 1**).

Unexpectedly, we also detected bacterial species harboring A34-tRNAs cognate for amino acids other than arginine (A34-tRNA diversity > 1). For example, we found a gene coding for A34-tRNA<sup>Ile</sup> in the genome of the Tenericute *Mycoplasma bovis*, and A34-tRNAs cognate for isoleucine, serine and threonine in the Firmicute *Oenococcus oeni* (**Figure 1, Table 1**). At first sight, the disperse distribution of bacterial species with increased diversity of A34-tRNAs may suggest that this increase took place independently in different bacterial orders.

The distribution of A34-tRNA genes among eukaryotic species also displayed significant heterogeneity. We first divided eukaryotes into four major groups: Metazoa, Fungi, Plantae, and the rest of Eukarya (hereinafter “Protists”), and found that the vast majority contain A34-tRNAs coding for TAPSLIVR (A34-tRNA diversity = 8). However, we noticed that the A34-tRNA gene diversity in Protists was extremely variable among the Harosa or SAR group (Stramenopiles, Alveolata, and Rhizaria) (**Figure 1**). For example, we identified A34-tRNA genes cognate for only four amino acid (Ala, Ser, Arg and Ile) in the algae *Nannochloropsis gaditana*, while we detected a full complement of A34-tRNAs cognate for TAPSLIVR in *Tetrahymena thermophila* (**Table 1**). This observation suggested that SAR may contain examples of intermediate stages of the expansion of A34-tRNAs in eukaryotes. For this reason, we decided to analyze the group SAR as an additional set of eukaryotic species, in parallel with the canonical division of protists, plants, fungi, and metazoans.



**Figure 1:** Standard phylogeny based on NCBI taxonomy (NCBI). Each tip represents an organism that is coloured based on their tRNA diversity (see Table 1, note that colour orange here corresponds to colour maroon in Table 1). Following the order of the phylogeny, it is depicted the tRNA A34 that are present (tRNA ANN), the Adatness for ADAT2 in Eukarya and TadaA in bacteria (Adatness) and the ADAT-stretch enrichment for each organism. Missing values means that are not present in the analysis.

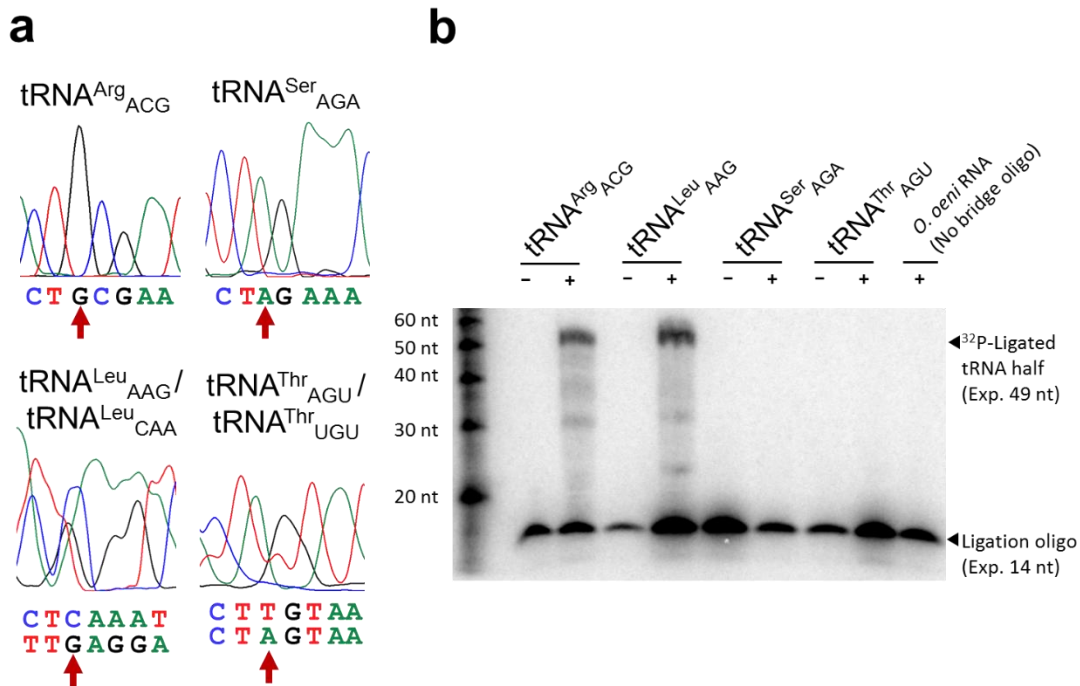
**Table 1:** tRNA dichotomy. tRNA gene copy number for A34 and G34 tRNA isoacceptors for TAPS and LIVR amino acids. The colour legend is based on the number of different A34 tRNA genes that are present for each organism (green numbers where the tRNA gene copy number is  $\geq 1$ ).

Organism	Class	Domain	A34 diversity	ANN anticodon								GNN anticodon							
				Thr	Ala	Pro	Ser	Leu	Ile	Val	Arg	Thr	Ala	Pro	Ser	Leu	Ile	Val	Arg
				agu	agc	agg	aga	aag	aa	aac	acg	gg	ggc	ggg	gga	gag	gau	gac	gcg
<i>Homo sapiens</i>	Metazoa	Eukarya	8-7	10	34	10	10	11	15	11	7	0	0	0	0	0	3	0	0
<i>Arabidopsis thaliana</i>	Plantae	Eukarya		10	16	16	37	12	17	15	11	1	0	0	3	1	0	2	0
<i>Saccharomyces cerevisiae</i>	Fungi	Eukarya		11	11	2	11	0	13	14	7	0	0	0	0	1	1	0	0
<i>Acanthamoeba castellanii</i>	Protist	Eukarya		19	22	11	8	21	10	13	16	0	0	0	0	0	1	0	0
<i>Tetrahymena thermophila</i>	SAR	Eukarya		22	32	20	20	19	28	21	8	0	0	0	0	0	0	0	0
<i>Blastocystis hominis</i>	SAR	Eukarya		6	9	6	5	3	7	0	7	0	0	0	0	0	1	4	0
<i>Phytophthora pinifolia</i>	SAR	Eukarya		2	6	1	1	1	0	0	4	0	0	0	0	0	2	0	1
<i>Nannochloropsis gaditana</i>	SAR	Eukarya		0	1	0	1	0	1	0	1	0	0	0	0	0	3	1	0
<i>Stramenopiles sp.</i>	SAR	Eukarya	2	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	
<i>Oenococcus oeni</i>	Firmicutes	Bacteria	>2	1	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0
<i>Lactobacillus casei</i>	Firmicutes	Bacteria		0	0	0	0	1	0	0	2	1	1	0	1	0	3	0	0
<i>Mycoplasma bovis</i>	Tenericutes	Bacteria	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	
<i>Escherichia coli</i>	Gammaprot.	Bacteria	1	0	0	0	0	0	0	3	2	2	1	1	1	4	2	0	
<i>Aquifex aeolicus</i>	Aquificae	Bacteria		0	0	0	0	0	0	0	1	1	1	1	1	1	2	1	0
<i>Staphylococcus aureus</i>	Firmicutes	Bacteria	0	0	0	0	0	0	0	2	0	0	0	1	2	1	0	0	
<i>Mycoplasma pneumoniae</i>	Tenericutes	Bacteria	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	1	
<i>Spirochaeta thermophila</i>	Spirochaetes	Bacteria	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	
<i>Chloroflexus aggregans</i>	Chloroflexi	Bacteria		0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	
<i>Thermotoga maritima</i>	Thermotogae	Bacteria		0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	

## Experimental analysis of A34-deamination in *O. oeni* and *T. thermophila*

Our computational analysis revealed bacterial species with A34-tRNA diversity higher than 1, whereas eukaryotic organisms in the SAR group displayed high heterogeneity in this regard with an A34-tRNA diversity between 0 and 8. We set to experimentally determine the modification status of the A34-tRNAs found in the bacteria *Oenococcus oeni* and the SAR *Tetrahymena thermophila*.

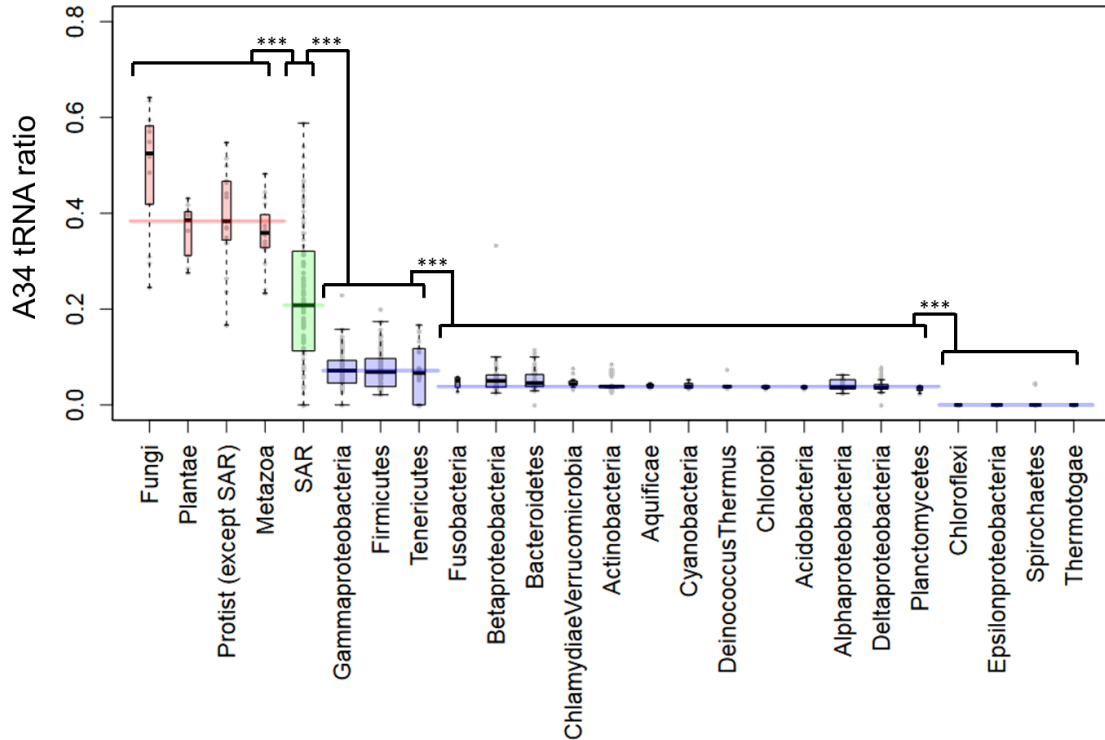
First, we assessed the possibility that the four genes coding for A34-tRNAs found in *O. oeni* were transcribed and modified to I34. This information is important because it could shed light on the potential for bacterial organisms to synthesize I34 in tRNAs other than tRNA<sup>Arg</sup><sub>ACG</sub>, and on the order of events required to achieve this substrate expansion. We purified total RNA from *O. oeni* and sequenced its tRNAs to determine the presence of A34 or I34 in tRNAs cognate for TAPSLIVR. I34 is read as G by reverse transcriptase, resulting in a substitution that can be detected by DNA sequencing. This analysis revealed that all four *O. oeni* A34-tRNAs are transcribed, and showed that tRNA<sup>Arg</sup><sub>ACG</sub> and tRNA<sup>Leu</sup><sub>AAG</sub> are modified to I34 (Figure 2a). We then verified this result by the splinted ligation method, which unequivocally determines the presence of inosine at position 34 in specific tRNAs (Figure 2b) (Torres et al., to be submitted). Again, we found that both tRNA<sup>Arg</sup><sub>ACG</sub> and tRNA<sup>Leu</sup><sub>AAG</sub> are modified to I34 in *O. oeni*, while tRNA<sup>Ser</sup><sub>AGA</sub> and tRNA<sup>Thr</sup><sub>AGT</sub> are not modified. This demonstrates that some bacteria have increased their



**Figure 2:** (a) Sequencing spectra of RT-PCR amplicons derived from *O. oeni* tRNA<sup>Arg</sup> (ACG), tRNA<sup>Ser</sup> (AGA), tRNA<sup>Leu</sup> (AAG), and tRNA<sup>Thr</sup> (AGU). The anticodon ‘wobble’ position is indicated (red arrows). Inosine at this position should be detected as a Guanosine, as opposed to the unmodified residue which is detected as Adenosine. Note that due to sequence similarities, sequencing of RT-PCR amplicons derived from tRNA<sup>Leu</sup> (AAG) are partially masked by amplicons derived from tRNA<sup>Leu</sup> (CAA); and those derived from tRNA<sup>Thr</sup> (AGU) are partially masked by amplicons derived from tRNA<sup>Thr</sup> (UGU). (b) Evaluation of I34 presence/absence on *O. oeni* for the same tRNA substrates in (a) by the splinted-ligation method for I34 detection. A reaction containing *O. oeni* total RNA without a bridge oligo to capture tRNA sequences was used as a negative control.

repertoire of I34-tRNAs. This suggests that this process of expansion begins with the emergence of A34-tRNA genes that are initially not modified to I34.

We then explored whether we could find similar situations of coexisting I34-and A34-tRNAs in unicellular eukaryotes, which could be the reflection of intermediate evolutionary stages in the process of increasing A34-tRNA diversity from bacteria to eukaryotes. To answer this question we purified and sequenced total RNA from *T. thermophila*, a protist that contains genes coding for A34-tRNAs for TAPSLIVR. This initial analysis indicated the presence of I34 in *T. thermophila* tRNAs cognate for Thr, Ser, Leu, Ile, Val, and Arg (**Figure S1a**). Sequencing results for the tRNAs cognate for Ala and Pro were inconclusive, but further experiments using the splinted ligation technique demonstrated that *T. thermophila* also modifies these tRNAs to I34 (**Figure S1b**). Thus, full modification of all A34-tRNAs for TAPSLIVR exists among SAR species. Although we cannot rule out that other species in this group contain unmodified A34-tRNAs, these results suggest that the full substrate expansion from TadA to ADAT took place early in the evolution of eukaryotic organisms



**Figure 3:** A34 tRNA ratio boxplot for eukaryotic kingdoms (red), the eukaryotic superphylum Heterokonta (green) and different bacterial phyla (blue). The y-axis represents the fraction of A34 tRNAs among all the isoacceptors for TAPS and LIVR amino acids. The width of each boxplot is proportional to the number of organisms analyzed. The horizontal colored lines represent the median values for the sets they highlight.

### Analysis of the genomic enrichment in A34-tRNA genes

Our initial analysis confirmed that A34-tRNA diversity increases gradually from bacteria to metazoans (Figure 1). Genomic analysis indicated that ancestral bacterial phyla might lack A34-tRNAs completely, while other bacterial species may contain A34-tRNAs cognate to amino acids other than Arg. Among eukaryotes, some unicellular species of the SAR group contain a limited number of genes for A34-tRNAs while others, like *T. thermophila*, host A34-tRNA genes cognate for TAPSLIVR and modify them all to I34. We have previously shown that the number of genes coding for A34-tRNAs (A34-tRNA gene copy number) dramatically increased relative to other isoacceptor tRNAs in eukaryotic genomes, a process that coincided with the increase of A34-tRNA diversity and the emergence of ADAT in these species (Novoa, Pavon-Eternod et al. 2012). We propose that the emergence of I34-tRNAs for TAPSLIVR in eukaryotes generated a selective pressure in favor of increasing the respective A34-tRNA gene copy numbers.

We asked whether the gradual variation in A34-tRNA diversity that we identified in bacteria and SAR is linked to an increase in the abundance of A34-tRNA genes relative to their respective isoacceptors. To answer this question we quantified the total set of

tRNA genes cognate for TAPSLIVR in each of the genomes of our dataset, and we calculated the ratio of A34-tRNA genes to C34-, U34-, and G34-tRNA gene isoacceptors of the same 4-box family (note that Ser, Arg and Leu have 2 extra tRNA isoacceptors that have been included too) (A34-tRNA ratio). **Figure 3** shows a boxplot analysis where the values of A34-tRNA ratios are clustered by phyla. Logically, A34-tRNA ratios are minimal in the bacterial phyla where A34-tRNA diversity is 0. Among the rest of bacterial groups, the A34-tRNA ratios increase minimally from phyla that mostly contain species with tRNA<sup>Arg</sup><sub>GCA</sub> to species that contain genes for more than one type of A34-tRNA isoacceptor. Thus, it does not appear that the increase in A34-tRNA diversity leads to a significant increase in A34-tRNA genes ratio in bacteria (**Figure 3**). The modiest increase detected is due to the appearance of new A34-tRNAs cognate for amino acids other than arginine (**Figure S2**).

As previously reported, the genomes of Fungi, Plants, Protists, and metazoans display sharp rise in A34-tRNA ratios coinciding with the expansion in A34-tRNA diversity. The eukaryotic group of unicellular eukaryotes SAR also displays the largest internal variability in A34-tRNA ratios, with a median value below the other eukaryotic phyla, again suggesting that SAR represent an intermediate step in the evolution of I34-tRNAs from bacteria to metazoans (**Figure 3**).

We then plotted the fraction of tRNAs that code for TAPSLIVR normalized by all the tRNA genes found in each organism, and found that the proportion between these two values remains constant across all groups analyzed, suggesting that ADAT does not directly affect to the proteome composition (**Figure S3**).

### Identification of TadA and ADAT genes, and phylogenetic analysis

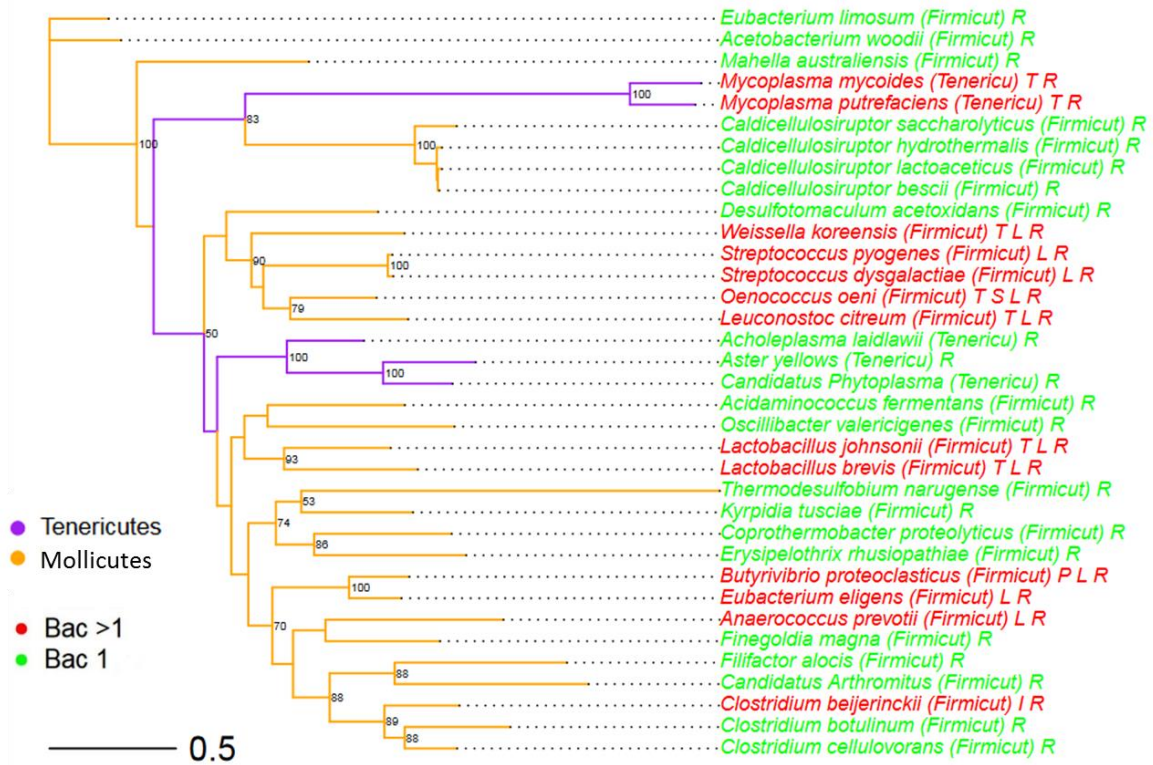
Based on sequence and structure comparisons it has been proposed that TadA evolved from the superfamily of cytidine deaminases (CDA) (Gerber and Keller 1999), a relationship reflected in the conservation of important structural and functional motifs in the sequence of both enzymes. ADAT evolved later through duplication of a TadA gene, and it retains a detectable sequence identity with CDA. We asked whether the variation in A34-tRNA diversity among different bacterial and eukaryotic phyla would be reflected in the evolution of their corresponding deaminases, TadA and ADAT respectively. To answer this question, we set up to analyze the evolution of the two enzymes using molecular phylogenetics.

A required initial step for the phylogenetic analyses of TadA and ADAT is the reliable identification of the genes coding for these enzymes, and the exclusion of CDA sequences in the library of genes used for the analysis. To this end, we developed a search pipeline designed to identify, for each available bacterial or eukaryotic genome, the sequence more likely to correspond to TadA, ADAT2 or ADAT3 (**Figure S4**). In this analysis, each putative TadA or ADAT sequence identified is compared to *bona fide* sequences from the CDA superfamily that contain a similar deaminase domain, including cytosine deaminases, cytidine deaminases, deoxycytidine deaminases, adenosine deaminases, C to U RNA deaminases and guanine deaminases.

The application of this search strategy to the genomes used in this study revealed that the vast majority of species whose genome contains one or more genes coding for A34-tRNAs also contain genes that are more similar to adenosine deaminases than to cytidine deaminases, and these were accepted as either TadA, ADAT2, or ADAT3. In those bacterial species that lack A34-tRNAs genes this trend is reversed, and we could only detect genes that display a higher sequence similarity for CDA than for TadA. This suggests that all bacterial species that lack A34-tRNAs also lack TadA, again reinforcing the possibility that ancestral bacteria entirely lacked the machinery to introduce I34 into tRNAs (**Figure 1** Adatness and **Figure S5**).

Once a reliable dataset of bacterial TadA, and eukaryotic ADAT2 and ADAT3 sequences were obtained, we proceeded to align them using the conserved structural and functional sequence motifs as references to guide the alignment (**Figure S6**). These alignments were then used to construct molecular phylogenies by the maximum likelihood method. The sequence identification process, and the ensuing phylogenetic analysis of TadA sequences, revealed that the TadA sequences obtained from species with A34-tRNA diversity higher than 1 do not form a monophyletic group. This rules out the possibility that the existence of different extant bacterial groups with more than one A34-tRNA is due to the lateral gene transfer of a TadA gene coding for an enzyme with increased tRNA specificity. We confirmed this observation by determining the phylogenetic relationships of a restricted set of TadA sequences from Firmicute and Tenericute species that contain one, or more, A34-tRNA genes (**Figure 4**). For example, *O. oeni* TadA (a species with four different A34-tRNAs) does not differ significantly from enzymes of related species that only contain tRNA<sup>Arg</sup><sub>AGC</sub>. Thus, TadA can increase its substrate

# TadA phylogeny

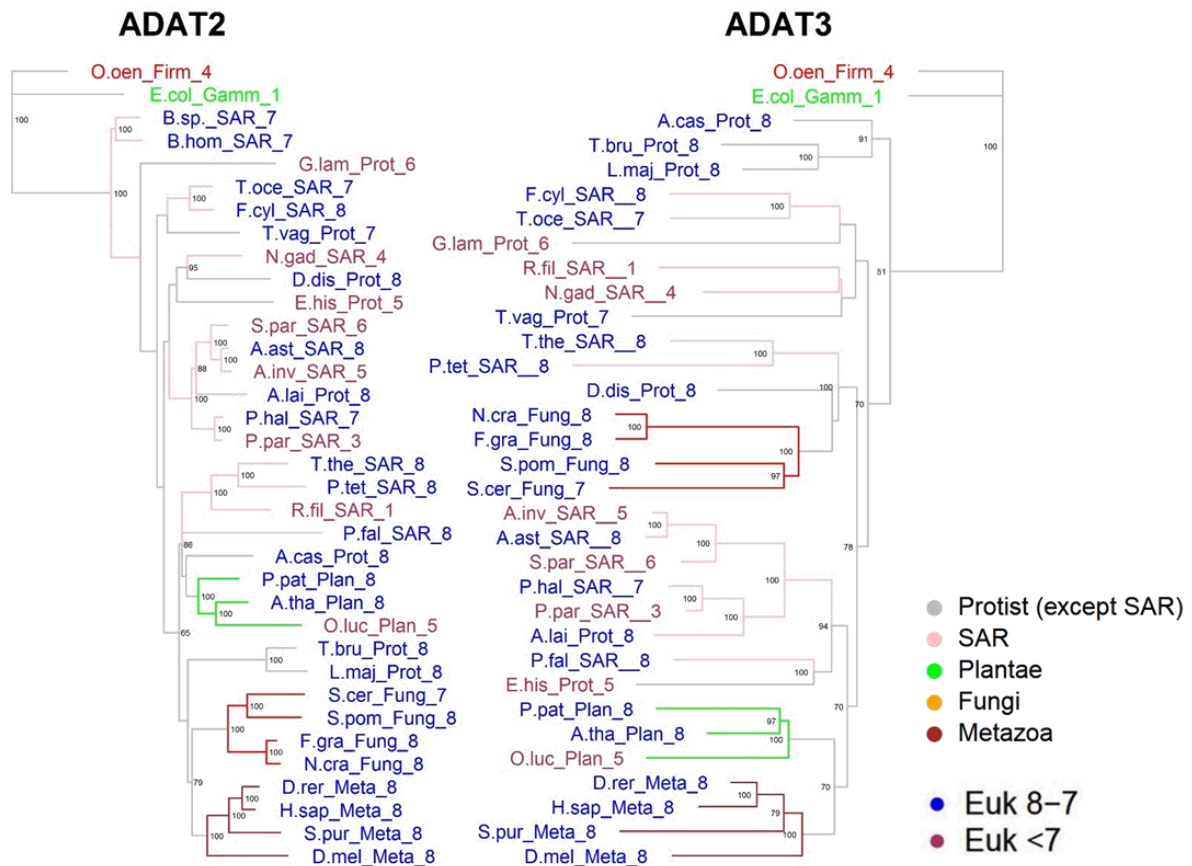


**Figure 4:** Phylogenetic tree of TadA for Firmicutes and Mollicutes phyla based on a 100 bootstrapped ML-tree. Branches are coloured according to the phyla and nodes according to the tRNA<sub>gcn</sub> (see the legends). Numbers between two branches correspond to bootstrap values higher than 50, otherwise they are not plotted. For more details see Materials and Methods.

range without major sequence or structural rearrangements, and the increase in A34-tRNA genes took place independently in different bacterial phyla (**Figure 4**).

The phylogenetic analysis of ADAT2 and ADAT3 are generally consistent with the canonical distribution of eukaryotic species (**Figure 5**), suggesting that the duplication that gave rise to these two enzymes is ancestral to extant eukaryotic phyla. Those SAR species whose genomes contain reduced numbers of A34-tRNA genes also contain genes coding for ADAT2 and ADAT3, indicating that the gene duplication that gave rise to the extant heterodimeric enzyme preceded the expansion of eukaryotic A34-tRNA genes to all TAPSLIVR isoacceptor tRNAs.



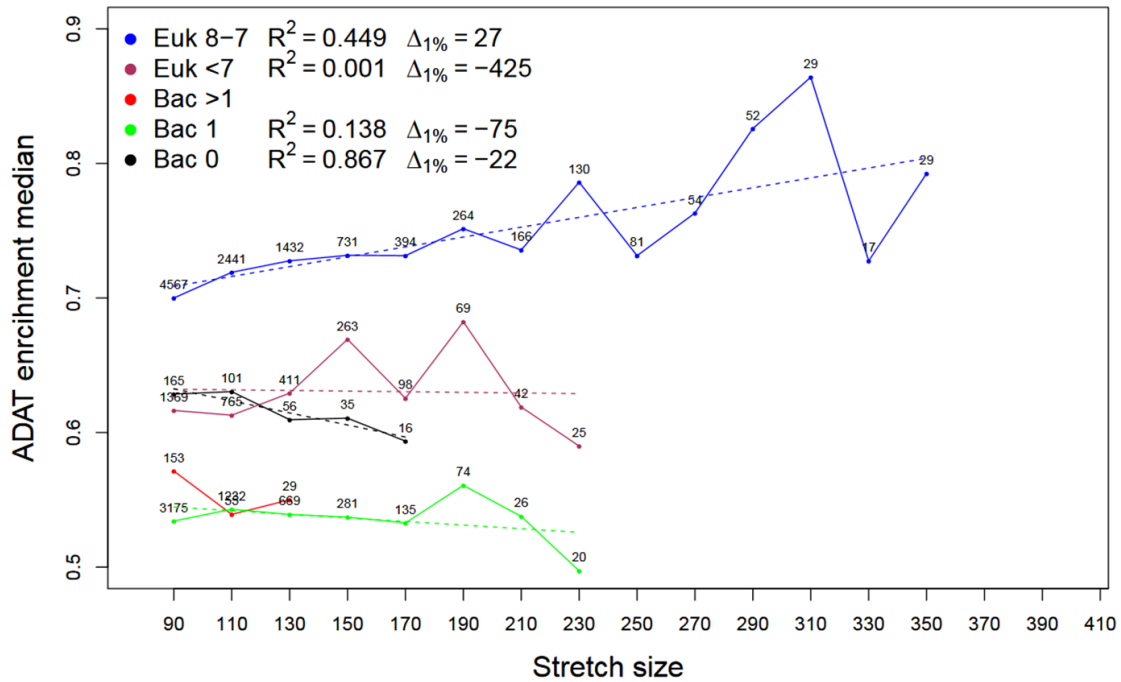


**Figure 5:** ADAT2 and ADAT3 phylogeny based on a 100 bootstrapped ML-tree. Branches are colored according to the kingdoms and nodes according to the tRNA<sub>A</sub>G<sub>C</sub> (see the legends). Numbers between two branches correspond to bootstrap values higher than 50, otherwise they are not plotted. For more details see Materials and Methods.

### TAPSLIVR-rich protein abundance and I34-sensitive codon usage.

We have shown that eukaryotic genes coding for TAPSLIVR-rich proteins predominantly use codons recognized by I34-tRNAs, while such codon usage bias (Rafels-Ybern, Torres et al. 2017) is absent in bacteria. We have proposed that this codon preference in eukaryotes is due to a functional advantage instilled by I34-tRNAs, which allows eukaryotic cells to more efficiently translate longer TAPSLIVR-rich stretches. This advantage resulted in an increase in the abundance and lengths of such proteins in eukaryotic proteomes (Rafels-Ybern, Torres et al. 2017).

We have shown here that A34-tRNA diversity divides bacteria into three groups, and that unicellular eukaryotes of the group SAR tend to have lower A34-tRNA diversity values



**Figure 6:** ADAT stretch distribution. Median values of ADAT-sensitive codon enrichment by length of ADAT amino acid stretches plotted for different groups based on tRNA colour legend described in Table 1 (see legend). Linear regression was calculated for each kingdom with more than 5 dots (colored dashed lines and legend). Numbers close to the dots correspond to the number of samples for each interval. Intervals with less than 15 samples were omitted.  $\Delta_{1\%}$  represents the number of codons needed to increase the ADAT enrichment by 1% based on the slope of the linear model.

than the rest of eukaryotes. We asked whether this heterogeneity was also reflected in the proteomes of these groups of species in terms of their composition of TAPSLIVR-rich proteins and the codon bias in the genes coding for such proteins. Using algorithms previously applied to the analysis of bacterial and eukaryotic proteomes (Rafels-Ybern, Attolini et al. 2015), we divided the transcriptomes of our species dataset in five groups, namely: eukaryotic species with A34-tRNA diversity equal to 8-7 and lower than 7, and bacterial species with A34-tRNA diversity higher than 1, equal to 1, and equal to 0.

First, we compiled different measures of abundance, length, and codon usage for the total sets of TAPSLIVR-rich proteins identified in each group (Table 2). This initial analysis revealed clear differences between bacterial and eukaryotic organisms in terms of stretch composition of their proteomes. In general, eukaryotic species are strongly enriched in terms of stretch abundance, length, and I34-dependent codon composition. While stretches are present in all eukaryotic proteomes, only 26% of bacterial species that lack A34-tRNAs present these sequences. The average number of such protein sequences in eukaryotic proteomes ranges from 87 (fungi) to 245 (plants), whereas bacterial proteomes only contain 4 to 8 TAPSLIVR stretches. Correcting the number of

detected stretches by the total number of CDS in each species reveals a maximum 8-fold difference in the abundance of stretches between eukaryotes and bacteria.

A comparison of the complete sets of TAPSLIVR stretches found in the five groups of species shows a gradual increase in codon bias for ADAT codons from bacteria with A34-tRNA diversity of 1 to eukaryotes with A34-tRNA diversity higher than 7 (**Figure 6 and 7**).

Interestingly, we detect a small number of stretches in bacteria lacking A34-tRNAs accompanied by a significant increase in I34-dependent codon bias. The analysis of the internal distribution of stretches among these species revealed that they are absent from the vast majority of species in this set (75%) (**Table 2**), concentrate in Chloroflexi and are completely absent in Thermotogae (**Figure 1**). Thus, unlike in the other groups, TAPSLIVR stretches in bacteria with A34-diversity equal to 0 are rare, and present in a restricted group of species.

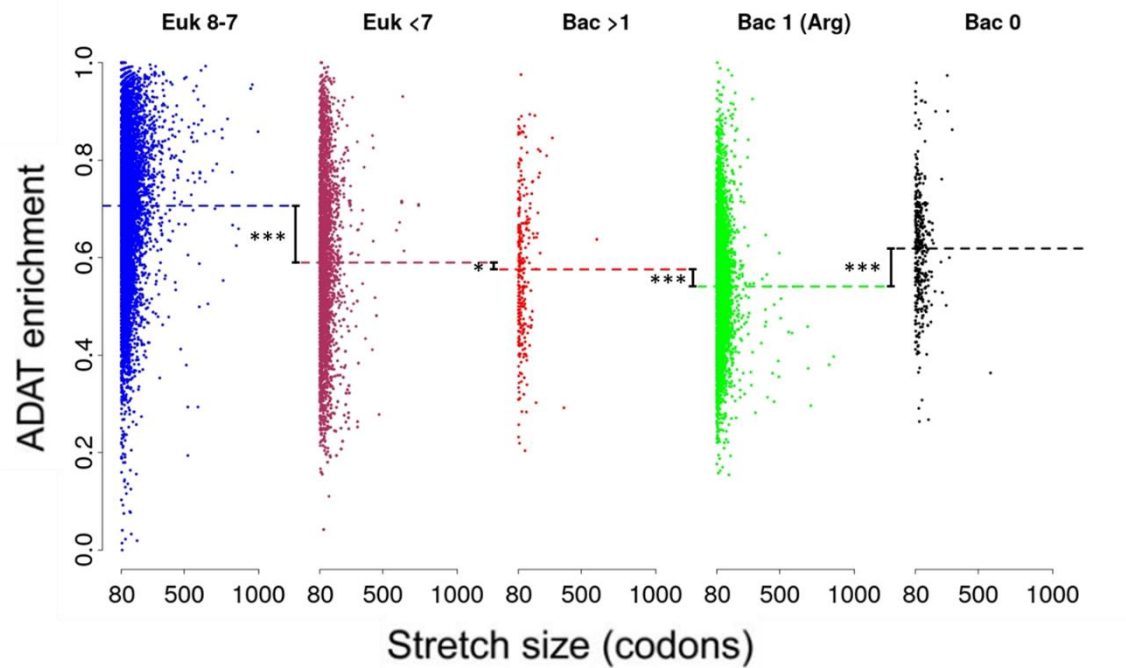
In addition, a positive correlation between stretch length I34-dependent codon usage was detected in eukaryotes with tRNA diversity values of (**Figure 6**). Interestingly, proteomes from bacteria that lack A34-tRNAs seem to contain a significant increase in the codon bias favoring codons ended in C, A or U.

A clear preference for codons recognized by I34-tRNAs is only detectable in eukaryotes, where this preference increases with the length of the TAPSLIVR stretch. This indicates that the I34 modification influences the length of TAPSLIVR stretches and the codon composition of the corresponding genes in eukaryotes.

## Discussion

The process of molecular evolution of the bacterial deaminase TadA that gave rise to eukaryotic ADAT was accompanied by a remarkable series of genomic transformations that modified the structure of eukaryotic genomes at three levels: tRNA gene diversity, tRNAome composition, and codon usage of the whole transcriptome (Novoa, Pavon-Eternod et al. 2012). Our understanding of the chain of events involved in this process is poor and, until recently, limited to the observation that duplication of a bacterial gene coding for a monomeric deaminase with a single tRNA substrate (tRNA<sup>Arg</sup><sub>ACG</sub>) was the source of the two genes that, in eukaryotes, code for the heterodimeric enzyme capable of deaminating eight different tRNA isoacceptors cognate for TAPSLIVR.

This fragmented understanding of the evolution of TadA and ADAT begs many questions, including where did TadA originate?, Is TadA functionally limited to



**Figure 7:** Characterization of proteins enriched in ADAT amino acids for different groups based on tRNA colour legend described in Table 1. Each dot represents a stretch of amino acids of length >80 with a composition of ADAT amino acids > 83%. Stretches higher than 1000 codons were not considered.

tRNA<sub>Arg</sub><sup>ACG</sup>?, How did ADAT increase its substrate repertoire from TadA?, What was the selective pressure driving the emergence of ADAT in eukaryotes?

Here we show that the evolutionary origin of TadA is probably nested within the bacterial kingdom. We show that ancestral bacterial groups such as Thermotogales and Chloroflexi, typically do not contain A34-tRNAs, nor a gene coding for TadA. The analysis of the phylogenetic distribution of all species lacking these genes reveals that a secondary loss probably also took place in isolated groups of bacterial species that underwent genome reduction events. However, the widespread absence of TadA and A34-tRNA coding genes in groups of ancestral, thermophilic bacteria supports the notion that these genes were absent at the root of the bacterial kingdom. These species do contain a gene coding for cytidine deaminase, which has been previously proposed to be the ancestral enzyme to TadA. We should like to propose that the duplication of a CDA gene leading to the emergence of TadA took place after the divergence of the most ancient bacterial clades.

Until now it has been generally accepted that the duplication that gave rise to TadA generated an enzyme functionally limited to tRNA<sub>Arg</sub><sup>ACG</sup>, and that this tRNA was the only I34-tRNA found in prokaryotes. We now report that several bacterial species contain additional A34-tRNA coding genes and that other I34-tRNAs are present in bacteria, as

**Table 2:** Statistics from Bacteria and Eukarya ADAT stretches divided in different groups of interest.

	Eukarya					Bacteria		
	Fungi	Metazoa	Plantae	Protist	SAR	black	green	red
Num. organisms	13	16	9	15	36	86	794	76
with streches	13 (100%)	16 (100%)	9 (100%)	15 (100%)	36 (100%)	22 (26%)	466 (59%)	33 (43%)
Num. stretch	1131	2733	2208	3174	6489	417	6194	295
Mean stretch size	117±61	124±68	115±48	120±64	116±51	114±43	106±40	110±48
Mean stretch enrich.	0.761±0.10	0.762±0.12	0.619±0.16	0.7±0.15	0.63±0.15	0.619±0.11	0.541±0.11	0.576±0.14
Mean CDS enrich.	0.706±0.04	0.692±0.03	0.603±0.07	0.667±0.08	NA	0.686±0.05	0.630±0.08	0.712±0.06
Num. CDS	1.25E+05	4.10E+05	2.40E+05	2.26E+05	6.29E+05	1.83E+05	2.93E+06	1.76E+05
Stretch/CDS	0.0090	0.0067	0.0092	0.0140	0.0103	0.0023	0.0021	0.0017
Stretch/organisms	87	171	245	212	180	5	8	4

we showed for *O. oeni*. Indeed we have detected several genes coding for A34-tRNA<sup>Ile</sup>, A34-tRNA<sup>Leu</sup>, A34-tRNA<sup>Ser</sup>, A34-tRNA<sup>Thr</sup>, and A34-tRNA<sup>Pro</sup>, and we have demonstrated that in *O. oeni* tRNA<sup>Arg</sup><sub>ACG</sub> and tRNA<sup>Leu</sup><sub>AAG</sub>, but not tRNA<sup>Ser</sup><sub>AGA</sub> and tRNA<sup>Thr</sup><sub>AGU</sub>, are modified to I34. Although at this stage we cannot rule out the possibility that the I34 modification in *O. oeni* tRNA<sup>Leu</sup><sub>AAG</sub> is catalyzed by a different enzyme than TadA, our results indicate that the increase in the repertoire of I34-modified tRNAs probably begins with the appearance of A34-tRNAs that are initially not modified and evolve into deaminase substrates possibly through a process of enzyme-substrate co-evolution.

Such process of co-evolution possibly took place in eukaryotes with regards to ADAT's tRNA substrates, which must have logically increased in diversity from a reduced number of original substrates. Our analysis of eukaryotic genomes reveals that the SAR group of nucleated organisms (Stramenopiles, Alveolates, and Rhizaria) holds a high apparent heterogeneity in terms of A34-tRNA diversity that is greatly reduced in the other sets of eukaryotes analyzed in this study. However, our analysis of the A34-tRNA set in the alveolate *Tetrahymena thermophila* demonstrates that this species holds a full complement of I34-modified tRNAs cognate for TAPSLIVR. Thus, the genomic analysis using tRNA-Scan suggests that SAR species exist with lower number of ADAT substrates, but our experimental results show that other species within have acquired a full complement of modified I34-tRNAs. We cannot rule out the possibility tRNA-Scan is unable to identify existing A34-tRNAs in SAR genomes but, barring this possibility, our data points at the SAR group as the most likely set of eukaryotes where preliminary stages in ADAT evolution may be discovered.

The evolution of the deamination activity from TadA to ADAT, and the concomitant increase in A34-tRNA diversity, were accompanied by a dramatic increase in A34-tRNA ratios of eukaryotic genomes. Indeed A34-tRNA coding genes that are mostly absent in bacteria became the most abundant among genes coding for TAPSLIVR isoacceptor tRNAs. We extended this analysis to our genome set and discovered that bacteria with

increased A34-tRNA diversity do not follow this trend of A34-tRNA gene enrichment. Thus, the selective drive that drove this enrichment in eukaryotic genomes did not apply to bacteria that acquired additional I34-tRNAs.

In eukaryotes, the SAR group once again displayed an intermediate situation between the average A34-tRNA ratios found in bacteria and those found in fungi, plants, metazoans, and the rest of protists. Again, this situation is suggestive of SAR representing an intermediate state in the evolution of eukaryotic genomes. The calculation of A34-ratios depends upon the ability of tRNA-SCAN to identify the relevant tRNAs; this interpretation suffers from the same caveat described above.

We have shown in the past that the emergence of ADAT in eukaryotes also correlates with an increase in the abundance of TAPSLIVR-rich proteins in these organisms, and with a shift in the codon usage of the transcripts of these same proteins that favor the utilization of I34-tRNAs for their translation. TAPSLIVR-rich proteins are, in average, increased 11-fold in eukaryotic genomes after their numbers are corrected by genome size. The average length of such TAPSLIVR-rich stretches is also increased by 20% in eukaryotes, and the codon usage in favor of I34-tRNAs increases gradually with the length of eukaryotic stretches, but not of bacterial ones.

These observations prompted us to hypothesize that the selective advantage driving the increase of tRNA substrates of ADAT, and their genomic enrichment, was the ability that I34-tRNAs confer to translate highly TAPSLIVR-repetitive transcripts. We envisage that I34-tRNAs represented a functional improvement over pre-existing translational machinery, in a manner comparable to the advent of EF-P as a factor to facilitate the translation of consecutive proline codons (Doerfel, Wohlgemuth et al. 2013; Ude, Lassak et al. 2013; Lassak, Wilson et al. 2016), or to the modifications in the tRNAome of arthropod salivary glands that allow the synthesis of the protein components of silk (Chevallier and Garel 1982; Li, Ye et al. 2015; Ribas de Pouplana, Torres et al. 2017).

We have characterized the transcriptomes of the genomes analyzed in this work in terms of the abundance of TAPSLIVR-rich sequences, and their codon usage with regards to I34-tRNA-dependent codons. As previously described, we find the abundance of TAPSLIVR-rich proteins to be much higher in eukaryotes than in bacteria. In bacteria, TAPSLIVR-rich proteins are found predominantly in species with a single A34-tRNA gene (tRNA<sup>Arg</sup><sub>GCA</sub>), where 60% of all the genomes analyzed contain genes with TAPSLIVR stretches. This frequency is reduced to 45% in genomes with A34-tRNA diversity higher than 1, and to 25% in bacterial species that lack A34-tRNAs ([Table 2](#)). It should be noted that among species lacking A34-tRNAs, TAPSLIVR stretches are

persistently found in the Chloroflexi phylum, whereas are completely absent in Thermotogales. The fact that we cannot detect any correlation between codon usage preference for I34-tRNAs and the length of transcripts coding for bacterial TAPSLIVR-rich proteins suggests that I34 does not influence the ability of prokaryotes to synthesize these polypeptides.

In eukaryotes, on the other hand, TAPSLIVR-rich proteins are universally distributed and up to 20 fold more abundant. Also, the codon usage in eukaryotic sequences for stretch regions is biased towards triplets decoded by I34-tRNAs, and this bias increases with the length of the TAPSLIVR stretch. These observations support the idea that I34 in eukaryotes promotes the abundance of proteins that contain TAPSLIVR stretches. Our results indicate that the evolutionary advantage conferred by ADAT, and the deamination of A34 in eukaryotic tRNAs cognate for TAPSLIVR, was the optimization of translating genes encoding for proteins highly enriched in these amino acids.

The existence of species-specific translation machinery adaptations acquired to synthesize proteins with low complexity sequences has been demonstrated in arthropods, and offers a possible explanation for the expansion of I34 tRNAs in eukaryotes. In this context, I34-tRNAs may result in more efficient translation elongation of sequences that are extremely biased in TAPSLIVR codons. The functional characteristics of such proteins may, in turn, confer new properties to eukaryotic cells, and would in turn drive the enrichment in this types of polypeptides that we have detected. Thus, we would like to propose that, in opisthokonta eukaryotes, the modification of A34 to inosine in the anticodons of tRNAs led to an increased efficiency in the synthesis of TAPSLIVR-rich proteins, and that the selective advantage of such polypeptides drove their enrichment in eukaryotic proteomes, the modification of codon usage of their corresponding genes, and the enrichment of A34-tRNA coding genes.

## Materials and methods

### Definitions

ADAT amino acids were defined as those amino acids charged to tRNAs that were modified by ADAT in most Eukarya, and corresponds to Thr, Ala, Pro, Ser, Leu, Ile, Val, and Arg (TAPSLIVR). ADAT codons were defined as those codons that code for ADAT amino acids and corresponds to 37 codons (**Table S1**, blue). ADAT-sensitive codons were defined as the subset of ADAT codons susceptible to be recognized by modified I34 tRNAs and corresponds to the 24 codons (**Table S1**, dark blue only). ADAT stretches

were defined as those regions within the CDSs that were particularly enriched in ADAT codons and were found using the 'Running Windows' method with the same parameters used in (Rafels-Ybern, Attolini et al. 2015). For each ADAT stretch, we defined the ADAT enrichment as the fraction of ADAT-sensitive codons / ADAT codons. In **Figure 1**, Adatness is defined as the e-value for ADAT2 (or TadA in bacteria) divided by the corresponding e-value for the CDA superfamily. In **Figure 6**,  $\Delta_{1\%}$  represents the number of codons needed to increase the ADAT-enrichment by 1% based on the slope of the corresponding linear model.

## Genomic dataset retrieval

We have analysed the transcriptome, genome and proteome of 150 eukaryotic species and 956 bacterial species. We analyse a total of 3.6 Gb of bacterial genomes and 18.8 Gb of eukaryotic genomes. We analyse a total of 3.28 million of bacterial CDSs and 1.79 million of eukaryotic CDSs. Only CDSs with a start codon, a stop codon and a number of nucleotides multiple of 3 were used in the analysis. 12% of eukaryotic CDSs and 20% of bacterial CDSs were discarded because they did not fulfil one of the three requisites. Eukaryotic sequences were downloaded from Ensembl website except for those sequences from the SAR clade. Bacterial sequences and SAR sequences, were downloaded from NCBI website. Proteins were analysed in NR database from NCBI.

## Identification of stretches by the 'Running Windows' method

To study in more detail the presence of ADAT stretches, the Running Windows Method was applied, based on similar methodologies (McDonald 1996; Hutter, Vilella et al. 2006; Librado and Rozas 2009). For each CDS from each organism, a window (fragment of the sequence with a fixed length) slides down codon by codon from the beginning to the end of the sequence. For each position, the percentage of ADAT codons was calculated and represented with respect its location (Rafels-Ybern, Attolini et al. 2015). We fixed the window size to 80 codons, and we considered that a window was enriched in ADAT codons when they represent more than 83% (i.e. >67 codons from the window were ADAT codons) based on our previous analysis of the human proteome (Rafels-Ybern, Attolini et al. 2015). We defined an ADAT stretch as those regions corresponding to a window or a set of consecutive windows, enriched in ADAT codons. Two or more windows were considered consecutive if their intersection in the CDS was not void.

## Statistical analyses

Significant differences observed in **Figures 3, 7** and **S3** were obtained using one-tail Wilcoxon test with confidence level at 0.95, where multiple testing correction were taken



into account, using the *wilcox.test()* function from *R* language (Novoa and Ribas de Pouplana 2012; Team 2014).

In order to perform robust statistics in **Figure 6** we discarded data from those intervals with less than 15 samples (full data can be shown in **Figure S7**) and linear regression was made for groups with 5 or more points. Linear regression was fitted using the *lm()* function from *R* language.

## Gene and protein identification methods

tRNA genes were predicted with *tRNA scan-SE* software (Fichant and Burks 1991; Nawrocki and Eddy 2013). Options *-B* for bacterial genomes and *-G* for eukaryotic genomes were used. Pseudogenes and undetermined tRNAs from standard output were discarded from the analysis. tRNA gene copy number for each organism was inferred based on the number of different A34 tRNA genes that are present for each organism.

In order to extract the TadA protein sequence from each bacterial organism (the procedure was analogous for ADAT2 and ADAT3 for each eukaryotic organism) we followed the schema in **Figure S4**. First, we ran a protein-protein BLAST (pBLAST) using the *BLAST+* software (Camacho, Coulouris et al. 2009) of a known set of TadA proteins against the non-redundant protein sequences database (NR) (**Figure S4**, step 1) obtaining a list (**Figure S4**, Hits A) with 120000 hits including all kind of organisms from NCBI. NR was downloaded from (website) and has a size of 110 Gb. The results were filtered for our 956 bacterial organisms taking the hit with the best e-value (**Figure S4**, step 2). A putative TadA sequence were found in 728 out of 956 bacterial organisms (76%) with e-values always lower than  $10^{-6}$  (**Figure S4**, Hits B). This method allowed to compare all the organisms and its different strains at once and compare its e-values in a single file simplifying the process. Afterwards, each putative TadA sequence was double checked by doing a second BLAST against a set of known cytidine deaminase proteins (CDAs) (**Figure S4**, step 3) because TadA (and ADAT) contain CDA-conserved motifs, suggesting that they were evolutionarily related (Gerber and Keller 2001; Rubio, Pastar et al. 2007). By doing this BLAST, we could identify a *bonafide* TadA sequence or reject it because it was closer to a CDA. The putative TadA sequence was accepted only if the e-value from TadA was higher than the e-value from CDA (i.e.  $Adatness > 1$ ) (**Figure S4**, step 4).

## Phylogenetic trees reconstruction

From *the bonafide* ADAT and TadA candidates ( $Adatness > 1$ ), we selected a representative set of 47 TadA sequences, 33 ADAT2 sequences and 33 ADAT3

sequences using the *T-coffee* software with the option *+trim\_n* (Notredame, Higgins et al. 2000). This option selects the best informative sequences maximizing their variability. The representative set of sequences were aligned using *T-coffee* software with the mode *expresso* (Armougom, Moretti et al. 2006). With this method, a BLAST ran for each sequence against the PDB dataset, finding a similar sequence (35% identity by default) that was used as a template for doing an structural MSA. Afterwards, >50% gapped columns were removed from the MSA matrix. Maximum likelihood (ML) trees were performed using *RaxML* software (Stamatakis 2014). 100 replicas were computed using the options *GAMMA* as a model of rate heterogeneity and *LG* as a model of substitutions. A consensus tree was build using the *-f b* option that provides the bootstrap support values on the best-known ML tree from all the replicas. ML trees were plotted using the libraries *ape*, *numDeriv*, *phytools* and *phangorn* from *R* software (Team 2014). The standard phylogenetic tree from **Figure 1** was build using the online tool *PhyloT* (Letunic 2015).

## Experimental procedures

Inosine 34 detection by the splinted ligation method was performed as follows: (Torres et al manuscript in preparation) (1<sup>st</sup>) 2 µg total RNA were digested with Endonuclease V (Thermo Scientific) overnight at 37 °C following the manufacturer's protocol. Endonuclease V reactions were then purified using the MiTotal RNA extraction kit (Viogene) and the obtained RNA was quantified using a Nanodrop ND-1000. (2<sup>nd</sup>) tRNA-specific Bridge oligo Working Solution was prepared in 10X Capture Buffer (100 mM Tris-HCl pH 7.5; 750 mM KCl) to a final concentration of 100 nM. Ligation oligo was 5'-end radiolabelled with T4 Polynucleotide Kinase (Takara) following the manufacturer's protocol. (3<sup>rd</sup>) 1.8 µg of Endonuclease V digested RNA in 9 µL water was hybridized with 1.1 µL tRNA-specific Bridge oligo [100 nM] and 1 µL 32P-labelled Ligation oligo. Samples were heated at 95 °C for 5 minutes and were left at room temperature overnight to slowly cool down to obtain efficient RNA/DNA hybridization. 1 µL T4 DNA Ligase (Fermentas), 1.5 µL T4 DNA Ligase Buffer 10X (Fermentas) and 2.5 µL water was added to each reaction and was incubated for 1 hour at 37 °C. After incubation, 1 µL CIP (New England Biolabs) was added to each reaction to remove the 5'-end 32P from free ligation oligo. Samples were heated at 95°C for 5 minutes to inactivate the enzymes and were stored at -20° until ready to use. (4<sup>th</sup>) 4 µL of tRNA loading buffer (8M urea, 30% glycerol; 20% formamide; bromophenol blue/xylene cyanol) was added to each reaction and 20 µL of sample was loaded into a 12 % polyacrylamide gel containing 8 M urea. Samples were run for 1.5 hour at 120V; and the gels were directly exposed to a Typhoon Film for 30 minutes. Radioactive signals were detected with a Typhoon Scanner.

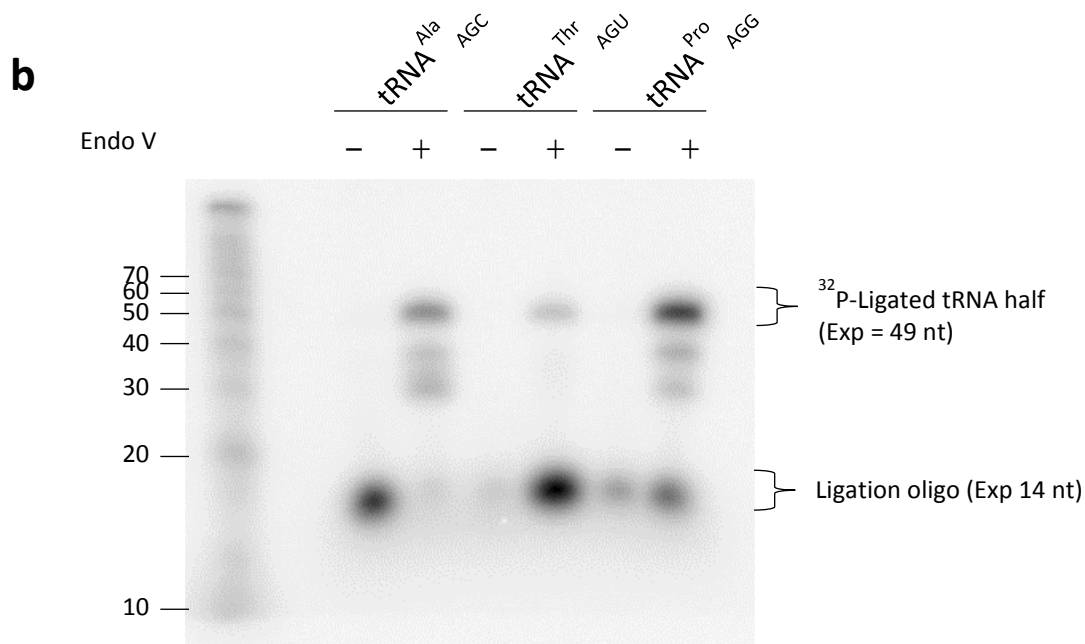
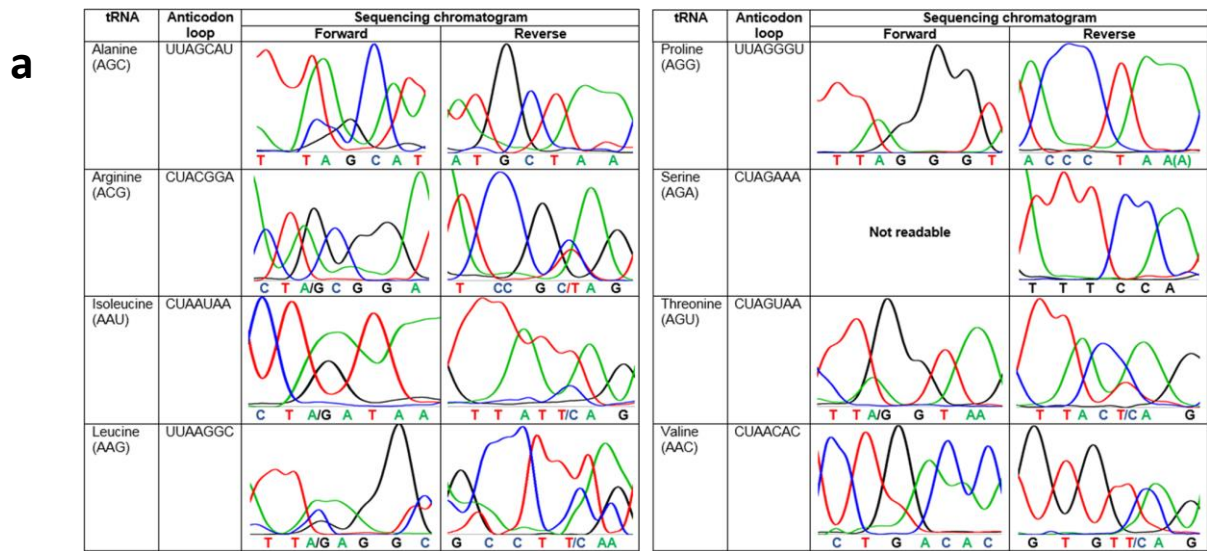
Specific A34-encoded tRNA isoacceptors were PCR amplified using the oligos in [Table S2](#), and amplicons were subjected to Sanger sequencing.

## References

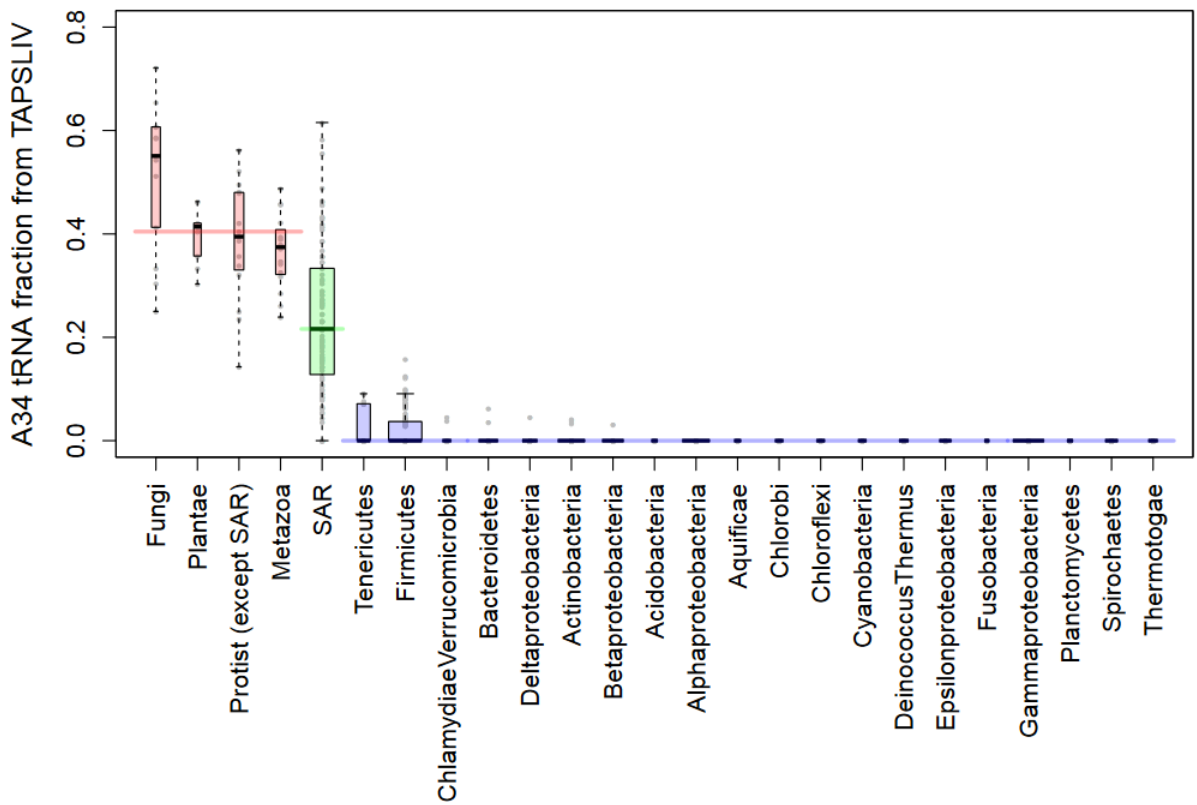
- Armougom, F., S. Moretti, et al. (2006). "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee." *Nucleic Acids Res* **34**(Web Server issue): W604-608.
- Auxilien, S., P. F. Crain, et al. (1996). "Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of inosine 34 in the anticodon of transfer RNA." *J Mol Biol* **262**(4): 437-458.
- Camacho, C., G. Coulouris, et al. (2009). "BLAST+: architecture and applications." *BMC Bioinformatics* **10**: 421.
- Crick, F. H. (1966). "Codon--anticodon pairing: the wobble hypothesis." *J Mol Biol* **19**(2): 548-555.
- Chevallier, A. and J. P. Garel (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**(3): 477-482.
- Doerfel, L. K., I. Wohlgemuth, et al. (2013). "EF-P is essential for rapid synthesis of proteins containing consecutive proline residues." *Science* **339**(6115): 85-88.
- El Yacoubi, B., M. Bailly, et al. (2012). "Biosynthesis and function of posttranscriptional modifications of transfer RNAs." *Annu Rev Genet* **46**: 69-95.
- Elias, Y. and R. H. Huang (2005). "Biochemical and structural studies of A-to-I editing by tRNA:A34 deaminases at the wobble position of transfer RNA." *Biochemistry* **44**(36): 12057-12065.
- Fichant, G. A. and C. Burks (1991). "Identifying potential tRNA genes in genomic DNA sequences." *J Mol Biol* **220**(3): 659-671.
- Gaston, K. W., M. A. Rubio, 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**(20): 6740-6749.
- Gerber, A. P. and W. Keller (1999). "An adenosine deaminase that generates inosine at the wobble position of tRNAs." *Science* **286**(5442): 1146-1149.
- Gerber, A. P. and W. Keller (2001). "RNA editing by base deamination: more enzymes, more targets, new mysteries." *Trends Biochem Sci* **26**(6): 376-384.
- Grosjean, H., S. Auxilien, et al. (1996). "Enzymatic conversion of adenosine to inosine and to N1-methylinosine in transfer RNAs: a review." *Biochimie* **78**(6): 488-501.
- Grosjean, H., M. Breton, et al. (2014). "Predicting the minimal translation apparatus: lessons from the reductive evolution of mollicutes." *PLoS Genet* **10**(5): e1004363.
- Hutter, S., A. J. Vilella, et al. (2006). "Genome-wide DNA polymorphism analyses using VariScan." *BMC Bioinformatics* **7**: 409.
- Jackman, J. E. and J. D. Alfonzo (2013). "Transfer RNA modifications: nature's combinatorial chemistry playground." *Wiley Interdiscip Rev RNA* **4**(1): 35-48.
- Lassak, J., D. N. Wilson, et al. (2016). "Stall no more at polyproline stretches with the translation elongation factors EF-P and IF-5A." *Mol Microbiol* **99**(2): 219-235.
- Letunic, I. (2015). "phyloT : Phylogenetic Tree Generator. [online] Phylot.biobyte.de. ."
- Li, J. Y., L. P. Ye, et al. (2015). "Comparative proteomic analysis of the silkworm middle silk gland reveals the importance of ribosome biogenesis in silk protein production." *J Proteomics* **126**: 109-120.
- Librado, P. and J. Rozas (2009). "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data." *Bioinformatics* **25**(11): 1451-1452.

- Liu, H., Q. Wang, et al. (2016). "Genome-wide A-to-I RNA editing in fungi independent of ADAR enzymes." *Genome Res* **26**(4): 499-509.
- Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." *Nucleic Acids Res* **25**(5): 955-964.
- Maas, S. and A. Rich (2000). "Changing genetic information through RNA editing." *Bioessays* **22**(9): 790-802.
- Marck, C. and H. Grosjean (2002). "tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features." *RNA* **8**(10): 1189-1232.
- McDonald, J. H. (1996). "Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence." *Mol Biol Evol* **13**(1): 253-260.
- Nawrocki, E. P. and S. R. Eddy (2013). "Infernal 1.1: 100-fold faster RNA homology searches." *Bioinformatics* **29**(22): 2933-2935.
- Nishikura, K. (2010). "Functions and regulation of RNA editing by ADAR deaminases." *Annu Rev Biochem* **79**: 321-349.
- Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." *J Mol Biol* **302**(1): 205-217.
- Novoa, E. M., M. Pavon-Eternod, et al. (2012). "A role for tRNA modifications in genome structure and codon usage." *Cell* **149**(1): 202-213.
- Novoa, E. M. and L. Ribas de Pouplana (2012). "Speeding with control: codon usage, tRNAs, and ribosomes." *Trends Genet* **28**(11): 574-581.
- Phizicky, E. M. and A. K. Hopper (2010). "tRNA biology charges to the front." *Genes Dev* **24**(17): 1832-1860.
- Rafels-Ybern, A., C. S. Attolini, et al. (2015). "Distribution of ADAT-Dependent Codons in the Human Transcriptome." *Int J Mol Sci* **16**(8): 17303-17314.
- Rafels-Ybern, A., A. G. Torres, et al. (2017). "Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla." *RNA Biol*: 1-8.
- Ribas de Pouplana, L., A. G. Torres, et al. (2017). "What Froze the Genetic Code?" *Life (Basel)* **7**(2).
- Rubio, M. A., I. Pastar, et al. (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**(19): 7821-7826.
- Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* **30**(9): 1312-1313.
- Team, R. C. (2014). "R: A Language and Environment for Statistical Computing."
- Torres, A. G., E. Batlle, et al. (2014). "Role of tRNA modifications in human diseases." *Trends Mol Med* **20**(6): 306-314.
- Torres, A. G., D. Pineyro, et al. (2014). "A-to-I editing on tRNAs: biochemical, biological and evolutionary implications." *FEBS Lett* **588**(23): 4279-4286.
- Torres, A. G., D. Pineyro, et al. (2015). "Inosine modifications in human tRNAs are incorporated at the precursor tRNA level." *Nucleic Acids Res* **43**(10): 5145-5157.
- Towns, W. L. and T. J. Begley (2012). "Transfer RNA methyltransferases and their corresponding modifications in budding yeast and humans: activities, predications, and potential roles in human health." *DNA Cell Biol* **31**(4): 434-454.
- Tsutsumi, S., R. Sugiura, et al. (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**(46): 33459-33465.
- Ude, S., J. Lassak, et al. (2013). "Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches." *Science* **339**(6115): 82-85.
- website, N. from <https://www.ncbi.nlm.nih.gov/>.
- Weisburg, W. G., J. G. Tully, et al. (1989). "A phylogenetic analysis of the mycoplasmas: basis for their classification." *J Bacteriol* **171**(12): 6455-6467.

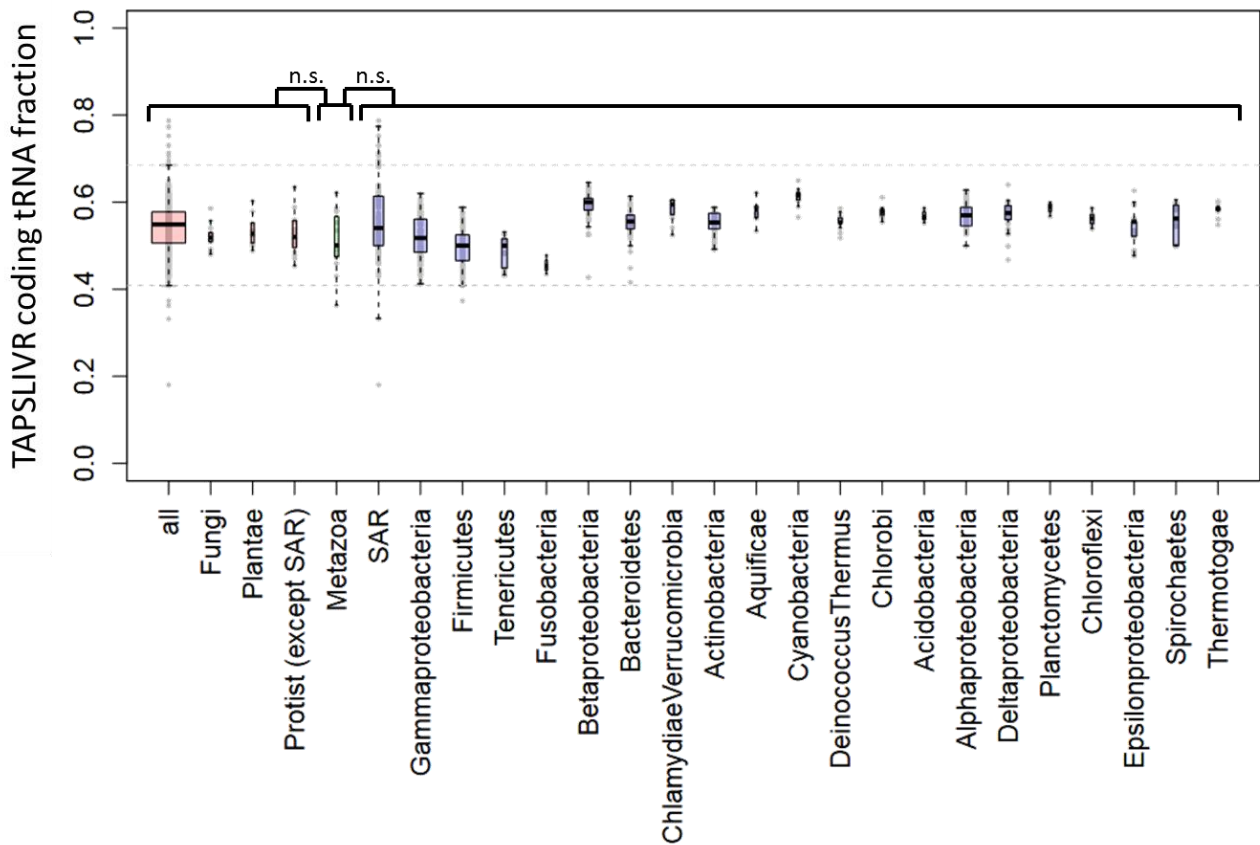
- Wolf, J., A. P. Gerber, et al. (2002). "tadA, an essential tRNA-specific adenosine deaminase from *Escherichia coli*." EMBO J **21**(14): 3841-3851.
- Wulff, B. E. and K. Nishikura (2010). "Substitutional A-to-I RNA editing." Wiley Interdiscip Rev RNA **1**(1): 90-101.
- Zhou, W., D. Karcher, et al. (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**(4): 1985-1997.



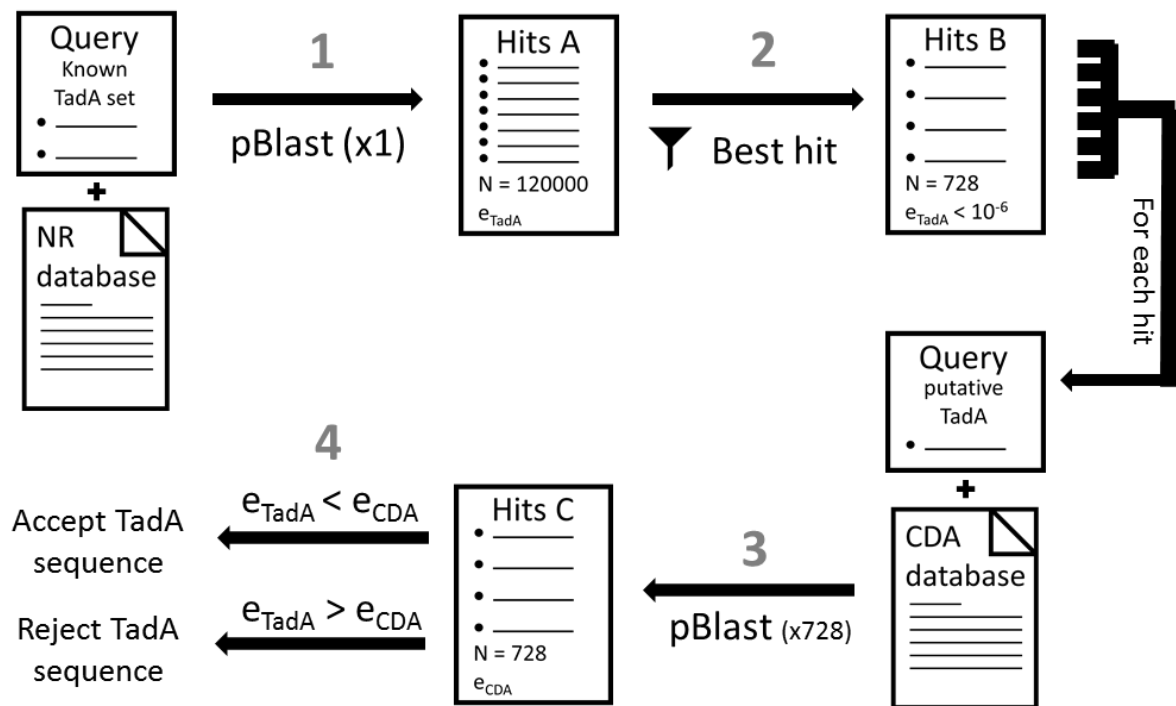
**Figure S1:** (a) Sequencing spectra of RT-PCR amplicons derived from *T. thermophila* tRNA<sup>Ala</sup><sub>AGC</sub>, tRNA<sup>Arg</sup><sub>AGG</sub>, tRNA<sup>Ile</sup><sub>AAU</sub>, tRNA<sup>Leu</sup><sub>AAG</sub>, tRNA<sup>Pro</sup><sub>AGG</sub>, tRNA<sup>Ser</sup><sub>AGA</sub>, tRNA<sup>Thr</sup><sub>AGU</sub>, and tRNA<sup>Val</sup><sub>AAC</sub>. Sequencing spectra using Forward and Reverse primers are shown. Inosine at the anticodon 'wobble' position should be detected as a Guanosine (Cytosine in the reversed strand), as opposed to the unmodified residue which is detected as Adenosine (Thymine in the reversed strand). (b) Evaluation of I34 presence/absence on *T. thermophila* tRNA<sup>Ala</sup><sub>AGC</sub>, tRNA<sup>Thr</sup><sub>AGU</sub>, and tRNA<sup>Pro</sup><sub>AGG</sub> by the splinted-ligation method for I34 detection.



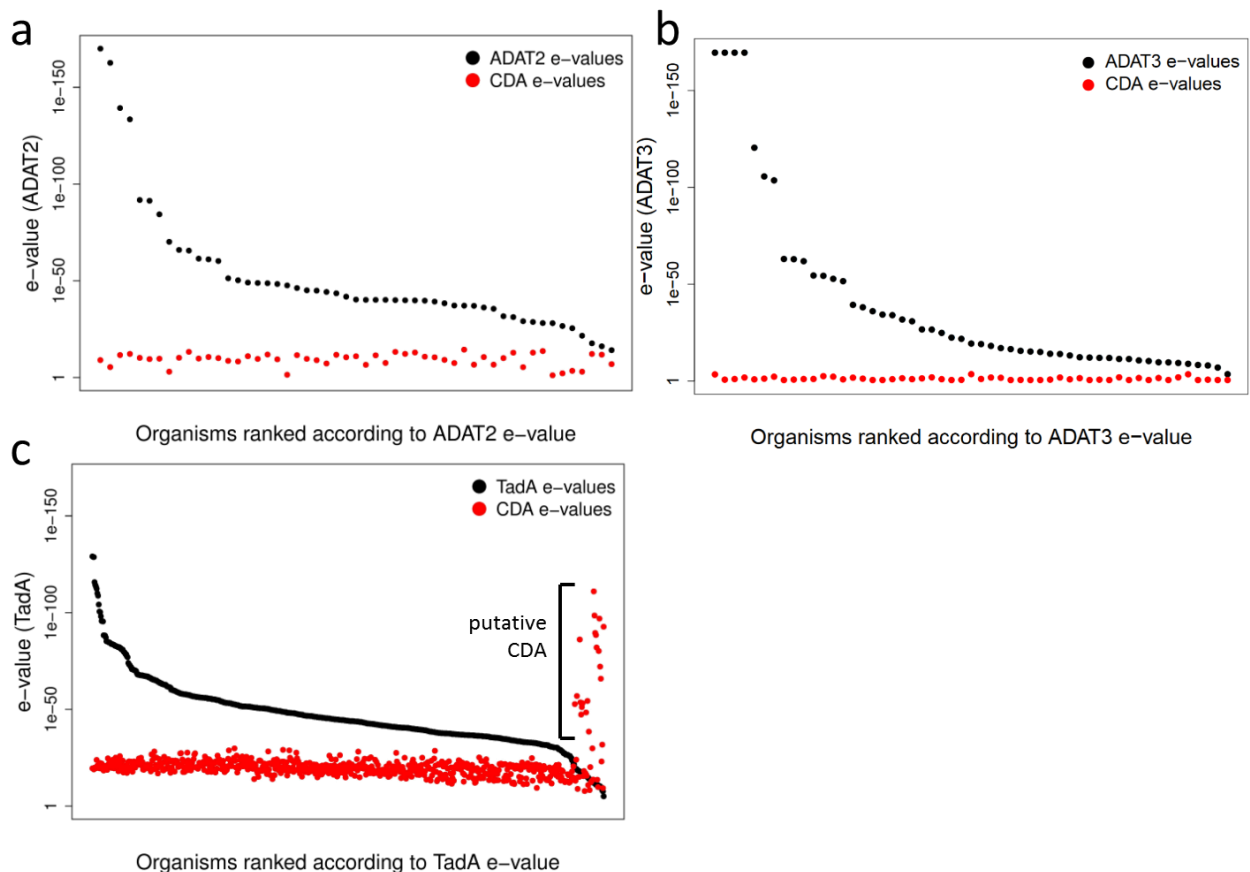
**Figure S2:** A34 tRNA ratio boxplot for eukaryotic kingdoms (red), the eukaryotic superphylum *Heterokonta* (green) and different bacterial phyla (blue). The y-axis represents the fraction of A34 tRNAs among all the isoacceptors for TAPS and LIVR amino acids. The width of each boxplot is proportional to the number of organisms analyzed. The horizontal colored lines represent the median values for the sets they highlight.



**Figure S3:** tRNA quantification boxplot for eukaryotic kingdoms (red), the eukaryotic superphylum *Heterokonta* (green) and different bacterial phyla (blue). The y-axis represents the fraction of all the isoacceptors for TAPS and LIVR amino acids among all the isoacceptors for all the amino acids. The width of each boxplot is proportional to the number of organisms analyzed. 'all' boxplot represents all the data together. Horizontal dashed lines are placed in Q1 = 0.41 and Q4 = 0.68 from 'all' boxplot.

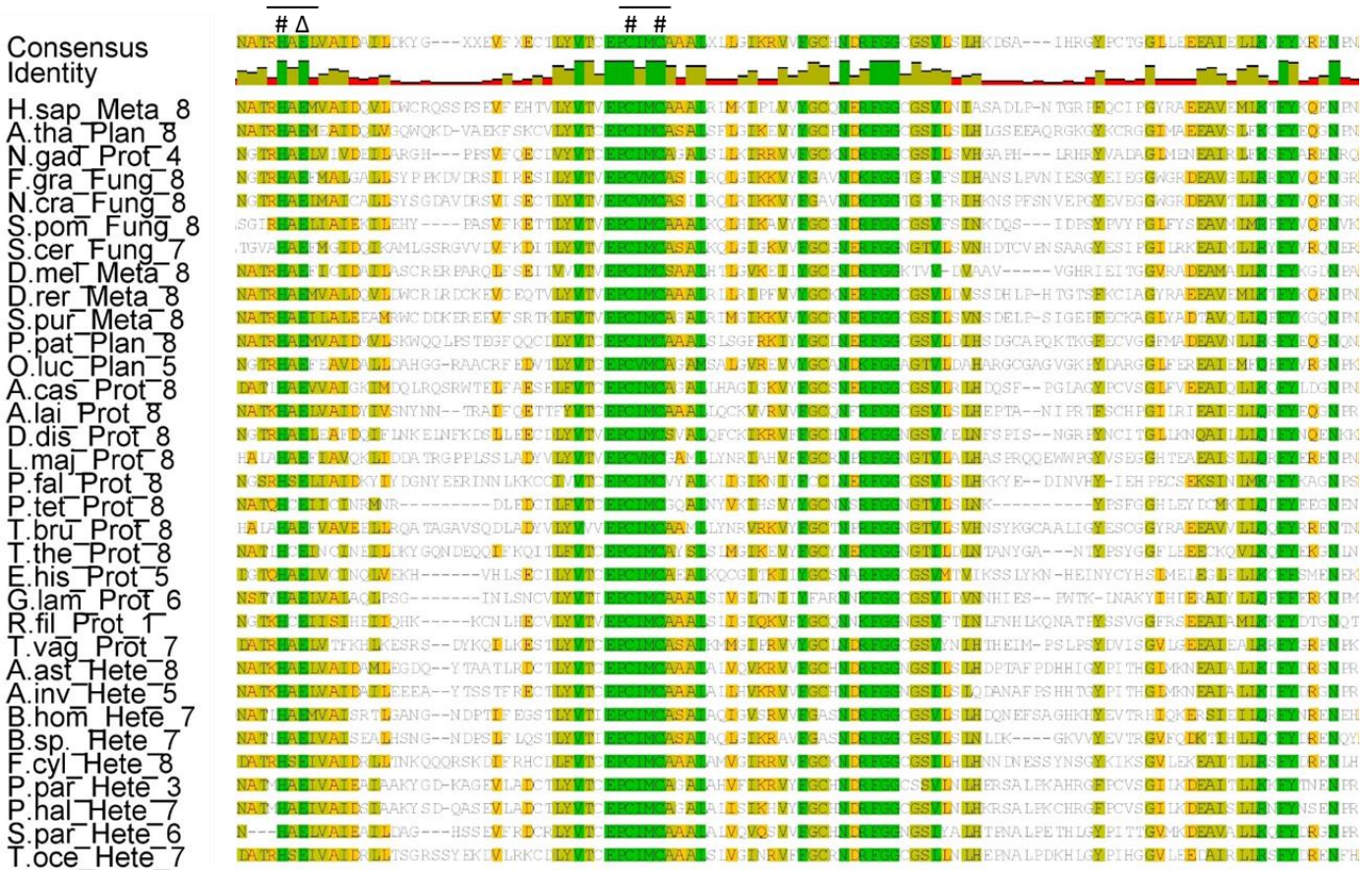


**Figure S4:** Identification of TadA sequences scheme. NR: non-redundant NCBI database.  $e_x$ : BLAST e-value from X query. CDA: cytidine deaminase. pBLAST: protein BLAST. N: number of hits. The scheme is analogous to find the ADAT2 and ADAT3 protein sequences.

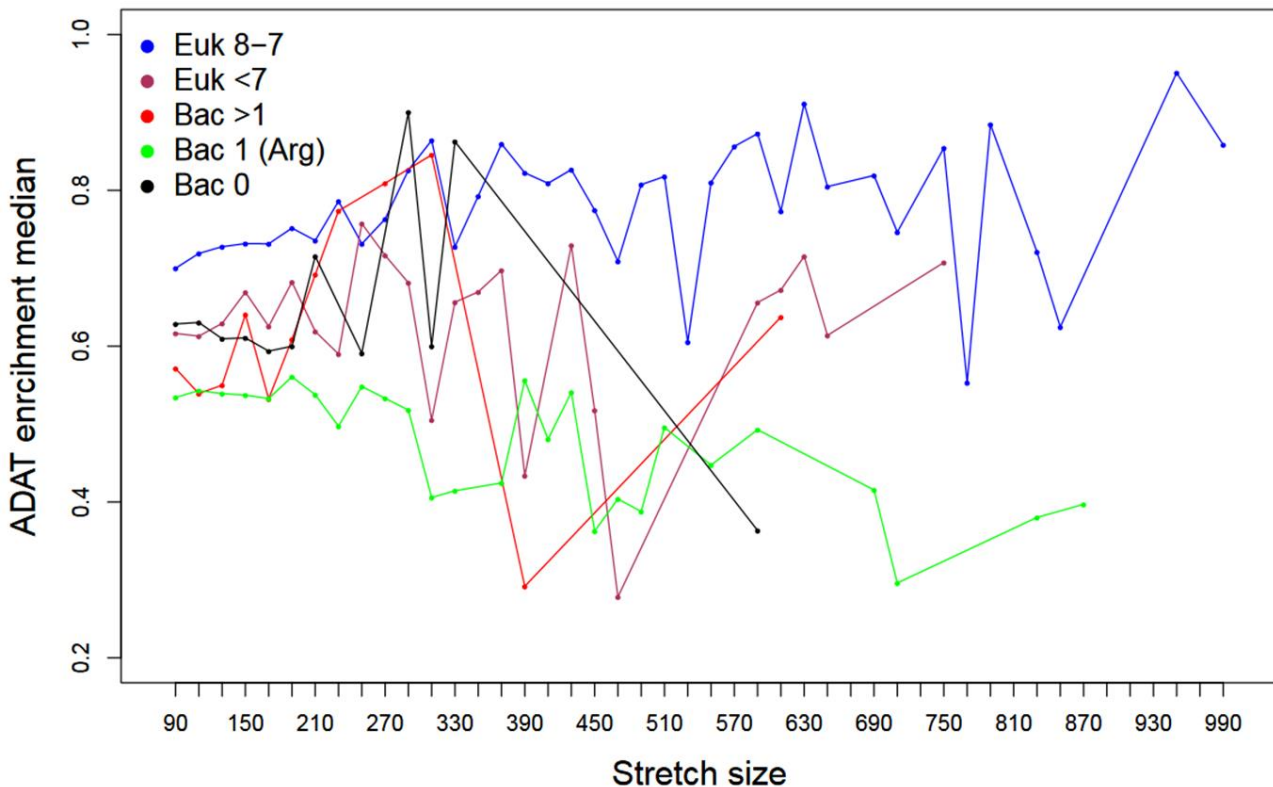


**Figure S5:** comparison of e-values between the Adenosine deaminase (black circles) and the cytidine deaminases (grey circles). The values are ranked for ADAT2 (a), ADAT3 (b) and TadA (c).





**Figure S6:** Multiple sequence alignments of ADAT2 deaminase domains. Residues with 100%, 80-100%, 60-80% and <60% of similarity are respectively framed with green, olive, orange and white. The deaminase domains are overlined, where the residues that coordinates with Zn (#) and the glutamic acid that allows the nucleophilic attack ( $\Delta$ ) are depicted. The organisms names are displayed as *X.yyy\_zzzz\_n* where *X* is the first letter of the Genre, *yyy* the first 3 letters of the specie, *zzzz* the 4 first letters of the phylum and *n* the number of different A34 tRNAs.



**Figure S7:** ADAT stretch distribution. Median values of ADAT-sensitive codon enrichment by length of ADAT amino acid stretches plotted for different groups based on tRNAGcn (see legend).

**Table S1:** The Universal Genetic Code. ADAT codons are colored in blue (dark and soft). ADAT-sensitive codons are colored in dark blue. For each codon, the amino acid that decodes is depicted in one-letter and three-letter convention.

		Second Codon Letter								
		U		C		A		G		
First Codon Letter	U	F	Phe	S	Ser	Y	Tyr	C	Cys	U
		F	Phe	S	Ser	Y	Tyr	C	Cys	C
		L	Leu	S	Ser	Stop		Stop		A
		L	Leu	S	Ser	Stop		W	Trp	G
	C	L	Leu	P	Pro	H	His	R	Arg	U
		L	Leu	P	Pro	H	His	R	Arg	C
		L	Leu	P	Pro	Q	Gln	R	Arg	A
		L	Leu	P	Pro	Q	Gln	R	Arg	G
	A	I	Ile	T	Thr	N	Asn	S	Ser	U
		I	Ile	T	Thr	N	Asn	S	Ser	C
		I	Ile	T	Thr	K	Lys	R	Arg	A
		M	Met	T	Thr	K	Lys	R	Arg	G
	G	V	Val	A	Ala	D	Asp	G	Gly	U
		V	Val	A	Ala	D	Asp	G	Gly	C
		V	Val	A	Ala	E	Glu	G	Gly	A
		V	Val	A	Ala	E	Glu	G	Gly	G

**Table S2:** oligos used for *T. thermophila* and *O. oeni* PCR amplification.

Name	Sequence	Details
oTFW_94	gaatgtcataagcgCCAAGCGAGCGCTCTACCATTG	Bridge oligo T. the/H. sapiens tRNA AlaAGC - Splinted Ligation
oAGT_230	gaatgtcataagcgCCAGGCGAATGCTCTAACCCTG	Bridge oligo T. the tRNA ThrAGT - Splinted Ligation
oAGT_231	gaatgtcataagcgCCAACGAGAATCATGCCACTAG	Bridge oligo T. the tRNA ProAGG - Splinted Ligation
oAGT_312	gaatgtcataagcgGCAGTCAGACGCTCTATCCAATTG	Bridge oligo O. oeni tRNA ArgACG - Splinted Ligation
oAGT_313	gaatgtcataagcgTCAATCTGGCGCTCTGCCAATTC	Bridge oligo O. oeni tRNA LeuAAG - Splinted Ligation
oAGT_314	gaatgtcataagcgCCAGACCGACCCCTTCAGCCAC	Bridge oligo O. oeni tRNA SerAGA - Splinted Ligation
oAGT_315	gaatgtcataagcgCCAGTGAAGTGCTCTAGCCAAC	Bridge oligo O. oeni tRNA ThrAGT - Splinted Ligation
oTFW-31	AGCTTAATACGACTCACTATAGGGGATCTAGCTCA	FWD primer T. thermophila PCR Ala
oTFW-36	GATCCACATGTTGGTGGAGAACCTGGGCATT	RVR primer T. thermophila PCR Ala
oTFW-37	AGCTTAATACGACTCACTATAGGGGTGATGG	FWD primer T. thermophila PCR Arg
oTFW-42	GATCCCCTGGCGAGATGAGCAGGACTCGAAC	RVR primer T. thermophila PCR Arg
oTFW-43	AGCTTAATACGACTCACTATAGCTCGGGTAGCTCAG	FWD primer T. thermophila PCR Ile
oTFW-48	GATCCCCTGGTGTCCGGGAGGGGCTTGAAC	RVR primer T. thermophila PCR Ile
oTFW-49	AGCTTAATACGACTCACTATAGATGAAGTGGCCGAG	FWD primer T. thermophila PCR Leu
oTFW-54	GATCCCCTGGTGTGAAGGCGAGATTCGAACT	RVR primer T. thermophila PCR Arg
oTFW-55	AGCTTAATACGACTCACTATAGGGTGTGGTGC	FWD primer T. thermophila PCR Pro
oTFW-60	GATCCCCTGGGGGTCGTCGAGAAATCGA	RVR primer T. thermophila PCR Pro
oTFW-61	AGCTTAATACGACTCACTATAGACAATTTGTCCGAG	FWD primer T. thermophila PCR Ser
oTFW-66	GATCCCCTGGCGACAACATGCAGGATTCTGA	RVR primer T. thermophila PCR Ser
oTFW-67	AGCTTAATACGACTCACTATAGCGCTTTAGCTC	FWD primer T. thermophila PCR Thr
oTFW-72	GATCCCCTGGAGCCACTTGGCGGGATTG	RVR primer T. thermophila PCR Thr
oTFW-73	AGCTTAATACGACTCACTATAGATTCCTTAGTG	FWD primer T. thermophila PCR Val
oTFW-78	GATCCCCTGGTGATTCTCCGAGGTTGA	RVR primer T. thermophila PCR Val
oAGT_296	CAGGAACAGCTATGACCCACCATTAGCGCAATTGG	FWD primer tRNA Arg ACG Oenococcus oeni with M13-RP adaptor for sequencing
oAGT_297	GCACCATGTAGGAGTCGAAC	RVR primer tRNA Arg ACG Oenococcus oeni
oAGT_298	CAGGAACAGCTATGACCGACGTGGCGGAATTGGCAG	FWD primer tRNA Leu AAG Oenococcus oeni with M13-RP adaptor for sequencing
oAGT_299	GGCGATGGGAGTCGAACCCATAC	RVR primer tRNA Leu AAG Oenococcus oeni
oAGT_300	CAGGAACAGCTATGACCGATGGATACCCAAGTGGC	FWD primer tRNA Ser AGA Oenococcus oeni with M13-RP adaptor for sequencing
oAGT_301	GAGAGATTCGAACTCTCG	RVR primer tRNA Ser AGA Oenococcus oeni
oAGT_302	ATTTAGGTGACACTATAGAATAGCTCAGTTGGCTAGAGCAC	FWD primer tRNA Thr AGT Oenococcus oeni with Sp6 adaptor for sequencing
oAGT_303	TGCCGACTAGAGGATTTCG	RVR primer tRNA Thr AGT Oenococcus oeni



## 6.5 Publication 4

Ribas de Pouplana, L., A. G. Torres and **A. Rafels-Ybern** (2017). “What Froze the Genetic Code?” *Life (Basel)* 7(2).

CiteScore (2016, Scopus): 2.95



Concept Paper

# What Froze the Genetic Code?

Lluís Ribas de Pouplana <sup>1,2,\*</sup>, Adrian Gabriel Torres <sup>1</sup> and Àlbert Rafels-Ybern <sup>1</sup>

<sup>1</sup> Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldori Reixac, 10, 08028 Barcelona, Spain; adriangabriel.torres@irbbarcelona.org (A.G.T.); albert.rafels@irbbarcelona.org (À.R.-Y.)

<sup>2</sup> Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

\* Correspondence: lluis.ribas@irbbarcelona.org; Tel.: +34-934034868

Academic Editor: Koji Tamura

Received: 2 March 2017; Accepted: 3 April 2017; Published: 5 April 2017

**Abstract:** The frozen accident theory of the Genetic Code was a proposal by Francis Crick that attempted to explain the universal nature of the Genetic Code and the fact that it only contains information for twenty amino acids. Fifty years later, it is clear that variations to the universal Genetic Code exist in nature and that translation is not limited to twenty amino acids. However, given the astonishing diversity of life on earth, and the extended evolutionary time that has taken place since the emergence of the extant Genetic Code, the idea that the translation apparatus is for the most part immobile remains true. Here, we will offer a potential explanation to the reason why the code has remained mostly stable for over three billion years, and discuss some of the mechanisms that allow species to overcome the intrinsic functional limitations of the protein synthesis machinery.

**Keywords:** translation; evolution; speciation; protein folds; tRNA; ribosome

## 1. The Limits of the Genetic Code

The race to identify the structure of the Genetic Code was intense. However, the literature of the time suggests that it was, nevertheless, a collaborative exercise enriched by an intense academic debate that tried to offer explanations to the many questions that kept popping up.

In his seminal paper ‘The origin of the Genetic Code’, Francis Harry Compton Crick offered a good example of this dynamic as he and Leslie Orgel published their respective views on this topic in back-to-back papers [1,2]. In his paper Crick used the term ‘frozen accident’ to refer to the apparent inability of the code to accept new variations, and he contrasted this hypothesis with an alternative possibility: the stereochemical theory for the origin of the Genetic Code.

In the forty-nine years that have passed since the publication of this paper, we have advanced very significantly in our understanding of the molecular mechanisms that govern the Genetic Code. However, many fundamental questions regarding the origin and evolution of the code remain open, and chief among them is the reason why the system stopped incorporating new amino acids despite the obvious availability of codon sequences.

Nevertheless, progress has been made. The remarkable advances in the structural analysis of ribosomes, tRNAs, and aminoacyl-tRNA synthetases (ARS) have led to several important conclusions regarding the central roles of RNA in the early Genetic Code, which persist today in the functions of transfer RNAs and the ribosome, among others [3–5]. We now have strong support for the notion that extant proteomes functionally replaced a preceding RNA world where most, if not all, biological catalysis was performed by RNA molecules [6].

It is generally accepted that a primitive Genetic Code, using a limited number of amino acids or groups of related amino acids under a single identity, expanded through the generation of new tRNA

identities that increased the number of residues being used, while allowing for a better discrimination between similar amino acid sidechains [7]. The remarkable clustering of chemically-related amino acids that can be seen in the Genetic Code possibly reflects the process of establishment of the different codon and tRNA identities, and is the basis for the coevolution theory of Wong [8,9].

It is reasonable to expect that the expansion of tRNA identities was accompanied by the evolution of tRNA-associated polypeptides (ancestors of extant ARS). Indeed, both the distribution of amino acids in the Genetic Code, as well as the structural features of tRNA, are closely mirrored by the organization of the two ARS classes [10].

It is possible that the initial interaction between primitive tRNAs and the ancestral forms of ARS was in a complex of tRNA molecules bound by a heterodimer from which the two families of ARS later would emerge [11]. It has also been proposed that these two ancestral ARS domains could be coded by complementary strands and, as such, be under tightly coupled selection [12]. This hypothesis can explain the broad internal organization of the two ARS classes, the intriguing distribution of amino acid specificities that can be seen within these same classes, and the many unexplained similarities in identity elements found between tRNAs that are aminoacylated by ARS of different classes [13,14].

## 2. Why Did the Genetic Code Freeze?

Given the extraordinary chemical diversity of biological amino acids, and the potential for a three-base code based on four bases to theoretically incorporate up to sixty-three amino acids, it is a priori unclear why the universal Genetic Code includes only twenty amino acids. This is even more puzzling if one considers that several additional amino acids, such as selenocysteine and pyrrolysine, are used for protein synthesis. Chemical modifications of side chains are widespread, suggesting that cells could use a larger repertoire of residues within the canonical Genetic Code. Thus, what drove the arrest in the emergence of new tRNA identities and the expansion of the Genetic Code?

Although faithful amino acid recognition is an essential feature of the Genetic Code, it is unlikely that it was a limiting factor in the growth of the system because the recognition is limited to the interactions with ARS active sites, which are extremely adaptable and supported by editing domains that can discriminate between similar side chains [15]. On the other hand, the recognition of tRNAs is a much harder challenge because the three-dimensional structures of all tRNAs are very similar, their chemical composition before modifications is more uniform, and the number of required specific interactions with protein components of the translation apparatus is much larger.

We have proposed that a functional boundary exists with regards to the ability of the translation apparatus to successfully discriminate different tRNA identities. This boundary is determined by the overall capacity of the tRNA structure to incorporate different recognition elements. The incorporation of a new amino acid (hence a new tRNA identity) greatly increases the combinatorial problem faced by the translation machinery to specifically recognize individual tRNAs. This problem applies to modification enzymes, transport systems, ARS, elongation factors, ribosomes, etc. All tRNA identity elements need to coexist in a short RNA sequence whose structure is necessarily similar among all tRNAs in the cell. Additional constraints on tRNA evolution emerging from its non-canonical functions can also be envisaged. Our proposal is that this complex recognition network reaches a limit beyond which the incorporation of new tRNA identities is impossible without generating a recognition conflict with a pre-existing tRNA [16].

We have demonstrated that the saturation of structural and identity signals in a tRNA can prevent this molecule from incorporating other identities in evolution. We investigated the reasons for the intriguing lack of tRNA<sup>Gly</sup><sub>ACC</sub> in eukaryotic genomes and showed that pre-existing features of the tRNA<sup>Gly</sup> anticodon loop are incompatible with the presence of an adenosine at position 34, explaining why an A34-containing tRNA could not evolve and become enriched in eukaryotes [16].

At the genomic level, we observed that species with low numbers of tRNA genes have significantly more nucleotide differences between their orthologous tRNA pairs than closely related species with a larger number of tRNA genes. This is consistent with the notion that an increase in complexity

of tRNA populations leads to a higher conservation of tRNA sequences. Conversely, it would be expected that tRNA sequences would evolve faster in genomes with smaller numbers of tRNA genes. This situation is evident, for example, in mitochondria whose genomes have low numbers of tRNA genes [17]. Mitochondrial genomes display abundant deviations from the canonical Genetic Code, and contain the highest known variability in the structure and identity elements of tRNAs [18–20].

### 3. Evolutionary Strategies to Expand the Functional Boundaries of the Translation Apparatus

The study of globular protein folds has shown that the extant universe of proteins covers a minimal area of the vast potential number of protein structures. It is likely that extant protein structures evolved from the repetition of simpler domains that were assembled gradually through mechanisms of genetic recombination [21,22].

The synthesis of proteins generated through multiple repetitions of simple sequences may encounter difficulties due to the physicochemical characteristics of such repetitive peptides, or to the inability of tRNAs to maintain fidelity and reading frame when low complexity mRNA sequences are encountered. A number of adaptations have emerged to overcome some of these limitations. For example, the structure of the mammalian mitochondrial ribosome reveals that its polypeptide exit channel has been remodeled to allow not only the synthesis of the hydrophobic proteins that constitute the mitochondrial respiratory chain, but also their insertion into the mitochondrial membrane [23]. Also, EF-P (or eIF5A in eukaryotes), is a universally distributed elongation factor required for the translation of stretches of poly-proline codons [24]. Finally, translation of the extremely codon-biased mRNA transcripts coding for sericin and fibroin (protein components of silk) in the salivary glands of some arthropods requires a unique and highly skewed pool of cellular tRNAs, specifically selected to favor the translation of these mRNAs [25,26]. Thus, certain sequence combinations are a priori inaccessible to the translation apparatus, and functional improvements are needed to translate them.

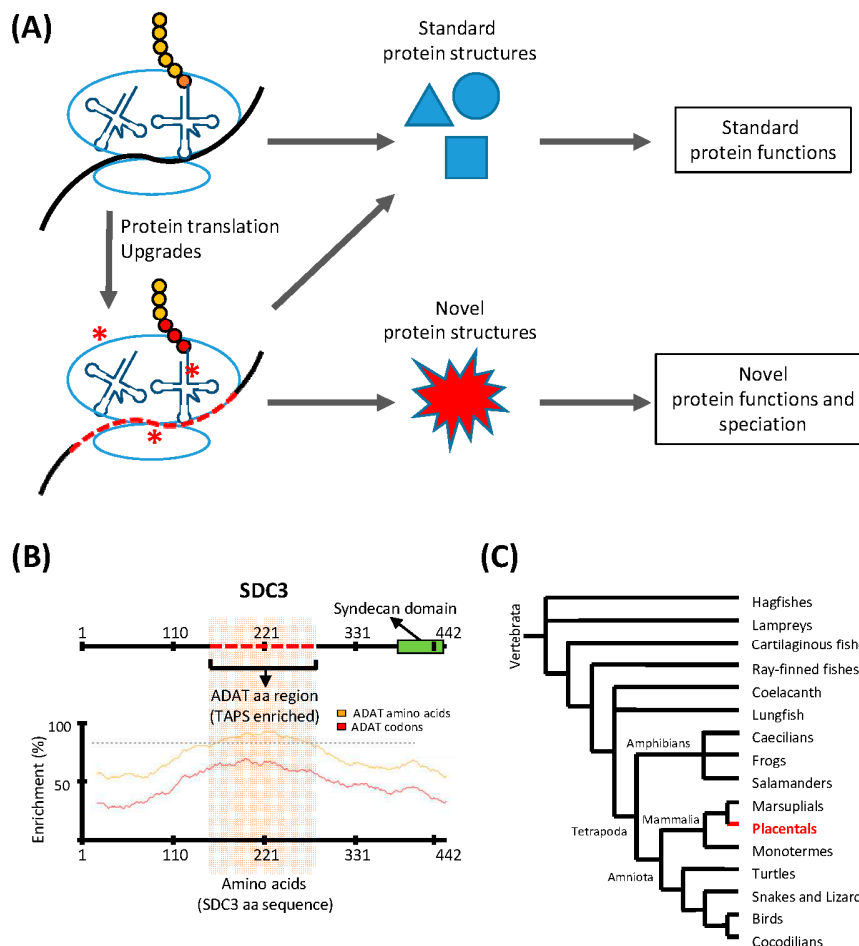
The existence of species-specific adaptations of the translation apparatus indicate that some species have access to protein structures that are inaccessible to others [27]. We envisage that adaptations of the protein synthesis apparatus that allowed a given species to assemble new types of proteins would provide these organisms with the opportunity to evolve novel and unique functions (in the aforementioned example, the production of silk resulted in a novel mechanism for the growth and development of certain arthropods). We believe that this evolutionary process could start with a simple modification of the translation apparatus, which would allow species to increase their proteome diversity and drive speciation in a punctuated manner (Figure 1A).

Two important parameters that differentiate the translation apparatuses of the three domains of life are their genomic composition of tRNA genes and the set of base modifications in their mature tRNAs [28,29]. We have shown that the divergence of eukaryotic and bacterial genomes in terms of tRNA composition is tightly linked to the evolution of different base modifications in the two kingdoms [30]. In eukaryotic genomes, a remarkable enrichment in genes coding for A34-containing tRNA isoacceptors coincided with the appearance of heterodimeric adenosine deaminases acting on tRNAs (ADAT). This enzyme deaminates A34 to inosine (I34) in tRNAs decoding for eight different amino acids [31]. The activity of this enzyme allows the tRNA pool in eukaryotic cells to match the codon composition of their genomes [30].

In the human transcriptome, codons recognized by ADAT-modified tRNAs are significantly more abundant than those that do not require these modified tRNAs, and this preference is greater in proteins that are highly enriched in the eight amino acids that can be decoded by ADAT-modified tRNAs. We have shown that, in the human proteome, the polypeptides that display the highest preference for these ADAT-modified tRNAs contain extremely biased stretches of the amino acids threonine, alanine, proline, and serine (TAPS) [32]. Figure 1B shows an example of such proteins, Syndecan 3 (SDC3), a member of a proteoglycan family unique to placentals (Figure 1C). This observation suggests that the emergence of TAPS-enriched proteins in eukaryotes was facilitated by the evolutionary emergence of ADAT, which caused an ‘upgrade’ of the translation machinery through the modification of the



composition and the codon-pairing capacity of their tRNA pool. We propose that the capacity of bacterial- and archaeal-type translation machineries to synthesize polypeptides highly enriched in TAPS, is limited by the functional characteristics of their tRNA pools, which may be either inefficient during the elongation phase of these transcripts (causing ribosomal stalling), or prone to decoding errors in these circumstances (causing deleterious levels of mutations in the resulting polypeptides).



**Figure 1.** Translation upgrades may lead to novel protein structures and drive speciation. (A) The translation machinery is capable of synthesizing a finite number of standard protein structures, and translation ‘upgrades’ (red asterisks) such as codon usage adaptations, or modulation of the tRNA pool, allow the translation machinery to synthesize proteins with novel structures and functions. This process may drive speciation; (B) An example of a gene (SDC3) containing a region with a sequence highly enriched in ADAT-related amino acids (red dashed line; upper panel). The codon composition of the DNA coding for this domain is highly biased towards triplets recognized by tRNAs modified by ADAT. The lower panel shows the enrichment in ADAT-related amino acids (yellow line) and ADAT-dependent codons (red line) across the whole sequence of SDC3. The dashed line marks an enrichment level of ADAT-dependent codons of 80%; (C) Consensus phylogeny for Vertebrata. SDC3 belongs to the syndecan proteoglycan family found solely in placentals (highlighted region). The activity of ADAT may have contributed to the emergence of SDC3-type domains in placentals.

The number of known species-specific features of the translation apparatus continues to grow and already includes the composition and regulation of several tRNA modifications, alterations to the ribosomal structure, the differential functionality of translation factors, and the protein and RNA composition of ribosomes, among others. Some of these adaptations may have resulted in translation machinery upgrades that allowed the synthesis of proteins with novel structures and functionality.

The extent to which each of these differential features contributed to the divergence of proteomes is still unknown. However, a comparative analysis of the regions of the protein universe that are occupied by the proteomes of archaeal, bacterial, and eukaryotic organisms could shed light on this question.

In conclusion, the frozen accident that Francis Crick proposed with his characteristic genius may have been the result of the intrinsic limitations imposed by tRNA recognition, but translation has learned to overcome some of these initial limitations through additional functional adaptations that allow species to increase the range and roles of their proteins.

**Acknowledgments:** This work was supported by the Spanish Ministry of Economy and Competitiveness [FPDI-2013-17742] to AGT; [BES2013-064551] to AR-Y; and [BIO2015-64572] to LRdP.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Crick, F.H. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379. [[CrossRef](#)]
2. Orgel, L.E. Evolution of the genetic apparatus. *J. Mol. Biol.* **1968**, *38*, 381–393. [[CrossRef](#)]
3. Di Giulio, M. The origin of the tRNA molecule: implications for the origin of protein synthesis. *J. Theor. Biol.* **2004**, *226*, 89–93. [[CrossRef](#)] [[PubMed](#)]
4. Noller, H.F. *On the Origin of the Ribosome: Coevolution of Subdomains of tRNA and rRNA, in The RNA World*; Gesteland, R.F., Atkins, J.A., Eds.; Cold Spring Harbor Laboratory Press: New York, NY, USA, 1993; pp. 137–156.
5. Petrov, A.S.; Gulen, B.; Norris, A.M.; Kovacs, N.A.; Bernier, C.R.; Lanier, K.A.; Fox, G.E.; Harvey, S.C.; Wartell, R.M.; Hud, N.V.; et al. History of the ribosome and the origin of translation. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 15396–15401. [[CrossRef](#)] [[PubMed](#)]
6. Pressman, A.; Blanco, C.; Chen, I.A. The RNA World as a Model System to Study the Origin of Life. *Curr. Biol.* **2015**, *25*, R953–R963. [[CrossRef](#)] [[PubMed](#)]
7. Grosjean, H.; Westhof, E. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res.* **2016**, *44*, 8020–8040. [[CrossRef](#)] [[PubMed](#)]
8. Wong, J.T. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA* **1975**, *72*, 1909–1912. [[CrossRef](#)] [[PubMed](#)]
9. Wong, J.T. Coevolution theory of the genetic code at age thirty. *Bioessays* **2005**, *27*, 416–425. [[CrossRef](#)] [[PubMed](#)]
10. Ribas de Pouplana, L.; Schimmel, P. Aminoacyl-tRNA synthetases: Potential markers of genetic code development. *Trends Biochem. Sci.* **2001**, *26*, 591–596. [[CrossRef](#)]
11. Ribas de Pouplana, L.; Schimmel, P. Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem. *Cell* **2001**, *104*, 191–193. [[CrossRef](#)]
12. Pham, Y.; Li, L.; Kim, A.; Erdogan, O.; Weinreb, V.; Butterfoss, G.L.; Kuhlman, B.; Carter, C.W., Jr. A minimal TrpRS catalytic domain supports sense/antisense ancestry of class I and II aminoacyl-tRNA synthetases. *Mol. Cell* **2007**, *25*, 851–862. [[CrossRef](#)] [[PubMed](#)]
13. Giege, R.; Sissler, M.; Florentz, C. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res.* **1998**, *26*, 5017–5035. [[CrossRef](#)] [[PubMed](#)]
14. Beuning, P.J.; Musier-Forsyth, K. Transfer RNA recognition by aminoacyl-tRNA synthetases. *Biopolymers* **1999**, *52*, 1–28. [[CrossRef](#)]
15. Martinis, S.A.; Boniecki, M.T. The balance between pre- and post-transfer editing in tRNA synthetases. *FEBS Lett.* **2010**, *584*, 455–459. [[CrossRef](#)] [[PubMed](#)]
16. Saint-Leger, A.; Bello, C.; Dans, P.D.; Torres, A.G.; Novoa, E.M.; Camacho, N.; Orozco, M.; Kondrashov, F.A.; de Pouplana, L.R. Saturation of recognition elements blocks evolution of new tRNA identities. *Sci. Adv.* **2016**, *2*, e1501860. [[CrossRef](#)] [[PubMed](#)]
17. Gray, M.W.; Burger, G.; Lang, B.F. The origin and early evolution of mitochondria. *Genome Biol.* **2001**, *2*, REVIEWS1018. [[CrossRef](#)] [[PubMed](#)]
18. Chihade, J.W.; Brown, J.R.; Schimmel, P.R.; De Pouplana, L.R. Origin of mitochondria in relation to evolutionary history of eukaryotic alanyl-tRNA synthetase. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 12153–12157. [[CrossRef](#)] [[PubMed](#)]

19. Sengupta, S.; Higgs, P.G. Pathways of Genetic Code Evolution in Ancient and Modern Organisms. *J. Mol. Evol.* **2015**, *80*, 229–243. [[CrossRef](#)] [[PubMed](#)]
20. Sengupta, S.; Yang, X.; Higgs, P.G. The mechanisms of codon reassignments in mitochondrial genetic codes. *J. Mol. Evol.* **2007**, *64*, 662–688. [[CrossRef](#)] [[PubMed](#)]
21. Alva, V.; Soding, J.; Lupas, A.N. A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **2015**, *4*, e09410. [[CrossRef](#)] [[PubMed](#)]
22. Alva, V.; Remmert, M.; Biegert, A.; Söding, J. A galaxy of folds. *Protein Sci.* **2010**, *19*, 124–130. [[CrossRef](#)] [[PubMed](#)]
23. Greber, B.J.; Boehringer, D.; Leitner, A.; Bieri, P.; Voigts-Hoffmann, F.; Erzberger, J.P.; Leibundgut, M.; Aebersold, R.; Ban, N. Architecture of the large subunit of the mammalian mitochondrial ribosome. *Nature* **2014**, *505*, 515–519. [[CrossRef](#)] [[PubMed](#)]
24. Lassak, J.; Wilson, D.N.; Jung, K. Stall no more at polyproline stretches with the translation elongation factors EF-P and IF-5A. *Mol. Microbiol.* **2016**, *99*, 219–235. [[CrossRef](#)] [[PubMed](#)]
25. Chevallier, A.; Garel, J.P. 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.* **1982**, *124*, 477–482. [[CrossRef](#)] [[PubMed](#)]
26. Li, J.Y.; Ye, L.P.; Che, J.Q.; Song, J.; You, Z.Y.; Yun, K.C.; Wang, S.H.; Zhong, B.X. Comparative proteomic analysis of the silkworm middle silk gland reveals the importance of ribosome biogenesis in silk protein production. *J. Proteomics* **2015**, *126*, 109–120. [[CrossRef](#)] [[PubMed](#)]
27. Caetano-Anolles, G.; Wang, M.; Caetano-Anollés, D.; Mittenthal, J.E. The origin, evolution and structure of the protein world. *Biochem. J.* **2009**, *417*, 621–637. [[CrossRef](#)] [[PubMed](#)]
28. Chan, P.P.; Lowe, T.M. GtRNADB: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **2009**, *37*, D93–D97. [[CrossRef](#)] [[PubMed](#)]
29. Grosjean, H.; de Crecy-Lagard, V.; Marck, C. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett.* **2010**, *584*, 252–264. [[CrossRef](#)] [[PubMed](#)]
30. Novoa, E.M.; Pavon-Eternod, M.; Pan, T.; Ribas de Pouplana, L. A role for tRNA modifications in genome structure and codon usage. *Cell* **2012**, *149*, 202–213. [[CrossRef](#)] [[PubMed](#)]
31. Torres, A.G.; Piñeyro, D.; Rodríguez-Escribà, M.; Camacho, N.; Reina, O.; Saint-Léger, A.; Filonava, L.; Batlle, E.; de Pouplana, L.R. Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Res.* **2015**, *43*, 5145–5157. [[CrossRef](#)] [[PubMed](#)]
32. Rafels-Ybern, A.; Attolini, C.S.; Ribas de Pouplana, L. Distribution of ADAT-Dependent Codons in the Human Transcriptome. *Int. J. Mol. Sci.* **2015**, *16*, 17303–17314. [[CrossRef](#)] [[PubMed](#)]



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

## 7. Summary of Results

We first characterized how ADAT-dependent codons are distributed in human transcriptomes. This initial analysis is required to identify sections of the transcriptome potentially more dependent on ADAT activity levels. Our approach has been, first, to screen the complete human transcriptome and classify its genes based on the abundance of ADAT amino acids using two different methods to determine the distribution: the *half-gene* analysis, and the *running-window* approach. Using this initial curation, we identified those transcripts significantly enriched in ADAT codons, and we used this subset of sequences to determine the relative enrichment for each amino acid, and the variation of ADAT-sensitive codons (**Publication 1**).

Our results show, first, that the human transcriptome is biased towards proteins enriched in ADAT amino acids. Interestingly, the majority of these proteins rich in ADAT amino acids are specifically enriched in TAPS but not in LIVR. We also obtain an enrichment in TAPS but not in LIVR in all ADAT stretches from Bacteria and Eukarya analysed in **Publication 2**. Physicochemical parameters probably explain why, among all ADAT amino acids, only TAPS can reach higher relative frequencies. At the same time, functional features of proteins rich in TAPS must have driven the selection of these extremely biased protein sequences (**Publication 1**).

Our data also shows that, from a breakpoint concertation of 64%, the more enriched in ADAT amino acids a region is, the higher its tendency to use ADAT-sensitive codons instead of other codons that could not be recognized by I34-tRNAs. Therefore, transcripts containing ADAT stretches are composed preferentially of ADAT-sensitive codons, and this enrichment increases with the quality of the stretch. This observation confirms the hypothesis that ADAT-modified tRNAs are preferred in eukaryotic translation, and suggests that this preference increases with the frequency of ADAT amino acids (**Publication 1**).

Within the list of human proteins with the highest bias towards the use of ADAT-sensitive codons, the family of mucins is, by far, the most represented (**Table 7.1, data not published**). Similarly, Syndecan 3 (SDC3) is also enriched in ADAT-sensitive codons with the peculiarity that SDC3 is a member of the proteoglycan family unique to placentals (**Publication 4**). These results suggest that the expression of this protein candidates might be specially regulated by ADAT activity.

## Summary of Results

**Table 7.1:** List of the 32 human proteins with the largest ADAT stretches. The size of the stretch (size), its concentration in ADAT amino acids (ADAT aa) and its ADAT enrichment (ADAT enrich.) are depicted. The protein family of Mucins are in bold. The protein Syndecan 3 (SDC3) is in red. Proteins that are repeated in the table (e.g. MUC17 is present 7 times) is due to the presence of different regions with ADAT-stretches in the same sequence.

	protein	size	ADAT aa	ADAT enrich.			protein	size	ADAT aa	ADAT enrich.
1	<b>MUC5B</b>	719	0.86	0.79		17	AMOT	216	0.88	0.88
2	<b>MUC5B</b>	606	0.86	0.78		18	<b>MUC17</b>	214	0.84	0.77
3	<b>MUC17</b>	571	0.85	0.75		19	<b>MUC4</b>	210	0.85	0.93
4	SRCAP	560	0.88	0.77		20	SRRM2	208	0.88	0.62
5	<b>MUC5B</b>	460	0.86	0.76		21	<b>MUC4</b>	201	0.85	0.95
6	<b>MUC5B</b>	456	0.86	0.77		22	<b>MUC17</b>	200	0.84	0.74
7	<b>MUC4</b>	366	0.84	0.93		23	<b>MUC17</b>	191	0.84	0.75
8	<b>MUC17</b>	349	0.85	0.75		24	MYPOP	179	0.84	0.69
9	<b>MUC5B</b>	341	0.86	0.79		25	LDB3	178	0.84	0.81
10	<b>MUC17</b>	335	0.85	0.82		26	SH3BP1	178	0.85	0.72
11	SRRM2	325	0.92	0.70		27	C11orf24	175	0.85	0.71
12	<b>MUC6</b>	270	0.86	0.81		28	SELV	175	0.89	0.65
13	SRRM2	257	0.84	0.73		29	<b>MUC17</b>	173	0.85	0.76
14	BCORL1	251	0.88	0.81		30	PRDM2	172	0.85	0.91
15	PRRC2C	247	0.90	0.94		31	SON	172	0.86	0.57
16	<b>MUC7</b>	236	0.88	0.97		32	<b>SDC3</b>	170	0.85	0.72

Subsequently, we extended our initial analysis to a series of eukaryotic and bacterial organisms, spanning the whole tree of life. In Eukarya, ADAT stretches are longer and more frequent than in Bacteria. Eukaryotic genes coding for stretches of ADAT amino acids are generally biased towards ADAT-sensitive codons, with a positive correlation with respect to their length. Conversely, in Bacteria, the presence of stretches is 4-fold lower, and the corresponding genes are generally depleted of ADAT-sensitive codons. These results likely reflect the differences in I34 dependence between Eukarya and Bacteria and is consistent with the substrate-specificity differences found between ADAT and TadA (**Publication 2**).

Among Eukarya, only animals and plants have a positive correlation between lengths of stretches and its enrichment in ADAT-sensitive codons. Fungi and Protists show no such correlation. Moreover, Metazoa contains the longest ADAT stretches and the highest enrichment in ADAT-sensitive codons among Eukarya. These results indicate a preference for ADAT-sensitive codons and longer stretches in kingdoms with embryonic development where multicellularity is prevalent. In this regard, a connection between variation in codon usage and the establishment of multicellularity was already proposed (Ikemura 1985) (**Publication 2**).

Table 7.2: A34 and G34 num-tRNA for 66 Stramenopiles. *N. gaditana* is highlighted in red.

	Thr	Ala	Pro	Ser	Leu	Ile	Val	Arg	Thr	Ala	Pro	Ser	Leu	Ile	Val	Arg
	agu	agc	agg	aga	aag	aau	aac	acg	ggg	ggc	ggg	gga	gag	gau	gac	gcg
<i>Achlya hypogyna</i>	1	3	2	0	1	2	0	2	0	0	0	0	0	0	2	0
<i>Albugo candida</i>	1	1	1	1	2	4	1	2	0	0	0	0	0	1	0	1
<i>Aphanomyces astaci</i>	1	2	12	5	2	1	1	2	0	0	0	0	0	0	6	0
<i>Aphanomyces invadans</i>	0	0	1	7	5	1	0	3	0	0	0	0	0	1	2	0
<i>Asterionella formosa</i>	2	4	1	2	2	3	3	1	0	0	0	0	0	3	0	0
<i>Aurantiochytrium</i> sp. T66	2	4	1	2	4	0	0	4	0	0	0	0	0	0	3	0
<i>Anreococcus anophagefferens</i> bet	1	0	3	1	3	2	0	4	0	0	0	0	0	0	1	0
<i>Blastocystis hominis</i>	6	9	6	5	3	7	0	7	0	0	0	0	0	1	4	0
<i>Cladosiphon okamuranus</i>	2	3	2	2	3	0	2	9	9	8	7	10	7	17	9	1
<i>Ectocarpus siliculosus</i> bet	9	15	9	7	8	0	0	7	0	0	0	0	1	12	7	0
<i>Fistulifera solaris</i>	2	4	1	2	1	4	2	0	0	0	0	0	0	0	0	0
<i>Fragilariopsis cylindrus</i> CCMP1102	6	7	2	4	3	5	4	2	0	0	0	0	0	0	0	0
<i>Halocafeteria seosinensis</i>	1	2	2	1	2	1	4	3	0	0	0	0	0	0	1	0
<i>Heterococcus</i> sp. DN1	2	2	0	1	0	0	0	0	0	0	0	0	0	2	0	0
<i>Hyaloperonospora arabidopsidis</i> Emoy2	21	3	3	1	2	4	2	2	0	0	0	0	0	0	0	2
<i>Lagenidium giganteum</i>	1	0	3	1	1	1	1	0	0	0	0	0	0	5	0	1
<i>Nannochloropsis gaditana</i> CCMP526	0	1	0	1	0	1	0	1	0	0	0	0	0	3	1	0
<i>Nannochloropsis limnetica</i>	0	1	1	0	1	1	0	2	0	0	0	1	0	2	1	0
<i>Nannochloropsis oceanica</i> OZ-1	0	1	1	1	1	1	0	2	0	0	0	0	0	1	1	0
<i>Nannochloropsis salina</i> CCMP1776	0	1	0	1	0	1	0	2	0	0	0	0	0	2	1	0
<i>Peronospora effusa</i>	0	0	1	1	0	0	1	0	0	0	0	0	0	1	0	1
<i>Peronospora tabacina</i>	0	3	2	2	0	1	1	1	0	0	0	0	0	0	0	0
<i>Phytophthora agathidicida</i>	2	2	2	1	1	2	2	3	0	0	0	0	0	1	0	1
<i>Phytophthora cactorum</i>	0	0	2	4	0	3	4	5	0	0	0	0	0	2	0	1
<i>Phytophthora cambivora</i>	2	2	1	2	1	2	0	0	0	0	0	0	0	0	0	0
<i>Phytophthora capsici</i> LT1534	17	27	6	5	14	17	5	32	0	0	0	0	0	2	0	0
<i>Phytophthora cinnamomi</i>	1	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1
<i>Phytophthora colocasiae</i>	16	38	7	2	4	11	3	17	0	0	0	0	0	0	0	3
<i>Phytophthora cryptogea</i>	0	3	0	1	0	0	1	0	0	0	0	0	0	0	0	0
<i>Phytophthora fragariae</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>Phytophthora lateralis</i> MPF4	0	0	2	0	1	1	1	3	3	2	2	1	2	2	2	1
<i>Phytophthora megakarya</i>	0	1	0	2	0	2	2	0	0	0	0	0	0	2	0	2
<i>Phytophthora multivora</i>	1	1	2	1	2	1	2	0	0	0	0	0	0	1	0	1
<i>Phytophthora nicotianae</i>	23	17	67	0	53	34	1	48	0	0	0	0	0	0	1	0
<i>Phytophthora parasitica</i> CJ01A1	0	0	1	0	1	1	0	0	1	0	0	0	1	0	1	0
<i>Phytophthora pinifolia</i>	2	6	1	1	1	0	0	4	0	0	0	0	0	2	0	1
<i>Phytophthora pisi</i>	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>Phytophthora plurivora</i>	1	3	1	2	2	1	2	2	0	0	0	0	0	1	0	1
<i>Phytophthora pluvialis</i>	1	3	1	1	1	1	2	2	0	0	0	0	0	4	0	1
<i>Phytophthora ramorum</i>	2	3	3	2	4	1	1	3	0	0	0	0	0	0	0	0
<i>Phytophthora rabi</i>	0	3	1	1	2	0	2	1	0	0	0	0	0	2	0	1
<i>Phytophthora sojae</i>	28	71	70	60	63	23	61	34	0	0	0	0	0	0	0	0
<i>Phytophthora taxon totara</i>	2	1	0	2	1	2	3	3	0	0	0	0	0	1	0	1
<i>Pilasporangium apinajurcum</i>	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
<i>Plasmopara halstedii</i>	3	2	2	2	1	5	1	0	0	0	0	0	0	2	0	1
<i>Plasmopara viticola</i>	2	8	5	5	8	5	6	4	0	0	0	0	0	3	0	1
<i>Proteromonas lacertae</i>	2	6	1	3	1	5	4	3	0	0	0	0	0	2	0	0
<i>Pseudo-nitzschia multistriata</i>	8	11	5	7	9	8	7	0	0	0	0	0	0	0	0	0
<i>Pseudoperonospora cubensis</i>	0	2	0	1	0	0	2	14	5	3	3	4	4	13	5	1
<i>Pythium aphanidermatum</i> DAOM BR444	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0
<i>Pythium arrhenomanes</i> ATCC 12531	0	0	0	0	0	0	2	0	0	0	0	0	0	11	0	1
<i>Pythium insidiosum</i>	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0
<i>Pythium irregulare</i> DAOM BR486	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>Pythium inwayamai</i> DAOM BR242034	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1
<i>Pythium oligandrum</i>	0	0	2	0	2	1	0	0	0	0	0	0	0	0	0	0
<i>Pythium periplocum</i>	0	1	0	0	1	1	0	0	0	0	0	0	0	2	0	2
<i>Pythium ultimum</i> DAOM BR144	10	10	4	6	7	6	4	17	0	0	0	0	0	8	0	2
<i>Pythium vexans</i> DAOM BR484	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>Saccharina japonica</i>	10	14	10	12	11	0	0	13	5	4	1	13	1	31	13	1
<i>Saprolegnia diclina</i> VS20	0	0	0	0	0	2	0	0	0	0	0	0	0	0	9	0
<i>Saprolegnia parasitica</i> CBS 223.65	7	5	5	10	5	2	0	0	0	0	0	0	0	1	5	1
<i>Schizochytrium</i> sp. CCTCC M209059	6	8	1	4	5	5	0	8	0	0	0	0	0	1	10	0
<i>Sclerospora graminicola</i>	74	311	171	44	186	226	59	131	0	0	0	0	0	70	0	64
<i>Stramenopiles</i> sp. TOS.AG23-2	1	1	2	1	0	1	2	2	0	0	0	0	0	0	0	0
<i>Thalassiosira oceanica</i>	8	5	0	4	5	2	4	2	4	0	7	0	0	4	0	0
<i>Thalassiosira pseudonana</i> CCMP1335	2	4	2	2	1	3	2	0	0	0	0	0	0	0	0	0
<i>Thraustotheca clavata</i>	2	1	3	1	1	1	0	1	0	0	0	0	0	1	2	1

## Summary of Results

We have discovered that in the bacterial phyla of Firmicutes and Cyanobacteria genes encoding for proteins rich in ADAT amino acids are also enriched in ADAT-sensitive codons. This is initially surprising because in Bacteria only tRNA<sup>Arg</sup> (ACG) is known to be deaminated by TadA, and arginine is not significantly enriched in the ADAT-stretched proteins that we have identified. This apparent contradiction may be resolved by our discovery that the genomes of several Firmicutes code for previously unnoticed A34-containing tRNAs for amino acids threonine, proline, serine, leucine and isoleucine. These tRNAs could potentially be deaminated to I34. Thus, it is possible that the enrichment in ADAT-sensitive codons in these species is, at least partially, due to their use of additional I34-modified tRNAs. Cyanobacteria genomes do not contain this type of tRNAs, their stretch length is shorter than in Firmicutes, and the frequency of ADAT-sensitive codons does not increase with stretch length (**Publication 2**).

We continued our analysis by characterizing the distribution of genes coding for A34-tRNAs among the eukaryotic and bacterial organisms analyzed before. This analysis provided a first approximation to the distribution of I34 in the phylogenetic tree. We discover that some bacteria code for A34-tRNAs other than tRNA<sup>Arg</sup>, whereas in eukaryotes, a division can be established between protists, which display large variability in their genomic contents of A34-tRNA genes, and fungi, plants and animals, which are extremely uniform in the same regard (**Publication 3**).

We experimentally determined that the genome of the firmicute *Oenococcus oeni* has an expanded repertoire of A34-tRNAs that includes tRNA<sup>Arg</sup>, tRNA<sup>Leu</sup>, tRNA<sup>Ser</sup> and tRNA<sup>Thr</sup>, but only tRNA<sup>Arg</sup> and tRNA<sup>Leu</sup> are modified to I34 under standard culture conditions. Conversely, despite the variability seen in protists, we experimentally determined that the tRNAome of the protozoan *Tetrahymena thermophila* contains I34-tRNAs coding for TAPSLIVR (the same as for metazoans), showing that a fully functional ADAT evolved early in eukaryotic evolution (**Publication 3**).

We then developed a pipeline to identify *bona fide* TadA and ADAT proteins. We applied this pipeline to determine the phylogenetic distribution of TadA and ADAT proteins in bacteria and eukaryotes, respectively. Our findings indicate that several ancestral bacterial groups lack both TadA and A34-tRNAs, suggesting that these species never developed the machinery to generate I34-modified tRNAs. On the other hand, limited sets of bacterial species have either lost the system secondarily, or expanded it to additional tRNA substrates (**Publication 3**).

Finally, we characterized the ADAT stretches composition of bacterial and eukaryotic groups in terms of its A34-tRNA diversity. We find that there is a gradual increase in the codon bias for ADAT codons from bacteria with A34-tRNA diversity of 1 to eukaryotes with A34-tRNA diversity higher than 7. Intriguingly we also detect a small number of stretches in the group of bacteria with A34-tRNA diversity equal to 0, which is also accompanied by a significant increase in the codon bias favoring codons ended in C, A, or U. A more detailed analysis of the stretches found in these species revealed that they are absent from the vast majority of species in this set, and concentrate almost exclusively in bacteria of the genus *Chloroflexi*. Thus, unlike in the other groups, TAPSLIVR stretches in bacteria with A34-diversity equal to 0 are extremely rare, and only present in a very restricted group of species (**Publication 3**).

Within the SAR clade in Eukarya, the phyla of Stramenopiles have an increased variation in A34 gene diversity that ranges from 0 to 8 (**Table 7.2**). However, such tRNA A34 gene reduction is not always correlated with the appearance of the correspondent G34 tRNA genes, so the translation of many C-ended codons is difficult to explain for these organisms (**Table 7.2**). We tried to experimentally prove that the *Nannochloropsis gaditana*, a Stramenopile with tRNA A34 genes predicted only for Ala, Ser, Ile and Arg, effectively lacks the A34 tRNAs genes for Thr, Pro, Leu and Val, which tRNAscan-SE did not found. We purified the RNA from *N. gaditana* and performed small size RNA sequencing. Unfortunately, the amount of data that aligned with our tRNA gene templates was very low leading to us inconclusive results.



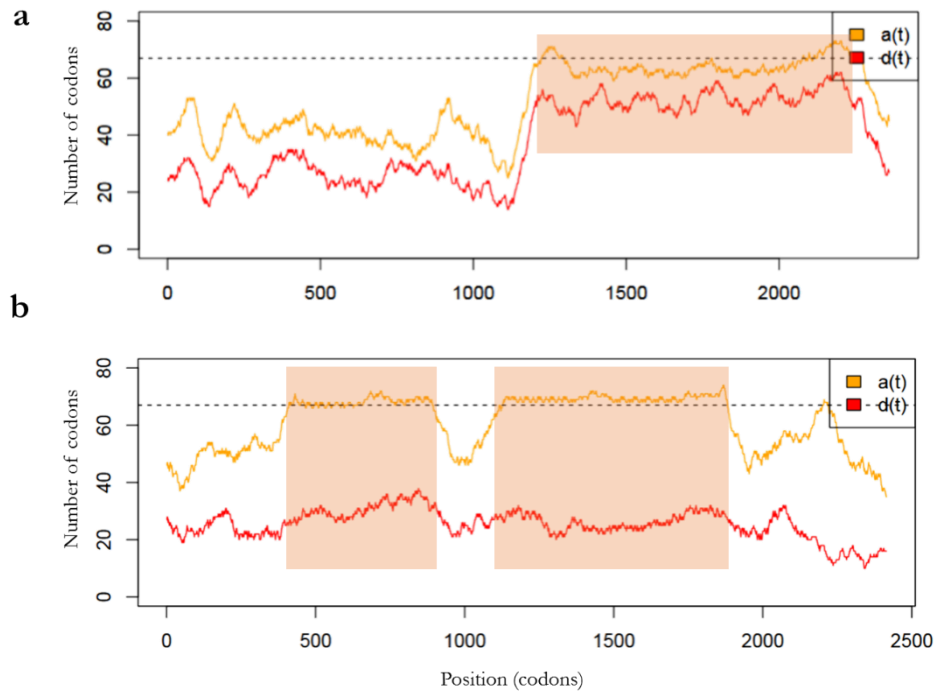
## Summary of Results

## 8. Discussion

I34-tRNAs modified by ADAT and TadA are essential for cell survival (Gerber and Keller 1999; Wolf, Gerber et al. 2002; Torres, Pineyro et al. 2015). They regulate the process of translation by increasing tRNA pairing ability to synonymous codons ended in C, U or A, and avoiding those ended in G. It is clear that inosine is important to balance codon usage and tRNA gene copy number in Eukaryotes, and that highly translated genes in these species tend to be enriched in ADAT-sensitive codons (Novoa and Ribas de Pouplana 2012). For these reasons, we intended to investigate whether I34, and therefore ADAT, could be involved in the regulation of gene expression levels.

The protein enrichment with TAPS instead of LIVR is universal in Eukarya and Bacteria. In Eukarya those proteins tend to be translated by ADAT-dependent codons whereas in Bacteria, this correlation is absent. The translation of TAPS-enriched proteins seems to be independent of the codons that translate them, suggesting that ADAT (and TadA) does not affect the amino acid composition of proteins. The mechanisms that promotes this TAPS-bias remains unknown, but we venture to say that physicochemical interactions between amino acids must promote TAPS enrichment and avoid LIVR. TAPS are small and not charged amino acids, T and S are polar whereas A and P are non-polar. LIVR instead, have larger residues, LIV are non-polar and R is very large and positively charged. As example, poly-Arg peptides higher than 15 residues at 1  $\mu$ M, cause cell death to cortical neuronal cultures (Meloni, Brookes et al. 2015). Interestingly, all the codons that code for TAPS have the structure  $N_1C_2N_3$ , sharing a common cytidine at the second position where  $N_1$  and  $N_3$  represents any of the nucleotides, with the exception of Ser AGU and AGC (**Publication 1, Figure 5**). Each amino acid from TAPS form a 4-box family where  $N_3$  is not important for the amino acid determination. Moreover, some literature demonstrates that tRNAs with A34 are able to pair with any base at the third position of the codon (Auxilien, Crain et al. 1996; Alkatib, Scharff et al. 2012). Thus, the selectivity between these four codon sets depends only upon the recognition of the first codon base. Under these circumstances, it is possible that the proposed higher selectivity of I over A may be preferred to minimize the possibility of decoding errors, particularly in highly repetitive transcript regions such as the TAPS-enriched ADAT stretches identified in our analysis.

While is true that the amino acid composition of ADAT stretches is similar between Eukarya and Bacteria, the frequency and the size of ADAT stretches are clearly higher in Eukarya. We hypothesize that the extension of I34 modified tRNA, allowed the translation machinery



**Figure 8.1:** Concentrations for ADAT codons (yellow) and ADAT-sensitive codons (red) along the proteins human MUC6 (a) and a bacterial hemagglutinin from *Burkholderia ambifaria* (b). Concentrations are calculated for a window of 80 residues. ADAT-stretches are highlighted in orange.

to deal with highly repetitive TAPS-enriched protein regions by turning their translation towards the ADAT sensitive codons. It is similarly conceivable that certain highly repetitive transcript sequences may be inaccessible to ribosome processing unless new functional improvements that increase efficiency or selectivity can be found. The selection of modified bases, such as inosine, that possibly allow species to synthesize proteins previously unavailable may be a major driving force in speciation. Several examples of codon and amino acid compositions are known that impair ribosomal functional and, in some cases, require additional cofactors to allow the ribosome to progress through these regions. For example, the structure of the mammalian mitochondrial ribosome reveals that its polypeptide exit channel has been remodeled to allow not only the synthesis of the hydrophobic proteins that constitute the mitochondrial respiratory chain but also their insertion into the mitochondrial membrane (Greber, Boehringer et al. 2014). Also, elongation factor P (EF-P) (or its eukaryotic and archaeal homolog, initiation factor 5A (eIF5A), is a universally distributed elongation factor required for the translation of stretches of poly-proline codons (Lassak, Wilson et al. 2016). Finally, translation of the extremely codon-biased mRNA transcripts coding for sericin and fibroin (protein components of silk) in the salivary glands of some

**Table 8.1:** num-tRNA for A34 tRNAs for some bacterial Firmicutes.

Organism	Thr agu	Ala agc	Pro agg	Ser aga	Leu aag	Ile aau	Val aac	Arg acg
Oenococcus_oeni	1	0	0	1	1	0	0	1
Lactobacillus_amylovorus	1	0	0	0	1	0	0	2
Lactobacillus_brevis	1	0	0	0	1	0	0	2
Lactobacillus_crispatus	1	0	0	0	1	0	0	2
Lactobacillus_delbrueckii	2	0	0	0	2	0	0	4
Leuconostoc_C2	1	0	0	0	1	0	0	2
Leuconostoc_carnosum	1	0	0	0	1	0	0	2
Weissella_koreensis	1	0	0	0	1	0	0	2
Butyrivibrio_proteoclasticus	0	0	1	0	1	0	0	1
Anaerococcus_prevotii	0	0	0	0	1	0	0	1
Clostridium_lentocellum	0	0	0	0	2	0	0	2
Clostridium_phytofermentans	0	0	0	0	1	0	0	1
Clostridium_saccharolyticum	0	0	0	0	2	0	0	2
Eubacterium_eligens	0	0	0	0	1	0	0	1
Eubacterium_rectale	0	0	0	0	2	0	0	1
Lactobacillus_acidophilus	0	0	0	0	1	0	0	2
Lactobacillus_casei	0	0	0	0	1	0	0	2
Lactobacillus_rhamnosus	0	0	0	0	1	0	0	2
Lactobacillus_ruminis	0	0	0	0	2	0	0	2
Lactobacillus_salivarius	0	0	0	0	1	0	0	3
Lactococcus_garvieae	0	0	0	0	2	0	0	2
Lactococcus_lactis	0	0	0	0	2	0	0	2
Pediococcus_clausenii	0	0	0	0	1	0	0	2
Pediococcus_pentosaceus	0	0	0	0	1	0	0	2
Roseburia_hominis	0	0	0	0	1	0	0	1
Streptococcus_dysgalactiae	0	0	0	0	1	0	0	3
Streptococcus_equi	0	0	0	0	1	0	0	3
Streptococcus_galloyticus	0	0	0	0	1	0	0	2
Bacillus_coagulans	0	0	1	0	0	0	0	1
Clostridium_beijerinckii	0	0	0	0	0	1	0	1
Desulfosporosinus_meridiei	0	0	0	1	0	0	0	2
Desulfosporosinus_meridiei	0	0	0	1	0	0	0	2

arthropods requires a unique and highly skewed pool of cellular tRNAs, specifically selected to favor the translation of these mRNAs (Chevallier and Garel 1982; Li, Ye et al. 2015). Thus, certain sequence combinations are *a priori* inaccessible to the translation apparatus, and functional improvements are needed to translate them. The species-specific adaptations of the translation apparatus indicate that some species have access to protein structures that are inaccessible to others (Caetano-Anolles, Wang et al. 2009). We envisage that adaptations of the protein synthesis apparatus that allowed a species to assemble new types of proteins would provide these organisms with the opportunity to evolve novel and unique functions (in the aforementioned example, the production of silk resulted in a novel mechanism for growth and development of certain arthropods). We believe this evolutionary process could start with a simple modification of the ribosomal apparatus, but would open up the possibility of qualitative physiological changes and could drive speciation in a punctuated manner.

## Discussion

The family of mucins are probably the best example of highly TAPS-repetitive proteins in humans. **Figure 8.1** shows the enrichment in ADAT codons (yellow) and ADAT-sensitive codons (red) of human mucin MUC6 compared with a bacterial hemagglutinin (cause red blood cells agglutination) that is also enriched in ADAT codons. We can observe how the codon usage strategies are different, because while the ADAT-stretched regions of human MUC6 are enriched in ADAT-sensitive codons, the bacterial protein avoid them. These differences, probably allowed the translation of such long and TAPS-repetitive family of mucins only in Eukarya. Mucins are large glycosylated proteins present in epithelial tissues of the respiratory and gastrointestinal tracts and major components of gel-like secretions. Alterations in mucins production and glycosylation patterns are related to disease. Mucus hypersecretion, airway obstruction and mucociliary clearance impairment found in certain respiratory conditions like asthma, chronic obstructive pulmonary disease (COPD) or cystic fibrosis (CF), have been associated to mucins overexpression (Rose and Voynow 2006; Kreda, Davis et al. 2012). Moreover, invasive proliferation of some cancers is linked to mucins overexpression or aberrant glycosylation (Kufe 2009). Due to the putative need for I34-modification in mucins expression, the inhibition of ADAT activity would be a promising strategy to block its overexpression in the context of these pathologies. Thus, the development of selective ADAT inhibitors might constitute a promising therapeutic strategy with a novel mechanism of action for the treatment of these diseases. Another ADAT-stretched protein in human is Syndecan 3 (SDC3) (**Table 7.1**), a protein that belongs to proteoglycan family unique to placentals. This observation suggests that the emergence of TAPS-enriched proteins in eukaryotes was facilitated by the evolutionary emergence of ADAT, which caused an ‘upgrade’ of the translation machinery through the modification of the composition and the codon-pairing capacity of their tRNA pool. Experimental unpublished data gave us promising results about the dependence on ADAT for the correct expression levels of SDC3.

Identification of ADAT-stretched protein functions in all the organisms (except for model organisms such as human) is a tedious task that remains to be done because the lack of genome annotations in most of these species. It is also not possible to carry on analyses of gene ontology.

The I34 expansion that is observed in *Oenococcus oeni* provides new information about the evolution of TadA substrate recognition patterns. This bacterial Firmicute, has expanded its A34-tRNA gene repertoire to those coding for Arg, Leu, Ser, and Thr. However, only  $\text{tRNA}_{\text{ICG}}^{\text{Arg}}$  and  $\text{tRNA}_{\text{IAG}}^{\text{Leu}}$  are modified with I34, whereas  $\text{tRNA}_{\text{AGA}}^{\text{Ser}}$  and  $\text{tRNA}_{\text{AGU}}^{\text{Thr}}$

remains unmodified in standard culture conditions. This experiment evidences for the first time a bacterial organism with an I34-modified tRNA other than Arg.

The presence of unmodified A34-tRNAs have been described several times in the literature, suggesting a relaxation of the wobble decoding rules for these specific cases. In bacteria, an unmodified tRNA<sup>Thr</sup><sub>AGU</sub> have been shown for the Mollicute *Mycoplasma capricolum* (Andachi, Yamao et al. 1987) and an unmodified tRNA<sup>Pro</sup><sub>AGG</sub> was found in *Salmonella typhimurium* (Chen, Qian et al. 2002). In Eukarya, an unmodified tRNA<sup>Arg</sup><sub>ACG</sub> was purified from the mitochondria of the nematode *Ascaris suum* (Watanabe, Tsurui et al. 1997) as well as the cytoplasm of several higher plants. Interestingly, the tRNA<sup>Arg</sup><sub>ACG</sub> is imported from the cytoplasm to the mitochondria of higher plants where is deaminated by a mitochondrial-specific deaminase (Aldinger, Leisinger et al. 2012).

Wolf et al. demonstrated that RNA minisubstrates with the anticodon loop derived tRNA<sup>Arg</sup><sub>ACG</sub> are deaminated by TadA in *E. coli*. They also demonstrated that, tRNA<sup>Arg</sup><sub>ACG</sub> from yeast can be deaminated by TadA from *E. coli*, and moreover, that tRNA<sup>Asp</sup> from yeast, whose anticodon loop was substituted for those from tRNA<sup>Arg</sup><sub>ACG</sub> (the sequence of the anticodon is the same for *E. coli* and yeast) is also deaminated. However, TadA from *E. coli* could not deaminate other A34 tRNA such as tRNA<sup>Ser</sup><sub>AGA</sub> from yeast or tRNA<sup>Ala</sup><sub>AGC</sub> from *B. mori* (Auxilien, Crain et al. 1996).

These results suggest that TadA substrates would be selected passively by exclusion of tRNAs that do not contain the anticodon ACG. However, the discovery of tRNA<sup>Leu</sup><sub>IAG</sub> in *O. oeni* invalidate this hypothesis as long as it is proven that the TadA found in *O. oeni* is responsible for the deamination of tRNA<sup>Arg</sup><sub>ACG</sub> and tRNA<sup>Leu</sup><sub>IAG</sub>. To demonstrate this, TadA from *O. oeni* was cloned, purified and incubated with synthetic A34 tRNA for Arg and Leu. The results were not conclusive, although suggest that both tRNAs were effectively modified by TadA. The existence of A34 unmodified tRNAs in this organism lead us to reconsider the enzyme-substrate recognition between TadA and A34 tRNAs.

Interestingly, the phyla of Firmicutes have several organisms with A34 tRNAs genes other than tRNA<sup>Arg</sup><sub>ACG</sub>, where tRNA<sup>Leu</sup><sub>IAG</sub> and tRNA<sup>Thr</sup><sub>AGU</sub> are the most common (**Table 8.1**). Based on these results, it could be interesting to answer whether tRNA<sup>Thr</sup><sub>AGU</sub> could be modified in any of these organisms, or if tRNA<sup>Leu</sup><sub>IAG</sub> is modified in other Firmicutes where is found? These questions can be answered with relative ease by sequencing their tRNAs or analyzing them using the splint ligation methodology.



## 9. Conclusions

- In eukaryotic proteomes, particularly in human, there is a tendency to translate proteins using ADAT-sensitive codons specifically for those repetitive regions enriched with TAPS amino acids.
  - These results indicate a dependence on I34 modified tRNAs for translation of ADAT-stretched proteins previously unavailable and open the possibility that the enrichment in ADAT-sensitive codons is an adaptation of the translational apparatus to improve the synthesis of proteins enriched in TAPS. Also suggest that ADAT may be playing a role in translation regulation as well as, open the possibility to use this enzyme as a novel drug target for the treatment of some diseases such as those related with mucin alterations.
- The appearance of new I34 tRNA substrates is tightly related with the appearance of ADAT-stretches enriched in ADAT-sensitive codons.
  - In Metazoa and Plantae the correlation between stretch sizes and their ADAT enrichment is high. Fungi does not preserve such correlation, but the ADAT enrichment is still high. These three kingdoms precisely have fixed its I34 tRNA variability to 7-8 different substrates.
  - On the other hand, in Protists, there is no such correlation between the ADAT stretch length and their ADAT enrichment, and the prevalence on I34 tRNA substrates is very variable with several organisms having less than 7 substrates, specially on the monophyletic group of SAR.
  - Most of Bacteria, only have 1 tRNA modified with I34, coinciding with a drop in the creation of TAPS-enriched protein regions in this kingdom. Moreover, bacterial ADAT-stretches are shorter and their enrichment in ADAT-sensitive codons is lower, suggesting that the translation of long TAPS-enriched protein regions is prevalent only in organisms with a broad range of I34 tRNAs.
  - We discovered some bacterial organisms code for previously unnoticed A34-tRNAs other than tRNA<sup>Arg</sup>. This substrate extension is specially prevalent in



## Conclusions

the Firmicutes, which surprisingly, corresponds to the phylum with the highest overall enrichment in ADAT-sensitive codons compared with the rest of bacteria. These results open the possibility that the use of additional I34 tRNAs in these species have promoted a codon bias throughout the ADAT-sensitive codons.

- Conversely, we also detect some bacterial phyla that lack all the A34-tRNA gene-encoded, and in agreement, they also lack the enzyme TadA. Based on taxonomic classification of these phyla we suggest that Thermotogae and Spirochaetes never developed the machinery to generate I34-modified tRNAs whereas Epsilonbacteria lost the system secondarily. The frequency of ADAT stretches in these bacterial organisms is one of the lowest among bacteria, and most of them do not even have the presence of ADAT-stretched regions in their proteome. Again, these results suggest that the absence of I34 tRNAs avoid the formation of TAPS-enriched regions in the proteome. Intriguingly, we also detect a small number of stretches partially enriched in ADAT-sensitive codons mostly present in the phylum of Chloroflexi, suggesting that for these punctual cases, unknown decoding systems must have evolved to deal with the translation of this regions.

## 9. References

- Agris, P. F. (2004). "Decoding the genome: a modified view." *Nucleic Acids Res* **32**(1): 223-238.
- Agris, P. F., F. A. Vendeix, et al. (2007). "tRNA's wobble decoding of the genome: 40 years of modification." *J Mol Biol* **366**(1): 1-13.
- Akashi, K., M. Takenaka, et al. (1998). "Coexistence of nuclear DNA-encoded tRNA<sup>Val</sup>(AAC) and mitochondrial DNA-encoded tRNA<sup>Val</sup>(UAC) in mitochondria of a liverwort *Marchantia polymorpha*." *Nucleic Acids Res* **26**(9): 2168-2172.
- Alberts, B. (1998). *Essential cell biology: an introduction to the molecular biology of the cell*. New York, Garland Pub.
- Aldinger, C. A., A. K. Leisinger, et al. (2012). "The absence of A-to-I editing in the anticodon of plant cytoplasmic tRNA (Arg) ACG demands a relaxation of the wobble decoding rules." *RNA Biol* **9**(10): 1239-1246.
- Alkatib, S., L. B. Scharff, et al. (2012). "The contributions of wobbling and superwobbling to the reading of the genetic code." *PLoS Genet* **8**(11): e1003076.
- Andachi, Y., F. Yamao, et al. (1987). "Occurrence of unmodified adenine and uracil at the first position of anticodon in threonine tRNAs in *Mycoplasma capricolum*." *Proc Natl Acad Sci U S A* **84**(21): 7398-7402.
- Arnez, J. G. and D. Moras (1997). "Structural and functional considerations of the aminoacylation reaction." *Trends Biochem Sci* **22**(6): 211-216.
- Ashraf, S. S., G. Ansari, et al. (1999). "The uridine in "U-turn": contributions to tRNA-ribosomal binding." *RNA* **5**(4): 503-511.
- Ashraf, S. S., E. Sochacka, et al. (1999). "Single atom modification (O<sup>-</sup>->S) of tRNA confers ribosome binding." *RNA* **5**(2): 188-194.
- Astrom, S. U. and A. S. Bystrom (1994). "Rit1, a tRNA backbone-modifying enzyme that mediates initiator and elongator tRNA discrimination." *Cell* **79**(3): 535-546.
- Atkins, J. F., R. F. Gesteland, et al. (2011). *RNA worlds: from life's origins to diversity in gene regulation*. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.
- Auxilien, S., P. F. Crain, et al. (1996). "Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of inosine 34 in the anticodon of transfer RNA." *J Mol Biol* **262**(4): 437-458.
- Baldwin, A. N. and P. Berg (1966). "Transfer ribonucleic acid-induced hydrolysis of valyladenylate bound to isoleucyl ribonucleic acid synthetase." *J Biol Chem* **241**(4): 839-845.
- Baranov, P. V., R. F. Gesteland, et al. (2002). "Recoding: translational bifurcations in gene expression." *Gene* **286**(2): 187-201.
- Barrell, B. G., A. T. Bankier, et al. (1979). "A different genetic code in human mitochondria." *Nature* **282**(5735): 189-194.
- Barta, A., G. Steiner, et al. (1984). "Identification of a site on 23S ribosomal RNA located at the peptidyl transferase center." *Proc Natl Acad Sci U S A* **81**(12): 3607-3611.
- Belhassen, E., B. Domme, et al. (1991). "Complex determination of male sterility in *Thymus vulgaris* L.: genetic and molecular analysis." *Theor Appl Genet* **82**(2): 137-143.
- Benne, R. and P. Sloof (1987). "Evolution of the mitochondrial protein synthetic machinery." *Biosystems* **21**(1): 51-68.
- Betat, H. and M. Morl (2015). "The CCA-adding enzyme: A central scrutinizer in tRNA quality control." *Bioessays* **37**(9): 975-982.
- Bjork, G. R., J. M. Durand, et al. (1999). "Transfer RNA modification: influence on translational frameshifting and metabolism." *FEBS Lett* **452**(1-2): 47-51.
- Bjork, G. R., B. Huang, et al. (2007). "A conserved modified wobble nucleoside (mcm5s2U) in lysyl-tRNA is required for viability in yeast." *RNA* **13**(8): 1245-1255.
- Bjork, G. R., K. Jacobsson, et al. (2001). "A primordial tRNA modification required for the evolution of life?" *EMBO J* **20**(1-2): 231-239.
- Caetano-Anolles, G., M. Wang, et al. (2009). "The origin, evolution and structure of the protein world." *Biochem J* **417**(3): 621-637.
- Cavalier-Smith, T. (1985). *The Evolution of genome size*. Chichester West Sussex ; New York, J. Wiley.
- Cavalli-Sforza, L. L. and A. W. Edwards (1967). "Phylogenetic analysis. Models and estimation procedures." *Am J Hum Genet* **19**(3 Pt 1): 233-257.
- Cavarelli, J., G. Eriani, et al. (1994). "The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction." *EMBO J* **13**(2): 327-337.
- Comeron, J. M. (2004). "Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence." *Genetics* **167**(3): 1293-1304.
- Cramer, F., U. Englisch, et al. (1991). "Aminoacylation of tRNAs as critical step of protein biosynthesis." *Biochimie* **73**(7-8): 1027-1035.
- Crick, F. (1970). "Central dogma of molecular biology." *Nature* **227**(5258): 561-563.
- Crick, F. H. (1966). "Codon-anticodon pairing: the wobble hypothesis." *J Mol Biol* **19**(2): 548-555.
- Cusack, S. (1997). "Aminoacyl-tRNA synthetases." *Curr Opin Struct Biol* **7**(6): 881-889.
- Chen, P., Q. Qian, et al. (2002). "A cytosolic tRNA with an unmodified adenosine in the wobble position reads a codon ending with the non-complementary nucleoside cytidine." *J Mol Biol* **317**(4): 481-492.
- Chevallier, A. and J. P. Garel (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**(3): 477-482.
- Christian, T. and Y. M. Hou (2007). "Distinct determinants of tRNA recognition by the TrmD and Trm5 methyl transferases." *J Mol Biol* **373**(3): 623-632.
- Dale, T. and O. C. Uhlenbeck (2005). "Amino acid specificity in translation." *Trends Biochem Sci* **30**(12): 659-665.

## References

- Dalluge, J. J., T. Hashizume, et al. (1996). "Quantitative measurement of dihydrouridine in RNA using isotope dilution liquid chromatography-mass spectrometry (LC/MS)." *Nucleic Acids Res* **24**(16): 3242-3245.
- de Crecy-Lagard, V., C. Brochier-Armanet, et al. (2010). "Biosynthesis of wyosine derivatives in tRNA: an ancient and highly diverse pathway in Archaea." *Mol Biol Evol* **27**(9): 2062-2077.
- Delannoy, E., M. Le Ret, et al. (2009). "Arabidopsis tRNA adenosine deaminase arginine edits the wobble nucleotide of chloroplast tRNA<sup>Arg</sup>(ACG) and is essential for efficient chloroplast translation." *Plant Cell* **21**(7): 2058-2071.
- Dittmar, K. A., J. M. Goodenbour, et al. (2006). "Tissue-specific differences in human transfer RNA expression." *PLoS Genet* **2**(12): e221.
- Dittmar, K. A., E. M. Mobley, et al. (2004). "Exploring the regulation of tRNA distribution on the genomic scale." *J Mol Biol* **337**(1): 31-47.
- Doerfel, L. K., I. Wohlgemuth, et al. (2013). "EF-P is essential for rapid synthesis of proteins containing consecutive proline residues." *Science* **339**(6115): 85-88.
- Donovan, J. and P. R. Copeland (2010). "The efficiency of selenocysteine incorporation is regulated by translation initiation factors." *J Mol Biol* **400**(4): 659-664.
- Dufresne, A., L. Garczarek, et al. (2005). "Accelerated evolution associated with genome reduction in a free-living prokaryote." *Genome Biol* **6**(2): R14.
- Durant, P. C., A. C. Bajji, et al. (2005). "Structural effects of hypermodified nucleosides in the Escherichia coli and human tRNA<sup>Lys</sup> anticodon loop: the effect of nucleosides s2U, mcm5U, mcm5s2U, mnm5s2U, t6A, and ms2t6A." *Biochemistry* **44**(22): 8078-8089.
- Duret, L. (2002). "Evolution of synonymous codon usage in metazoans." *Curr Opin Genet Dev* **12**(6): 640-649.
- Edmonds, C. G., P. F. Crain, et al. (1991). "Posttranscriptional modification of tRNA in thermophilic archaea (Archaeobacteria)." *J Bacteriol* **173**(10): 3138-3148.
- El Yacoubi, B., M. Bailly, et al. (2012). "Biosynthesis and function of posttranscriptional modifications of transfer RNAs." *Annu Rev Genet* **46**: 69-95.
- Elias, Y. and R. H. Huang (2005). "Biochemical and structural studies of A-to-I editing by tRNA:A34 deaminases at the wobble position of transfer RNA." *Biochemistry* **44**(36): 12057-12065.
- Eriani, G., M. Delarue, et al. (1990). "Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs." *Nature* **347**(6289): 203-206.
- Ernst, J., P. Kheradpour, et al. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types." *Nature* **473**(7345): 43-49.
- Felsenstein, J. (1973). "Maximum-likelihood estimation of evolutionary trees from continuous characters." *Am J Hum Genet* **25**(5): 471-492.
- Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." *J Mol Evol* **17**(6): 368-376.
- Francklyn, C. and P. Schimmel (1989). "Aminoacylation of RNA minihelices with alanine." *Nature* **337**(6206): 478-481.
- Fraser, T. H. and A. Rich (1975). "Amino acids are not all initially attached to the same position on transfer RNA molecules." *Proc Natl Acad Sci U S A* **72**(8): 3044-3048.
- Gerber, A. P. and W. Keller (1999). "An adenosine deaminase that generates inosine at the wobble position of tRNAs." *Science* **286**(5442): 1146-1149.
- Greber, B. J., D. Boehringer, et al. (2014). "Architecture of the large subunit of the mammalian mitochondrial ribosome." *Nature* **505**(7484): 515-519.
- Grosjean, H. (2005). *Fine-tuning of RNA functions by modification and editing*. Berlin ; New York, Springer.
- Grosjean, H., V. de Crecy-Lagard, et al. (2010). "Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes." *FEBS Lett* **584**(2): 252-264.
- Gupta, P. R., Dey A, Vijan A, Gartia B. (2017). "In Silico Structure Modeling and Characterization of Hypothetical Protein YP\_004590319.1 Present in *Enterobacter aerogens*." *J Proteomics Bioinform* **10**: 152-170.
- Hirokawa, G., N. Demeshkina, et al. (2006). "The ribosome-recycling step: consensus or controversy?" *Trends Biochem Sci* **31**(3): 143-149.
- Holley, R. W. (1965). "Structure of an alanine transfer ribonucleic acid." *JAMA* **194**(8): 868-871.
- Hou, Y. M. and P. Schimmel (1988). "A simple structural feature is a major determinant of the identity of a transfer RNA." *Nature* **333**(6169): 140-145.
- Hou, Y. M. and P. Schimmel (1989). "Modeling with in vitro kinetic parameters for the elaboration of transfer RNA identity in vivo." *Biochemistry* **28**(12): 4942-4947.
- Ibba, M., S. Morgan, et al. (1997). "A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases." *Science* **278**(5340): 1119-1122.
- Ikemura, T. (1981). "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system." *J Mol Biol* **151**(3): 389-409.
- Ikemura, T. (1985). "Codon usage and tRNA content in unicellular and multicellular organisms." *Mol Biol Evol* **2**(1): 13-34.
- Jackman, J. E. and J. D. Alfonzo (2013). "Transfer RNA modifications: nature's combinatorial chemistry playground." *Wiley Interdiscip Rev RNA* **4**(1): 35-48.
- Jackson, R. J., C. U. Hellen, et al. (2010). "The mechanism of eukaryotic translation initiation and principles of its regulation." *Nat Rev Mol Cell Biol* **11**(2): 113-127.
- Johansson, M. J., A. Esberg, et al. (2008). "Eukaryotic wobble uridine modifications promote a functionally redundant decoding system." *Mol Cell Biol* **28**(10): 3301-3312.
- Kamenski, P., O. Kolesnikova, et al. (2007). "Evidence for an adaptation mechanism of mitochondrial translation via tRNA import from the cytosol." *Mol Cell* **26**(5): 625-637.

- Kanaya, S., Y. Yamada, et al. (2001). "Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis." *J Mol Evol* **53**(4-5): 290-298.
- Kanaya, S., Y. Yamada, et al. (1999). "Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis." *Gene* **238**(1): 143-155.
- Kapp, L. D. and J. R. Lorsch (2004). "The molecular mechanics of eukaryotic translation." *Annu Rev Biochem* **73**: 657-704.
- Kim, S. H., F. L. Suddath, et al. (1974). "Three-dimensional tertiary structure of yeast phenylalanine transfer RNA." *Science* **185**(4149): 435-440.
- Kim, S. H., J. L. Sussman, et al. (1974). "The general structure of transfer RNA molecules." *Proc Natl Acad Sci U S A* **71**(12): 4970-4974.
- Kreda, S. M., C. W. Davis, et al. (2012). "CFTR, mucins, and mucus obstruction in cystic fibrosis." *Cold Spring Harb Perspect Med* **2**(9): a009589.
- Kufe, D. W. (2009). "Mucins in cancer: function, prognosis and therapy." *Nat Rev Cancer* **9**(12): 874-885.
- Kuratani, M., R. Ishii, et al. (2005). "Crystal structure of tRNA adenosine deaminase (TadA) from *Aquifex aeolicus*." *J Biol Chem* **280**(16): 16002-16008.
- Lagerkvist, U. (1986). "Unconventional methods in codon reading." *Bioessays* **4**(5): 223-226.
- Lassak, J., D. N. Wilson, et al. (2016). "Stall no more at polyproline stretches with the translation elongation factors EF-P and IF-5A." *Mol Microbiol* **99**(2): 219-235.
- Lee, W. H., Y. K. Kim, et al. (2007). "Crystal structure of the tRNA-specific adenosine deaminase from *Streptococcus pyogenes*." *Proteins* **68**(4): 1016-1019.
- Leidel, S., P. G. Pedrioli, et al. (2009). "Ubiquitin-related modifier Urm1 acts as a sulphur carrier in thiolation of eukaryotic transfer RNA." *Nature* **458**(7235): 228-232.
- Levitt, M. (1969). "Detailed molecular model for transfer ribonucleic acid." *Nature* **224**(5221): 759-763.
- Lewin, B. (1994). *Genes V*. Oxford ; New York, Oxford University Press.
- Li, J. Y., L. P. Ye, et al. (2015). "Comparative proteomic analysis of the silkworm middle silk gland reveals the importance of ribosome biogenesis in silk protein production." *J Proteomics* **126**: 109-120.
- Li, W.-H. (1997). *Molecular evolution*. Sunderland, Mass., Sinauer Associates.
- Li, Z. and M. P. Deutscher (1996). "Maturation pathways for *E. coli* tRNA precursors: a random multienzyme process in vivo." *Cell* **86**(3): 503-512.
- Lim, V. I. and J. F. Curran (2001). "Analysis of codon:anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure." *RNA* **7**(7): 942-957.
- Losey, H. C., A. J. Ruthenburg, et al. (2006). "Crystal structure of *Staphylococcus aureus* tRNA adenosine deaminase TadA in complex with RNA." *Nat Struct Mol Biol* **13**(2): 153-159.
- Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." *Nucleic Acids Res* **25**(5): 955-964.
- Luo, M. and V. L. Schramm (2008). "Transition state structure of *E. coli* tRNA-specific adenosine deaminase." *J Am Chem Soc* **130**(8): 2649-2655.
- Madore, E., C. Florentz, et al. (1999). "Effect of modified nucleotides on *Escherichia coli* tRNA<sup>Glu</sup> structure and on its aminoacylation by glutamyl-tRNA synthetase. Predominant and distinct roles of the mnm5 and s2 modifications of U34." *Eur J Biochem* **266**(3): 1128-1135.
- Mager-Heckel, A. M., N. Entelis, et al. (2007). "The analysis of tRNA import into mammalian mitochondria." *Methods Mol Biol* **372**: 235-253.
- Mahlab, S., T. Tuller, et al. (2012). "Conservation of the relative tRNA composition in healthy and cancerous tissues." *RNA* **18**(4): 640-652.
- McClain, W. H. and K. Foss (1988). "Changing the identity of a tRNA by introducing a G-U wobble pair near the 3' acceptor end." *Science* **240**(4853): 793-796.
- Meloni, B. P., L. M. Brookes, et al. (2015). "Poly-arginine and arginine-rich peptides are neuroprotective in stroke models." *J Cereb Blood Flow Metab* **35**(6): 993-1004.
- Moazed, D. and H. F. Noller (1989). "Interaction of tRNA with 23S rRNA in the ribosomal A, P, and E sites." *Cell* **57**(4): 585-597.
- Moore, V. G., R. E. Atchison, et al. (1975). "Identification of a ribosomal protein essential for peptidyl transferase activity." *Proc Natl Acad Sci U S A* **72**(3): 844-848.
- Nagel, G. M. and R. F. Doolittle (1991). "Evolution and relatedness in two aminoacyl-tRNA synthetase families." *Proc Natl Acad Sci U S A* **88**(18): 8121-8125.
- Nissen, P., J. Hansen, et al. (2000). "The structural basis of ribosome activity in peptide bond synthesis." *Science* **289**(5481): 920-930.
- Noma, A., Y. Kirino, et al. (2006). "Biosynthesis of wybutosine, a hyper-modified nucleoside in eukaryotic phenylalanine tRNA." *EMBO J* **25**(10): 2142-2154.
- Noma, A. and T. Suzuki (2006). "Ribonucleome analysis identified enzyme genes responsible for wybutosine synthesis." *Nucleic Acids Symp Ser (Oxf)* **50**: 65-66.
- Novoa, E. M., M. Pavon-Eternod, et al. (2012). "A role for tRNA modifications in genome structure and codon usage." *Cell* **149**(1): 202-213.
- Novoa, E. M. and L. Ribas de Pouplana (2012). "Speeding with control: codon usage, tRNAs, and ribosomes." *Trends Genet* **28**(11): 574-581.
- Nuttall, G. H. and O. Inchley (1904). "An improved Method of measuring the amount of Precipitum in connection with Tests with Precipitating Antisera." *J Hyg (Lond)* **4**(2): 201-206.

## References

- Pfzinger, H., J. H. Weil, et al. (1990). "Codon recognition mechanisms in plant chloroplasts." *Plant Mol Biol* **14**(5): 805-814.
- Phizicky, E. M. and A. K. Hopper (2010). "tRNA biology charges to the front." *Genes Dev* **24**(17): 1832-1860.
- Polycarpo, C., A. Ambrogelly, et al. (2003). "Activation of the pyrrolysine suppressor tRNA requires formation of a ternary complex with class I and class II lysyl-tRNA synthetases." *Mol Cell* **12**(2): 287-294.
- Prokopowich, C. D., T. R. Gregory, et al. (2003). "The correlation between rDNA copy number and genome size in eukaryotes." *Genome* **46**(1): 48-50.
- Ribas de Pouplana, L. and P. Schimmel (2001). "Operational RNA code for amino acids in relation to genetic code in evolution." *J Biol Chem* **276**(10): 6881-6884.
- Ribas de Pouplana, L. and P. Schimmel (2001). "Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem." *Cell* **104**(2): 191-193.
- Rich, A. and U. L. RajBhandary (1976). "Transfer RNA: molecular structure, sequence, and properties." *Annu Rev Biochem* **45**: 805-860.
- Richmond, R. C. (1970). "Non-Darwinian evolution: a critique." *Nature* **225**(5237): 1025-1028.
- Rinehart, J., B. Krett, et al. (2005). "Saccharomyces cerevisiae imports the cytosolic pathway for Gln-tRNA synthesis into the mitochondrion." *Genes Dev* **19**(5): 583-592.
- Robertus, J. D., J. E. Ladner, et al. (1974). "Structure of yeast phenylalanine tRNA at 3 Å resolution." *Nature* **250**(467): 546-551.
- Rodin, S., A. Rodin, et al. (1996). "The presence of codon-anticodon pairs in the acceptor stem of tRNAs." *Proc Natl Acad Sci U S A* **93**(10): 4537-4542.
- Rose, M. C. and J. A. Voynow (2006). "Respiratory tract mucin genes and mucin glycoproteins in health and disease." *Physiol Rev* **86**(1): 245-278.
- Rould, M. A., J. J. Perona, et al. (1989). "Structure of E. coli glutamyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 Å resolution." *Science* **246**(4934): 1135-1142.
- Rubio, M. A., I. Pastar, et al. (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**(19): 7821-7826.
- Rudinger, J., C. Florentz, et al. (1994). "Histidylolation by yeast HisRS of tRNA or tRNA-like structure relies on residues -1 and 73 but is dependent on the RNA context." *Nucleic Acids Res* **22**(23): 5031-5037.
- Ruff, M., S. Krishnaswamy, et al. (1991). "Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp)." *Science* **252**(5013): 1682-1689.
- Salinas, T., A. M. Duchene, et al. (2008). "Recent advances in tRNA mitochondrial import." *Trends Biochem Sci* **33**(7): 320-329.
- Sanger, F. (1952). "The arrangement of amino acids in proteins." *Adv Protein Chem* **7**: 1-67.
- Sanger, F. and E. O. Thompson (1952). "The amino-acid sequence in the glycol chain of insulin." *Biochem J* **52**(1): iii.
- Schaub, M. and W. Keller (2002). "RNA editing by adenosine deaminases generates RNA and protein diversity." *Biochimie* **84**(8): 791-803.
- Schimmel, P. and E. Schmidt (1995). "Making connections: RNA-dependent amino acid recognition." *Trends Biochem Sci* **20**(1): 1-2.
- Schneider, A. and L. Marechal-Drouard (2000). "Mitochondrial tRNA import: are there distinct mechanisms?" *Trends Cell Biol* **10**(12): 509-513.
- Shinozaki, K., M. Ohme, et al. (1986). "The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression." *EMBO J* **5**(9): 2043-2049.
- Sibler, A. P., G. Dirheimer, et al. (1986). "Codon reading patterns in Saccharomyces cerevisiae mitochondria based on sequences of mitochondrial tRNAs." *FEBS Lett* **194**(1): 131-138.
- Smith, D. W. and A. L. McNamara (1971). "Specialization of rabbit reticulocyte transfer RNA content for hemoglobin synthesis." *Science* **171**(3971): 577-579.
- Smith, D. W., V. N. Meltzer, et al. (1974). "A comparison of rabbit liver and reticulocyte transfer RNA: evidence of unique species in reticulocytes." *Biochim Biophys Acta* **349**(3): 366-375.
- Sorensen, M. A. and S. Pedersen (1991). "Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate." *J Mol Biol* **222**(2): 265-280.
- Spears, J. L., M. A. Rubio, et al. (2011). "A single zinc ion is sufficient for an active Trypanosoma brucei tRNA editing deaminase." *J Biol Chem* **286**(23): 20366-20374.
- Sprinzel, M., C. Horn, et al. (1998). "Compilation of tRNA sequences and sequences of tRNA genes." *Nucleic Acids Res* **26**(1): 148-153.
- Suzuki, T., A. Nagao, et al. (2011). "Human mitochondrial tRNAs: biogenesis, function, structural aspects, and diseases." *Annu Rev Genet* **45**: 299-329.
- Tisne, C., M. Rigourd, et al. (2000). "NMR and biochemical characterization of recombinant human tRNA(Lys)3 expressed in Escherichia coli: identification of posttranscriptional nucleotide modifications required for efficient initiation of HIV-1 reverse transcription." *RNA* **6**(10): 1403-1412.
- Tomita, K. and A. M. Weiner (2001). "Collaboration between CC- and A-adding enzymes to build and repair the 3'-terminal CCA of tRNA in Aquifex aeolicus." *Science* **294**(5545): 1334-1336.
- Torres, A. G., E. Batlle, et al. (2014). "Role of tRNA modifications in human diseases." *Trends Mol Med* **20**(6): 306-314.
- Torres, A. G., D. Pineyro, et al. (2015). "Inosine modifications in human tRNAs are incorporated at the precursor tRNA level." *Nucleic Acids Res* **43**(10): 5145-5157.
- Tsutsumi, S., R. Sugiura, et al. (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**(46): 33459-33465.
- Tuller, T., A. Carmi, et al. (2010). "An evolutionarily conserved mechanism for controlling the efficiency of protein translation." *Cell* **141**(2): 344-354.

- Urbonavicius, J., Q. Qian, et al. (2001). "Improvement of reading frame maintenance is a common function for several tRNA modifications." *EMBO J* **20**(17): 4863-4873.
- Urrutia, A. O. and L. D. Hurst (2001). "Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection." *Genetics* **159**(3): 1191-1199.
- Urrutia, A. O. and L. D. Hurst (2003). "The signature of selection mediated by expression on human genes." *Genome Res* **13**(10): 2260-2264.
- Watanabe, Y., H. Tsurui, et al. (1997). "Primary sequence of mitochondrial tRNA(Arg) of a nematode *Ascaris suum*: occurrence of unmodified adenosine at the first position of the anticodon." *Biochim Biophys Acta* **1350**(2): 119-122.
- Watson, J. D. (1987). *Molecular biology of the gene*. Menlo Park, Calif., Benjamin/Cummings.
- Watson, J. D. and F. H. Crick (1953). "The structure of DNA." *Cold Spring Harb Symp Quant Biol* **18**: 123-131.
- Wende, S., S. Bonin, et al. (2015). "The identity of the discriminator base has an impact on CCA addition." *Nucleic Acids Res* **43**(11): 5617-5629.
- Wolf, J., A. P. Gerber, et al. (2002). "tadA, an essential tRNA-specific adenosine deaminase from *Escherichia coli*." *EMBO J* **21**(14): 3841-3851.
- Yamamoto, H., Y. Qin, et al. (2014). "EF-G and EF4: translocation and back-translocation on the bacterial ribosome." *Nat Rev Microbiol* **12**(2): 89-100.
- Yokobori, S., A. Kitamura, et al. (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**(13): 6531-6543.
- Zhou, W., D. Karcher, et al. (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**(4): 1985-1997.
- Zuckerlandl, E., R. T. Jones, et al. (1960). "A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis." *Proc Natl Acad Sci U S A* **46**(10): 1349-1360.
- Zuckerlandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." *J Theor Biol* **8**(2): 357-366.

## References