EVOLUTION OF DIFFERENTIAL CODON USAGE PREFERENCES AND SUBFUNCTIONALISATION IN PARALOGOUS GENES: THE SHOWCASE OF POLYPYRIMIDINE TRACT BINDING PROTEINS

Jérôme Bourret[1, †, *], Fanni Borvető[1, †], and Ignacio G. Bravo[1]

[1]Centre National de la Recherche Scientifique (CNRS), Laboratoire MIVEGEC (CRNS IRD UM), Montpellier, France

[†]These authors contributed equally to this work

ABSTRACT

Gene paralogs are copies of a same gene that appear after gene or full genome duplication. Redundancy generated by gene duplication may release certain evolutionary pressures, allowing one of the copies to access novel gene functions. Here we focused on role of codon usage preferences (CUPrefs) during the evolution of the polypyrimidine tract binding protein (*PTBP*) splicing regulator paralogs.

PTBP1-3 show high identity at the amino acid level (up to 80%), but display different nucleotide composition, divergent CUPrefs and distinct tissue-specific expression levels. Phylogenetic inference differentiates the three orthologs and suggests that the three *PTBP1-3* lineages predate the basal diversification within vertebrates. We identify a distinct substitution pattern towards GC3-enriching mutations in *PTBP1*, with a trend for the use of common codons and for a tissue-wide expression. Genomic context analysis shows that GC3-rich nucleotide composition for *PTBP1*s is driven by local mutational processes. In contrast, *PTBP2*s are enriched in AT-ending, rare codons, and display tissue-restricted expression. Nucleotide composition and CUPrefs of *PTBP2* are only partly driven by local mutational forces, and could have been shaped by selective forces. Interestingly, trends for use of UUG-Leu codon match those of AT-ending codons.

Our interpretation is that a combination of mutation and selection has differentially shaped CUPrefs of *PTBPs* in Vertebrates: GC-enrichment of *PTBP1* is linked to the strong and broad tissue-expression, while AT-enrichment of *PTBP2* and *PTBP3* is linked to rare CUPrefs and specialized spatio-temporal expression. Our model is compatible with a gene subfunctionalisation process by differential expression regulation associated to the evolution of specific CUPrefs.

Keywords Codon usage bias, codon usage preferences, gene duplication, paralog, ortholog, evolution, mutation-selection, nucleotide composition, tissue-specific expression

[*]Corresponding author. email : jerome.bourret@ird.fr

1 Introduction

During translation, ribosomes assemble proteins by specific amino acid linear polymerisation guided by the successive reading of mRNA nucleotide triplets called codons. Each time a codon is read, it is chemically compared to the set of available tRNAs' anticodons. Upon codon-anticodon sequence match the ribosome loads the tRNA and adds the associated amino acid to the nascent protein. The main 20 amino acids are decoded by 61 codon-anticodon combinations, so that multiple codons are associated to the same amino acid and are named synonymous codons (Nirenberg and Matthaei, 1961; Khorana et al., 1966). Codon Usage Preferences (CUPrefs) refer to the differential usage of synonymous codons, between species, or between genes and genomic regions in the same genome (Grantham et al., 1980; Carbone et al., 2003). Mutation and selection are the two main forces shaping CUPrefs (Duret, 2002; Chamary et al., 2006; Plotkin and Kudla, 2011). Mutational biases relate to directional mechanistic biases during genome replication (Reijns et al., 2015; Apostolou-Karampelis et al., 2016), during genome repair (Lujan et al., 2012), or during recombination (Pouyet et al., 2017), preferentially introducing one nucleotide over others or inducing recombination and maintaining genomic regions depending on their composition. Mutational biases are well known in prokaryotes and eukaryotes, ranging from simple molecular preferences towards 3'A-ending in the *Taq* polymerase (Clark, 1988) to the complex GC-biased gene conversion in vertebrates (Pouyet et al., 2017). Selective forces shaping CUPrefs are often described as translational selection. This notion refers to the ensemble of mechanistic steps and interactions during translation that are affected by the particular CUPrefs of the mRNA, so that the choice of certain codons at certain positions may actually enhance the translation process and can be subject to selection (Bulmer, 1991). Translational selection covers thus codon-mediated effects acting on mRNA maturation, secondary structure and overall stability (Presnyak et al., 2015; Novoa and Ribas de Pouplana, 2012), subcellular localisation, programmed frameshifts, translation speed and accuracy, or protein folding (Caliskan et al., 2015; Mordstein et al., 2020; Spencer and Barral, 2012). Translational selection has been demonstrated in unicellular prokaryotes and eukaryotes (Satapathy et al., 2016; Percudani et al., 1997; Duret and Mouchiroud, 1999; Whittle and Extavour, 2016), often in the context of tRNA availability (Ikemura, 1981). However, its very existence in multi-cellular eukaryotes remains highly debated (Pouyet et al., 2017; Galtier et al., 2018).

Homologous genes share a common origin either by speciation (orthology) or by duplication events (paralogy) (Sonnhammer and Koonin, 2002). Upon gene (or full genome) duplication, the new genome will contain two copies of the original gene, referred to as in-paralogs. After speciation, each daughter cell will inherit one couple of paralogs, *i.e.* one copy of each ortholog (Koonin, 2005). The emergence of paralogs by gene duplication releases the evolutionary constraints on the individual genes. Evolution can thus potentially lead to function specialisation, such as evolving a particular substrate preferences, or engaging each paralog on specific enzyme activity preferences in the case of promiscuous enzymes (Copley, 2020). Gene duplication can also allow one paralog to explore broader sequence space and to evolve radically novel functions, while the remaining counterpart can assure the original function.

The starting point for our research are the experimental observations by Robinson and coworkers reporting differential expression of the polypyrimidine tract binding protein (*PTBP*) human paralogs as a function of their nucleotide com-

position (Robinson et al., 2008). Vertebrates genomes encode for three in-paralogous versions of the *PTBP* genes, all of them fulfilling similar functions in the cell: they form a class of hnRNP RNA-Binding Proteins that are involved in the modulation of mRNAs alternative splicing (Pina et al., 2018). Within the same genome the three paralogs display high amino-acid sequence similarity, around 70% in humans and with similar overall values in vertebrates (Pina et al., 2018).

Despite the high resemblance at the protein level, the three *PTBP* paralogs sharply differ in nucleotide composition, CUPrefs and tissue expression pattern. In humans, *PTBP1* is enriched in GC-ending synonymous codons and is widely expressed in all tissues, while *PTBP2* and *PTBP3* are AT3-rich and display an enhanced expression in the brain and in hematopoietic cells respectively (Supplementary Material S1). Robinson and coworkers studied the expression in human cells of all three human *PTBP* paralogous genes placed under the control of the same promoter. They showed that the GC-rich paralog *PTBP1* was more highly expressed than the AT-rich ones, and that the expression of the AT-rich paralog *PTBP2* could be enhanced by synonymous codons recoding towards the use of GC-rich codons (Robinson et al., 2008). Here we have built on the evolutionary foundations of this observation and extended the analyses of CUPrefs to *PTBP* paralogs to vertebrate genomes. Our results suggest that paralog-specific directional changes in CUPrefs in mammalian *PTBP* concurred with a process of subfunctionalisation by differential tissue pattern expression of the three paralogous genes.

2 Material and Methods

Sequence retrieval

We assembled a dataset of DNA sequences from 47 mammalians and 27 non-mammalians Vertebrates and 3 protostomes using the BLAST function on the nucleotide database of NCBI (NCBI Resource Coordinators, 2018) using the human *PTBP* paralogs as references (see supplementary Material S2 for accession numbers). We could identify the corresponding three ortholog genes in all Vertebrates species screened except for the European rabbit *Oryctolagus cuniculus*, lacking *PTBP1* and from the rifleman bird *Acanthisitta chloris*, lacking *PTBP3* (Supplementary Material S2). The final vertebrate dataset contained 75 *PTBP1*, 76 *PTBP2* and 75 *PTBP3* sequences. As outgroups for the analysis, we retrieved the orthologous genes in three protostomes genomes, which contained a single *PTBP* homolog per genome (Supplementary Material S3). From the original dataset, we identified a subset of nine mammalian and six non-mammalian vertebrates species with a good annotation of the *PTBP* chromosome context, and we retrieved compositional information on the flanking regions and on the intron composition (Supplementary Material S3). Because of annotation hazards, intronic and flanking regions information were missing for some *PTBP*s in the African elephant *Loxodonta africana*, Schlegel's Japanese Gecko *Gekko japonicus* and the whale shark *Rhincodon typus* assemblies. For these 15 species the values for codon adaptation index (CAI) (Sharp and Li, 1987) and codon usage similarity index (COUSIN) (Bourret et al., 2019) were calculated using the COUSIN server (available at https://cousin.ird.fr).

Clustering PTBPs by their CUPrefs

For each *PTBP* paralog we calculated codon composition and CUPrefs analyses via the COUSIN tool (Bourret et al., 2019). For each *PTBP* gene we constructed a vector of 59 positions with the relative frequencies of all synonymous codons. As tools for information dimension reduction to analysis CUPrefs we applied on the 229 59-dimension vectors: i) a k-means clustering; ii) a hierarchical clustering; and iii) a principal component analysis (PCA).

Alignment and phylogenetic analyses

To generate robust alignments without introducing artefacts due to large evolutionary distances between in-paralogs we proceeded stepwise, as follows: i) we aligned separately at the amino acid level each set of *PTBP* paralog sequences of mammals and non-mammalians Vertebrates; ii) for each *PTBP* paralog we merged the alignments for mammals and for non mammals, obtaining the three *PTBP1*, *PTBP2* and *PTBP3* alignments for all Vertebrates; iii) we combined the three alignments for each paralog into a single one; iv) we aligned the outgroup sequences to the global Vertebrate *PTBP*s alignment. All alignments steps were performed using MAFFT (Katoh et al., 2002). The final amino acid alignment was back-translated to obtain the codon-based nucleotide alignment. The codon-based alignment was trimmed using Gblocks (Castresana, 2000).

Phylogenetic inference was performed at the amino acid and at the nucleotide level using RAxML v8.2.9 and bootstrapping over 1000 cycles (Stamatakis, 2014). For nucleotides we used codon-based partitions and applied the GTR+G4 model while for amino acids we applied the LG+G4 model. For the 79 species used in the analyses we retrieved a species-tree from the TimeTree tool (Kumar et al., 2017). Distances between phylogenetic trees were computed using the Robinson-Foulds index, which accounts for differences in topology (Robinson and Foulds, 1981), and the K-tree score, which accounts for differences in topology and in branch length (Soria-Carrasco et al., 2007). After phylogenetic inference we computed marginal ancestral states for the respectively most recent common ancestors at the nucleotide level of each paralog using RAxML. Using these ancestral sequences we estimated the number of synonymous and non-synonymous mutations of each extant sequence to the corresponding most recent common ancestor.

Statistical analyses

Correlation between matrices was assessed via the Mantel test. Non-parametric comparisons were performed using the Wilcoxon-Mann-Whitney test for population medians and the Wilcoxon signed rank test for paired comparisons. Statistical analyses were performed using the *ape* and *ade4* R packages and JMP v14.3.0.

3 Results

Vertebrate PTBP paralogs differ in nucleotide composition

In order to understand the evolutionary history of *PTBP* genes we performed first a nucleotide composition and CUPrefs analysis on the three paralogs in 79 species. Overall, *PTBP1* are GC-richer than *PTBP2* and *PTBP3* (respective mean percentages 55.9, 42.3 and 44.9 for GC content and 69.5, 33.4 and 38.3 for GC3 content; Figure 1, Supplementary Material S2). In addition, *PTBP1* show a difference in GC3 between mammalian and non-mammalian gene (respectively 79.8 against 59.9 mean percentages). A linear regression model followed by a Tukey's honest significant differences analysis for GC3 using as explanatory levels paralog (*i.e. PTBP1-3*), taxonomy (*i.e.* mammalian or non-mammalian) and their interaction identifies three mains groups of *PTBP*s (Table 1): a first one corresponding to mammalian *PTBP1*, a second one grouping non-mammalian *PTBP1* and a third one spanning all *PTBP2* and *PTBP3*. The largest explanatory factor for GC3 was the paralog *PTBP1-3*, accounting alone for 65% of the variance, while the interaction between the levels taxonomy and paralog captured around 15% of the remaining variance (Table 1). These trends are confirmed when performing paired comparisons between paralogs present in the same mammalian genome, with significant differences in GC3 content in the following order: *PTBP1* > *PTBP3* > *PTBP2* (Wilcoxon

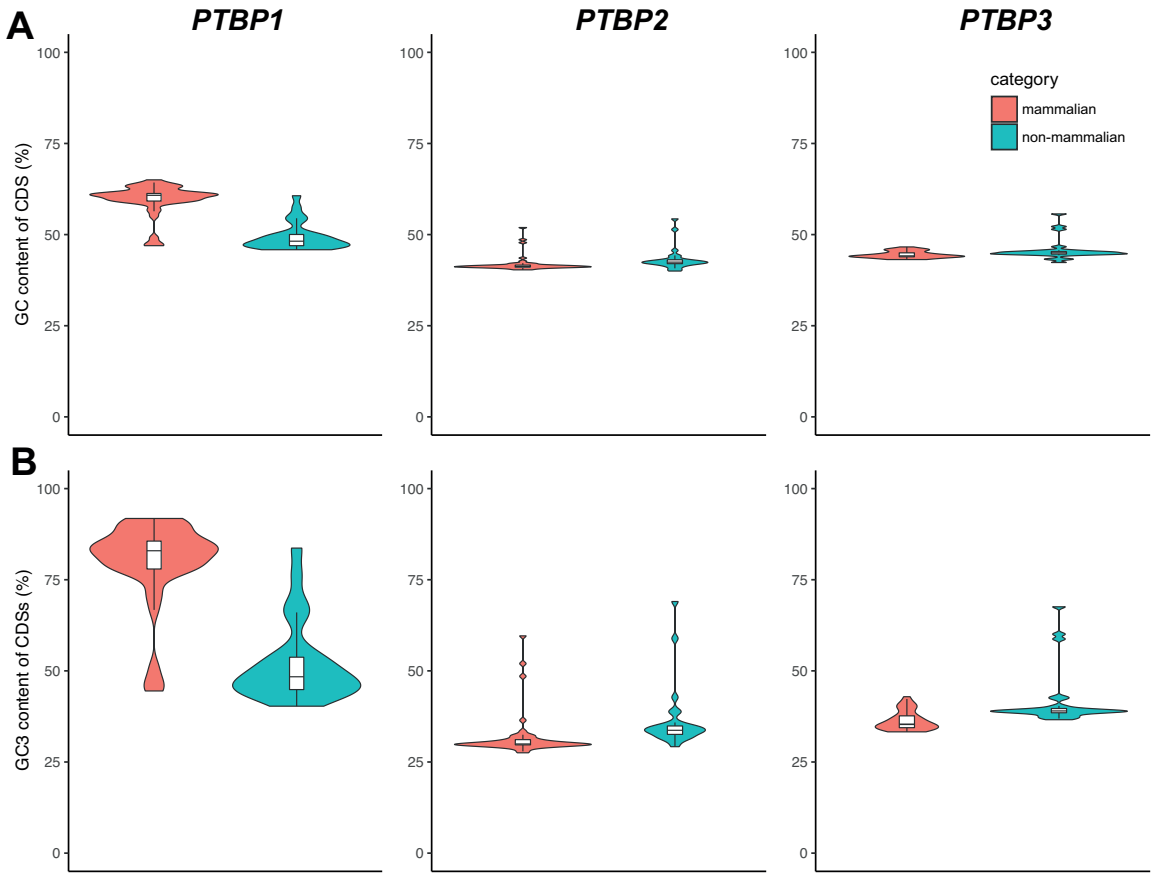

Figure 1: **GC content (A) and GC3 content (B) of Vertebrates *PTBP*s.** Violin plots display the overall distribution while box and whiskers display median, quartiles and 95% of the corresponding values for mammalian (red) and non-mammalian (blue) individual genomes.

signed rank test: *PTBP1* vs *PTBP2*, mean diff=48.0, S=539.50, p-value <0.0001; *PTBP1* vs *PTBP3*, mean diff=43.5, S=517.50, p-value <0.0001; *PTBP3* vs *PTBP2*, mean diff=4.5, S=406.50, p-value <0.0001). Note that even if all of them significantly different, the mean paired differences in GC3 between *PTBP1* and *PTBP2-3* are ten times larger than the corresponding mean paired differences between *PTBP2* and *PTBP3*.

The distribution of the residuals between observed and expected values after our model fit to the data allows to identify a number of outliers species with interesting taxonomical patterns in compositional deviation (Table 2). For non mammals, the three *PTBP* paralogs in the rainbow trout *Oncorhynchus mykiss* genome display high GC3 content (between 67% and 76%), all of them significantly higher than model-predicted values (expected values between 36% and 51%). A similar case occurs for the zebrafish *Danio rerio* genome: the three paralogs display GC3 values around 58%, which for *PTBP2* and *PTBP3* paralogs is significantly higher than predicted by the model (expected values around 38%). Very interestingly, for the monotrema platypus *Ornithorhynchus anatinus* as well as for the three marsupials in the dataset the Tasmanian devil *Sarcophilus harrisii*, the koala *Phascolarctos cinereus* and the grey short-tailed opossum *Monodelphis domestica* their *PTBP1* genes present similar GC3 content around 47%, which is significantly lower than predicted by the model (expected values around 79%).

In many vertebrate species, strong compositional heterogeneities are observed along chromosomes often referred to as "isochores". To explore the influence of the genomic environment on the nucleotide composition of *PTBP*s, for 15 species with well-annotated genomes we analyzed the correlation of paralog GC3 with two local compositional variables of the corresponding gene (GC content of intronic and flanking regions) and with three global compositional variables for the corresponding genomes (global GC3 in the complete genomic ORFome, global GC content in all introns, and global GC content in all flanking regions) (Table 3 and Figure 2). First, for *D. rerio* the GC3 composition of *PTBP2* and *PTBP3* is clearly different from the rest, in line with the outlier results presented in Table 2. We have thus excluded the zebra fish values and performed an individual as well as a stepwise linear fit to explain the variance in GC3 composition by the variance in the local and global compositional variables mentioned above (Table 3). For all three *PTBP*s the local GC content explains best the corresponding GC3 content, but with strong differences between paralogs: while variation in the local composition captures almost perfectly variation in the GC3 content in the case of *PTBP1* (R^2=0.97) and strongly in the case of for *PTBP3* (R^2=0.78), the fraction of variance explained by the local composition significantly drops in the case of *PTBP2* (R^2=0.46).

Vertebrate *PTBP* paralogs differ in CUPrefs

For each *PTBP* coding sequence we extracted the relative frequencies of synonymous codons and performed different approaches to reduce information dimension and visualise CUPrefs trends. The results of a principal component analysis (PCA) are shown in Figure 3. The first PCA axis captured 68.9% of the variance, far before the second and

Table 1: Global linear regression model and post-hoc Tukey's honest significant differences (HSD) test for GC3 composition as explained variable and the explanatory levels paralog (*PTBP1-3*), taxonomy (*i.e.* mammalian or non-mammalian) and their interactions. Overall goodness of the fit: Adj Rsquare=0.83; F ratio=205.7; Prob > F: <0.0001.Individual effects for the levels: i) paralog: F ratio=274.3; Prob > F: <0.0001; ii) taxonomy: F ratio=27.2; Prob > F: <0.0001; iii) interaction paralog*taxonomy: F ratio=87.9; Prob > F: <0.0001.

Level	**Least Sq. Mean (GC3%)**	**Standard error**	**Tukey's HSD group**
	Paralog		
PTBP1	65.87	1.00	A
PTBP3	39.00	1.01	B
PTBP2	34.03	1.00	C
	Taxonomy		
mammalian	49.32	0.70	A
non-mammalian	43.28	0.92	B
	Paralog*Taxonomy		
PTBP1, mammalian	79.81	1.22	A
PTBP1, non-mammalian	51.93	1.59	B
PTBP3, non-mammalian	41.64	1.62	C
PTBP3, mammalian	36.36	1.22	C, D
PTBP2, non-mammalian	36.27	1.59	C, D
PTBP2, mammalian	31.79	1.20	D

the third axes (respectively 6.7% and 3.2%). In codon families with multiplicity two, the two codons are necessarily symmetrically related in the PCA, creating a redundancy. We thus simplified the analysis by performing again a PCA using only the codon families of multiplicity four and six, obtaining similar results (Supplementary Material S5 B). Codons segregate in the first axis by their GC3 composition, the only exception being the UUG-Leu codon, which grouped together with AT-ending codons. This first axis differentiates mammalian *PTBP1*s on the one hand and *PTBP2*s and *PTBP3*s on the other hand. Non-mammalian *PTBP1*s scatter between mammalian *PTBP1*s and *PTBP3*s, along with the protostoma *PTBP*s. In the second PCA axis the only obvious (but nevertheless cryptic) codon-structure trends are: i) the split between C-ending and G-ending codons, but not between A-ending and U-ending codons; and ii) the large contribution in opposite directions to this second axis of the AGA and AGG-Arginine codons. This second PCA axis differentiates *PTBP2*s from *PTBP3*s paralogs, consistent with these composition trends, a paired-comparison confirms that *PTBP3*s are richer in C-ending codons than *PTBP2*s, respectively 21.7% against 15.4% (Wilcoxon signed rank test: mean diff=6.2, S=1184.0, p-value <0.0001).

As an additional way to identify groups of genes with similar CUPrefs we applied a hierarchical clustering and a k-means clustering. Both analyses mainly aggregate *PTBP* genes by their GC3 richness. The *PTBP* dendrogram resulting of the hierarchical clustering (rows in clustering in Figure 3) shows five main clades that cluster the paralogs with a good match to the following groups: mammalian *PTBP1*s, non-mammalian *PTBP1*s, *PTBP2*s, *PTBP3*s and a fifth group containing the protostomata *PTBP*s and a few individuals of all three paralogs (Kappa-Fleiss consistency score = 0.76). Regarding codon clustering, the hierarchical stratification sharply splits GC-ending codons from AT-ending codons, with the only exception again of the UUG-Leu codon, which consistently groups within the AT-ending

Table 2: Individual genes with outlier values with respect to the linear regression expected values for the levels paralog (*PTBP1-3*), taxonomy (mammalian or non-mammalian) and their interactions.

Species	**paralog**	**observed GC3 (%)**	**expected GC3 (%)**	**deviation GC3 (%)**
		mammalian		
Desmodus rotundus	*PTBP2*	59.60	31.79	27.81
Miniopterus natalensis	*PTBP2*	48.52	31.79	16.72
Monodelphis domestica	*PTBP1*	44.49	79.81	-35.32
Ornithorhynchus anatinus	*PTBP1*	51.14	79.81	-28.67
Ornithorhynchus anatinus	*PTBP2*	52.00	31.79	20.21
Phascolarctos cinereus	*PTBP1*	47.53	79.81	-32.28
Sarcophilus harrisii	*PTBP1*	45.44	79.81	-34.37
		non-mammalian		
Danio rerio	*PTBP2*	58.89	36.27	22.62
Danio rerio	*PTBP3*	60.08	41.64	18.44
Lepisosteus oculatus	*PTBP3*	58.73	41.64	17.10
Oncorhynchus mykiss	*PTBP1*	76.27	51.93	24.34
Oncorhynchus mykiss	*PTBP2*	69.03	36.27	32.76
Oncorhynchus mykiss	*PTBP3*	67.58	41.64	25.95
Pogona vitticeps	*PTBP1*	83.68	51.93	31.75

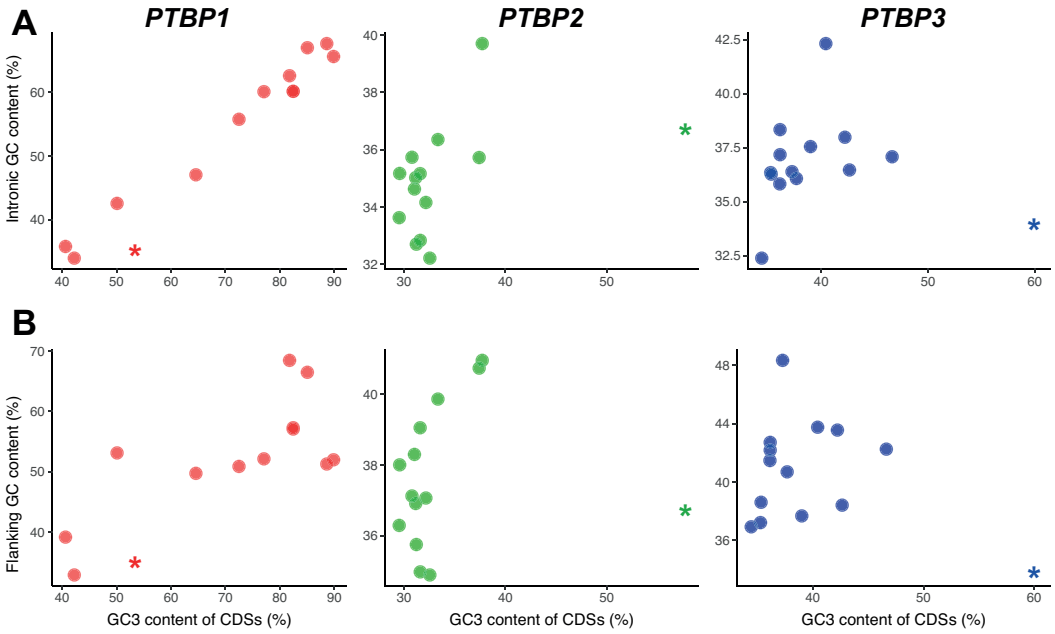

Figure 2: **Variation in GC3 content of *PTBP*s (x-axis) and in the GC content of the corresponding introns (A, y axis) or flanking regions (B, y axis)**. Each dot represents one of the 15 individual used for the genomic context analysis. The asterisk indicates the values for the species *Danio rerio*, which shows peculiar results for *PTBP2* and *PTBP3*, consistent with its outlier behaviour in the global model.

codons. The elbow approach of k-means clustering identifies an optimal number of four clusters and separates the paralog genes with a good match as following: *PTBP1*, *PTBP2*, *PTBP3* and a group containing the protostoma and individuals from all paralogs (Kappa-Fleiss consistency score = 0.75).

Overall, k-means clustering and hierarchical clustering, both based on the 59-dimensions vectors of the CUPrefs, are congruent with one another (Kappa-Fleiss consistency score = 0.83), and largely concordant with the PCA results. CUPrefs define thus groups of *PTBP* genes consistent with their orthology and taxonomy. It is interesting to note that for some species the *PTBP* paralogs display unique distributions of CUPrefs, such as an overall similar CUPrefs in the three *PTBP* genes of the whale shark *Rhincodon typus*, or again some shifts in nucleotide composition between paralogs in the Natal long-fingered bat *Miniopterus natalensis*.

In order to characterise the directional CUPrefs bias of the different paralogs, we have analysed for the 15 species with well-annotated genomes described above, the match between each individual *PTBP* and the average CUPrefs of the corresponding genome (Table 4). Our results highlight strong differences for mammalian paralogs: *PTBP1*s display COUSIN values above 1 while *PTBP2*s display COUSIN values below zero. Given the interpretation of COUSIN values (Bourret et al., 2019) these results mean that in mammals *PTBP1*s are enriched in commonly used codons in a higher proportion that the average in the genome, while *PTBP2*s are enriched in rare codons so that their CUPrefs go in the opposite direction to the average in the genome.

Phylogenetic reconstruction of *PTBP*s

We explored the evolutionary relationships between *PTBP*s by phylogenetic inference at the amino acid and at the nucleotide level (4, Supplementary Material S?). Our final dataset contained 74 *PTBP* sequences from mammals (47 species within 39 families) and non mammal vertebrates (27 species within 24 families). We used the *PTBP* genes from three protostome species as outgroups. Both amino acid and nucleotide phylogenies rendered three main clades grouping the *PTBP*s by orthology. In both topologies, *PTBP1* and *PTBP3* orthologs cluster together, although the protostome outgroups are linked to the tree by very a long branch making it difficult the proper identification of the Vertebrate *PTBP* tree root. Amino acid and nucleotide subtrees are largely congruent (see topology and branch length comparisons in Table5). The apparently large nodal and split distance values between nucleotide and amino acid *PTBP2* trees stem from disagreements in very short branches, as evidenced by the lowest K-tree score for this ortholog

Table 3: Results for an individual or for a sequential least squares regression for explaining variation in GC3 composition of *PTBP*s genes, by variation of different local or of global compositional variables in 14 well-annotated vertebrate genomes. For each gene, individual variables are ordered according to their contribution to the sequentially better model. Variables labelled with N.S. (not significant) do not contribute with significant additional explanatory power when added to the sequential model. BIC, Bayesian information content.

PTBP1				
	Individual contribution		**Sequential contribution**	
Parameter	$\mathbf{R^2}$	**BIC**	$\mathbf{R^2}$	**BIC**
Local intronic GC	0.96	74.42	0.96	74.42
Global intronic GC	0.03	111.98	0.97	71.23
Global flanking GC	0.05	111.70	0.98 (N.S.)	72.26
Global exomic GC3	0.62	100.71	0.98 (N.S.)	74.27
Local flanking GC	0.55	112.66	0.98 (N.S.)	76.55
PTBP2				
	Individual contribution		**Sequential contribution**	
Parameter	$\mathbf{R^2}$	**BIC**	$\mathbf{R^2}$	**BIC**
Local flanking GC	0.46	60.12	0.46	60.12
Global flanking GC	0.03	67.66	0.49 (N.S.)	61.86
Local intronic GC	0.37	61.95	0.49 (N.S.)	64.38
Global exomic GC3	0.09	66.75	0.49 (N.S.)	66.89
Global intronic GC	0.05	67.38	0.50 (N.S.)	69.35
PTBP3				
	Individual contribution		**Sequential contribution**	
Parameter	$\mathbf{R^2}$	**BIC**	$\mathbf{R^2}$	**BIC**
Local intronic GC	0.78	78.11	0.78	78.11
Global intronic GC	0.12	96.38	0.80 (N.S.)	79.56
Global exomic GC3	0.02	97.73	0.82 (N.S.)	80.66
Local flanking GC	0.38	91.77	0.84 (N.S.)	81.70
Global flanking GC	0.02	97.77	0.84 (N.S.)	84.27

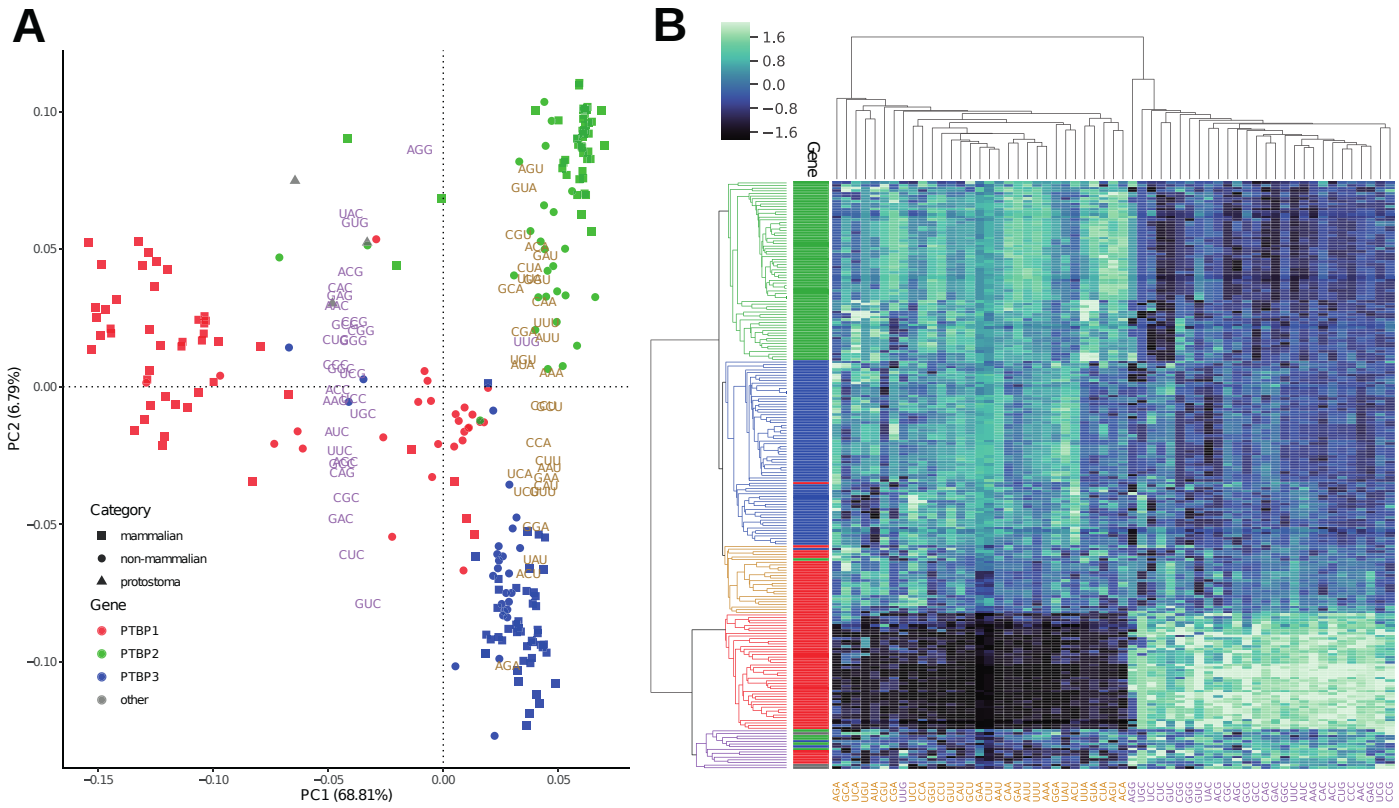

Figure 3: **CUPrefs analysis of *PTBP*s**. A) Plot of the two first dimensions of a PCA analysis based on the codon usage preferences of *PTBP1*s (red), *PTBP2*s (green), *PTBP3*s (blue) and protostoma (grey) individuals. Taxonomic information is included as mammals (squares), non-mammals (circles) and protostomates (triangles). The PCA was created using as variables the vectors of 59 positions (representing the relative frequencies of the 59 synonymous codons) for each individual gene. The eigenvalues of the individual codon variables are given by their position on the graph. Each codon variable is identified by its name and by a colour code, purple for GC-ending codons and orange for AT-ending codons. The percentage of the total variance explained by each axis is shown in parenthesis. B) Heatmap of *PTBP*s individuals (rows) and synonymous codons (columns). Left dendrogram represents the hierarchical clustering of *PTBP*s based on their CUPrefs with colour codes that stand for the clusters created from this analysis. Side bars give information on heatmap individuals regarding i) their origin : *PTBP1* (red), *PTBP2* (green), *PTBP3* (blue) or protostoma (grey). Note the position of the UUG-Leu codon, in both the PCA and the codon dendrogram, as the sole GC-ending codon clustering with all other AT-ending codons)

(as a reminder, the Robinson-Foulds index exclusively regards topology while the K-tree score combines topological and branch-length dependent distance between trees, see Material and Methods). In all three cases, internal structure of the ortholog trees essentially recapitulates species taxonomy at the higher levels (Table5). Some of the species identified by the mathematical model as displaying a largely divergent nucleotide composition present accordingly long branches in the phylogenetic reconstruction, such as *PTBP3* for *O. mykiss*.

We have then analysed the correspondence between nucleotide-based and amino acid-based pairwise distances. We observe a good correlation between both reconstructions for all paralogs, except for mammalian *PTBP2*s, which display

extremely low divergence at the amino acid level (Figure 5 B, Supplementary Material S8 B). For mammalian *PTBP1*s, the plot allows to clearly differentiate a cloud with the values corresponding to the monotremes+marsupial mammals,

Table 4: Global linear regression model and post-hoc Tukey's honest significant differences (HSD) test, the explained variable being the COUSIN value of the each *PTBP* gene against the average of the corresponding genome and the explanatory levels paralog (*PTBP1-3*), taxonomy (*i.e.* mammalian or non-mammalian) and their interactions. Overall goodness of the fit: Adj Rsquare=0.82; F ratio=36.84; Prob > F: <0.0001.Individual effects for the levels: i) paralog: F ratio=40.72; Prob > F: <0.0001; ii) taxonomy: F ratio=10.87; Prob > F: =0.0021; iii) interaction paralog*taxonomy: F ratio=28.11; Prob > F: <0.0001.

Level	**Least Sq. Mean (COUSIN)**	**Standard error**	**Tukey's HSD group**
	Paralog		
PTBP1	1.45	0.11	A
PTBP3	0.29	0.11	B
PTBP2	0.19	0.11	B
	Taxonomy		
mammalian	0.44	0.080	A
non-mammalian	0.85	0.098	B
	Paralog*Taxonomy		
PTBP1, mammalian	1.90	0.14	A
PTBP1, non-mammalian	0.99	0.17	B
PTBP2, non-mammalian	0.81	0.17	B
PTBP3, non-mammalian	0.75	0.17	B
PTBP3, mammalian	-0.16	0.14	C
PTBP2, mammalian	-0.43	0.14	C

Table 5: Comparison between species tree and subtrees of the nucleotide based maximum likelihood tree. Each subtree corresponds to a paralog. The K-tree score compares topological and pairwise distances between trees after re-scaling overall tree length, with higher values corresponding to more divergent trees. The Robinson-Foulds score compares only topological distances between trees, the values shown corresponding to the fraction of divergent nodes between trees.

Reference tree	**Comparison tree**	**K-tree score**	**Robinson-Foulds score**
	Nucleotide tree VS species tree		
PTBP1	Species tree	0.759	42
PTBP2	Species tree	0.762	24
PTBP3	Species tree	1.700	28
	Nucleotide tree VS Amino acid tree		
PTBP1-AA	*PTBP1*-NT	0.149	78
PTBP2-AA	*PTBP2*-NT	0.129	110
PTBP3-AA	*PTBP3*-NT	0.380	40

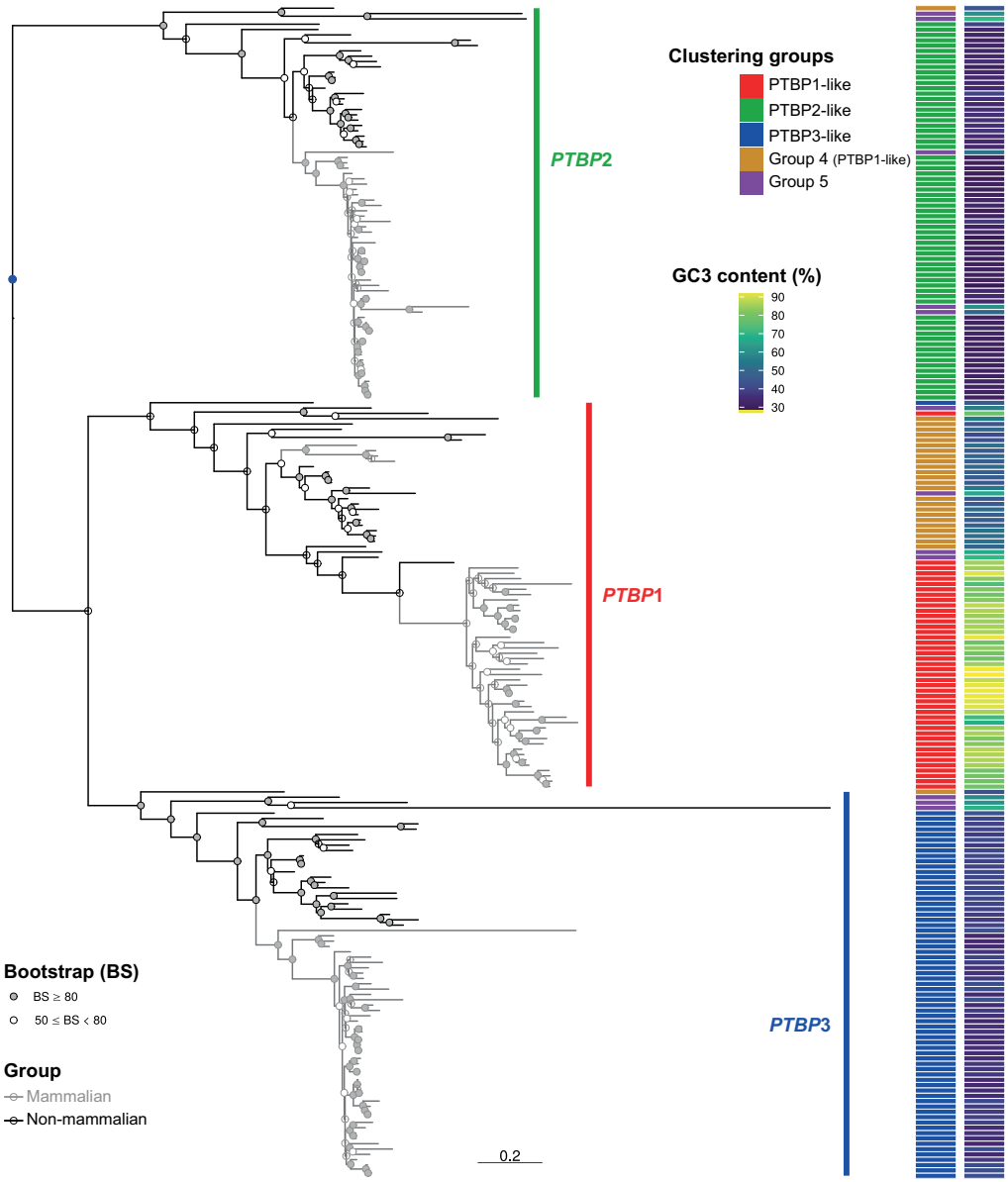

Figure 4: **Maximum-likelihood nucleic acid phylogeny of *PTBP*s genes.** The phylogram depicts *PTBP2*s (green side bar), *PTBP1*s (red side bar) and *PTBP3*s (blue side bar) clades. The outgroup genes from protostomata are not shown to focus on the scale for vertebrate *PTBP*s, but their placement on the tree and the polarity they provide for vertebrate *PTBP*s is given by the blue dot. Gray branches indicate mammalian *PTBP*s, while black branches indicate non-mammalian species. Note the lack of monophyly for mammals for *PTBP1*s. Filled dots on nodes indicate bootstrap values above 80, and empty dots indicate lower support values. Side bar on the left identifies the classification of each gene into the five groups identified by the hierarchical clusters, with the colour code in the inset. Side bar on the right displays GC3 content of the corresponding genes, with the gradient for the colour code ranging from 0 (blue) to 100% (yellow).

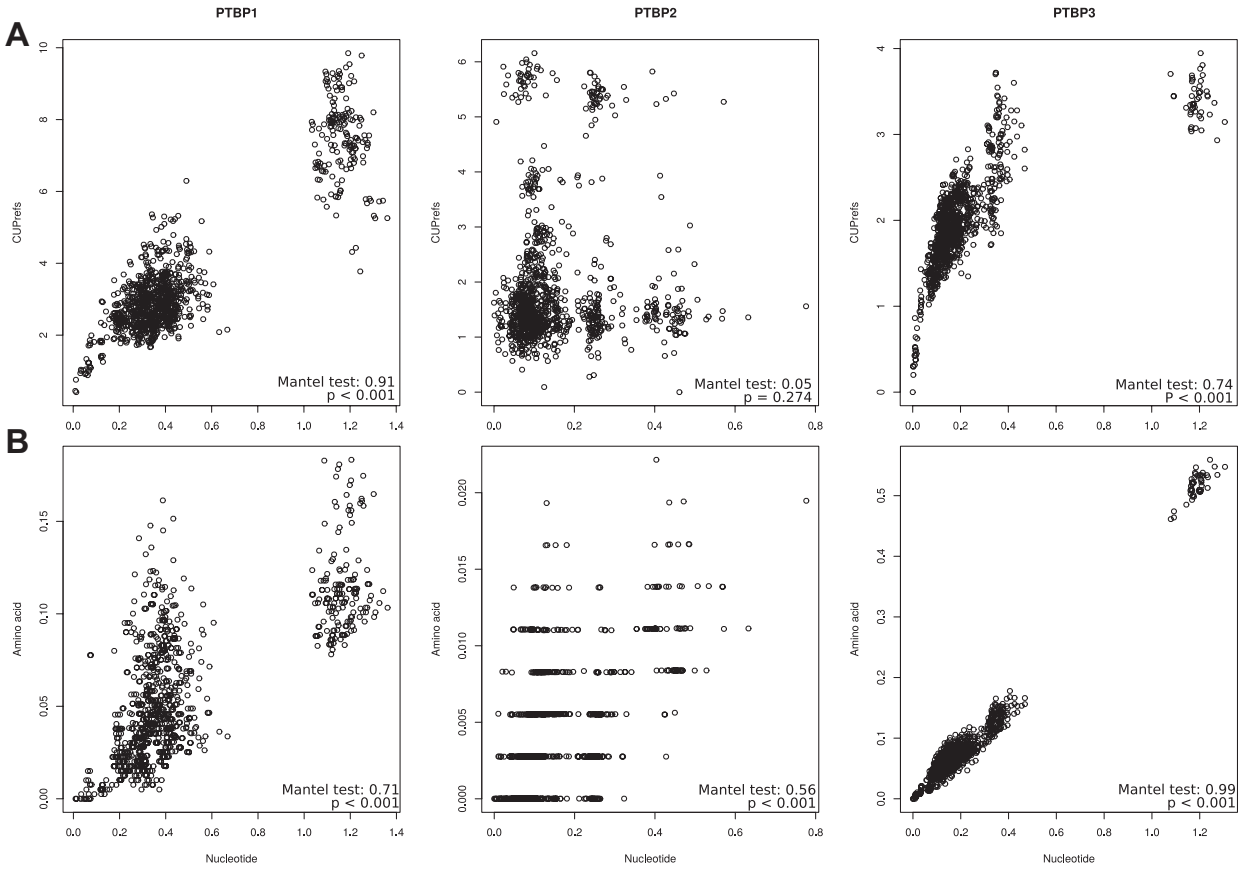

Figure 5: **Nucleotide-based pairwise distances against A) CUPrefs and B) amino-acid based pairwise distances for the different mammalian *PTBP* orthologs.** The results for a Mantel test assessing the correlation between the corresponding matrices are shown in the inset.

split apart from placental mammals in terms of both amino acid and nucleotide distances. This distribution matches well the fact that monotremes+marsupials do not cluster together with placental mammals in *PTBP1* phylogeny (see grey branches not being monophyletic for *PTBP1* in Figure 4). The same holds true for the platypus *PTBP3*, extremely divergent from the rest of the mammalian orthologs. For mammalian paralogs, the plots allow to see the increased number of overall mutations in general and of non-synonymous mutations in particular in *PTBP3*s compared with *PTBP1*. The precise mutational patterns are analysed in detail below. The histograms describing the accumulation of synonymous and non-synonymous mutations confirm that mammalian *PTBP1*s have selectively accumulated the largest number of synonymous mutations compared to non-mammalian *PTBP1*s and to other orthologs.

We have finally analysed the connection between nucleotide-based evolutionary distances within *PTBP* paralogs and CUPrefs-based distances (Figure 5A, Supplementary Material S8 A). A trend showing increased differences in CUPrefs as evolutionary distances increase is evident only for *PTBP1*s and *PTBP3*s in mammals. For mammalian *PTBP1*s the plot clearly differentiates a cloud with the values corresponding to the monotremes+marsupials splitting apart from placental mammals in terms of both evolutionary distance and CUPrefs. For mammalian *PTBP2*s the plot captures the divergent CUPrefs of the platypus and of the bats *M. natalensis* and *Desmodus rotundus*, while for non-

mammalian *PTBP2*s the divergent CUPrefs of the rainbow trout are obvious. Finally, for mammalian *PTBP3*s the large nucleotide divergence of the platypus paralog is evident. Importantly, all these instances of divergent behaviour (except for the platypus *PTBP3*) are consistent with the deviations described above from the expected composition by the mathematical modelling of the ortholog nucleotide composition.

Mammalian PTBP1s accumulate GC-enriching synonymous substitutions

We have shown that *PTBP1* genes are GC-richer and specifically GC3-richer than the *PTBP2* and *PTBP3* paralogs in the same genome, and that this enrichment is of a larger magnitude in placental *PTBP1*s. We have thus assessed whether a directional mutational pattern underlies this enrichment, especially regarding synonymous mutations. For this we have inferred the ancestral sequences of the respective most recent common ancestors of each *PTBP* paralogs, recapitulated synonymous and non-synonymous mutations between extant sequences and these ancestors, and constructed the corresponding mutation matrices (table S10). The two first axes of a principal component analysis using these mutational matrices capture, with a similar share, 66.95% of the variance between individuals (Figure 6). The first axis of the PCA separates synonymous from non-synonymous substitutions. Intriguingly though, while T<->C transitions are associated to synonymous mutations, as expected, G<->A transitions are associated to non-synonymous mutations. The second axis separates substitutions by their effect on nucleotide composition: GC-stabilizing/enriching on one direction, AT-stabilizing/enriching on the other one. Strikingly, the mutational spectrum of mammalian *PTBP1*s sharply differs from the rest of the paralogs. Substitutions in mammalian *PTBP1* towards GC-enriching changes, in both synonymous and non-synonymous compartments, are the main drivers of the second PCA axis. In contrast, synonymous mutations in *PTBP3* as well as all mutations in *PTBP2* tend to be AT-enriching. Finally, the mutational trends for *PTBP1* in mammals are radically different from those in non-mammals, while for *PTBP2* and *PTBP3*s the substitution patterns are similar in mammals and non-mammals for each of the compartments synonymous and non-synonymous.

4 Discussion

The non equal use of synonymous codons has puzzled biologists since first described. It has allowed for fruitful (and unfruitful) controversies between defenders of *all-is-neutralism* and defenders of *all-is-selectionism*, and has opened the door to the quest for embedded codes and signals behind CUPrefs patterns. The main questions around CUPrefs are twofold. On the one hand, their origin: to what extent they are the result of fine interplay between mutation and selection processes. On the other hand, their functional implications: whether and how particular CUPrefs can be linked to specific gene expression regulation processes, by modifying the kinetics and dynamics of DNA transcription, mRNA maturation and stability, mRNA translation, or protein folding and stability. In the present work we have built on the experimental results presented by Robinson and coworkers about the differential expression of the *PTBP* human gene paralogs as a function of their CUPrefs (Robinson et al., 2008). From this particular example, we have aimed at exploring by inductive thinking the general nature of the connection between paralogous gene evolution and CUPrefs. Our results show that the three *PTBP* paralogous genes of Vertebrates, which display divergent expression patterns, also have divergent nucleotide composition and CUPrefs. We propose here that this evolutionary pattern is compatible with a phenomenon of phenotypic evolution by sub-functionalisation (in this case specialisation in tissue-specific expression levels), associated to genotypic evolution by association to specific CUPrefs patterns.

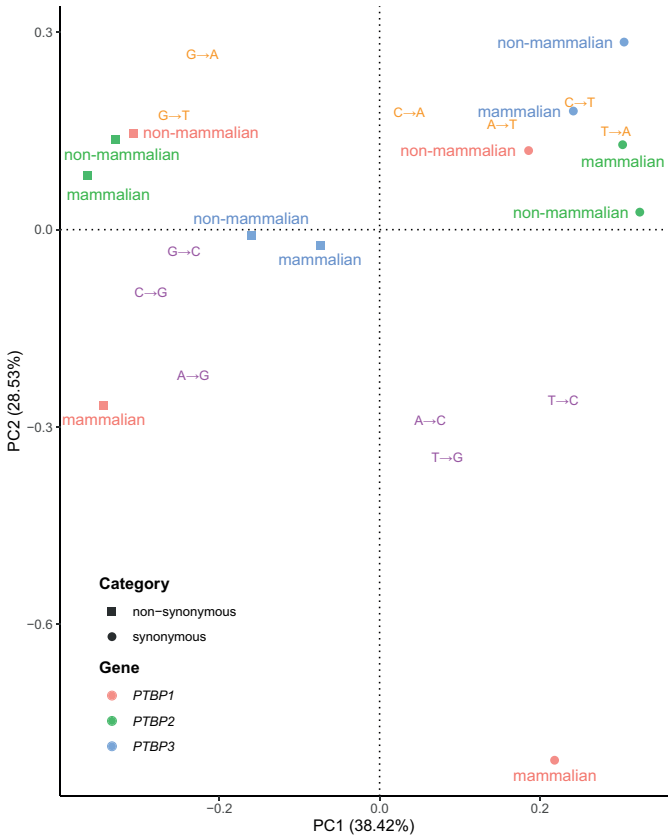

Figure 6: **Mutational spectra of synonymous and non-synonymous substitutions for *PTBP*s.** This principal component analysis (PCA) has been built using the observed nucleotide synonymous and non-synonymous substitution matrices for each *PTBP* paralog, inferred after phylogenetic inference and comparison of extant and ancestral sequences. The variables in this PCA are the types of substitution (*e.g.* A->G), identified by a colour code as GC-enriching / stabilizing substitutions (purple) or AT-enriching / stabilizing substitutions (orange). Variables are plotted according to their eigenvalues. Individuals in this PCA are the mutation categories in *PTBP* genes, stratified by their nature (synonymous or non-synonymous), by orthology (colour code for the different *PTBP*s is given in the inset) and by their taxonomy (mammals, or non-mammals).

We have reconstructed the phylogenetic relationships and analysed the evolution and diversity of CUPrefs among *PTBP* paralogs within 74 vertebrate species. The phylogenetic reconstruction shows that the genome of ancestral vertebrates already contained the three extant *PTBP* paralogs. This is consistent with the ortholog and paralog identification in the databases ENSEMBLE or ORTHOMAM (Yates et al., 2020; Scornavacca et al., 2019; Pina et al., 2018). Although our results suggest that *PTBP1* and *PTBP3* are sister lineages, the distant relationship of the vertebrate genes with the protostomate outgroup precludes the inference of a clear polarity between vertebrate *PTBP*s. We do not identify any instance of replacement between paralogs, and the evolutionary histories of the different *PTBP*s comply well with those of the corresponding species. The most blatant mismatch between gene and species trees is the polyphyly of mammalian *PTBP1*, with genes in monotremes and marsupials constituting a monophyletic clade, but not being the basal to and monophyletic with placental mammals. Multiple findings in our results point in this direction: i) the

excess of accumulation of synonymous mutations in mammalian *PTBP1*s for a similar total number of mutations (Figure 5 B); ii) the larger differences in CUPrefs between genes with a similar total number of nucleotide changes in the case of *PTBP1*s in mammals (Figure 5 A); iii) the explicitly different mutational spectrum of synonymous mutations in *PTBP1*s, enriched in A->C, T->G and T->C substitutions (Figure 6); iv) the sharp difference of CUPrefs between *PTBP1*s, and *PTBP2-3*s; and v) the clustering of *PTBP1* genes in monotremes and marsupials together with *PTBP1* genes in non-mammals according to their CUPrefs (Figure3 A). Overall, the particular nucleotide composition and the associated CUPrefs in mammalian *PTBP1* genes are most likely associated to specific mutational biases.

While GC3-rich nucleotide composition and CUPrefs of mammalian *PTBP1*s are dominated by local mutational biases, this is not the case for mammalian *PTBP2*s, overall AT3-richer. In vertebrates, nucleotide composition varies strongly along chromosomes, so that long stretches, historically named "isochores", appear enriched in GC or in AT nucleotides and present particular physico-chemical profiles (Caspersson et al., 1968). Local mutational biases underlying such heterogeneity, are the strongest evolutionary force shaping local nucleotide composition, so that the physical location of gene along the chromosome largely shapes its CUPrefs (Holmquist, 1989). In agreement with this mutational bias hypothesis, variation in GC3 composition of *PTBP1*s is almost totally explained by the variation in local GC composition (Table 3), suggesting that a same mutational bias has shaped the GC-rich composition of the flanking, intronic and coding regions of *PTBP1*s. The same trend, but to a lesser degree holds also true for *PTBP3*s. GC-biased gene conversion is often invoked as a powerful mechanism underlying such local GC-enrichment processes, leading to the systematic replacement of the alleles with the lowest GC composition by their GC richer homologs (Marais, 2003). It has been proposed that gene expression during meiosis facilitates GC-biased gene conversion during meiotic recombination (Pouyet et al., 2017), and in humans expression of *PTBP1*, GC3-enriched, is indeed documented during meiosis in the ovocite germinal line. Nevertheless, this line of reasoning does not holds true for *PTBP2*s. On the one hand, variations in local GC composition account barely for half of the variation in the *PTBP2* GC3 composition (Table 3). On the other hand, expression of *PTBP2*, AT3-enriched, is essential during spermatogenic meiosis (Zagore et al., 2015; Hannigan et al., 2017). Overall, GC3-enrichment in mammalian *PTBP2*s is compatible with GC-biased gene conversion events driving local mutational biases, but the AT3-enrichment of mammalian *PTBP2*s requires probably additional mechanisms to be explained, other than basal polymerase-related mutational biases for AT-enrichment, which acts as a background on the full genome (Hershberg and Petrov, 2010; Glémin et al., 2015; Petrov and Hartl, 1999).

In mammals, global GC-enriching genomic biases strongly impact CUPrefs, so that the most used codons in average tend to be GC-richer (Hershberg and Petrov, 2009). For this reason, in mammals GC3-rich *PTBP1*s match better the average genomic CUPrefs than AT3-richer *PTBP2*, which actually display CUPrefs in the opposite direction to the average of the genome. In the case of humans, *PTBP1* presents a COUSIN value of 1.747, consisting with an enrichment in preferentially-used codons, while on the contrary, the COUSIN value of -0.477 for *PTBP2* clearly points towards an enrichment in rare codons (Supplementary Material S4). Indeed, the poor match between human *PTBP2* CUPrefs and the human average CUPrefs results in poor expression of this gene in different human and murine cell lines, otherwise capable of expressing at high levels *PTBP1* and *PTBP3* (Robinson et al., 2008). The barrier to *PTBP2* expression seems to be the translation process, as *PTBP2* codon-recoding towards GC3-richer codons results in strong protein production in the same cellular context, without significant changes in the corresponding mRNA levels (Robinson et al., 2008). Such codon recoding strategy towards preferred codons has become indeed a standard practice

for gene expression engineering, despite our lack a comprehensive understanding of the impact and interaction on gene expression of local and global gene composition, nucleotide CUPrefs or mRNA structure (Brule and Grayhack, 2017).

The poor expression ability of human *PTBP2* in human cells, the large increase in protein production by the simple introduction of common codons and the lack of power of mutational biases to explain *PTBP2* nucleotide composition and CUPrefs, all raise the question of the adaptive value of the poor CUPrefs for this paralog. Specific tissue-dependent or cell-cycle dependent gene expression regulation patterns have been invoked to explain the codon usage-limited gene expression for certain human genes, such as *TLR7* or *KRAS* (Newman et al., 2016; Lampson et al., 2013; Fu et al., 2018). In humans, the expression levels of the three *PTBP* paralogs are tissue-dependent (Supplementary Material S1), and these differences are conserved through mammals (Keppetipola et al., 2012). In the case of the duplicated genes, subfunctionalisation through specialisation in spatio-temporal gene expression has often been proposed as the main evolutionary force driving conservation of paralogous genes (Ferris and Whitt, 1979). Such differential gene expression regulation in paralogs has actually been documented for a number of genes at very different taxonomic levels (Donizetti et al., 2009; Guschanski et al., 2017; Freilich et al., 2006). Specialised expression patterns in time and space can result in antagonistic presence/absence of the paralogous proteins (Adams et al., 2003). This is precisely the case of *PTBP1* and *PTBP2* during central nervous system development: in non-neuronal cells, *PTBP1* represses *PTBP2* expression by the skip of the exon 10 during *PTBP2* mRNA maturation, while during neuronal development, the micro RNA miR124 downregulates *PTBP1* expression, which in turn leads to upregulation of *PTBP2* (Keppetipola et al., 2012; Makeyev et al., 2007). Further, despite the high level of amino acid similarity between both proteins, *PTBP1* and *PTBP2* seem to perform complementary activities in the cell and to display different substrate specificity, so that they are not directly inter-exchangeable by exogenous manipulation of gene expression patterns (Vuong et al., 2016).

In a different subject, we want to drive the attention of the reader towards the puzzling trend of the UUG-Leu codon in our CUPrefs analyses. This UUG codon is the only GC-ending codon systematically clustering with AT-ending codons in all our analyses, and does not show the expected symmetrical behaviour with respect to UUA (see Figure 3). Such behaviour for UUG has been depicted, but not discussed, in other analyses of CUPrefs in mammalian genes (see figure 7 in Laurin-Lemay et al. (2018)), as well as for AGG-Arg and GGG-Gly in a global study of codon usages across the tree of life (see figure 1 in (Novoa et al., 2019)). The reasons underlying the clustering of UUG with AT-ending codons are unclear. A first line of thought could be functional: the UUG-Leu codon is particular because it can serve as alternative starting point for translation (Peabody, 1989). However, other codons such as ACG or GUG act more efficiently than UUG as translation initiation, and do not display any noticeable deviation (Ivanov et al., 2011). A second line of thought could be related to the tRNA repertoire, but both UUG and UUA are decoded by similar numbers of dedicated tRNAs in the vast majority of genomes (*e.g.* respectively six and seven tRNA genes in humans (Palidwor et al., 2010)). Finally, another line of thought suggests that UUG and AGG could be disfavoured if mutational pressure towards GC is very high, despite being GC-ending codons (Palidwor et al., 2010). Indeed, the series of synonymous transitions UUA->UUG->CUG for Leucine and the substitution chain AGA->AGG->CGG for Arginine are expected to lead to a depletion of UUG and of AGG codons when increasing GC content. Both UUG and ACG codons would this way display a non-linear, non-monotonic response to GC-mutational biases (Palidwor et al., 2010). In our dataset, however, AGG maps with the rest of GC-ending codons, symmetrically opposed to AGA as expected, and strongly contributing to the second PCA axis. Thus, only UUG presents frequency use patterns similar

to those of AT-ending codons. We humbly admit that we do not find a satisfactory explanation for this behaviour and invite researchers in the field to generate alternative explanatory hypotheses.

We have presented here an evolutionary analysis of the *PTBP* paralogs family, as a paradigm of evolution upon gene duplication. Our results show that CUPrefs in *PTBP*s have evolved in parallel with specific gene expression regulation patterns. In the case of *PTBP1*, the most tissue-wise expressed of the paralogs, we have identified compositional, mutational biases as the driving force leading to strong enrichment in GC-ending codons. In contrast, for *PTBP2* the enrichment in AT-ending codons is rather compatible with selective forces related to specific spatio-temporal gene expression pattern, antagonistic to those of *PTBP1*. Our results suggest that the systematic study of composition, genomic location and expression patterns of paralogous genes can contribute to understanding the complex mutation-selection interplay shaping CUPrefs in multicellular organisms.

5 Acknowledgments

J.B. is the recipient of a PhD fellowship from the French Ministry of Education and Research. This study was supported by the European Union's Horizon 2020 research and innovation program under the grant agreement CODOVIREVOL (ERC-2014-CoG-647916) to I.G.B. The authors acknowledge the CNRS and the IRD for additional (meagre) intra-mural support. The computational results presented have been achieved in part using the IRD Bioinformatic Cluster itrop.

6 Data Availability Statement

All data required to reproduce our findings is provided in the tables in the main text or in the Supplementary Material section.

References

Adams KL, Cronn R, Percifield R, Wendel JF. 2003, April. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proceedings of the National Academy of Sciences of the United States of America. 100(8):4649–4654.

Apostolou-Karampelis K, Nikolaou C, Almirantis Y. 2016, August. A novel skew analysis reveals substitution asymmetries linked to genetic code GC-biases and PolIII a-subunit isoforms. DNA research: an international journal for rapid publication of reports on genes and genomes. 23(4):353–363.

Bourret J, Alizon S, Bravo IG. 2019, December. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. Genome Biology and Evolution. 11(12):3523–3528. Publisher: Oxford Academic.

Brule CE, Grayhack EJ. 2017. Synonymous Codons: Choose Wisely for Expression. Trends in genetics: TIG. 33(4):283–297.

Bulmer M. 1991, November. The selection-mutation-drift theory of synonymous codon usage. Genetics. 129(3):897–907.

Caliskan N, Peske F, Rodnina MV. 2015, May. Changed in translation: mRNA recoding by 1 programmed ribosomal frameshifting. Trends in Biochemical Sciences. 40(5):265–274.

Carbone A, Zinovyev A, Képès F. 2003, November. Codon adaptation index as a measure of dominating codon bias. Bioinformatics (Oxford, England). 19(16):2005–2015.

Caspersson T, Farber S, Foley GE, Kudynowski J, Modest EJ, Simonsson E, Wagh U, Zech L. 1968, January. Chemical differentiation along metaphase chromosomes. Experimental Cell Research. 49(1):219–222.

Castresana J. 2000, April. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution. 17(4):540–552.

Chamary JV, Parmley JL, Hurst LD. 2006, February. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Reviews. Genetics. 7(2):98–108.

Clark JM. 1988, October. Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. Nucleic Acids Research. 16(20):9677–9686.

Copley SD. 2020, April. Evolution of new enzymes by gene duplication and divergence. The FEBS journal. 287(7):1262–1283.

Donizetti A, Fiengo M, Minucci S, Aniello F. 2009, October. Duplicated zebrafish relaxin-3 gene shows a different expression pattern from that of the co-orthologue gene. Development, Growth & Differentiation. 51(8):715–722.

Duret L. 2002, December. Evolution of synonymous codon usage in metazoans. Current Opinion in Genetics & Development. 12(6):640–649.

Duret L, Mouchiroud D. 1999, April. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proceedings of the National Academy of Sciences. 96(8):4482–4487. Publisher: National Academy of Sciences Section: Biological Sciences.

Ferris SD, Whitt GS. 1979, April. Evolution of the differential regulation of duplicate genes after polyploidization. Journal of Molecular Evolution. 12(4):267–317.

Freilich S, Massingham T, Blanc E, Goldovsky L, Thornton JM. 2006. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. Genome Biology. 7(10):R89.

Fu J, Dang Y, Counter C, Liu Y. 2018. Codon usage regulates human KRAS expression at both transcriptional and translational levels. The Journal of Biological Chemistry. 293(46):17929–17940.

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, Bierne N, Duret L. 2018, May. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Molecular Biology and Evolution. 35(5):1092–1103.

Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015, August. Quantification of GC-biased gene conversion in the human genome. Genome Research. 25(8):1215–1228. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980, January. Codon catalog usage and the genome hypothesis. Nucleic Acids Research. 8(1):r49–r62.

Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. Genome Research. 27(9):1461–1474.

Hannigan MM, Zagore LL, Licatalosi DD. 2017, June. Ptbp2 controls an alternative splicing network required for cell communication during spermatogenesis. Cell reports. 19(12):2598–2612.

Hershberg R, Petrov DA. 2009, July. General rules for optimal codon choice. PLoS genetics. 5(7):e1000556.

Hershberg R, Petrov DA. 2010, September. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genetics. 6(9).

Holmquist GP. 1989, June. Evolution of chromosome bands: Molecular ecology of noncoding DNA. Journal of Molecular Evolution. 28(6):469–486.

Ikemura T. 1981, September. 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. Journal of Molecular Biology. 151(3):389–409.

Ivanov IP, Firth AE, Michel AM, Atkins JF, Baranov PV. 2011, May. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Research. 39(10):4220–4234.

Katoh K, Misawa K, Kuma Ki, Miyata T. 2002, July. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research. 30(14):3059–3066.

Keppetipola N, Sharma S, Li Q, Black DL. 2012, August. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. Critical Reviews in Biochemistry and Molecular Biology. 47(4):360–378.

Khorana HG, Büchi H, Ghosh H, Gupta N, Jacob TM, Kössel H, Morgan R, Narang SA, Ohtsuka E, Wells RD. 1966. Polynucleotide synthesis and the genetic code. Cold Spring Harbor Symposia on Quantitative Biology. 31:39–49.

Koonin EV. 2005. Orthologs, Paralogs, and Evolutionary Genomics. Annual Review of Genetics. 39(1):309–338. _eprint: https://doi.org/10.1146/annurev.genet.39.073003.114725.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution. 34(7):1812–1819.

Lampson BL, Pershing NLK, Prinz JA, Lacsina JR, Marzluff WF, Nicchitta CV, MacAlpine DM, Counter CM. 2013, January. Rare codons regulate KRas oncogenesis. Current biology: CB. 23(1):70–75.

Laurin-Lemay S, Rodrigue N, Lartillot N, Philippe H. 2018. Conditional Approximate Bayesian Computation: A New Approach for Across-Site Dependency in High-Dimensional Mutation-Selection Models. Molecular Biology and Evolution. 35(11):2819–2834.

Lujan SA, Williams JS, Pursell ZF, Abdulovic-Cui AA, Clark AB, McElhinny SAN, Kunkel TA. 2012, October. Mismatch Repair Balances Leading and Lagging Strand DNA Replication Fidelity. PLOS Genetics. 8(10):e1003016. Publisher: Public Library of Science.

Makeyev EV, Zhang J, Carrasco MA, Maniatis T. 2007, August. The MicroRNA miR-124 Promotes Neuronal Differentiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing. Molecular cell. 27(3):435–448.

Marais G. 2003, June. Biased gene conversion: implications for genome and sex evolution. Trends in Genetics. 19(6):330–338. Publisher: Elsevier.

Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, Liss M, Taylor MS, Hurst LD, Kudla G. 2020, April. Codon Usage and Splicing Jointly Influence mRNA Localization. Cell Systems. 10(4):351–362.e8.

NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 46(D1):D8–D13.

Newman ZR, Young JM, Ingolia NT, Barton GM. 2016, March. Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. Proceedings of the National Academy of Sciences of the United States of America. 113(10):E1362–1371.

Nirenberg MW, Matthaei JH. 1961, October. THE DEPENDENCE OF CELL- FREE PROTEIN SYNTHESIS IN E. COLI UPON NATURALLY OCCURRING OR SYNTHETIC POLYRIBONUCLEOTIDES. Proceedings of the National Academy of Sciences of the United States of America. 47(10):1588–1602.

Novoa EM, Jungreis I, Jaillon O, Kellis M. 2019. Elucidation of Codon Usage Signatures across the Domains of Life. Molecular Biology and Evolution. 36(10):2328–2339.

Novoa EM, Ribas de Pouplana L. 2012, November. Speeding with control: codon usage, tRNAs, and ribosomes. Trends in genetics: TIG. 28(11):574–581.

Palidwor GA, Perkins TJ, Xia X. 2010, October. A general model of codon bias due to GC mutational bias. PloS One. 5(10):e13431.

Peabody DS. 1989, March. Translation initiation at non-AUG triplets in mammalian cells. The Journal of Biological Chemistry. 264(9):5031–5035.

Percudani R, Pavesi A, Ottonello S. 1997, May. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae11Edited by J. Karn. Journal of Molecular Biology. 268(2):322–330.

Petrov DA, Hartl DL. 1999, February. Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proceedings of the National Academy of Sciences. 96(4):1475–1479. Publisher: National Academy of Sciences Section: Biological Sciences.

Pina J, Ontiveros RJ, Keppetipola N, Nikolaidis N. 2018, April. A Bioinformatics Approach to Discover the Evolutionary Origin of the PTBP Splicing Regulators. The FASEB Journal. 32(1_supplement):802.16–802.16. Publisher: Federation of American Societies for Experimental Biology.

Plotkin JB, Kudla G. 2011, January. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics. 12(1):32–42.

Pouyet F, Mouchiroud D, Duret L, Sémon M. 2017. Recombination, meiotic expression and human codon usage. eLife. 6.

Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J. 2015, March. Codon optimality is a major determinant of mRNA stability. Cell. 160(6):1111–1124.

Reijns MAM, Kemp H, Ding J, Marion de Procé S, Jackson AP, Taylor MS. 2015, February. Lagging-strand replication shapes the mutational landscape of the genome. Nature. 518(7540):502–506. Number: 7540 Publisher: Nature Publishing Group.

Robinson DF, Foulds LR. 1981, February. Comparison of phylogenetic trees. Mathematical Biosciences. 53(1):131–147.

Robinson F, Jackson RJ, Smith CWJ. 2008, March. Expression of Human nPTB Is Limited by Extreme Suboptimal Codon Content. PLOS ONE. 3(3):e1801. Publisher: Public Library of Science.

Satapathy SS, Powdel BR, Buragohain AK, Ray SK. 2016, October. Discrepancy among the synonymous codons with respect to their selection as optimal codon in bacteria. DNA Research. 23(5):441–449. Publisher: Oxford Academic.

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019, April. OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. Molecular Biology and Evolution. 36(4):861–862. Publisher: Oxford Academic.

Sharp PM, Li WH. 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. 15(3):1281–1295.

Sonnhammer ELL, Koonin EV. 2002, December. Orthology, paralogy and proposed classification for paralog subtypes. Trends in genetics: TIG. 18(12):619–620.

Soria-Carrasco V, Talavera G, Igea J, Castresana J. 2007, November. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics (Oxford, England). 23(21):2954–2956.

Spencer PS, Barral JM. 2012, March. Genetic code redundancy and its influence on the encoded polypeptides. Computational and Structural Biotechnology Journal. 1.

Stamatakis A. 2014, May. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England). 30(9):1312–1313.

Vuong JK, Lin CH, Zhang M, Chen L, Black DL, Zheng S. 2016. PTBP1 and PTBP2 Serve Both Specific and Redundant Functions in Neuronal Pre-mRNA Splicing. Cell Reports. 17(10):2766–2775.

Whittle CA, Extavour CG. 2016, September. Expression-Linked Patterns of Codon Usage, Amino Acid Frequency, and Protein Length in the Basally Branching Arthropod Parasteatoda tepidariorum. Genome Biology and Evolution. 8(9):2722–2736. Publisher: Oxford Academic.

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugán JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, IIsley G, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P. 2020, January. Ensembl 2020. Nucleic Acids Research. 48(D1):D682–D688. Publisher: Oxford Academic.

Zagore LL, Grabinski SE, Sweet TJ, Hannigan MM, Sramkoski RM, Li Q, Licatalosi DD. 2015, December. RNA Binding Protein Ptbp2 Is Essential for Male Germ Cell Development. Molecular and Cellular Biology. 35(23):4030–4042.