A Role for tRNA Modifications in Genome Structure and Codon Usage

Eva Maria Novoa,[1] Mariana Pavon-Eternod,[2] Tao Pan,[2] and Lluís Ribas de Pouplana[1,3,*]
[1]Institute for Research in Biomedicine, c/ Baldiri Reixac 15-21, 08028 Barcelona, Catalonia, Spain
[2]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA
[3]Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, 08010 Barcelona, Catalonia, Spain
*Correspondence: lluis.ribas@irbbarcelona.org
DOI 10.1016/j.cell.2012.01.050

SUMMARY

Transfer RNA (tRNA) gene content is a differentiating feature of genomes that contributes to the efficiency of the translational apparatus, but the principles shaping tRNA gene copy number and codon composition are poorly understood. Here, we report that the emergence of two specific tRNA modifications shaped the structure and composition of all extant genomes. Through the analysis of more than 500 genomes, we identify two kingdom-specific tRNA modifications as major contributors that separated archaeal, bacterial, and eukaryal genomes in terms of their tRNA gene composition. We show that, contrary to prior observations, genomic codon usage and tRNA gene frequencies correlate in all kingdoms if these two modifications are taken into account and that presence or absence of these modifications explains patterns of gene expression observed in previous studies. Finally, we experimentally demonstrate that human gene expression levels correlate well with genomic codon composition if these identified modifications are considered.

INTRODUCTION

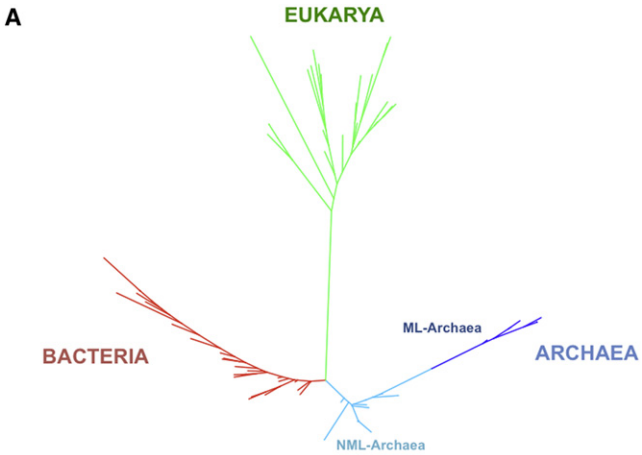
Transfer RNAs (tRNAs) are present in all living organisms, acting as adaptors that link amino acids to codons in messenger RNAs (mRNA). Based on their aminoacylation identity, all tRNAs are subdivided into 20 accepting groups (alloacceptors). Each group comprises several tRNAs (isoacceptors) that translate synonymous codons with the same amino acid thanks to synonymous anticodons that vary mostly at the third position. The redundancy of the genetic code is due to synonymous codons, and solved by isoacceptor tRNAs.

tRNA genes tend to be present in multiple copies in the genomes of most organisms, from prokaryotes to eukaryotes, but the number of gene copies for each tRNA species (tRNAs with the same anticodon) varies widely from species to species (Marck and Grosjean, 2002). For any actively dividing cell, the translation efficiency of a given codon is determined by the amount of tRNA in the cell (Ikemura, 1981; Bennetzen and Hall, 1982; Sharp et al., 1988; Man and Pilpel, 2007; Akashi, 2003; Elf et al., 2003; Dittmar et al., 2005). The concentration of each tRNA is determined by its number of gene copies in the genome (Tuller et al., 2010a). Thus, tRNA gene content determines relative tRNA isoacceptor abundances that, in turn, determine codon translation efficiency. Therefore, the study of tRNA gene content bias may help explaining codon usage biases in extant genomes.

Previous reports have shown that the number of genes coding for each tRNA is not conserved between kingdoms (Gerber and Keller, 2001; Marck and Grosjean, 2002). The variability in tRNA gene number is extreme in some cases: certain tRNA species are absent in entire branches of the phylogenetic tree, whereas others are clearly predominant (e.g., in *Homo sapiens* 29 out of the 43 $tRNA^{Ala}$ genes (68%) correspond to the isoacceptor $tRNA^{Ala}_{AGC}$). The factors that influence tRNA gene copy number within genomes have been studied mostly in individual species (Withers et al., 2006; Gonos and Goddard, 1990; Kanaya et al., 1999; Dong et al., 1996), but the principles that govern the evolution of tRNA gene populations remain unknown.

In addition to the variability in tRNA gene content, the diversity of tRNA populations is further increased by species-specific base modifications. Thus, the tRNA signature of each species, defined as the total set of mature tRNAs that results from tRNA gene transcription, tRNA maturation, and the action of modification enzymes, is a complex evolutionary trait. Little is known about the parameters that shape the tRNA signature of species in evolution.

Two enzymes are known to cause modifications in base 34 of the anticodon that increase codon-pairing ability: tRNA-dependent adenosine deaminases (ADATs) and tRNA-dependent uridine methyltransferases (UMs) (Agris et al., 2007). tRNA-adenosine deaminases are essential enzymes found in Bacteria and Eukarya that catalyze the conversion of adenine-34 to inosine-34 (A-to-I editing) (Wolf et al., 2002; Gerber and Keller, 1999; Maas and Rich, 2000). I34 is able to wobble with adenine, cytosine, and uridine (Gerber and Keller, 2001). Thus, INN anticodons are capable of pairing with three different codons. Unlike in Bacteria, where ADAT only modifies $tRNA^{Arg}$, in Eukarya a heterodimeric form of this enzyme (hetADAT) formed by Tad2p and Tad3p deaminates several tRNAs (Gerber and Keller, 1999). On the other hand, bacterial UMs, modify uridine to xo^5U_{34}, enabling its pairing with adenine, guanosine and uridine (Yokoyama et al., 1985). Two enzymes have been identified as responsible for the last step of xo^5U modifications: CmoA and CmoB (Näsvall et al., 2004).



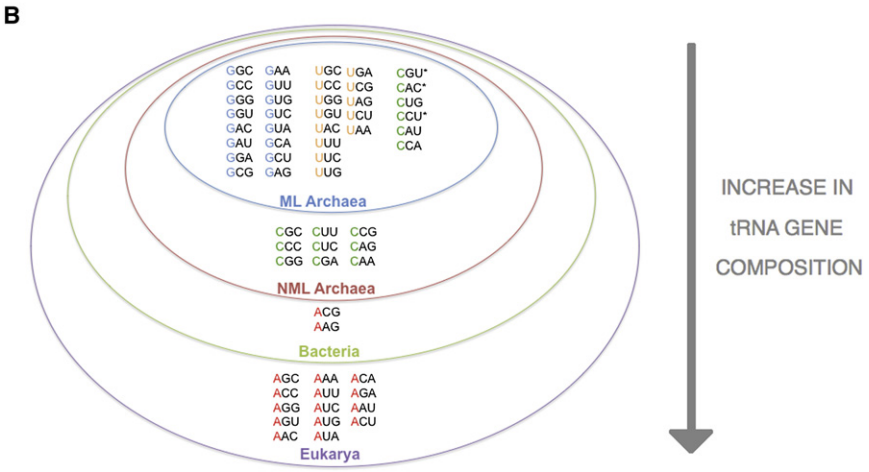


Figure 1. Genome Phylogeny Based on tRNA Gene Content

(A) Distance-based phylogeny based on tRNA gene content, performed with equal number of species of each kingdom. The four phylogenetic clusters have been labeled accordingly. The phylogeny performed with the whole set of 527 species is consistent with these results (see Figure S1).

(B) Diagram showing the increase in tRNA population complexity in the four main phylogenetic clusters found in this work (each tRNA is designated by its anticodon sequence). Each base at the wobble position is colored according to its chemical nature. Anticodons labeled with an asterisk (CGU, CAC, CCU) correspond to tRNA genes that are not found in all species comprising the ML-Archaea clade.

In this work, we have analyzed the distribution and abundance of all tRNA genes in more than 500 species across the three kingdoms of life. We first confirmed that tRNA gene composition can be considered a single trait that recapitulates the main evolutionary lines of the tree of life. Using principal component analysis, we identified those tRNA isoacceptors that became positively selected (increased in number) in Bacteria and Eukarya. Our results indicate that the appearance of UMs and hetADATs contributed to the divergence of eukaryal and bacterial genomes from their archaeal counterparts. The effect of the modifications caused by these enzymes increased the decoding capacity of modified tRNAs which, therefore, were positively selected during evolution. The diverse codon usage biases displayed by Bacteria and Eukarya are, at least partly, due to the different modification strategies used to improve translation efficiency, which are kingdom specific.

RESULTS

tRNA Gene Content as a Tool for Phylogenetic Analysis

The short sequence length of tRNAs, and their susceptibility to be transferred horizontally, limits the usefulness of their sequences for phylogenetic analysis. But tRNA gene content, defined as the set of tRNA genes used by a given organism to translate its genome, is unaffected by these limitations. In gene content-based phylogenies the evolutionary distance between species is calculated on the basis of acquisition or loss of genes. Gene content analyses using genome sequences (Snel et al., 1999; Iwasaki and Takagi, 2007; Fitz-Gibbon and House, 1999), protein domain content (Yang et al., 2005), and whole-proteome comparisons (Tekaia et al., 1999) have been previously reported.

Using tRNA gene content analysis, we have built a phylogenetic tree of more than 500 species that correctly identifies four known clades: (1) *Methanococcus*-like Archaea, (2) non-*Methanococcus*-like Archaea, (3) Bacteria, and (4) Eukarya (Figure 1A, see also Figure S1 available online). As can be seen in Figure 1A, tRNA gene content as a single trait follows the evolution of the whole tree of life, correctly clustering species into their corresponding kingdoms. Although this method is not powerful enough to correctly resolve the inner topology of individual clades, several outliers in tRNA signatures that have been previously reported (Man and Pilpel, 2007) are correctly identified by our approach. This indicates that kingdom-specific parameters drove the divergence of tRNA gene populations between the three kingdoms of life.

The four clades found in our gene-content analysis correspond to different levels of tRNA population complexity. Indeed, the tRNA gene populations of the clades vary from the relatively simple tRNA gene composition of Archaea, to an intermediate situation in Bacteria, and the most complex tRNA gene set found in Eukarya (Figure 1B). This increase in complexity implies that, along evolution, the number of tRNA species tended to increase through duplications or changes in anticodon specificity. Interestingly, the fact that *Methanococcus*-related species present

the simplest decoding strategy coincides with the proposed ancestral nature of this clade (Stetter, 1996; Brochier and Philippe, 2002).

To characterize the four identified phylogenetic clades, we quantified and analyzed the distribution of tRNA isoacceptor gene copy numbers within each of these four groups. As can be seen in Figure 2, each clade has different tRNA gene abundances and, more interestingly, unequal enrichment of certain tRNA isoacceptors. The archaeal clades are characterized by a relatively uniform distribution of tRNA gene copy numbers, with little variation between isoacceptors (e.g., all tRNA isoacceptors coding for alanine have similar gene frequencies). Thus, Archaea presents the simplest decoding scenario, with a minimal set of tRNA isoacceptors (Figure 1B) and uniform abundances of tRNA genes (Figure 2). In contrast Bacteria and Eukarya are more complex, both in terms of relative number of tRNA isoacceptors and in differences in the frequencies of tRNA gene copy numbers.

The loss of uniformity in tRNA gene abundances is not equivalent in Eukarya and Bacteria. For example, tRNAs with ANN anticodons tend to be absent both in archaeal and bacterial genomes, whereas in Eukarya they are the most abundant isoacceptors in four-codon (Pro, Ala, Val, Thr) and six-codon (Ser, Leu, Arg) tRNA sets (Figure 2). It is unclear, however, why should such selection act in a given kingdom and not in another. To try to answer this question, we first performed Principal Component Analysis (PCA) to statistically identify the tRNA isoacceptors that have been positively selected in each of the kingdoms.

Statistical Analysis of tRNA Gene Frequencies

PCA is a mathematical procedure that uses orthogonal transformation to reduce the dimensions of the data (correlated variables, in our case, tRNA gene frequencies), obtaining new variables (principal components, PCs) that are linear combinations of the original variables. Multivariate statistical analysis methods like PCA are particularly well adapted to the multidimensional nature of tRNA gene content data. If the original variables are correlated, most of the variance can be condensed in the two first PCs (PC1 and PC2). Analysis of our data shows that PC1 and PC2 account for 64.5% of the variance of tRNA gene content values, allowing us to analyze our results in two dimensions (Figure 3).

The *scores* plot—the transformed variable values (Figure 3A)—correctly clusters the species used in this analysis into their three respective kingdoms, and shows that PC1 is the principal component responsible for the separation of Bacteria, whereas PC2 is responsible for the separation of Eukarya (confirmed by t test, p values of 1e-5 and 2e-16, respectively). On the other hand, the *loadings* plot (Figure 3B) identifies which variables (tRNA isoacceptors) are contributing most to the differences between clusters. Top-ranked tRNA isoacceptors that are significantly associated to Bacteria and Eukarya are included inside an ellipse. The individual correlation values are listed in Table S1. Our data shows that eukaryal species present a positive selection of tRNA(ANN) isoacceptors belonging to four-codon families (Val, Pro, Ala, Thr), six-codon families (Leu, Ser) and split tRNA sets (Ile). On the other hand, bacterial species positively selected tRNA(UNN) isoacceptors for the same codon families.

The analysis of additional PCs was also performed to identify minor contributors to the differences between kingdom-specific tRNA gene populations (Figure S2). Interestingly, PC3 separates both Bacteria and Eukarya from Archaea due to the contribution of $tRNA^{Arg}$(ACG), confirming the importance of ANN isoacceptor tRNAs in the divergence of tRNA gene populations in the three kingdoms of life (r = 0.44, p value = 5.6e-27).

tRNA Modification as a Factor in Translational Efficiency

Translational efficiency is increased by optimized codons, i.e., those codons that correspond to the most abundant tRNA species (Hershberg and Petrov, 2008). Therefore, the positive selection of tRNA isoacceptors that we observe in our data could be due to the increased translational efficiency allowed by these tRNAs. As mentioned previously, kingdom-specific modifying enzymes exist that can increase the translational efficiency of tRNAs through modifications of the anticodon wobble base. We hypothesized that the selection of certain tRNAs over other isoacceptors, i.e., those identified in our analysis, may be due to their ability to incorporate anticodon modifications that increase their pairing repertoire (Figure S3).

If base modifications in the anticodon increase translational efficiency then those anticodons capable of accepting I_{34} and xo^5U_{34} modifications should be positively selected in the species where the corresponding modification enzymes exist. We first checked whether genes coding for tRNA(ANN) isoacceptors capable of being modified by hetADATs are overrepresented (Table 1) in species that contain these enzymes. This is exactly the case, indicating that the activity of hetADATs is exerting a selective force on the tRNA pool. We then checked whether genes coding for tRNA(UNN) isoacceptors modifiable by UMs are enriched among Bacteria. Indeed, UNN anticodons that are modified by UMs are enriched in bacterial genomes, indicating that the activity of UMs is associated with the tRNA composition of bacterial species toward U34 tRNAs (Table 1).

The analysis of further PCs supports the role of these two tRNA modifications in the divergence of tRNA gene populations. As mentioned above, PC3 clusters the bacterial and eukaryal kingdoms, and separates them from the archaeal species, mainly due to the contribution of $tRNA^{Arg}$(ACG). This tRNA isoacceptor is the only tRNA species deaminated by ADATs both in Bacteria (through TadA) and Eukarya (through Tad2/Tad3). Thus, our analysis indicates that the vast majority of the contributions to the segregation of extant tRNA gene populations are related to the activity of anticodon-modifying enzymes.

It should be noted that sequence modifications outside the anticodon can also have effects on codon:anticodon interactions (Geslain and Pan, 2010; Ledoux et al., 2009). However, to our knowledge, tRNA modifications outside the anticodon have not been found to expand the decoding capacity of tRNAs. The analysis of the full set of known tRNA anticodon modification enzymes (Table S2) reveals that only bacterial UMs and eukaryal hetADATs display phylogenetic distributions and sets of tRNA substrates fully compatible with the families of tRNAs found to be enriched in our study.

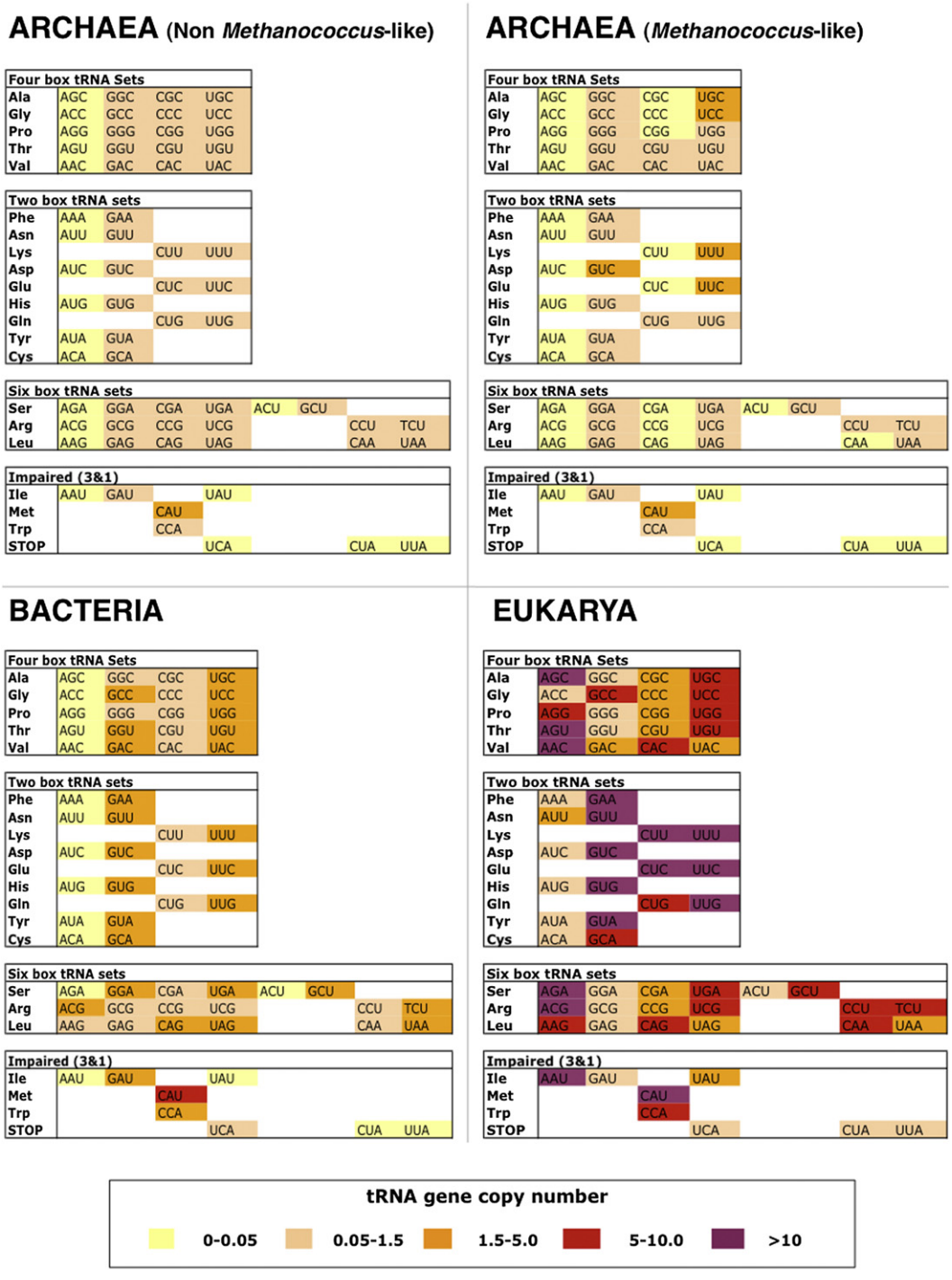


Figure 2. Unequal Enrichment of tRNA Isoacceptors Is Kingdom Specific

Mean tRNA abundances in the four phylogenetic clusters identified by gene content analysis: (1) *Methanococcus*-like Archaea, (2) non-*Methanococcus*-like Archaea, (3) Bacteria, and (4) Eukarya. Each tRNA anticodon is colored according to its average number of encoding tRNA genes. To deal with exceptional cases such as *Ferroplasma acidarmanus*, which is the sole archaea with a $tRNA^{Leu}$(AAG) gene (Marck and Grosjean, 2002), we have considered as absent those tRNA isoacceptors whose average tRNA gene copy number is between 0 and 0.05 (shown in yellow).

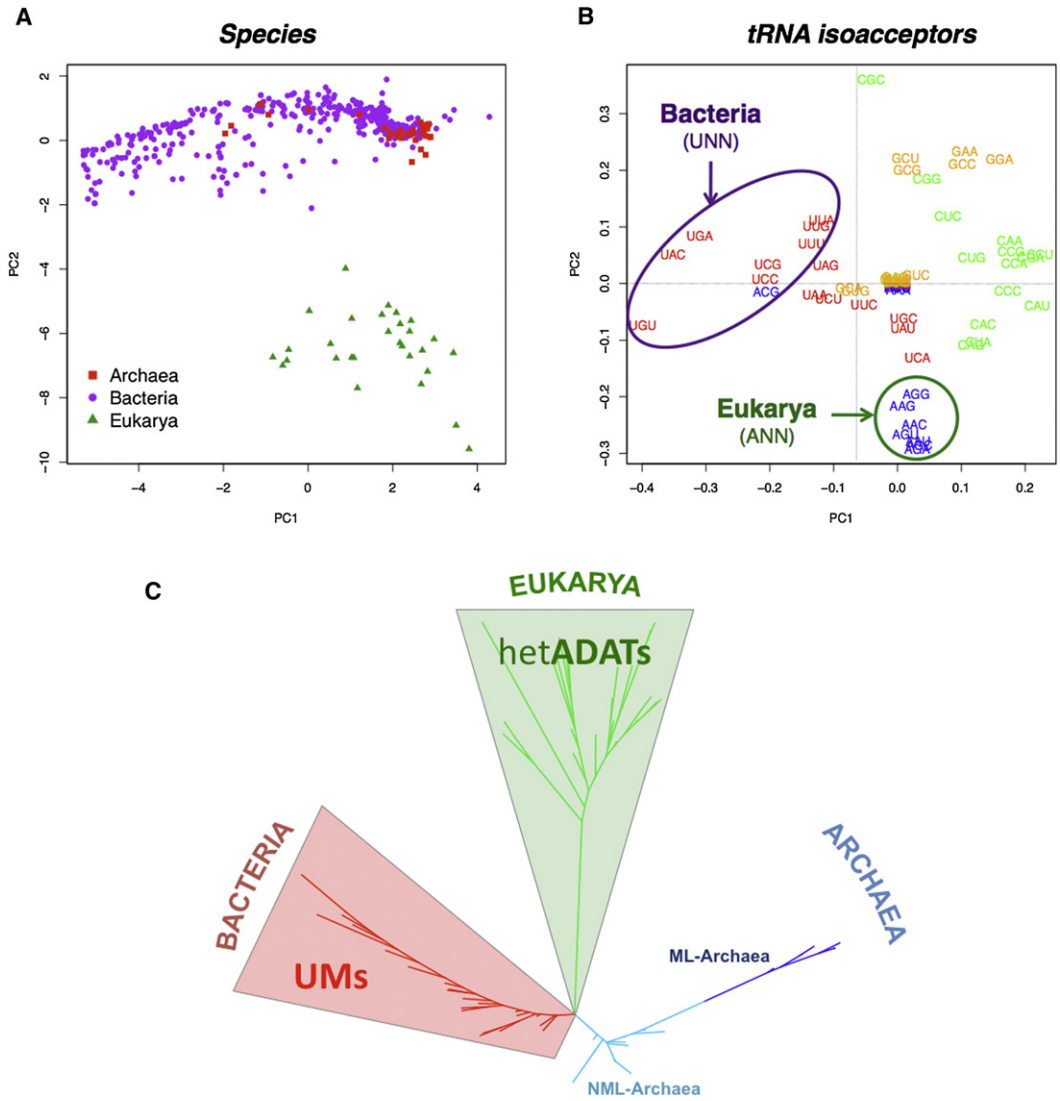


Figure 3. Identification and Quantification of Overrepresented tRNA Isoacceptors

(A) Biplot of the scores after performing Principal Component Analysis (PCA). Archaea (red), Bacteria (purple) and Eukarya (green) are distinguishable clusters using this analysis. The archaeal outliers correspond to *Methanococcus* species, which were already identified as a separate cluster using the tRNA gene content analysis.

(B) Biplot of the loadings, indicating the tRNA isoacceptors whose frequencies contribute the most to each of the clusters. Each anticodon has been colored depending on its wobble base. The ellipses surround those anticodons that are significantly associated to the PCs, either with PC1 negative values, which correspond to Bacteria (purple), or with PC2 negative values, which correspond to Eukarya (green) (see Table S1 for the individual correlation values). See also Figure S2 and Table S2.

(C) Genome phylogeny based on tRNA-gene content. The distributions of the two wobble base modification enzymes that act upon the tRNA isoacceptors identified in the PCA are shown. Uridine methyltransferases (UMs, labeled in red) are exclusively distributed across the bacterial kingdom. Heterodimeric adenosine deaminases (ADATs, labeled in green) are exclusively distributed in eukaryotes. Homodimeric forms of ADATs (TadA) are found in bacteria, but they only increase the decoding capacity of $tRNA^{Arg}$, and for simplicity, are not shown in the phylogeny.

Correlation between tRNA Gene Abundances and Codon Usage

Several studies performed on unicellular species have shown a correlation between tRNA abundance and codon usage (Ikemura, 1981; Ran and Higgs, 2010; Kanaya et al., 2001; Dong et al., 1996). In higher eukaryotes the search for this correlation has been less successful (Kanaya et al., 2001; dos Reis et al., 2004), and it has been proposed that in these species translation efficiency might not be the primary factor influencing codon usage (Kanaya et al., 2001). Studies in *Drosophila melanogaster* have concluded that in this organism selection acts to increase translation accuracy (Akashi, 1994; Moriyama and Powell, 1998), whereas other authors have linked codon usage in metazoans to several parameters, including average gene length

Table 1. Overrepresented tRNA Genes Correspond Exactly to Those Isoacceptors Modifiable at the Wobble Position by UMs and ADATs

	ADAT Gene	Anticodons Modified by ADATs	A34 Anticodons with RGF > 1.6[a]
Archaea			
Any species	—	—	none
Bacteria			
E. coli	*tadA*	ACG	ACG
Eukarya			
S. cerevisiae	*tad2p/tad3p*	AGA, AGG, AGU, AAC, AGA, ACG, AAU	AGA, AGG, AGU, AAC, AGA, ACG, AAU
H. sapiens	*tad2/tad3*	AGA, AGG, AGU, AAC, AGA, ACG, AAU, AAG	AGA, AGG, AGU, AAC, AGA, ACG, AAU, AAG
	UM Gene	Anticodons Modified by UMs	U34 Anticodons with RGF > 1.6[a]
Archaea			
Any species	—	—	none
Bacteria			
S. enterica	*cmoA/cmoB*	UGC, UGG, UGU, UAC, UGA, UAG	UGC, UGG, UGU, UAC, UGA, UAG
Eukarya			
Any species	—	—	none

[a]The RGF threshold was chosen such that the overrepresented tRNA isoacceptors also correspond to the most abundant isoacceptor among its tRNA codon family.

(Duret and Mouchiroud, 1999), cost of proofreading, or translation efficiency (Duret and Mouchiroud, 1999; Duret, 2000; Tuller et al., 2007, 2010b).

We analyzed the correlation between tRNA gene copy number and codon usage in more than 500 genomes using previously reported approaches. We first determined the set of *highly adapted* codons (those recognized by tRNAs coded by the most abundant tRNA genes) and compared them to the set of *highly abundant codons* (those with high relative synonymous codons usage [RSCU], determined from gene sequences of ribosomal proteins). Our results confirm that the most abundant codons (highest RSCU) in general correspond to the most adapted codons (61% match) (for four- and six- codon families, the two most abundant codons are included in the analysis). However, as previously reported, this correlation is not perfect, and it is poor in eukaryotic genomes. Indeed, when considering the top two tRNA isoacceptors, archaeal species present the best match (75%), whereas Bacteria and Eukarya show matches of 59% and 41%, respectively (Table S3).

Strikingly, the codons whose frequencies do not correlate well with tRNA gene content values are precisely those codons corresponding to tRNAs susceptible to be modified either by adenosine deaminases or uridine methyltransferases (Figure 4A, see also Figure S4). It is worth noting that hetADATs and UMs exclusively modify those previously nonmatching codons (Figure S4). We reclassified those codons in the correlation analysis to account for the increased pairing ability of anticodons modified by UMs and hetADATs. This new analysis provided quasiperfect correlations between RSCU values and tRNA gene copy numbers in Bacteria and Eukarya (95% match) (Figure 4A). Therefore, tRNA gene copy number is almost perfectly correlated with codon usage in all kingdoms, provided that tRNA modifications caused by hetADATs and UMs are considered. This implies that, in all kingdoms of life, translational efficiency seems to be a primary factor influencing codon usage.

To experimentally confirm that association between codon usage and tRNA abundance is enhanced by the inclusion of modification enzymes, we determined $tRNA^{Arg}$ isoacceptor concentrations in HeLa and Hek 293T cell lines. We chose $tRNA^{Arg}$ for this analysis because all five human arginine isoacceptors can be individually quantified thanks to isoacceptor-specific probes. We performed an association analysis for $tRNA^{Arg}$ expression and codon usage in the absence or presence of modification information. Only after the inclusion of hetADAT modification information in the calculations could a good correlation be found between tRNA abundance and codon usage (Pearson correlation: 0.86 and 0.81 for HeLa and 293T, respectively) (Figure 4B).

To further confirm these results we also analyzed published data on gene expression levels in other species. In a recent study, Kudla et al. synthesized a library of 154 genes coding for green fluorescent protein (GFP) that varied randomly at synonymous sites (Kudla et al., 2009). These genes were expressed in *Escherichia coli*, and GFP expression levels were obtained that varied 250-fold across the library. The initial analysis of this data failed to find a correlation between codon composition and gene expression (however, see Supek and Smuc, 2010; Navon and Pilpel, 2011). We wondered whether the inclusion of the activity of UMs in the model would improve the correlation between translation efficiency and codon composition. Thus, we tested whether codon composition correlated with protein production when the frequencies of UM- and hetADAT-modifiable anticodons (hereinafter named "preferred codons") and nonmodifiable anticodons (hereinafter named "nonpreferred codons") were taken into account. This was indeed the case, and we obtained quasiperfect correlations in the set of highly expressed GFP genes (94% match) (Figure S4).

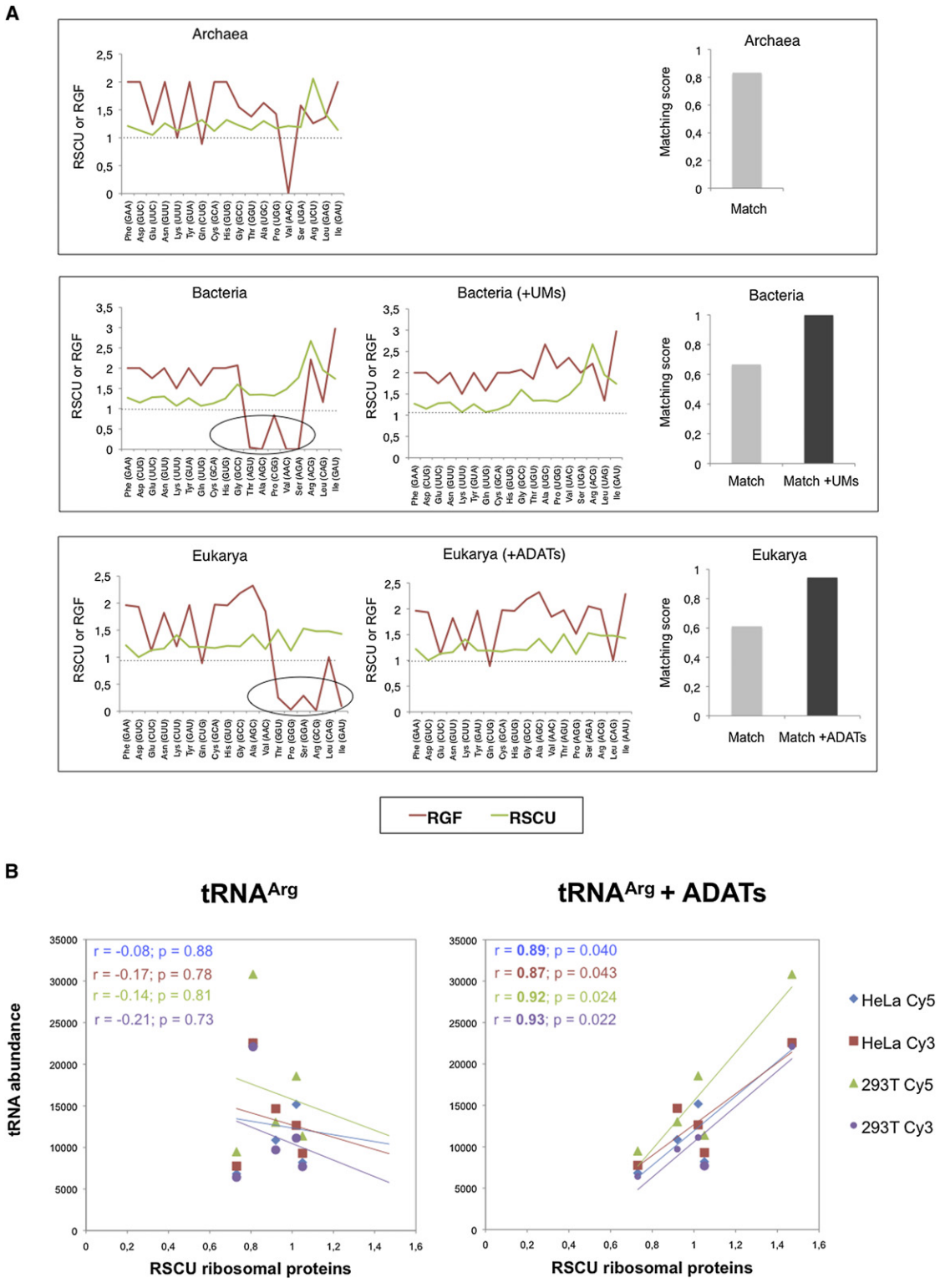


Figure 4. Match between Most Adapted Codons and Most Abundant Codons

(A) The match between the highest RSCU codon (green, most abundant codons) and the RGF value of its decoding tRNA (red, most adapted codons) is shown, for each kingdom, in the left column. The match after correcting the RGF values to account for the activity of UMs and ADATs is shown in the middle column. Archaea present neither ADATs nor UMs, and therefore the middle column is missing for this kingdom. The increase in the match score between RSCU and RGF after the correction is shown for each kingdom in the right histogram (except for Archaea).

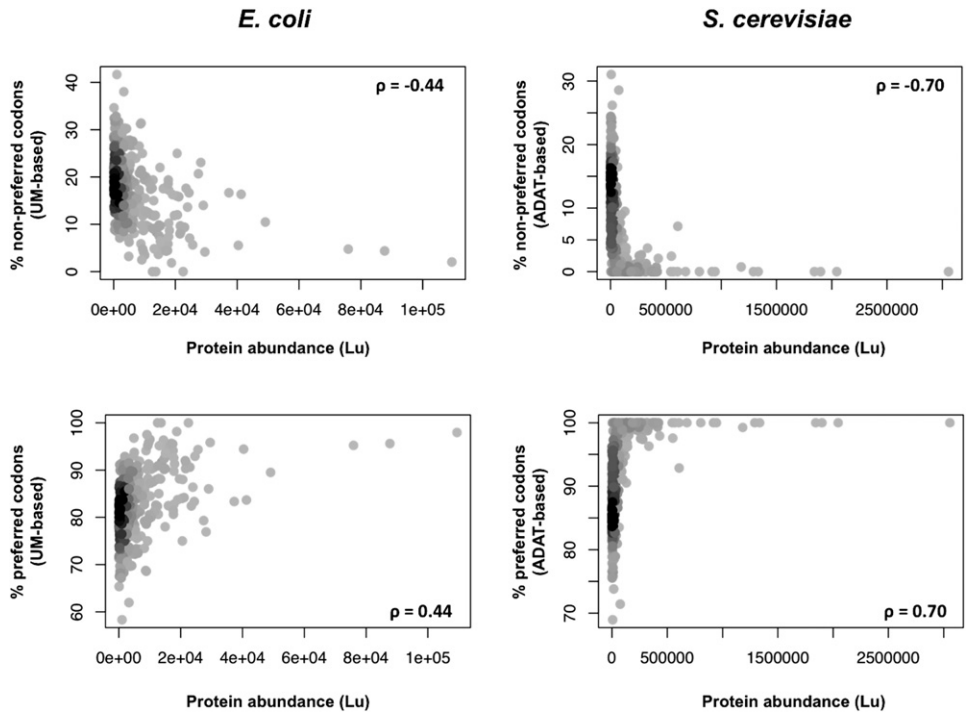


Figure 5. Correlation between Preferred Codons and Protein Abundance

In both *E. coli* and *S. cerevisiae*, the abundance of preferred codons in a gene correlates with protein abundance (Spearman correlation: 0.44 and 0.70, with p values of 9.7e-20 and 5.1e-52, respectively). Complementarily, the frequency of nonpreferred codons in genes decreases proportionally to protein abundance. The local density of data points in the graph is signified by their color (darker corresponding to more populated areas of the plot). See also Figure S5.

Analysis of the Influence of "Preferred Codons" in Protein Synthesis

Our results indicate that those transcripts whose codon composition is best adapted to anticodons modified by ADATs and UMs are the most efficiently translated. We therefore checked whether the relative abundance of preferred codons correlates with expression levels of any given gene. In this regard, genome-wide expression analyses (Lu et al., 2007; Ingolia et al., 2009; Ishihama et al., 2008; Ghaemmaghami et al., 2003; Taniguchi et al., 2010) provide experimental quantification of translational efficiency across a whole genome.

We examined the effect of UM and hetADAT modifications in published whole genome expression data obtained through the analysis of the *E. coli* and *Saccharomyces cerevisiae* transcriptomes. We found a good correlation between relative abundance of "preferred codons" of any given gene and its protein abundance in *E. coli* and *S. cerevisiae* ($\rho = -0.44$ and -0.70, respectively) (Figure 5, see also Figure S5). Different genome-wide expression data sets (Lu et al., 2007; Ishihama et al., 2008; Newman et al., 2006) produced similar correlations for both species ($\rho = -0.27$ and -0.74, respectively) (Figure S5). Moreover, an inverse correlation between protein abundance and nonpreferred codons was also detected, suggesting the existence of an upper maximum limit of nonpreferred codons per gene. Thus, the abundance of preferred codons possibly represents an additional level of translation control that needs to be considered in addition other mechanisms of posttranscriptional regulation (Mata et al., 2005).

DISCUSSION

Despite the central role of tRNAs in protein translation, the connections between tRNA gene population dynamics and genome evolution have rarely been explored. It is known that in unicellular organisms the most abundant codons are recognized by the most abundant tRNAs in the cell (Withers et al., 2006; Tuller et al., 2010a). However, we do not understand the reasons for the variability between tRNA pools of different species, nor the principles that determine tRNA gene abundances or genomic codon composition.

Our tRNA gene content analysis shows that genomic tRNA gene composition is an evolutionary trait that separates the main kingdoms of life. This separation is mainly due to the selection of tRNA genes containing anticodons modifiable by

(B) Correlation between human $tRNA^{Arg}$ isoacceptor abundance determined using tRNA microarrays and codon usage of ribosomal proteins (shown as RSCU), both for HeLa and HEK293T cell lines. The lack of correlation between these two parameters in the left plot is corrected in the right plot by the inclusion of the activity of ADATs.

See also Figure S4 and Tables S3–S6.

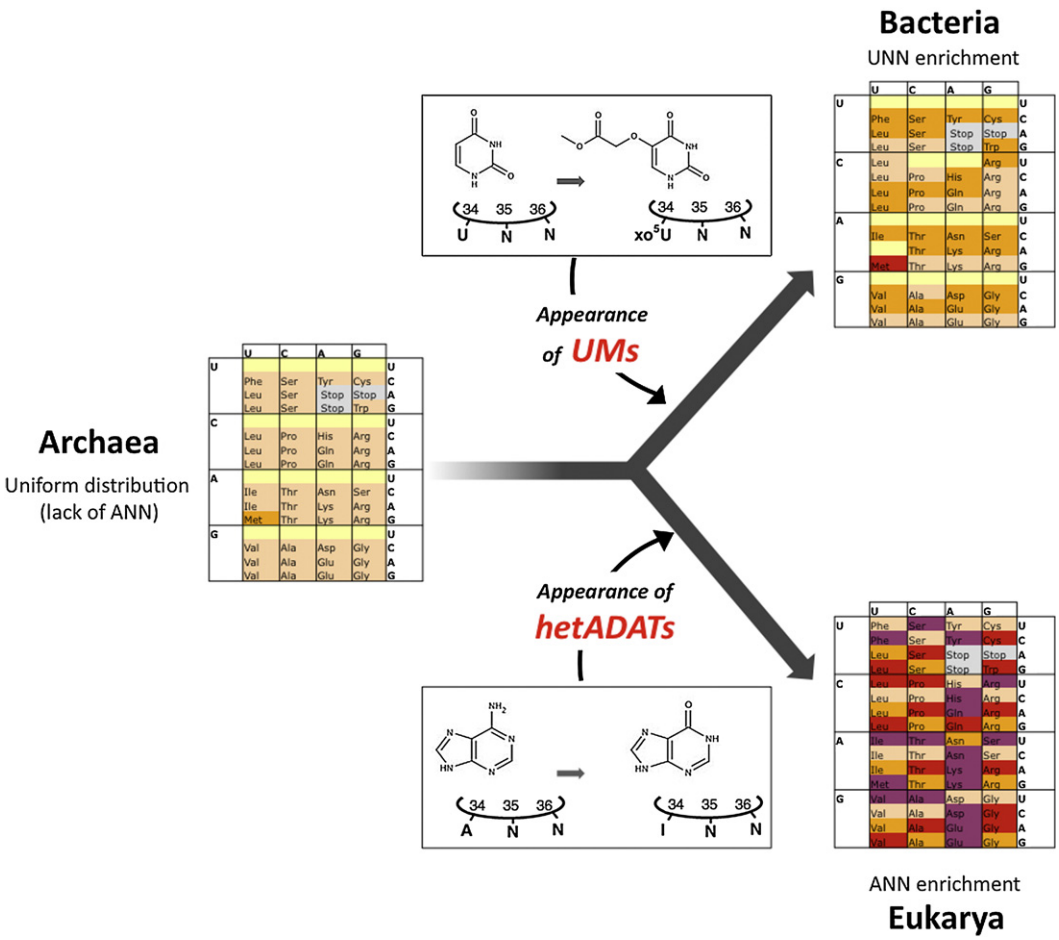


Figure 6. Model for the Role of Modification Enzymes in the Evolution of Genome Compositions

The emergence of the two tRNA modification enzymes (heterodimeric ADATs and UMs) was the main factor causing the divergence of decoding strategies between kingdoms. Archaea represents the most ancestral decoding strategy, where all isoacceptors are equally represented (and ANN anticodons are missing). ANN anticodons became overrepresented in eukaryotes due to the emergence of heterodimeric ADATs. Similarly, UNN anticodons became overrepresented in bacteria due to the appearance of UMs. Modification of the wobble position increased the decoding capacity of tRNAs, and consequently, translation efficiency. Thus, modifiable tRNAs were positively selected, causing a bias in tRNA gene content distribution which, in turn, caused the codon usage bias characteristic of the three main kingdoms.

kingdom-specific enzymes. This selection is likely driven by the improved decoding capacity that these modifications instill upon the modified tRNAs. A different solution to maximize tRNA decoding capacity was applied by Bacteria and Eukarya, thus contributing to the extant differences in tRNA pools and genome compositions.

Archaea would be the most ancestral kingdom in terms of decoding complexity (Figure 6). In Archaea neither ANN anticodons (Marck and Grosjean, 2002) nor ADATs are found (Mian et al., 1998). Therefore, the emergence of ADATs might be responsible for the appearance and selection of ANN-containing tRNAs that increased translation efficiency. In a similar fashion, the emergence of bacterial UMs would have driven the enrichment of tRNA genes with UNN anticodons in these organisms.

Several groups have demonstrated that preferred codon frequencies in highly expressed genes correlate with tRNA abundances within the cell (Withers et al., 2006; Tuller et al., 2010a). However, whether codon usage bias is caused by mutational bias or by natural selection has been a matter of controversy (Yang and Nielsen, 2008; Duret, 2002). In fast-growing organisms such as *E. coli* or *S. cerevisiae*, codon usage is generally thought to be under selective pressure (Sharp et al., 2005, 2010; Dong et al., 1996). On the other hand, in slowly growing organisms such as vertebrates, the existence of this selective pressure is controversial.

We have shown that the inclusion of modification data caused by ADATs and UMs in the definition of tRNA populations improves the codon usage-tRNA gene content correlation in Bacteria and Eukarya. Likely, the emergence of UMs and hetADATs in Bacteria and Eukarya allowed for the selection of new tRNAs that improved translation efficiency, and thus contributed to the evolution of genomic codon composition and tRNA gene content differences. Using published experimental data, we have shown that codons recognized by UM- and

hetADAT-modifiable anticodons are significantly enriched in highly expressed genes. Conversely, lowly expressed genes are enriched in codons recognized by nonmodifiable anticodons.

We have also shown that $tRNA^{Arg}$ populations in human cells do correlate well with genomic codon composition provided that anticodon modifications caused by hetADATs are considered in the definition of the different $tRNA^{Arg}$ isoacceptor concentrations. Thus, as previous studies have proposed for limited sets of species (Supek et al., 2010; Hershberg and Petrov, 2009; Drummond and Wilke, 2008), we conclude that translation efficiency influences tRNA gene populations in all kingdoms of life.

Several studies claim that the most significant parameter explaining codon bias differences among organisms is the level of GC content (Chen et al., 2004; Knight et al., 2001). Nevertheless, this observation does not explain codon bias variations within genomes, nor its correlation with gene expression levels. Anticodon modification strategies designed to improve translational efficiency could have evolved in parallel to the establishment of species-specific GC contents to ensure that tRNA gene populations were adapted to optimize translation. It should be noted that the triplet decoding strategies used by individual organisms have been determined (Marck and Grosjean, 2002; Grosjean et al., 2010). Each decoding strategy defines the minimum set of tRNAs needed to read all codons, and ranges from 25 up to 46 tRNAs. Interestingly, the defined minimal sets of eukaryotic and bacterial tRNAs conserve tRNA(ANN) and tRNA(UNN) isoacceptors respectively.

To summarize, Bacteria and Eukarya used two different tRNA modifications to increase the translational efficiency of their respective genomes. This phenomenon, in turn, contributed to the extant differences in tRNA gene populations and codon compositions of the main kingdoms of life. The discovery of kingdom-specific strategies to optimize translation efficiency opens new possibilities to further improve heterologous gene expression systems. Indeed, heterologous protein expression may be further improved if gene compositions are designed to match the mature tRNA gene population of the host species. In this regard, recent studies have started to analyze the potential of codon selection to tune translation efficiency (Cannarozzi et al., 2010; Tuller et al., 2010b) or protein folding (Zhang et al., 2009).

EXPERIMENTAL PROCEDURES

tRNA Sequence Retrieval

We have extracted, analyzed and compared over 53,000 sequences corresponding to cytoplasmatic nonorganellar tRNAs from 527 genomes distributed throughout the three kingdoms of life. All tRNA sequences have been downloaded from the GtRNAdb (http://gtrnadb.ucsc.edu), which uses the predictions made by the program tRNAscan-SE (Lowe and Eddy, 1997). Given that our analysis is based on average tRNA abundances, minor misannotations that may happen in tRNA genes using this prediction program are not statistically significant and thus should not affect the final results of this work.

Gene Content Analysis

Using the complete set of tRNA sequences we have built a distance-based phylogeny constructed on the basis of gene content. The similarity between two species is determined by the number resulting from dividing the number of tRNA genes that they have in common by the total number of gene types (isoacceptors). Using this method we have calculated a distance matrix that contains all pairwise distance values between the species analyzed. The distance matrix obtained has been used to cluster the sequences and build the phylogenetic tree, using the neighbor-joining method implemented in the program PHYLIP (Felsenstein, 1989). The program iTOL (Letunic and Bork, 2007) has been used for the visualization of the resulting phylogenetic tree.

Principal Component Analysis

A matrix consisting of the tRNA relative gene frequencies (RGF) for each anticodon and for all the analyzed species was used as input to perform PCA analysis (Jolliffe, 2002) using the program R (Team RDC, 2008, R: A Language and Environment for Statistical Computing, Vienna Austria R Foundation for Statistical Computing). The same software was used to obtain the resulting plots and to perform the t test and Wilcoxon test on the results. The significance of the association of the loadings with each principal component was computed using the FactoMineR package for R (Lê et al., 2008).

Retrieval of Coding Sequences and Codon Usage Estimation

All complete protein-coding sequences (CDS) for each of the selected 107 species were downloaded from the EMBLCDS database (http://www.ebi.ac.uk/embl/cds). For each species, a subset corresponding to ribosomal proteins was selected and visually inspected, and finally used as input to estimate the codon usage of highly expressed proteins using the GCUA software (McInerney, 1998).

Correlation between Codon Usage and tRNA Gene Content

For each species analyzed, the set of 18 preferred codons and preferred tRNA isoacceptors was computed (one for each amino acid, excluding Met and Trp). Initial correlations were computed by using the Watson-Crick base pairing rules (U:A; A:U; C:G; G:C), and extended correlations were computed including the extended wobble base pairing that result from the activities of ADATs (I:A; I:C; I:U) and UMs (xo^5U:A; xo^5U:G; xo^5U:U).

Correlation coefficients were computed as: $C = (\Sigma M / N) * 100$, where M is the number of codon-anticodon pairs for which there is a match (using Watson-Crick or extended wobble base pairing rules), and N is the number of codon-anticodon pairs considered in the analysis. We considered three different sets of matching codon-anticodon pairs. The simplest set (N = 8) includes the major tRNA isoacceptors with modifiable anticodons. A second set (N = 18) includes all major tRNA isoacceptors with the exception of methionine and tryptophan. Finally, a larger set (N = 27) was built by also considering the second most abundant tRNA isoacceptor from all four-, six-, and split (Ile) codon families.

The inclusion of modification data in our correlation analysis increases the number of acceptable codon-anticodon pairs, which could artificially increase correlation coefficients. To discard the possibility that the correlations that we obtain are simply the result of the increased number of acceptable pairs we tested the statistical significance of our data in both scenarios, i.e., with and without the inclusion of modification data. To that end, we approximated our data to a binomial distribution, computing for each set of data the expected distribution of random matches (Table S4). Our results show that the significance of our data is not due to the increased number of acceptable pairs caused by the inclusion of modification data (Tables S5 and S6). Using the same approach we confirmed that the statistical significance of our results is independent of the subset of tRNA isoacceptors analyzed.

tRNA Microarrays

tRNA abundance from HeLa and HEK293T cells was measured using a tRNA specific microarray method described previously (Dittmar et al., 2006; Pavon-Eternod et al., 2010). The standard tRNA microarray experiment consists of four steps starting from total RNA: (1) deacylation to remove remaining amino acids attached to the tRNA, (2) selective Cy3/Cy5 labeling of tRNA, (3) array hybridization, and (4) data analysis. The relative Cy3 or Cy5 fluorescent values from each tRNA probe of the same sample are used to determine the relative abundance of each tRNA in this sample, as described previously (Pavon-Eternod et al., 2010; Tuller et al., 2010a).

Protein Abundance and mRNA Levels

Protein abundance values and mRNA measurements of *E. coli* were taken from the work of Lu et al. (2007) and Ishihama et al. (2008); protein abundance values and mRNA levels of *S. cerevisiae* were taken from the work of Lu et al. (2007) and Newman et al. (2006). Correlation between protein expression levels and the abundance of preferred codons is shown in Figure 5 and Figures S4 and S5, and has been quantified using the Spearman's rank correlation coefficient.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and six tables and can be found with this article online at doi:10.1016/j.cell.2012.01.050.

ACKNOWLEDGMENTS

We thank Dr. M. Santos and Dr. V. de Crécy-Lagard for their critical analysis of the manuscript. We also thank E. Planet and D. Rossell for their help with the statistical analysis of the data. This work has been supported by grant BIO2009-09776 from the Spanish Ministry of Education and Science, and by grant MEPHITIS-223024 from the European Union. E.M.N. is supported by a La Caixa/IRB International Ph.D. Programme Fellowship. M.P.-E. was supported by a Ruth Kirshstein Pre-doctoral Fellowship from the NIH (1F31CA139968).

Received: September 20, 2011
Revised: November 23, 2011
Accepted: January 12, 2012
Published: March 29, 2012

REFERENCES

Agris, P.F., Vendeix, F.A., and Graham, W.D. (2007). tRNA's wobble decoding of the genome: 40 years of modification. J. Mol. Biol. *366*, 1–13.

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

Akashi, H. (2003). Translational selection and yeast proteome evolution. Genetics *164*, 1291–1303.

Bennetzen, J.L., and Hall, B.D. (1982). Codon selection in yeast. J. Biol. Chem. *257*, 3026–3031.

Brochier, C., and Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature *417*, 244.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., and Barral, Y. (2010). A role for codon order in translation dynamics. Cell *141*, 355–367.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., and McAdams, H.H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. USA *101*, 3480–3485.

Dittmar, K.A., Sørensen, M.A., Elf, J., Ehrenberg, M., and Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO Rep. *6*, 151–157.

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

Dong, H., Nilsson, L., and Kurland, C.G. (1996). Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J. Mol. Biol. *260*, 649–663.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. *32*, 5036–5044.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341–352.

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

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. *12*, 640–649.

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. USA *96*, 4482–4487.

Elf, J., Nilsson, D., Tenson, T., and Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. Science *300*, 1718–1722.

Felsenstein, J. (1989). PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics *5*, 164–166.

Fitz-Gibbon, S.T., and House, C.H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. *27*, 4218–4222.

Gerber, A.P., and Keller, W. (1999). An adenosine deaminase that generates inosine at the wobble position of tRNAs. Science *286*, 1146–1149.

Gerber, A.P., and Keller, W. (2001). RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem. Sci. *26*, 376–384.

Geslain, R., and Pan, T. (2010). Functional analysis of human tRNA isodecoders. J. Mol. Biol. *396*, 821–831.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Gonos, E.S., and Goddard, J.P. (1990). Human tRNAGlu genes: their copy number and organisation. FEBS Lett. *276*, 138–142.

Grosjean, H., de Crecy-Lagard, V., and Marck, C. (2010). Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. FEBS Lett. *584*, 252–264.

Hershberg, R., and Petrov, D.A. (2008). Selection on codon bias. Annu. Rev. Genet. *42*, 287–299.

Hershberg, R., and Petrov, D.A. (2009). General rules for optimal codon choice. PLoS Genet. *5*, e1000556.

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. *146*, 1–21.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M.J., and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics *9*, 102.

Iwasaki, W., and Takagi, T. (2007). Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics *23*, i230–i239.

Jolliffe, I.T. (2002). Principal Component Analysis (New York: Springer Series in Statistics).

Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T. (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*, 143–155.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., and Ikemura, T. (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*, 290–298.

Knight, R.D., Freeland, S.J., and Landweber, L.F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. *2*, RESEARCH0010.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. Science *324*, 255–258.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. J. Stat. Softw. *25*, 1–18.

Ledoux, S., Olejniczak, M., and Uhlenbeck, O.C. (2009). A sequence element that tunes Escherichia coli tRNA(Ala)(GGC) to ensure accurate decoding. Nat. Struct. Mol. Biol. *16*, 359–364.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. *25*, 117–124.

Maas, S., and Rich, A. (2000). Changing genetic information through RNA editing. Bioessays *22*, 790–802.

Man, O., and Pilpel, Y. (2007). Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat. Genet. *39*, 415–421.

Marck, C., and Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. RNA *8*, 1189–1232.

Mata, J., Marguerat, S., and Bähler, J. (2005). Post-transcriptional control of gene expression: a genome-wide perspective. Trends Biochem. Sci. *30*, 506–514.

McInerney, J.O. (1998). GCUA: general codon usage analysis. Bioinformatics *14*, 372–373.

Mian, I.S., Moser, M.J., Holley, W.R., and Chatterjee, A. (1998). Statistical modelling and phylogenetic analysis of a deaminase domain. J. Comput. Biol. *5*, 57–72.

Moriyama, E.N., and Powell, J.R. (1998). Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. *26*, 3188–3193.

Navon, S., and Pilpel, Y. (2011). The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. Genome Biol. *12*, R12.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840–846.

Näsvall, S.J., Chen, P., and Bjork, G.R. (2004). The modified wobble nucleoside uridine-5-oxyacetic acid in tRNAPro(cmo5UGG) promotes reading of all four proline codons in vivo. RNA *10*, 1662–1673.

Pavon-Eternod, M., Wei, M., Pan, T., and Kleiman, L. (2010). Profiling non-lysyl tRNAs in HIV-1. RNA *16*, 267–273.

Ran, W., and Higgs, P.G. (2010). The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol. Biol. Evol. *27*, 2129–2140.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., and Wright, F. (1988). Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res. *16*, 8207–8211.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., and Sockett, R.E. (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. *33*, 1141–1153.

Sharp, P.M., Emery, L.R., and Zeng, K. (2010). Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 1203–1212.

Snel, B., Bork, P., and Huynen, M.A. (1999). Genome phylogeny based on gene content. Nat. Genet. *21*, 108–110.

Stetter, K.O. (1996). Hyperthermophiles in the history of life. Ciba Found. Symp. *202*, 1–10, discussion 11–18.

Supek, F., and Smuc, T. (2010). On relevance of codon usage to expression of synthetic and natural genes in Escherichia coli. Genetics *185*, 1129–1134.

Supek, F., Skunca, N., Repar, J., Vlahovicek, K., and Smuc, T. (2010). Translational selection is ubiquitous in prokaryotes. PLoS Genet. *6*, e1001004.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533–538.

Tekaia, F., Lazcano, A., and Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. Genome Res. *9*, 550–557.

Tuller, T., Kupiec, M., and Ruppin, E. (2007). Determinants of protein abundance and translation efficiency in S. cerevisiae. PLoS Comput. Biol. *3*, e248.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010a). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344–354.

Tuller, T., Waldman, Y.Y., Kupiec, M., and Ruppin, E. (2010b). Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. USA *107*, 3645–3650.

Withers, M., Wernisch, L., and dos Reis, M. (2006). Archaeology and evolution of transfer RNA genes in the Escherichia coli genome. RNA *12*, 933–942.

Wolf, J., Gerber, A.P., and Keller, W. (2002). tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. EMBO J. *21*, 3841–3851.

Yang, S., Doolittle, R.F., and Bourne, P.E. (2005). Phylogeny determined by protein domain content. Proc. Natl. Acad. Sci. USA *102*, 373–378.

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

Yokoyama, S., Watanabe, T., Murao, K., Ishikura, H., Yamaizumi, Z., Nishimura, S., and Miyazawa, T. (1985). Molecular mechanism of codon recognition by tRNA species with modified uridine in the first position of the anticodon. Proc. Natl. Acad. Sci. USA *82*, 4905–4909.

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. *16*, 274–280.