The patterns of codon usage between chordates and arthropods are different but co-evolving with mutational biases

Ioanna Kotari[1,2], **Carolin Kosiol**[3], and **Rui Borges**[1*]

[1] *Institut für Populationsgenetik, University of Veterinary Medicine, Veterinärplatz 1, 1210 Vienna, Austria*
[2] *Vienna Graduate School of Population Genetics, Vienna, Austria*
[3] *Centre for Biological Diversity, University of St Andrews, St Andrews, Fife, UK*
[*] *corresponding author: ruiborges23@gmail.com*

Abstract

Different frequencies amongst codons that encode the same amino acid (i.e. synonymous codons) have been observed in multiple species. Studies focused on uncovering the forces that drive such codon usage bias showed that a combined effect of mutational biases and translational selection works to produce different frequencies of synonymous codons. However, only few have been able to measure and distinguish between these forces that may leave similar traces on the coding regions. Here, we have developed a codon model that allows the disentangling of mutation, selection on amino acids and synonymous codons, and GC–biased gene conversion (gBGC) which we employed on an extensive dataset of 415 chordates and 191 arthropods. We found that chordates need 15 more synonymous codon categories than arthropods to explain the empirical codon frequencies, which suggests that the extent of codon usage can vary greatly between animal phyla. Moreover, methylation at CpG sites seems to partially explain these patterns of codon usage bias in chordates but not in arthropods. Our findings also demonstrate that GC–rich codons are disfavoured in both phyla when mutations are GC–biased, and the opposite is true when mutations are AT–biased. This indicates that selection on the genomic coding regions might act primarily to stabilise its GC/AT content. Our study shows that the degree of synonymous codon usage varies considerably among animals, but is likely governed by a common underlying dynamic.

Keywords: synonymous codon usage; GC–biased gene conversion; translational selection; mutational bias; evolution; chordates; arthropods; non–model organisms

Significance statement

The reasons for the differential usage of codons encoding for the same amino acid has puzzled scientist for decades. By examining the frequencies of synonymous codons in different species, this study presents a novel model that sheds light on the underlying factors that drive differences in codon usage between chordates and arthropods. Our analysis unveiled more extensive codon usage patterns in chordates compared to arthropods. Moreover, the study highlights that genome–wide selection acts to balance mutational biases favouring G/C alleles. This research provides valuable insights into our understanding of the complex interplay between mutational biases and selection forces in shaping the variation at the synonymous sites, and has important implications for future studies of genome evolution and adaptation.

1 Introduction

Synonymous codons encode for the same amino acid and are expected to be neutral and used interchangeably in the genome. However, synonymous codons appear at different frequencies across protein–coding genes. This preferential usage of synonymous codons is called codon usage bias and is widely spread across the tree of life (Subramanian, 2008). In order to explain the differential use of synonymous codons, two main narratives have been proposed: mutational biases and translational selection (Duret, 2002; Hershberg and Petrov, 2008). Translational selection is expected to affect highly expressed genes and correlate well with tRNA abundance to increase translation efficiency, and has been found to bias codon usage in a number of taxa, e.g., *Drosophila* (Shields et al., 1988; Akashi, 1994; Bierne and Eyre-Walker, 2006), *C. elegans* (Duret and Mouchiroud, 1999), as well as in *E. coli* (Ikemura, 1981). On the other hand, mutational biases have been found to shape codon usage bias according to the base composition of the whole genome and GC–content in the third codon position (e.g., yeast (Sharp et al., 1995), and vertebrates (Urrutia and Hurst, 2001; Rao et al., 2011) including humans (Sueoka and Kawanishi, 2000)).

The exact causes of codon usage bias are still widely debated, but it is of common agreement that it arises from the interplay of selection, mutation, and genetic drift (Bulmer, 1991). Studies have found that mutational forces are acting alongside the well–established effect of translational selection on codon usage bias in *Drosophila* (Kliman and Hey, 1994). Similarly, a correlation between preferred codons and tRNA levels was discovered in several vertebrate species, showing that weak translational selection operates alongside mutations (Doherty and McInerney, 2013). Additionally, Galtier et al. (2018) recently described how GC–biased gene conversion (gBGC) – a meiotic recombination–associated mutation bias that favours GC– over AT–alleles (Marais, 2003) – has a widespread effect on codon preferences across animal species. Therefore, the debate has changed from identifying the forces responsible for codon usage bias to determining which of these forces are more significant.

Previous codon usage bias research was restricted to a few model organisms in animals and mainly

focused on highly expressed genes. Most extensive interspecific studies of codon usage are focused on yeast (LaBella et al., 2019) and bacteria (Sharp et al., 2005). With advancements in sequencing technology, whole genomic coding sequences from a wider range of species are now accessible. This led recent studies to shift their focus on inter-species variation in animals as well. Doherty and McInerney (2013) analysed a range of vertebrate species, while Galtier et al. (2018) expanded beyond vertebrates and included species across the whole animal kingdom. However, data now exists that allows a more detailed comparison of the patterns of codon usage across phyla.

The assessment of codon usage bias has traditionally relied on heuristic methods, e.g., codon usage indices. The most commonly used indices include the Codon Adaptation Index (CAI) (Sharp and Li, 1987), the Relative Synonymous Codon Usage (RSCU) values (e.g., Gerdol et al. (2015) and the GC–content at the third position of synonymous codons (e.g., Galtier et al. (2018)). These cluster the different codons in a small number of categories, usually as preferred and non-preferred, but fail to fully account for the effects of mutation and selection processes. To disentangle these effects, mechanistic approaches that rely on population genetics models have been proposed. A maximum likelihood approach includes the multi-allele model (Zeng, 2010), which attempts to categorise codons into four classes, and two Bayesian approaches, FMutSel (Yang and Nielsen, 2008) and ROC SEMPPR (Gilchrist et al., 2015), which quantify selection and mutation's impact on codon preference. However, none of these approaches incorporates gBGC as a factor in codon usage bias.

In this study, we devised a mechanistic model based on the mutation-selection Moran model (Moran, 1958), which explains the fixed differences between species based on population genetics forces. To estimate these forces, we developed a Bayesian estimator called DECUB (Disentangling the Effect of Codon Usage Bias) which quantifies the joint effect of mutations, selection and gBGC across the whole genome. We employed our model on coding sequences from over 600 species belonging to chordates and arthropods to (i) disentangle and evaluate the effects of the aforementioned confounding forces of codon usage bias, and (ii) compare the patterns of codon usage between the phyla. We found that codon usage bias is more extensive in chordates compared to arthropods, and that genome-wide codon usage in both taxa is co-evolving with mutational biases.

2 Results

2.1 Modelling the evolution of codon frequencies

In this study, we created a Moran model with mutations, GC-bias, and selection to model the codon frequencies along the genome. Mutations are modelled similarly to the general time–reversible (GTR) substitution model (Tavaré, 1986) and GC–bias is incorporated to capture the effects of GC–biased gene conversion (gBGC). Since both forces are mutational processes to their core, the joint effects of these on a given codon $I = i_1 i_2 i_3$ are summarised in the mutational coefficient of all three nucleotides $\beta_I = \beta_{i_1}\beta_{i_2}\beta_{i_3}$. Selection acting on said codon is modelled as a relative fitness coefficient Φ_I.

As we are interested in capturing the variation at the inter–species level, we used the stationary frequencies of the fixed sites (supplementary text S2), which can be described for each codon, as

$$\psi_I = \frac{1}{\sum_I \beta_I \Phi_I} \beta_I \Phi_I \ , \tag{1}$$

where the denominator is set such that the stationary frequencies add up to 1.

Using the stationary distribution in eq. (1), we developed a software called DECUB (Disentangling the Effects of Codon Usage Bias), which takes as input codon counts from a given taxon (e.g., population, species) and infers the mutational coefficient of each nucleotide and the fitness coefficients of each codon using a Bayesian framework. All mutational coefficients are normalised based on adenine (A) since the effect of gBGC does not influence mutational biases towards A. Similarly, as methionine is an essential amino acid encoded by a single codon and is therefore not confounded with codon usage bias, all fitness coefficients are normalised based on the codon ATG. The general model presented in (1) can potentially assume a fitness coefficient per codon. However, after performing extensive simulations, we found that the model is unidentifiable after 53 fitness coefficients (supplementary fig. S1).

2.2 Chordates have more pronounced patterns of codon usage than Arthropods

To characterise the patterns of codon usage in arthropods and chordates, we used codon counts from genome-wide coding sequences of 606 species: 415 chordates and 191 arthropods. All species are encoded as per the standard genetic code. We set up our inferences by using an amino acid mapping, which sets a fitness coefficient per amino acid, plus one per stop codon (total 23 Φ categories). Figure 1 shows the estimation of each mutational and fitness coefficient on a logarithmic scale. Although the estimates were distributed similarly, chordates were more homogeneous in their estimates across the species studied compared to arthropods.

The estimated mutational biases of thymine (β_T) are close to 0, therefore almost equal to β_A, the mutational effects of which were used to normalise all mutational coefficients. Moreover, as would be expected, we see greater variation in the mutation coefficients of GC alleles, which are driven by the combined effect of mutations and gBGC. The estimation of the fitness coefficients highlights the differential usage of each amino acid, and because we did not confine our analyses to the sense codons, it informs us on how deleterious the stop codons are in comparison to sense codons. Arthropods show a preference for TAA as their stop codon, despite its median fitness being 13 times lower than that of the lowest amino acid (arginine). On the other hand, chordates prefer TGA as their stop codon, with a median that is 7.3 times lower than that of the lowest amino acid (also arginine).

The amino acid mapping expresses the genetic code and is a natural approach to modelling fitness effects; however, it ignores variation between synonymous codons, thus disregarding codon usage bias. To identify codons needing their own fitness coefficient, we compared the error between the empirical codon frequencies and the predicted ones from our model. Figure 2 shows the percentage of species whose estimates deviate from each empirical codon count. In arthropods, most of the observed variation

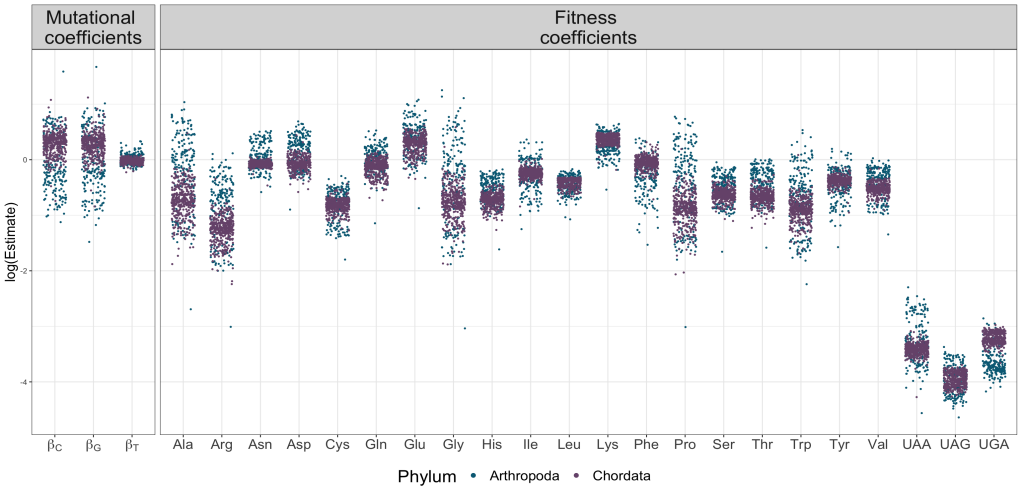

Figure 1: Estimation of mutational and fitness coefficients under an amino acid mapping. Logarithmic scale of each estimate which shows positive values if a coefficient is estimated higher than the reference one or negative if lower. The mutational coefficients were normalised with β_A which does not include the effects of GC-biased gene conversion (gBGC) and the fitness coefficients were normalised by the single codon for Methionine.

can be sufficiently explained with the amino acid mapping. However, more codon categories are clearly needed to account for the more extensive variation amongst the synonymous codons in chordates.

We proceeded to add more fitness coefficients in a step-wise manner to the codons where more than 20% of species do not fit the empirical distribution. First, we added a fitness coefficient to the codon of each amino acid that had the highest percentage of error. We proceeded to repeat this procedure until all variation was below the 20%, resulting in a final phylum-specific mapping (fig. 2). We note that chordates reached the identifiability threshold with 3 codons exceeding the 20% threshold; however, the errors of these were relatively low, not exceeding 25%. Hence, the final chordate mapping has 52 fitness coefficients compared to arthropods that are modelled with 37. Although chordates exhibit greater variation in synonymous codons, this pattern is comparatively more uniform across species than arthropods, who display less extensive bias but greater heterogeneity within their taxa (supp. fig. S2). Between arthropods and chordates, however, the final mappings show clear differences in patterns of codon usage, with chordates needing over twice the number of extra coefficients compared to arthropods (29 vs. 14 extra codon fitness coefficients).

2.3 Fitness coefficients balance mutational effects

We established that chordates and arthropods have different patterns of codon usage; however, the mutation and fitness coefficients vary considerably within these phyla (fig. 1). To understand these differences, we focused our analyses on the main representative chordate classes (mammals, birds, reptiles and amphibians, and fish), which comprise 97% of the dataset. Similarly for arthropods, the subsequent analyses focus on Diptera, Lepidoptera, and Hymenoptera (flies and mosquitoes, butterflies and moths, and ants, wasps and bees, respectively), which make up for 74% of the arthropod dataset,

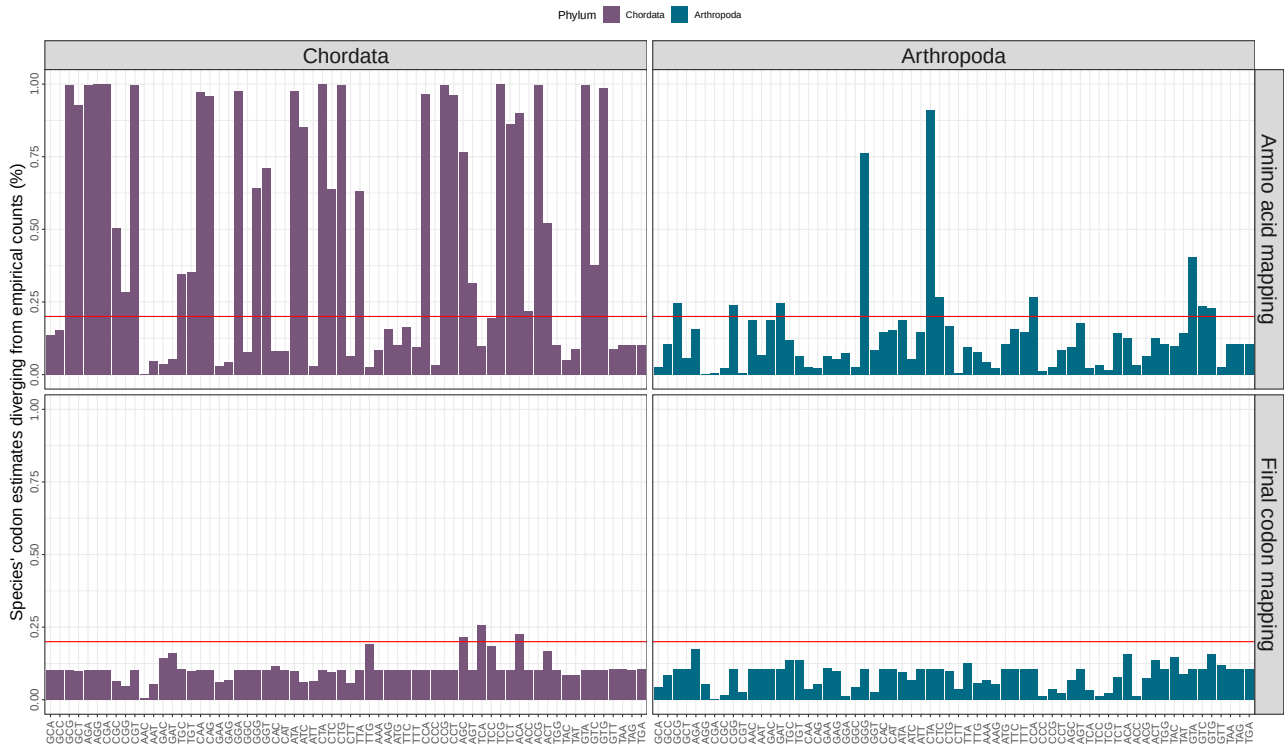

Figure 2: Error plots between predicted and empirical codon frequencies in chordates and arthropods, differences between the amino acid mapping and the final codon mapping. Assuming no variation amongst synonymous codons (amino acid mapping, i.e. one Φ per amino acid), showed a more extensive codon usage pattern in chordates compared to arthropods. In chordates (left, purple) 35 codons out of 64 had more than 20% of their species being outside the empirical distribution, whereas arthropods (right, blue) had overall smaller errors and only 10 codons exceeding the 20% threshold. We proceeded to minimise these errors per codon by adding more fitness coefficients. Each new mapping adds a new fitness coefficient per amino acid to the codon that had the highest error in the previous mapping. The final codon mapping (shown in the bottom plot) adds 29 extra coefficients to the chordates and 14 to arthropods, resulting in a total of 53 and 37 coefficients, respectively. We note that 3 codons still exceed the 20% threshold, as the model reached the identifiability threshold in chordates.

and more specifically more than 83% of the collected insect species (fig. 6, table S1).

In our model, mutational biases also express gBGC. We investigated its role in codon usage bias by correlating the GC–content of each codon with the bias of mutations towards GC alleles. Figure 3 shows the slope of the codon fitnesses with increasing GC–content for each codon against the difference between GC– and AT– mutational effects for all taxa studied in chordates and arthropods. Both chordates and arthropods have significant negative correlations (Spearman's ρ rank correlation of -0.976 and -0.934 respectively; p–value $< 2.2 \times 10^{-16}$). This indicates that the larger the mutational bias towards GC (compared to AT), the less favourable (i.e. the more negative the correlation coefficient) a codon is as its GC–content increases. Additionally, these patterns do not seem to be an artefact of the model, as upon integrating new coefficients, the newly added codons do not show an association with their nucleotide composition (supplementary fig. S3), and both GC–only and AT–only codons were chosen in both phyla.

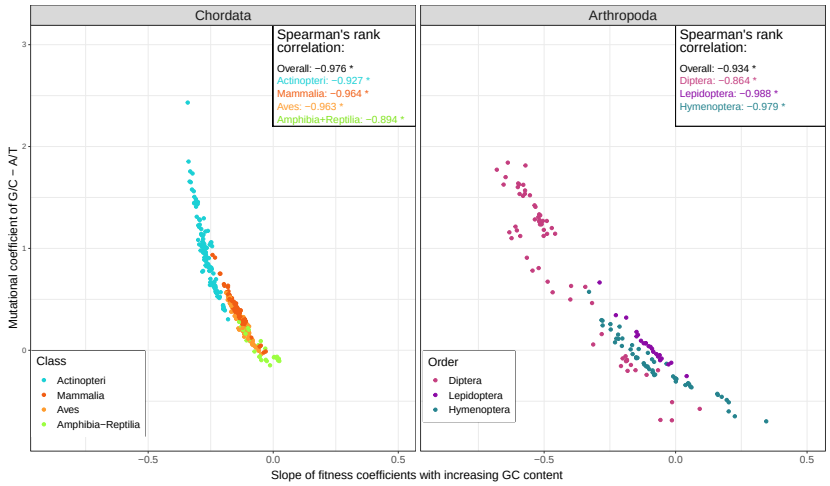

Figure 3: Negative correlation between mutational biases towards GC and increasing GC–content of codon fitness coefficients in chordates and arthropods. The larger the difference between the mutational coefficients of GC and AT nucleotides, the more negative the slope of fitnesses is, i.e. the less favourable a codon will be if it is richer in GC–content. In both chordates (left) and arthropods (right), this relationship is inverse with significant Spearman's rank correlation coefficients ρ, both overall and per inner taxa (box in the upper right corner of each graph). The colours represent the different taxa plotted, the classes fish (blue), mammals (red), birds (orange), and amphibians & reptiles (green) in chordates and the orders Diptera (flies & mosquitoes, pink), Lepidoptera (butterflies & moths, purple), and Hymenoptera (ants, bees & wasps, turquoise) in arthropods.

A negative slope between the mutation bias towards GC and the preference for GC–rich codons was also observed among chordate and arthropod groups (all p–values $< 1.7 \times 10^{-6}$). Fishes have the steepest negative slope (> -0.25), and mammals and birds show the highest correlation coefficients (-0.976 and -0.964 respectively, compared to -0.927 and -0.894 in fishes and reptiles), although all taxa are correlated to a similar degree. Finally, reptiles and amphibians have more homogeneous AT vs. GC mutation preferences, showing a slope very close to 0. Dipterans show some of the largest variation between the mutational differences (-1 to 2) with codon fitness slopes ranging from around 0 to -0.75. We observed in hymenopterans that AT mutations were equal or sometimes higher than GC, which is reflected in their fitnesses showing a rise (slope that reaches approximately 0.3) in their estimates as GC content increases.

To gain further insight into the relative importance of mutational biases and fitness coefficients in determining the patterns of codon usage in chordates and arthropods, we performed a sensitivity analysis. We used the Shannon entropy (Shannon, 1948), as this measure has been previously used to study patterns of codon usage (Suzuki et al., 2004) and is effective in assessing the likelihood of the codon frequencies changing due to variations in the population parameters. We varied the estimated mutation and fitness coefficients by $\pm 10\%$ and captured the difference in entropy caused to the stationary distribution of the model (see Methods for more details). Figure 4 shows the differences in entropy of each β and Φ coefficient. It is evident that the mutational biases had a much larger effect on the entropy, with β_T having the largest effect in chordates, as 10% variability results in 0.51% change on the stationary distribution. This effect is 18.6 times larger than the effect of codon fitnesses

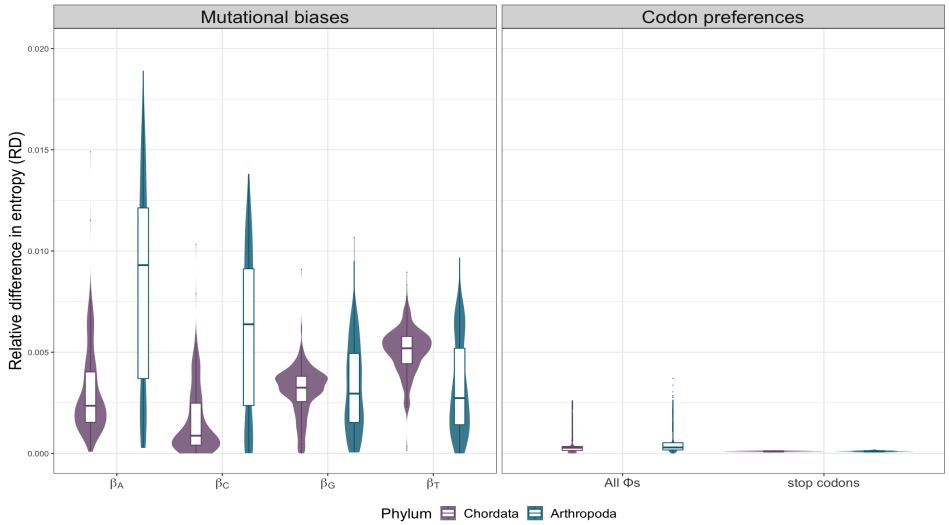

Figure 4: Relative difference in entropy measuring the evolutionary effect of each coefficient in arthropods and chordates. Violin plots with boxplots of the relative difference in entropy of every species, both in chordates (purple) and arthropods (blue). The entropy difference for the mutational biases (left) measures the effect a perturbation of 10% of each β has on the stationary distribution. For the fitness coefficients, the summarised effects of all sense and all stop codon fitness coefficients can be seen on the right. The effect of mutational effects are different between chordates and arthropods but both are multiple orders of magnitude higher than codon fitnesses (18.6× and 31.7× higher in chordates and arthropods, respectively)

with an average change of 0.027% (excluding the stop codons). In arthropods, this effect is even larger, with β_A affecting 31.7 times more compared to the effect of the fitness coefficients (0.9% vs 0.029%).

2.4 CpG/TpG sites in chordates affect codon usage patterns

CpG/TpG sites are known to affect codon usage patterns in humans (Scaiewicz et al., 2006). As we observed a strong relationship between the codon GC–content and their fitnesses, we aimed to test how much of it could be explained by CpG/TpG sites in chordates and arthropods. Methylated CpG dinucleotides are more likely to change into TpG dinucleotides compared to hypomethylated ones, therefore resulting in a depletion of CpG sites and a surplus of TpG (Miyahara et al., 2015; Simmen, 2008). We employed a one-sample Wilcoxon signed-rank test to test whether CpG (and TpG) codons are less (more) favourable than what is expected due to their GC–content. In chordates, we observed significant differences between the dinucleotides' coefficients and their expected GC content (p–value $< 2.2 \times 10^{-16}$), while arthropods show insignificant differences (fig. 5). After correcting for multiple testing (False Discovery Rate, FDR (Benjamini and Hochberg, 1995)), these results seem to not be driven by a single group, as significant p-values are recovered also for the chordate classes studied (apart from fish TpG surplus S2). In addition, all arthropod orders are insignificant (Supplementary table S2). Therefore, CpG codons seem to affect codon usage bias only in chordates, but not in arthropods.

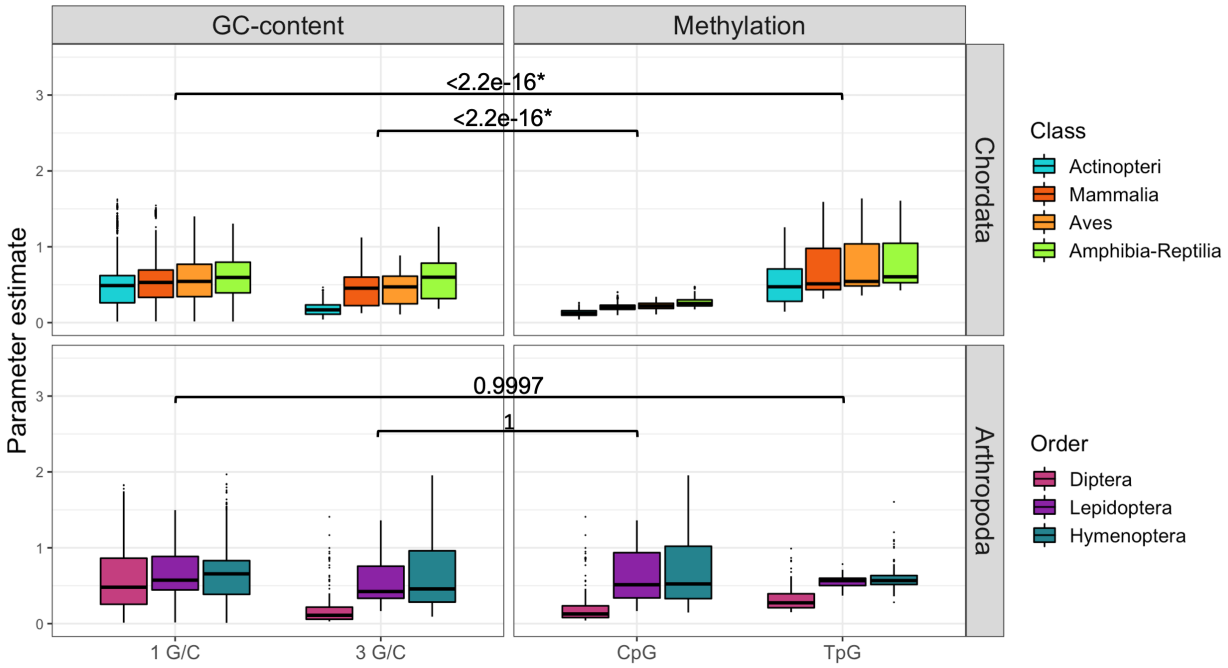

Figure 5: Differences in the fitness of codons with CpG sites compared to their expected GC content. CpG deficit and TpG surplus are expected due to methylation. The non-parametric Wilcoxon signed-rank test was used to compare CpG/TpG sites with their expected estimate due to GC content in chordates and arthropods. Significant results are shown with an , colours represent the different taxa.

3 Discussion

Our study introduces a mechanistic model of codon evolution that can distinguish between selection and mutational biases while accounting for GC–biased gene conversion. Previous research has introduced mechanistic models to estimate mutation and selection coefficients based on population dynamics (e.g., Zeng (2010); Sharp et al. (2005); Yang and Nielsen (2008)) or protein synthesis rates (Gilchrist et al., 2015). However, our model, DECUB, also takes into account the effects of gBGC. gBGC has been identified in vertebrates (Figuet et al., 2015) and other insect species (Kent et al., 2012; Wallberg et al., 2015), and there is increasing evidence that its impact is widespread across all metazoans (Galtier et al., 2018; Pessia et al., 2012). DECUB is modelled similarly to the FMutSel model introduced by Yang and Nielsen (2008). However, in addition to modelling gBGC, our model expands the parameter space to include fitness coefficients of the stop codons, which have not been accounted for in previous studies. This incorporation highlighted a higher fitness of the stop codon TGA in vertebrates, supporting previous studies on the effect of gBGC on the mammalian stop codons (Ho and Hurst, 2022).

While DECUB was utilised in this study to investigate phylum–specific patterns of codon usage, it has applications beyond this. It can be implemented on a more refined taxonomic level, such as species families or genera, or even on a per–species basis, enabling us to infer adaptations in codon usage specific to particular lineages. The latter application can be especially useful in taxa with variable

codon usage patterns. Additionally, DECUB has the potential to investigate sudden changes in the patterns of codon usage, such as those resulting from modifications in the genomic base composition or lineage–specific alterations to the genetic code.

Expanding previous research that studied a small number of representatives per phylum, DECUB was applied here to a vast dataset of over 600 species of chordates and arthropods. We found that codon usage is more extensive in chordates compared to arthropods (fig. 2). Behura and Severson (2012) showed that dipteran and hymenopteran insects have a low extent of codon usage, which is in agreement with our results. Additionally, Behura and Severson (2012) also suggested that insects have a pattern of codon usage that is unique to each species, explaining the variation we also observed within arthropods.

Although the extent of codon usage bias varies between chordates and arthropods, we showed that mutational biases have a more significant impact on shaping genome–wide patterns of codon usage than codon fitnesses (fig. 5). Corroborating these results, also in yeast the global codon patterns appear to be dominated by a combination of mutational biases towards AT and gBGC (LaBella et al., 2019). Mutations have been described previously as the main driver of codon usage in vertebrates (e.g., Doherty and McInerney (2013)), however, their effect on arthropod patterns is less extensive compared to natural selection (Kliman and Hey, 1994).

Our analysis revealed a strong negative correlation between the GC mutational biases – which also include the effects of gBGC – and selection forces acting on the GC–content of each codon (fig. 3). Indeed, despite an excess of GC to AT mutations, GC alleles are more likely to be fixed in mammals due to gBGC or GC–biased selection (Behura and Severson, 2013; Smith and Eyre-Walker, 2001). Birds are known to exhibit similar patterns of gBGC to mammals (Duret and Galtier, 2009; Figuet et al., 2015). However, there is also growing evidence that this phenomenon affects fish (Escobar et al., 2011), whereas its impact on reptiles appears to be less significant (Figuet et al., 2015). These studies support our estimations of mutational biases driven by mutations and gBGC, where reptiles and amphibians exhibited smaller biases towards GC compared to the other chordate classes. In insects, Vicario et al. (2007) found that most genes in *Drosophila* are under GC–biased selection, while Behura and Severson (2012) reported that GC–biases affect codon usage in Diptera, whereas AT–biases are more pronounced in Hymenoptera. This observation corroborates our analysis on arthropods, where mutational biases in Diptera were GC–biased, whereas Hymenoptera was the only taxon with increased AT–bias favouring codons with a higher GC–content.

Finally, our results showed a depletion of CpG and an excess of TpG codons compared to the expected GC–content in chordates but not in arthropods (fig. 5). Methylation in CpG dinucleotides has been shown to hypermutate into TpG through deamination of their cytosine, resulting in a depletion of CpG sites and a surplus of TpG (Miyahara et al., 2015; Simmen, 2008). Most CpG sites are methylated in vertebrates (Bird, 2002), however among arthropods, dipterans exhibit minimal to absent levels of methylation, while other holometabola species (i.e. lepidopterans and hymenopterans) show methylation, but in reduced levels in their protein-coding sequences (Provataris et al., 2018). In contrast,

Jabbari and Bernardi (2004) have suggested that differences in GC–content and not methylation may cause CpG shortages. However, our analysis actively compared CpG codons with their expected GC content and found significant differences in chordates, which showcases that methylation in CpG sites rather than GC–content is driving these differences between the two phyla.

In summary, despite differences between chordates and arthropods, in both, mutational biases have a significant impact on shaping genome–wide patterns of codon usage. In all taxa, as GC–mutational biases increase, GC–rich codons become less favourable and vice versa for AT–biased mutations. These contrasting patterns are highlighted in fishes where a strong GC–mutation bias has the most deleterious effect on GC–rich codons, while in some hymenopterans the opposite pattern is observed (fig. 3). This inverse relationship between the mutations towards GC and the fitness coefficients is not merely an artefact of the model (fig. 3), while it also cannot be explained by methylation in arthropods (fig. 5).

A possible explanation might be that stabilising selection is preventing an excess of GC or AT content in the coding regions of the genome, which would limit the occurrence of mRNAs that are either too GC- or AU-rich. Indeed, data from humans, chicken, and *Drosophila* show that GC–content in mRNAs ranges between 30%–70% (Courel et al., 2019). However, we must exercise caution when considering this hypothesis since, apart from methylation in CpG sites, we cannot differentiate selection for translational efficiency in our fitness coefficients from other mutational biases that cannot be captured by our model. Testing this hypothesis may require comparing the mutation rates between GC and AT alleles, gBGC, and codon substitutions rates across the chordate or arthropods phylogeny to determine whether these have co–evolved to stabilise GC content on the coding regions. However, this validation presents challenges and it necessitates polymorphic data as well as experimentally obtained mutation and gBGC rates, which are currently unavailable for most non–model organisms included in our analysis.

This study betters our understanding of the molecular mechanisms involved in the determination of codon composition in animals, the extent of which seems to vary considerably. It also provides insights on the variations in synonymous sites in light of these mechanisms. It is well known that failure to account for this variation breaks the assumption of neutral evolution of synonymous sites and can bias estimates of the ratio of nonsynonymous to synonymous substitution rates, i.e. $\omega = dN/dS$ (Spielman and Wilke, 2015; Goldman and Yang, 1994; Muse and Gaut, 1994), a parameter commonly used to detect natural selection acting on the protein. Therefore, our results highlight the importance for clade–specific approaches in the study of variation at synonymous sites and the detection of natural selection.

4 Methods

4.1 A model of codon evolution

To assess the evolutionary impact of the forces that govern codon usage bias, we devised a population genetic model on the 64 codons using a Moran model with reversible mutations and selection (sup-

plementary text S1). Mutations are reversible, biased, and modelled as the general time–reversible (GTR) substitution model (Tavaré, 1986), where the mutation rates are proportional to the stationary frequencies π between the four nucleotides. GC–biased gene conversion is incorporated as a selection coefficient γ favouring GC–alleles (Nagylaki, 1983). Genetic drift is modelled according to the Moran model (Moran, 1958) in a population of N individuals, where in each generation one individual is chosen to reproduce and one to die. Finally, selection acting on codons is modelled as a relative fitness coefficient ϕ. We further derived the stationary distribution of this model, which defines the frequencies of each of the 64 codons in terms of the aforementioned forces:

$$\psi_I \propto \pi_{i_1}\pi_{i_2}\pi_{i_3}\gamma^{(N-1)(\mathbf{1}_{G,C}(i_1)+\mathbf{1}_{G,C}(i_2)+\mathbf{1}_{G,C}(i_3))}\phi_I^{N-1} = \frac{1}{K}\beta_{i_1}\beta_{i_2}\beta_{i_3}\,\Phi_I = \frac{1}{K}\beta_I\,\Phi_I \ , \tag{2}$$

where $i_1 i_2 i_3$ are the three nucleotides of codon I and $\mathbf{1}_{G,C}(i)$ is the indicator function of nucleotide i. The indicator function guarantees that the γ parameter contributes to a given codon frequency only if it has a GC nucleotide, thus modelling for gBGC.

Using the stationary distribution of eq. (2), we developed DECUB, a Bayesian estimator that estimates the effects of different evolutionary forces on codon usage bias from codon counts. As the effects of mutation and gBGC are both mutational processes and we cannot disentangle them, we combined them into a single parameter β, representing mutational biases. Finally, the normalisation factor K is the sum of the stationary frequencies of all codons, $K = \sum_I \beta_I\,\Phi_I$, such that the stationary distribution adds up to 1.

We normalised each β by π_A, the stationary frequency of Adenine, which does not include the effects of gBGC. Similarly, the fitness coefficient $\Phi_I = \phi_I^{N-1}$ encompasses the effect of selection with genetic drift and is normalised by Φ_{ATG}, the fitness coefficient of methionine, as methionine is an essential amino acid and is encoded by a single codon, therefore it is not confounded with codon usage bias. In summary,

$$\beta_T = \frac{\pi_T}{\pi_A}, \quad \beta_{C/G} = \frac{\pi_{C/G}}{\pi_A}\gamma^{N-1} \quad \text{and} \quad \Phi_I = \left(\frac{\phi_I}{\phi_{\text{ATG}}}\right)^{N-1} \ . \tag{3}$$

4.2 Mapping

Mapping is the process of assigning a coefficient to be estimated by the model to a single codon or a group of codons. This can refer to an amino acid mapping, where every coefficient is placed per amino acid and assumes no variation within synonymous codons, or a more codon–usage–specific mapping. We started with an amino acid mapping that includes 23 fitness categories, one for each amino acid and one for each stop codon. Then, for each codon, we calculated the number of species where the predicted frequency is outside the 0.05-0.95 interval of the empirical codon frequency. In a step–wise manner, we added an extra fitness coefficient per amino acid to the codon with the highest error. We continued this process until all codons had an error smaller than 20%, which means that we are predicting its frequency correctly for 80% of the species. Using this procedure, we defined a phylum–specific codon mapping for chordates and arthropods, which we incorporated in DECUB, along with the amino acid

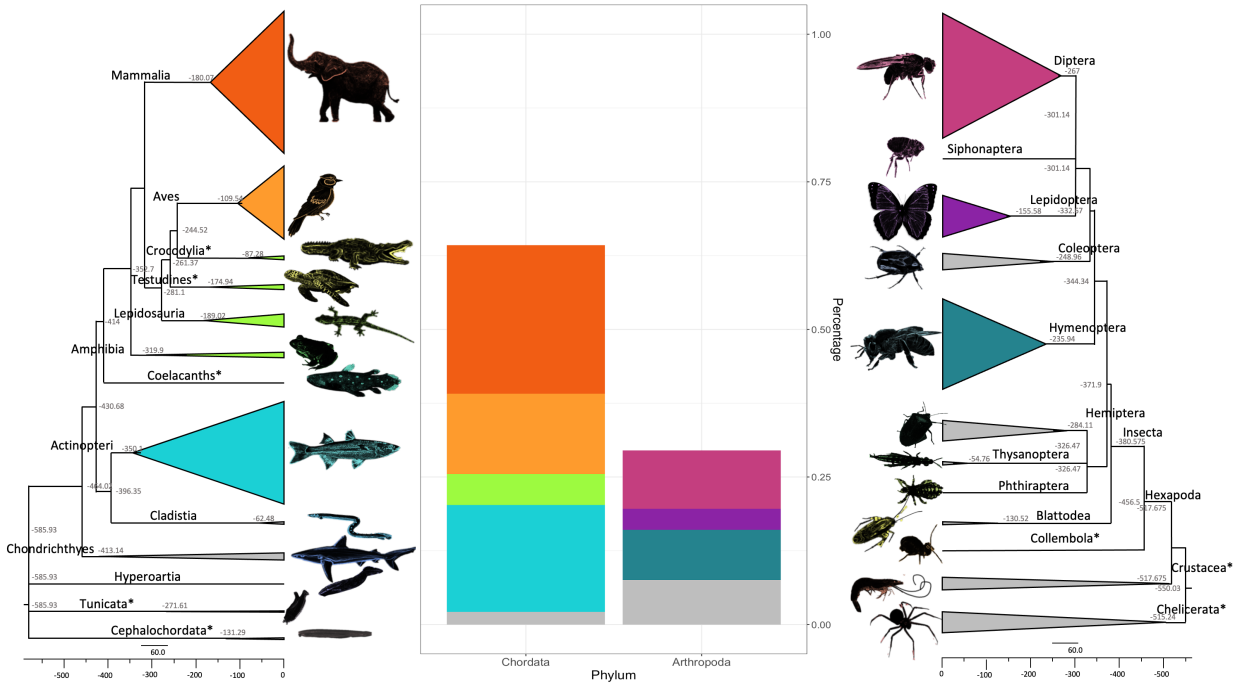

Figure 6: Phylogenetic reconstruction and relative counts of species in the dataset, belonging to Chordata and Arthropoda. The phylogenetic tree on the right represents the 415 chordates species spanning over 13 taxa, while the one on the left represents the arthropods, which summarises 191 species belonging to 12 taxa. For chordates, the taxa with an asterisk next to their names represent a class that was named conventionally by a more commonly known name (e.g. giving the name of the subphyla Tunicata and Cephalochordata to the correspondent class name), or the specific Class has not an official name yet on NCBI (Schoch et al., 2020), therefore was named by its most represented Order (e.g. Crocodylia, Testudines, Coelacanths). In arthropods, we recovered 10 main orders belonging to hexapods (note Collembola* is named here after the most recognised class name), and also summarised the orders found in the subphyla Chelicerata (spiders and scorpions) and Crustacea (crustaceans) which are shown on the phylogeny as a group named after their subphylum followed by an asterisk. The evolutionary time–scales were recovered using TimeTree (Kumar et al., 2022) and visualised with FigTree (Rambaut, 2018), with both trees sharing similar timescales (approximately diverged around 550 million years ago). Finally, the middle graph shows the proportions of the selected studied taxa.

one.

4.3 Dataset information

Codon counts were collected from the Codon Statistics Database (Subramanian et al., 2022) for a total of 606 species, 415 Chordata and 191 Arthropoda. All species are encoded using the standard genetic code (translation table 1 (Osawa et al., 1992)). The dataset includes counts for all 64 codons, including three stop codons along with the sense ones.

The dataset for chordates contains species from 13 taxa, most of those being shown at the Class level, with most subsequent analyses focusing on a subset of these classes, namely Mammals, Birds,

Fish, and all Reptiles and Amphibians combined, which represent almost 97% of all chordates dataset. Similarly, for arthropods, we have 12 taxa of species, the orders of which diverged at similar time points to chordates' classes. Here, we focused on the orders Diptera (flies and mosquitoes), Lepidoptera (butterflies and moths), and Hymenoptera (ants, bees, and wasps), which represent the majority of species in Hexapoda and comprise 74% of the arthropods dataset.

4.4 Entropy and sensitivity analysis

To measure the evolutionary significance of the mutation and selection bias in the codon frequencies, we calculated the impact of perturbing each of those parameters by $\pm$ 10% on the predicted frequencies. To summarise this effect, we used the Shannon entropy (Shannon, 1948):

$$H(\psi) = -\sum_{I=1}^{64} \psi_I \log(\psi_I), \tag{4}$$

where ψ_i is the predicted frequency of codon I. After obtaining the entropy we calculated the relative difference as:

$$RD = \frac{1}{H(\psi)} |H(\psi^+) - H(\psi^-)| \ , \tag{5}$$

where ψ^+ and ψ^- are the recalculated stationary distribution when we increased and decreased each parameter by 10%, respectively.

4.5 Computer statement, Software and Availability

All analyses were performed in python 3.9.13 (Van Rossum and Drake, 2009), R 4.2.0 (R Core Team, 2022) and C++11 (Josuttis, 2012) with a g++ compiler. DECUB is a free and open source software written in c++, available on GitHub along with a tutorial on how to use it. The link to the software can be found here: https://github.com/JaneK27/decub.

Acknowledgements

We thank all members of the Institute of Population Genetics for their feedback and support. This research was funded by the Austrian Science Fund (FWF) [P34524-B] awarded to Rui Borges.

References

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

S. K. Behura and D. W. Severson. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. *PLoS One*, 2012.

S. K. Behura and D. W. Severson. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biological Reviews*, 88(1):49–61, 2013.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

N. Bierne and A. Eyre-Walker. Variation in synonymous codon use and DNA polymorphism within the *Drosophila* genome. *Journal of Evolutionary Biology*, 19(1):1–11, 2006.

A. Bird. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21, 2002.

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

M. Courel, Y. Clément, C. Bossevain, D. Foretek, O. Vidal Cruchez, Z. Yi, M. Bénard, M.-N. Benassy, M. Kress, C. Vindry, et al. GC content shapes mRNA storage and decay in human cells. *eLife*, 8: e49708, 2019.

A. Doherty and J. O. McInerney. Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Molecular Biology and Evolution*, 30(10):2263–2267, 2013.

L. Duret. Evolution of synonymous codon usage in metazoans. *Current opinion in genetics & development*, 12(6):640–649, 12 2002.

L. Duret and N. Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10(1):285–311, 2009.

L. Duret and D. Mouchiroud. 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, 1999.

J. S. Escobar, S. Glémin, and N. Galtier. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. *Molecular Biology and Evolution*, 28(9):2561–2575, 2011.

E. Figuet, M. Ballenghien, J. Romiguier, and N. Galtier. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biology and Evolution*, 7(1): 240–250, 2015.

N. Galtier, C. Roux, M. Rousselle, J. Romiguier, E. Figuet, S. Glémin, N. Bierne, and L. Duret. 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, 01 2018.

M. Gerdol, G. De Moro, P. Venier, and A. Pallavicini. Analysis of synonymous codon usage patterns in sixty-four different bivalve species. *PeerJ*, 3:e1520, 2015.

M. A. Gilchrist, W.-C. Chen, P. Shah, C. L. Landerer, and R. Zaretzki. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. *Genome Biology and Evolution*, 7(6):1559–1579, 2015.

N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736, 9 1994.

R. Hershberg and D. A. Petrov. Selection on codon bias. *Annual Review of Genetics*, 42:287–299, 2008.

A. T. Ho and L. D. Hurst. Unusual mammalian usage of TGA stop codons reveals that sequence conservation need not imply purifying selection. *PLoS Biology*, 20(5):e3001588, 2022.

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

K. Jabbari and G. Bernardi. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333:143–149, 2004.

N. M. Josuttis. *The C++ standard library*. Addison-Wesley Educational, Boston, MA, 2 edition, Mar. 2012.

C. F. Kent, S. Minaei, B. A. Harpur, and A. Zayed. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proceedings of the National Academy of Sciences*, 109(44):18012–18017, 2012.

R. M. Kliman and J. Hey. The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics*, 137(4):1049–1056, 1994.

S. Kumar, M. Suleski, J. M. Craig, A. E. Kasprowicz, M. Sanderford, M. Li, G. Stecher, and S. B. Hedges. Timetree 5: An expanded resource for species divergence times. *Molecular Biology and Evolution*, 39(8):msac174, 2022.

A. L. LaBella, D. A. Opulente, J. L. Steenwyk, C. T. Hittinger, and A. Rokas. Variation and selection on codon usage bias across an entire subphylum. *PLoS Genetics*, 15(7):e1008304, 2019.

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

H. Miyahara, O. Hirose, K. Satou, and Y. Yamada. Factors to preserve CpG-rich sequences in methylated CpG islands. *BMC Genomics*, 16(1):1–15, 2015.

P. A. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958.

S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 9 1994.

T. Nagylaki. Evolution of a large population under gene conversion. *Proceedings of the National Academy of Sciences*, 80(19):5941–5945, 1983.

S. Osawa, T. H. Jukes, K. Watanabe, and A. Muto. Recent evidence for evolution of the genetic code. *Microbiological Reviews*, 56(1):229–264, 1992.

E. Pessia, A. Popa, S. Mousset, C. Rezvoy, L. Duret, and G. A. Marais. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution*, 4(7):675–682, 2012.

P. Provataris, K. Meusemann, O. Niehuis, S. Grath, and B. Misof. Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome Biology and Evolution*, 10(4):1185–1197, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

A. Rambaut. Figtree version 1.4. 4, 2018.

Y. Rao, G. Wu, Z. Wang, X. Chai, Q. Nie, and X. Zhang. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA Research*, 18(6):499–512, 2011.

V. Scaiewicz, V. Sabbía, R. Piovani, and H. Musto. CpG islands are the second main factor shaping codon usage in human genes. *Biochemical and Biophysical Research Communications*, 343(4):1257–1261, 2006.

C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020, 2020.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

P. M. Sharp and W.-H. Li. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3):1281–1295, 1987.

P. M. Sharp, M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. DNA sequence evolution: the sounds of silence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 349(1329):241–247, 1995.

P. M. Sharp, E. Bailes, R. J. Grocock, J. F. Peden, and R. E. Sockett. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research*, 33(4):1141–1153, 01 2005.

D. C. Shields, P. M. Sharp, D. G. Higgins, and F. Wright. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution*, 5(6): 704–716, 1988.

M. W. Simmen. Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. *Genomics*, 92(1):33–40, 2008.

N. G. Smith and A. Eyre-Walker. Synonymous codon bias is not caused by mutation bias in G + C-rich genes in humans. *Molecular Biology and Evolution*, 18(6):982–986, 2001.

S. J. Spielman and C. O. Wilke. The relationship between dN/dS and scaled selection coefficients. *Molecular Biology and Evolution*, 32(4):1097–1108, 2015.

K. Subramanian, B. Payne, F. Feyertag, and D. Alvarez-Ponce. The Codon Statistics Database: A database of codon usage bias. *Molecular Biology and Evolution*, 39(8):msac157, 2022.

S. Subramanian. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics*, 178(4):2429–2432, 2008.

N. Sueoka and Y. Kawanishi. DNA G + C content of the third codon position and codon usage biases of human genes. *Gene*, 261(1):53–62, 2000.

H. Suzuki, R. Saito, and M. Tomita. The 'weighted sum of relative entropy': a new index for synonymous codon usage bias. *Gene*, 335(1-2):19–23, 6 2004.

S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Some mathematical questions in biology / DNA sequence analysis edited by Robert M. Miura*, 1986.

A. O. Urrutia and L. D. Hurst. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, 159(3):1191–1199, 2001.

G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

S. Vicario, E. N. Moriyama, and J. R. Powell. Codon usage in twelve species of *Drosophila*. *BMC Evolutionary Biology*, 7(1):1–17, 2007.

A. Wallberg, S. Glémin, and M. T. Webster. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genetics*, 11(4):e1005189, 2015.

Z. Yang and R. Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, 25(3):568–579, 2008.

K. Zeng. A simple multiallele model and its application to identifying preferred–unpreferred codons using polymorphism data. *Molecular Biology and Evolution*, 27(6):1327–1337, 6 2010.