G3, 2021, 11(8), jkab191
DOI: 10.1093/g3journal/jkab191
Advance Access Publication Date: 2 June 2021
Investigation

Codon usage bias and dinucleotide preference in 29 *Drosophila* species

Prajakta P. Kokate , Stephen M. Techtmann , and Thomas Werner *

Department of Biological Sciences, Michigan Technological University, Houghton, MI 49931, USA

*Corresponding author: Email: twerner@mtu.edu

Abstract

Codon usage bias, where certain codons are used more frequently than their synonymous counterparts, is an interesting phenomenon influenced by three evolutionary forces: mutation, selection, and genetic drift. To better understand how these evolutionary forces affect codon usage bias, an extensive study to detect how codon usage patterns change across species is required. This study investigated 668 single-copy orthologous genes independently in 29 *Drosophila* species to determine how the codon usage patterns change with phylogenetic distance. We found a strong correlation between phylogenetic distance and codon usage bias and observed striking differences in codon preferences between the two subgenera *Drosophila* and Sophophora. As compared to the subgenus Sophophora, species of the subgenus *Drosophila* showed reduced codon usage bias and a reduced preference specifically for codons ending with C, except for codons with G in the second position. We found that codon usage patterns in all species were influenced by the nucleotides in the codon's 2nd and 3rd positions rather than the biochemical properties of the amino acids encoded. We detected a concordance between preferred codons and preferred dinucleotides (at positions 2 and 3 of codons). Furthermore, we observed an association between speciation, codon preferences, and dinucleotide preferences. Our study provides the foundation to understand how selection acts on dinucleotides to influence codon usage bias.

Keywords: codon usage bias; *Drosophila*; evolution; dinucleotide preference; synonymous codons

Introduction

Most amino acids are encoded by more than one codon due to the degeneracy of the genetic code (Lamolle *et al.* 2019). However, the synonymous codons for a particular amino acid are not necessarily used with equal frequency. This phenomenon, where specific codons are used more often than other synonymous codons, is called codon usage bias (CUB) (Heger and Ponting 2007). Currently, the widely accepted hypothesis proposes that CUB manifests due to the combined effects of three evolutionary forces: mutation, selection, and genetic drift (Guan *et al.* 2018).

The biological implications of CUB are well established (Quax *et al.* 2015), and the selective pressures acting on it are multifold. The codon usage pattern is known to influence mRNA folding, the translation elongation rate, and protein folding, thereby affecting gene expression (Quax *et al.* 2015). In prokaryotes, codons and codon pairs that resemble canonical Shine-Dalgarno sequences are avoided to prevent excessive ribosome pausing during translation (Shabalina *et al.* 2013). In another example, a synonymous change in the human *IRGM* gene alters the binding site for miR-196, causing tissue-specific dysregulation and predisposition to Crohn's disease (Brest *et al.* 2011). Although there are numerous examples of how selection may be acting on CUB, there are very few reports analyzing how codon usage patterns vary across species during evolution (LaBella *et al.* 2019; Lamolle *et al.* 2019).

Multiple reports mention that the codon usage patterns differ between different species (Sharp and Li 1986; Hershberg and Petrov 2008; Plotkin and Kudla 2011). However, a study analyzing CUB in 12 *Drosophila* species yielded contradicting results. Vicario *et al.* (2007) studied nine species from the subgenus Sophophora and three from the subgenus *Drosophila*, including one from the Hawaiian *Drosophila* radiation. The authors evaluated CUB using three methods; one specifically worth mentioning was the relative synonymous codon usage (RSCU) for the 10% highly biased genes based on their effective number of codons (ENC). Generally, the report found that the preferred set of codons was constant across the genus *Drosophila* in 11 of the 12 species studied. The only species that showed a different CUB was *D. willistoni*. Also, only serine showed a change in codon preference between species. The authors did not find any striking differences in the codon usage patterns, even between the two subgenera. Five years later, another group (Behura and Severson 2012) studied CUB in 22 insect genomes, 15 dipteran species (including the 12 *Drosophila* species reported previously), and seven hymenopteran species. They found differences in codon preferences between the two orders Diptera and Hymenoptera, as well as among the 12 *Drosophila* species. These contradicting reports warrant an extensive study in species within the genus *Drosophila* to test the hypothesis of whether each species prefers a different set of codons (Sharp and Li 1986).

Received: March 28, 2021. **Accepted:** May 13, 2021

The post-genomic era, along with improved computational methods, is an appropriate time to extend the study of CUB in the genus *Drosophila*. The NCBI genome database currently contains whole genomes, coding sequences, and translated coding sequences for 29 *Drosophila* species, providing a better representation of the subgenera Sophophora and *Drosophila* across the phylogenetic tree. In this study, we performed a CUB analysis in these 29 species to answer the following questions: (1) Is there a difference in CUB within the genus *Drosophila*? (2) How well does phylogenetic distance correlate with CUB? (3) What specific differences in CUB can be seen among closely versus distantly related species? (4) Does CUB depend on the biochemical properties of the amino acids encoded or the nucleotides at the dinucleotide$_{23}$ position of the codon? and (5) Is there a connection between codon preference, dinucleotide$_{23}$ preference, and speciation?

Here, we show that the species of the genus *Drosophila* show differences in CUB and that differences in CUB are strongly correlated with phylogenetic distance. We propose that nucleotides at the dinucleotide$_{23}$ position of the codon may influence CUB and establish, for the first time, an association between codon preference, dinucleotide$_{23}$ preference, and speciation in 29 *Drosophila* species.

Materials and methods

Data acquisition and ortholog identification

The genomes, coding sequences, and translated coding sequences of 29 *Drosophila* species (Supplementary Table S1) were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov). Whole-genome sequences and coding sequences were used to identify the GC content for each species. The latest version of OrthoFinder (Emms and Kelly 2019) that uses Diamond (Buchfink *et al.* 2015) to identify sequence similarities was applied to translated coding sequences of the 29 species to identify orthologous proteins. OrthoFinder identified 668 single-copy orthologous (SCO) genes present in all 29 species, which were selected for further analysis. The number of SCO genes is small compared to ~14,000 genes present in the *Drosophila* genome (Alberts *et al.* 2002). Also, SCO genes identified across multiple species have been reported to be highly conserved, duplication-resistant, and may be involved in essential metabolic processes (Han *et al.* 2014). The authors note that analyzing 668 highly conserved SCO genes may not necessarily reflect the entire genome of the *Drosophila* species as potential biases may be introduced based on the extent of conservation and the functional roles of these genes. However, studying SCO genes prevents the confounding effects of variations in gene length and expression levels that are known to influence CUB (Novoa *et al.* 2019). The authors also acknowledge that genes in the same genome showing differential expression between tissues may have different codon usage patterns (Payne and Alvarez-Ponce 2019) reaffirming the use of SCO genes when comparing CUB among multiple species.

Phylogenetic tree

OrthoFinder generated a phylogenetic tree for the 29 *Drosophila* species, based on gene trees inferred from 18,789 orthogroups, using the Markov Cluster Algorithm (Enright *et al.* 2002). The ape package (Paradis *et al.* 2004) was used to simulate the phylogenetic tree from the OrthoFinder results (Figure 1). This phylogenetic tree was used for Phylogenetic Generalized Least Squares (PGLS) regression analysis and calculating the phylogenetic signal.

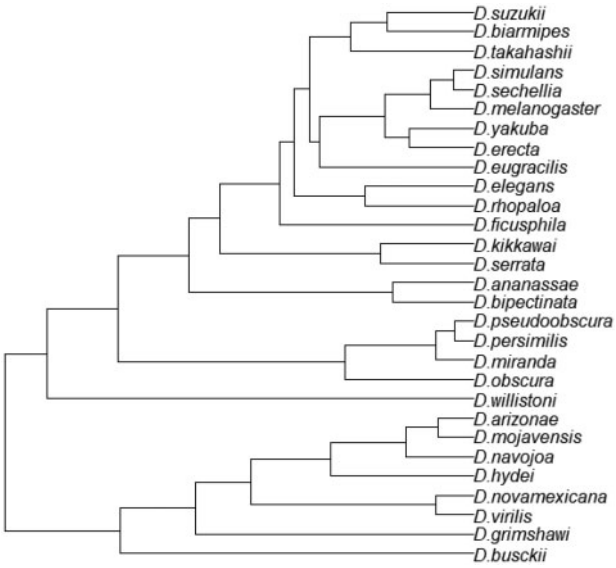

Figure 1 Phylogenetic tree of 29 *Drosophila* species generated using the results from Orthofinder and ape package in R.

tRNA abundance

To compare CUB with tRNA abundance, we estimated the number of tRNA genes in the genomes of each species using ARAGORN (Laslett and Canback 2004). It was observed that all species lacked certain tRNA genes (Supplementary Table S2), and multiple alloacceptor tRNA genes made the correlation between the tRNA gene abundance and CUB complicated (Sahyoun *et al.* 2015).

Codon usage analysis

Codon usage analysis was performed on the orthologous genes, using two R packages: seqinR (Charif and Lobry 2007) and coRdon (Elek *et al.* 2020). The seqinR package was used to estimate the overall GC content of the coding sequences, the SCO genes, and GC content at the third codon position (GC3). The ENC (Wright 1990) was calculated for all coding sequences and separately for 668 SCO genes, using the coRdon package. The ENC is a non-directional measure of CUB (Subramanian and Sarkar 2015), and its values can range from 20 (high CUB) to 61 (no CUB).

The seqinR software was used to identify the RSCU values and the codon counts. A codon with an RSCU value of >1.0 would indicate a preference for that codon. The codon counts were used to calculate the sENC-X value (Powell and Moriyama 1997) as a measure of the contribution of each amino acid toward CUB. This value provides the ENC for each amino acid, scaled from 0 to 1, for all amino acids, irrespective of the extent of redundancy, making the comparison between amino acids credible. A low sENC-X value would indicate a high CUB and *vice versa*.

Dinucleotide representation analysis

Dinucleotide representation analysis at the 2nd and 3rd nucleotide position of the codon (dinucleotide$_{23}$) was performed using the dinuq package (Lytras and Hughes 2020) in Python 3 (Van Rossum and Drake 2009). Because nucleotide changes at the 1st and 2nd nucleotide positions of the codons alter the amino acids encoded in most cases, they were not analyzed. The synonymous dinucleotide usage (SDU) values were calculated to identify whether certain dinucleotides were over-or underrepresented in the orthologous genes. An SDU value of 1 would indicate no CUB,

greater than 1 would indicate overrepresentation, and between 0 and 1 would indicate underrepresentation.

The mean RSCU, mean sENC-X, and mean SDU values of the 668 SCO genes were plotted using ggplot2 (Wickham 2016). The vhcub package (Anwar *et al.* 2019) was used to plot the ENC values versus the GC3 values of the SCO genes. Hierarchical clustering was performed using the Heatmaply package (Galili *et al.* 2018) on the mean RSCU values to generate a heatmap revealing codon usage patterns across the 29 *Drosophila* species.

Statistical analyses

The Mann–Whitney test was used to evaluate whether the mean ENC values and the mean RSCU values between the two subgenera *Drosophila* and Sophophora showed a statistically significant difference. The R packages ape (Paradis *et al.* 2004) and nlme (Pinheiro *et al.* 2021) were used for the PGLS regression analysis to study the correlation between the GC content and ENC values of the SCO genes. We calculated the Pagel's λ (Pagel 1999) to assess the degree of phylogenetic signal in codon preferences and dinucleotide$_{23}$ preferences, using the package phytools version 0.7-70 (Revell 2012).

Data availability

All code used for data processing, generating figures, and statistical analyses is available through GitHub (https://github.com/pkokate18/CUB/blob/main/data_processing). Relevant data required to execute the code are available in the supplemental material. Supplementary material is available at figshare: https://doi.org/10.25387/g3.14331020.

Results

The GC content positively correlates with CUB

Because the GC content is known to influence CUB (Behura and Severson 2012; Novoa *et al.* 2019), we calculated the GC contents of whole genomes, coding sequences, and SCO genes in 29 *Drosophila* species. The whole-genome GC contents ranged from 32 to 45%, with *D. grimshawi* showing the lowest and *D. persimilis* the highest GC contents (Table 1). The GC contents of the coding sequences were generally higher than those of the whole-genome sequences, ranging from 47 to 56%, which corroborated a previous report (Lamolle *et al.* 2019). The coding sequences of *D. willistoni* had the lowest GC content of 46.64%. We used PGLS to examine the relationship between GC content and ENC while correcting for phylogenetic signal. The PGLS regression analysis, which accounts for phylogenetic nonindependence between species, showed a negative correlation between the GC content and the ENC values of coding sequences as well as the SCO genes (Table 2), indicating that the GC content positively correlates with CUB.

Codon usage analysis

The extent of CUB differs between subgenera

To quantify the CUB in all 29 species, we used ENC (Wright 1990) values derived from coding sequences and 668 SCO genes (Table 1). When the mean ENC values of coding sequences were compared between the species of the subgenus *Drosophila* and the subgenus Sophophora, the difference was not statistically significant. However, the mean ENC values of the SCO genes were higher in the species of the subgenus *Drosophila*, as compared to those of the subgenus Sophophora, and this difference was statistically significant (Mann–Whitney test, $P < 0.01$). These results demonstrate that the subgenus *Drosophila* shows reduced CUB as compared to the subgenus Sophophora. Furthermore, these results establish the foundation of how the use of SCO genes may improve the evaluation and comparison of CUB among species.

Selection may play a substantial role in CUB

To assess the extent of influence of mutation bias and natural selection on CUB, we plotted the ENC values and the GC3 values of the SCO genes from each species using the vhcub package (Figure 2 and Supplementary Figure S1). The curve represents the null hypothesis that the bias at the synonymous position (GC3) is solely due to mutation. Genes plotted on or above the curve would suggest that mutation is the primary force acting on CUB, whereas genes with lower ENC values than the expected curve would indicate that natural selection substantially influences CUB (Ismail *et al.* 2019). As depicted in Figure 2 and Supplementary Figure S1, most of the SCO genes were below the curve, suggesting that selection may have a significant role in CUB observed in these genes.

Codon usage patterns may have changed with speciation

To identify differences in codon preference among the 29 *Drosophila* species, we analyzed and plotted the mean RSCU values (Sharp and Li 1986) for the 668 SCO genes (Figure 3 and Supplementary Figure S2). Table 3 presents the preferred codons in the two subgenera: *Drosophila* and Sophophora. We found that each subgenus preferred different codons for specific amino acids or showed a statistically significant difference in their preference for the same codons (Mann–Whitney test, $P < 0.001$, Table 3). Our data confirmed the results of our ENC analysis that between the two subgenera, species of the subgenus *Drosophila* generally showed a reduced CUB.

Correspondence analysis of the mean RSCU values of SCO genes provided further evidence of differences in codon preference between the two subgenera (Supplementary Figure S3). The first and second dimensions explained 75.9 and 16.9% of the variation between species, and the two subgenera showed noticeable segregation.

Differences in the codon preferences were evident even at the species group and subgroup levels (Figure 3, Supplementary Figure S2, and Table 3). For example, within the subgenus Sophophora, species from the obscura species group preferred AGC for serine, whereas species of the melanogaster species group preferred UCC slightly more than AGC. Species of the obscura species group also showed reduced CUB compared to species from the melanogaster species group for certain amino acids: histidine, phenylalanine, isoleucine, and threonine. Similarly, within the subgenus *Drosophila*, species from the virilis species group showed a slight preference for the codon CAU encoding the amino acid histidine. On the other hand, species from the repleta species group could be further divided into *D. hydei* from the hydei species subgroup that preferred CAU and species from the mulleri species subgroup that preferred CAC. A similar trend was seen for the amino acids phenylalanine and isoleucine. The observed differences in the codon usage pattern between subgenera, species groups, and species subgroups indicate a correlation between CUB and speciation. To confirm a correlation between CUB and speciation, we evaluated Pagel's λ (Pagel 1999) for mean RSCU values to assess the phylogenetic signal of codon preferences. The phylogenetic signal is a statistical approach to evaluate whether closely related species are more similar than species drawn randomly from the same tree (Blomberg and Garland 2002). Pagels's λ is a measure of the phylogenetic signal with values ranging from 0 to 1, where $\lambda = 0$

Table 1 Taxonomical classification, GC content, and ENC of 29 *Drosophila* species

Subgenus	Species group	Species subgroup	Species	Abbreviation	GC content (%)			ENC Mean (range)	
					Genomic	CDS	SCO genes	CDS	SCO genes
Drosophila			*D. busckii*	Dbus	38.56	50.39	50.04	50.18 (27.29–61)	50.51 (32.23–61)
Drosophila (Hawaiian)	grimshawi	grimshawi	*D. grimshawi*	Dgri	32.54	51.1	51.74	51.48 (26.26–61)	50.62 (33.65–61)
Drosophila	virilis		*D. virilis*	Dvir	40	52.23	52.92	50.2 (21.65–61)	49.19 (35.27–61)
Drosophila	virilis		*D. novamexicana*	Dnov	39.47	52.27	53.03	50.05 (21.45–61)	49.07 (33.82–61)
Drosophila	repleta	hydei	*D. hydei*	Dhyd	39.7	50.86	50.55	51.45 (19.35–61)	51.91 (33.93–61)
Drosophila	repleta	mulleri	*D. navojoa*	Dnav	33.14	52.85	52.37	49.58 (25.3–61)	49.83 (30.34–61)
Drosophila	repleta	mulleri	*D. mojavensis*	Dmoj	35.95	52.01	52.40	50.17 (24.94–61)	49.79 (31.41–61)
Drosophila	repleta	mulleri	*D. arizonae*	Dari	37.84	52.57	52.40	49.88 (23.63–61)	49.86 (30.48–61)
Sophophora	willistoni	willistoni	*D. willistoni*	Dwil	37.38	46.64	46.85	53.29 (25.68–61)	53.67 (33.65–61)
Sophophora	obscura	obscura	*D. obscura*	Dobs	40.03	54.74	55.81	48.66 (25.5–61)	46.44 (30.83–61)
Sophophora	obscura	pseudoobscura	*D. miranda*	Dmir	44.31	54.93	55.38	48.98 (23.86–61)	47.4 (29.39–61)
Sophophora	obscura	pseudoobscura	*D. persimilis*	Dper	45.15	54.79	55.39	48.84 (24.46–61)	47.4 (29.11–61)
Sophophora	obscura	pseudoobscura	*D. pseudoobscura*	Dpse	44.68	54.84	55.34	48.97 (23.86–61)	47.53 (29.54–61)
Sophophora	melanogaster	ananassae	*D. bipectinata*	Dbip	39.83	53.57	53.90	50.36 (24.14–61)	50.21 (27.9–61)
Sophophora	melanogaster	ananassae	*D. ananassae*	Dana	41.36	53.71	54.36	50.05 (23.05–61)	49.65 (27.19–61)
Sophophora	melanogaster	montium	*D. serrata*	Dser	37.08	53.68	55.07	50.6 (23.68–61)	48.62 (29.93–61)
Sophophora	melanogaster	montium	*D. kikkawai*	Dkik	42.11	54.68	56.51	48.92 (24.8–61)	46.13 (29.37–61)
Sophophora	melanogaster	ficusphila	*D. ficusphila*	Dfic	41.48	54.5	54.59	48.91 (22.52–61)	48.92 (25.86–61)
Sophophora	melanogaster	rhopaloa	*D. rhopaloa*	Drho	38.59	54.02	55.25	49.73 (24.58–61)	48.06 (30.27–61)
Sophophora	melanogaster	elegans	*D. elegans*	Dele	38.16	55.09	56.22	48.4 (25.06–61)	45.9 (29.41–61)
Sophophora	melanogaster	eugracilis	*D. eugracilis*	Deug	38.64	51.66	51.60	53.15 (25.05–61)	53.6 (31.23–61)
Sophophora	melanogaster	melanogaster	*D. erecta*	Dere	39.1	54.4	55.32	49.53 (22.79–61)	48.13 (26.3–61)
Sophophora	melanogaster	melanogaster	*D. yakuba*	Dyak	38.42	53.92	55.26	49.96 (24–61)	48.23 (27.86–61)
Sophophora	melanogaster	melanogaster	*D. melanogaster*	Dmel	38.19	53.89	54.73	50.13 (24.43–61)	48.93 (28.45–61)
Sophophora	melanogaster	melanogaster	*D. sechellia*	Dsec	39.93	53.82	55.16	50.17 (21.71–61)	48.19 (28.41–61)
Sophophora	melanogaster	melanogaster	*D. simulans*	Dsim	40.22	53.24	55.33	49.92 (21.02–61)	47.89 (27.8–61)
Sophophora	melanogaster	takahashii	*D. takahashii*	Dtak	39.35	54.79	57.13	47.95 (24.45–61)	44.01 (26.84–61)
Sophophora	melanogaster	suzukii	*D. biarmipes*	Dbia	38.74	56.18	57.91	46.68 (22.28–61)	43.34 (24.87–61)
Sophophora	melanogaster	suzukii	*D. suzukii*	Dsuz	34.72	53.83	55.80	49.49 (23.26–61)	46.76 (27.06–61)

indicates no phylogenetic signal (the trait has evolved independently of the phylogeny), whereas values close to 1 indicate a strong phylogenetic signal (similarities in the trait are proportional to the relatedness of the species (Molina-Venegas and Rodríguez 2017; James *et al.* 2020). The Pagel's λ for all but four codons (CCU, CUU, AGU, and AGA) was >0.95, with a $P < 0.05$ (Table 4), each coding for a sixfold degenerate amino acid, except CCU, which codes for proline. A strong phylogenetic signal for mean RSCU values of most codons indicates a strong correlation between codon preferences and speciation.

The nucleotides at the dinucleotide$_{23}$ position may influence CUB

The RSCU analysis was also useful to identify specific differences in CUB among the 29 *Drosophila* species. As mentioned previously, species of the subgenus *Drosophila* showed a reduced CUB. This difference was clearly evident for certain amino acids (histidine, tyrosine, phenylalanine, lysine, and isoleucine; Figure 3 and Supplementary Figure S2). On the contrary, species of the subgenus Sophophora showed a reduced CUB for aspartic acid and glycine (Figure 3B and Supplementary Figure S2C). In the case of amino acids having twofold degenerate codons with AC or AU in the dinucleotide$_{23}$ position, except for aspartic acid, species of the subgenus *Drosophila* either showed no preference (RSCU value < 1.0) or reduced preference (lower RSCU values) (Figure 3, A, C, and D and Table 3). When the two subgenera *Drosophila* and Sophophora were compared, species of the subgenus *Drosophila* showed reduced preference for codons ending with C, except for three amino acids. The codons CGC for arginine, AGC for serine, and GGC for glycine showed higher RSCU values in species from the subgenus *Drosophila* (Figure 4). It is noteworthy that, although species of the subgenus *Drosophila* usually show a reduced preference for C-ending codons, they make an exception when the codons have GC in their dinucleotide$_{23}$ position. These findings suggest that the nucleotides in the dinucleotide$_{23}$ position may have a substantial role to play in CUB.

Drosophila willistoni showed a distinct codon usage pattern. As seen in Figure 3, for all the twofold degenerate amino acids with codons ending with AU or AC in the dinucleotides$_{23}$ position, *D. willistoni* strongly preferred the AU-ending codons (Figure 3, A–D). *D. willistoni* showed either no preference or reduced preference for the other 5 twofold degenerate amino acids (cysteine, glycine, lysine, phenylalanine, and glutamine). Three of these amino acids have AG or AA in the dinucleotide$_{23}$ position (Figure 3, G–I).

Table 2 Correlation between GC content and ENC in coding sequences (CDS) and SCO genes

		Value	SE	t-value	P-value
CDS	Intercept	100.74	5.019	20.07	<0.0001
	GC content	−0.967	0.095	−10.15	<0.0001
SCO	Intercept	119.08	5.38	22.109	<0.0001
	GC content	−1.32	0.1	−13.088	<0.0001

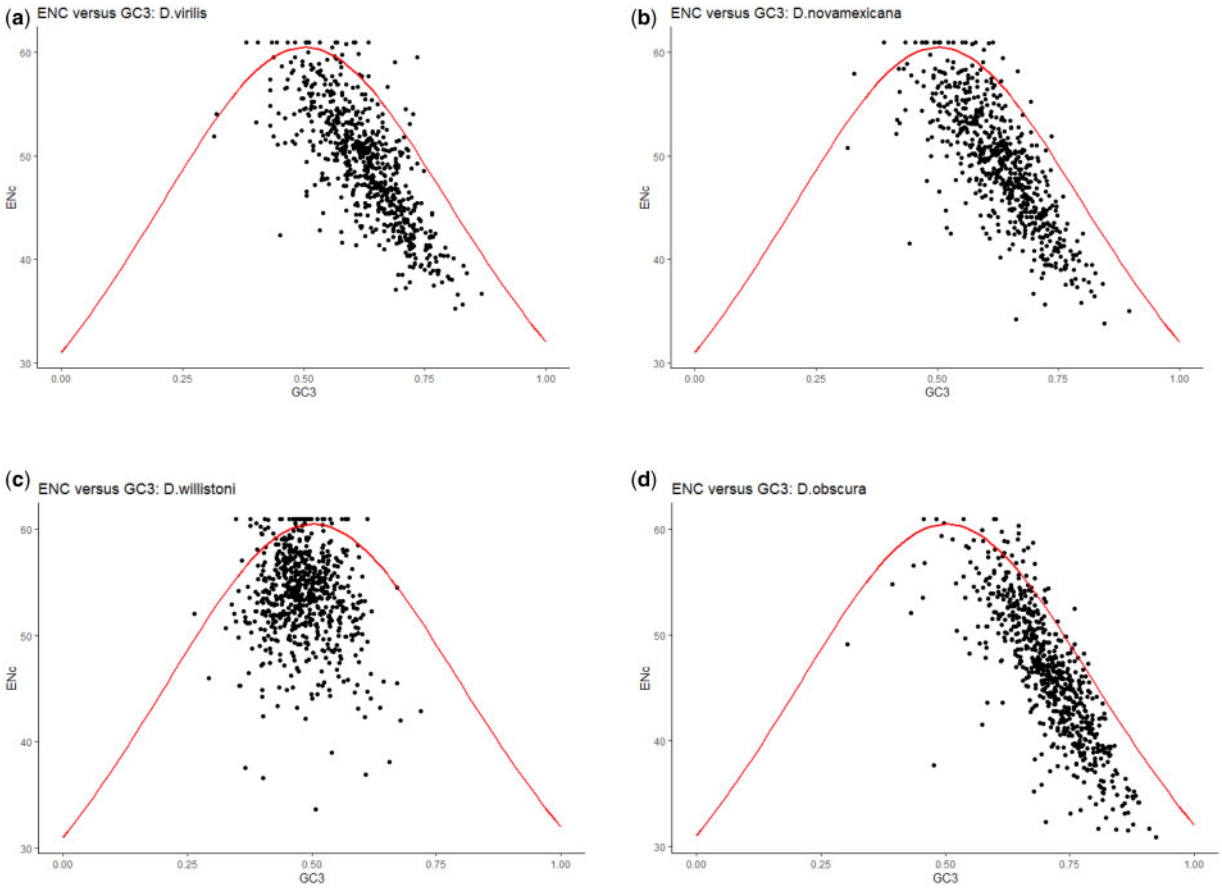

Figure 2 ENCGC3 plots of 668 SCO genes from 4 *Drosophila* species; (a) *D.* virilis, (b) *D.* novamexicana, (c) *D.* willistoni, and (d) *D.* obscura.

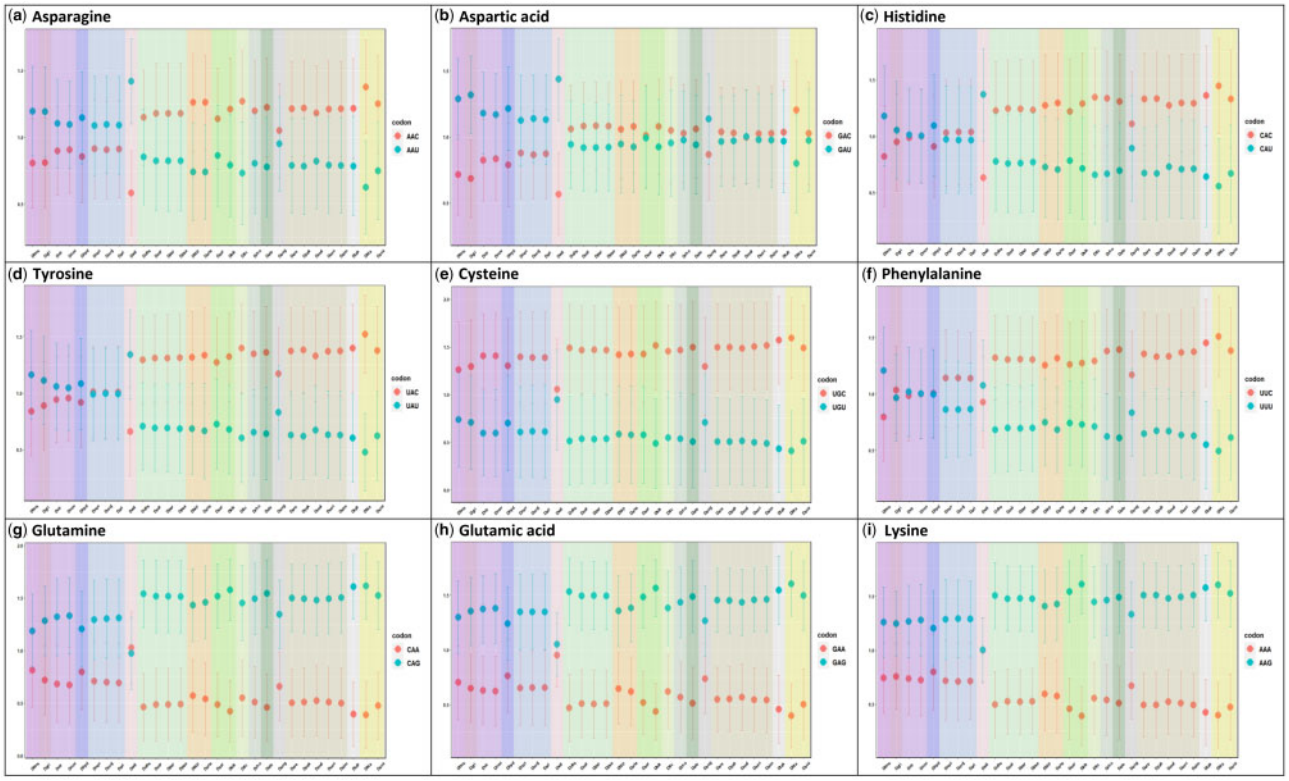

Figure 3 Plots showing mean RSCU values of synonymous codons for 2-fold degenerate amino acids in SCO genes (y axis) from 29 *Drosophila* species (x axis). The background shading indicates species from the same species sub-group. (a) Asparagine, (b) Aspartic acid, (c) Histidine, (d) Tyrosine, (e) Cysteine, (f) Phenylalanine, (g) Glutamine, (h) Glutamic acid, and (i) Lysine.

Table 3 Comparison of codon preference between subgenera *Drosophila* and Sophophora in SCO genes

Amino acid	Preferred codon				*Exceptions	*P-value* (Mann–Whitney test)
	Subgenus *Drosophila*		Subgenus *Sophophora*			
	Virilis species group	Repleta species group	Willistoni species group (only one species: *D. willistoni*)	Melanogaster and obscura species groups		
Asparagine	AAU	AAU	AAU	AAC		
Aspartic acid	GAU	GAU	GAU	GAC*	*D. eugracilis* prefers GAU	
Cysteine	UGC	UGC	UGC	UGC		6.76E-05
Glutamine	CAG	CAG	CAA*	CAG	*D. willistoni* shows very low CUB	2.25E-06
Glutamic acid	GAG	GAG	GAG*	GAG	*D. willistoni* shows very low CUB	0.0001486
Histidine	CAU	CAC*	CAU	CAC	*D. hydei* shows preference for CAU	
Lysine	AAG	AAG	no preference	AAG		2.25E-06
Phenylalanine	No preference*	UUC*	UUU	UUC	*D. virilis* shows slight preference for UUU; *D. hydei* shows no preference	
Tyrosine	UAU	no preference*	UAU	UAC	*D. hydei* shows slight preference for UAU	
Valine	GUG	GUG	GUG	GUG		0.002092
Alanine	GCC	GCC	GCC	GCC		2.25E-06
Glycine	GGC	GGC	GGC	GGC		9.01E-06
Proline	CCC	CCC*	No preference	CCC	*D. hydei* shows no preference	4.50E-06
Threonine	ACC/ACG	ACC/ACG*	ACA	ACC	*D. hydei* shows preference for ACA	
Arginine	CGC	CGC	CGU	CGC		
Leucine	CUG	CUG	UUG	CUG		
Serine	AGC	AGC	AGC/AGU/UCC	UCC*	obscura subgroup shows preference for AGC	
Isoleucine	AUU	AUC*	AUU	AUC	*D. hydei* shows preference for AUU	

Table 4 Pagel's λ estimate for the mean RSCU values of 59 codons

Codon	Pagel's λ	*P-value*	Codon	Pagel's λ	*P-value*	Codon	Pagel's λ	*P-value*
AGG	1.037037037	4.83E-19	UUC	1.014011999	2.99E-10	UGU	1.0094039	3.83E-06
UCC	1.037037037	5.70E-17	CAA	1.015947254	6.52E-10	UGC	1.0078439	4.65E-06
ACC	1.028807517	5.54E-16	CAG	1.01507155	7.92E-10	CUG	0.994765	6.70E-06
AUA	1.037037037	2.10E-15	CCA	1.011613927	1.11E-09	GUU	1.0073325	9.18E-06
ACA	1.017326513	1.17E-13	UCA	1.01040289	1.21E-09	GUC	1.037037	9.26E-06
CGG	1.03445007	1.18E-13	CCC	1.010885551	1.97E-09	CUA	0.9875778	1.58E-05
GCC	1.016616089	3.36E-13	UUA	1.020871003	7.67E-09	GGU	0.9976393	1.66E-05
GCA	1.018165385	3.42E-13	UUG	1.015222201	8.05E-09	GCU	1.0164908	2.02E-05
AUC	1.022260603	7.10E-13	CUC	1.036938511	9.38E-09	GUG	0.9998965	2.21E-05
UAU	1.015422856	1.88E-12	UUU	1.015015679	2.48E-10	CGA	1.0106425	6.90E-05
UAC	1.014524219	2.24E-12	AUU	0.997364498	1.28E-08	ACU	1.0155664	0.0002226
GGA	0.983707443	5.79E-12	GCG	1.028160385	2.89E-08	CCG	1.0187482	0.0005579
AAU	1.013265257	1.75E-11	GAU	0.982927412	8.09E-08	GUA	0.9818203	0.0013685
AAC	1.012300381	2.03E-11	AGC	1.006078819	8.69E-08	UCU	1.0009604	0.0028898
CAU	1.022170443	2.98E-11	GAC	0.981246733	9.20E-08	UCG	1.0208646	0.0132103
CAC	1.021410712	3.58E-11	GGC	0.95118194	1.47E-07	CCU	0.9970469	0.065568
CGU	1.01402525	8.94E-11	ACG	1.012412505	1.63E-06	AGA	0.2432681	0.1532303
AAA	1.017385707	1.07E-10	GAA	1.00069091	2.14E-06	CUU	0	0.9829608
GGG	1.029575014	1.20E-10	GAG	0.999345877	2.51E-06	AGU	0.0063535	1
AAG	1.016477846	1.31E-10	CGC	0.988323579	2.91E-06			

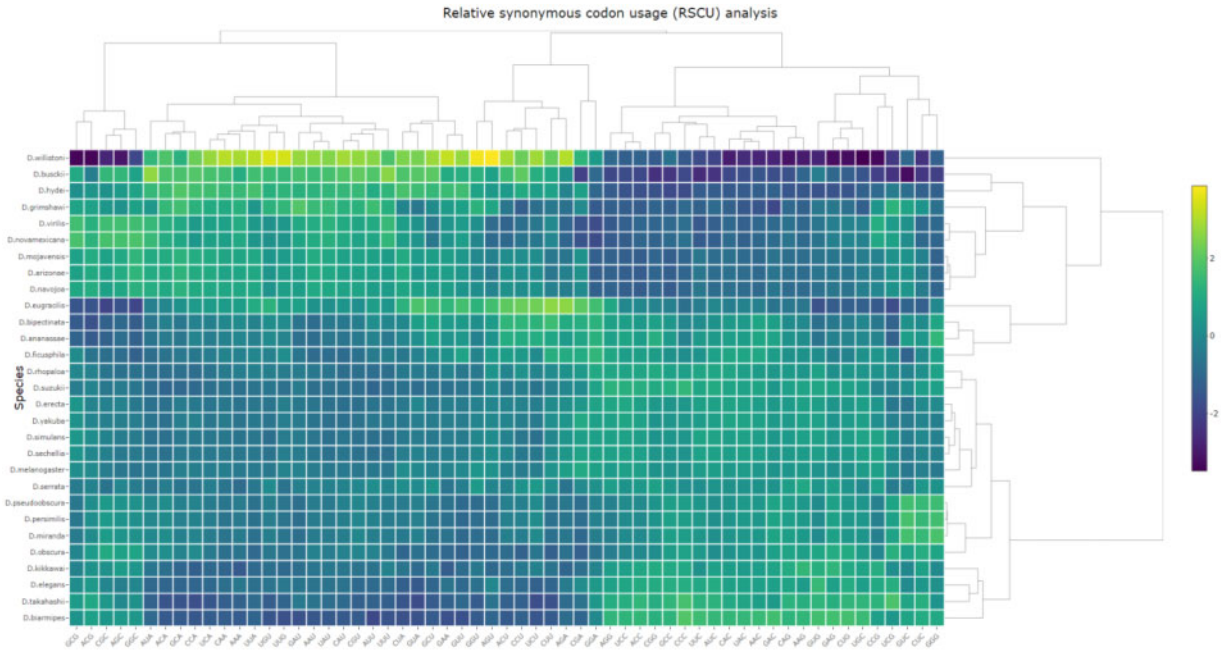

Figure 4 Hierarchical cluster mapping depicting differences in codon preferences between two subgenera: *Drosophila* and Sophophora. Species of subgenus *Drosophila* avoid C-ending codons except when the dinucleotide$_{23}$ is GC.

The results from *D. willistoni* provide further evidence to support our suggestion that nucleotides at the dinucleotide$_{23}$ position may have a considerable contribution toward CUB.

Amino acids with the same dinucleotide$_{23}$ have similar codon usage patterns

Comparing the two subgenera *Drosophila* and Sophophora, all two- and fourfold degenerate amino acids (except aspartic acid and glycine) showed high mean sENC-X values in species of the subgenus *Drosophila*, indicating reduced CUB (Figure 5). This observation is in agreement with our results from the ENC and the RSCU analysis, where certain amino acids showed a similar trend of CUB in each species from both subgenera. The Pagel's λ (Pagel 1999) for mean sENC-X values for each amino acid showed a strong phylogenetic signal ($\lambda > 0.85$, $P < 0.0001$, Table 5) for all amino acids except 2 fourfold degenerate amino acids (glycine and alanine) and all the sixfold degenerate amino acids (arginine, serine, and leucine). The amino acids, where the mean sENC-X values showed a strong phylogenetic signal, could be grouped further based on the dinucleotides$_{23}$ of the synonymous codons encoding them. Figure 5A depicts four amino acids that showed a consistent change in mean sENC-X values for each amino acid across the phylogenetic tree. For example, all four amino acids showed increased CUB, as evident by reduced mean sENC-X values in three species: *D. willistoni*, *D. ficusphila*, and *D. biarmipes*. The codons for these four amino acids had either AU or AC in their dinucleotide$_{23}$ position. The pattern was more conspicuous in Figure 5B, where three amino acids with synonymous codons ending with AA or AG showed a striking similarity in the codon usage pattern in each species.

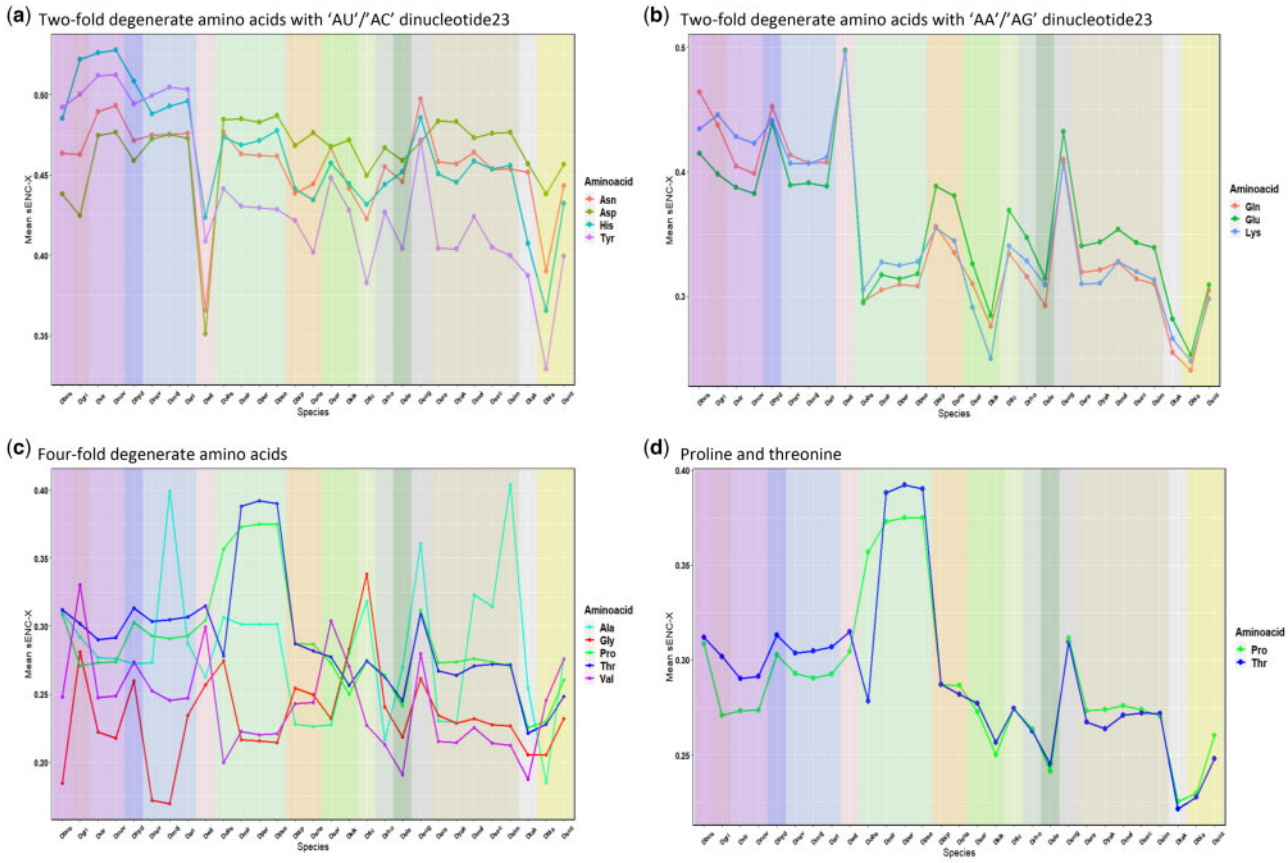

Figure 5 Mean sENC-X values for 2-fold and 4-fold degenerate amino acids in SCO genes from 29 *Drosophila* species. The background shading indicates species from the same species sub-group. a. Two-fold degenerate amino acids with 'AU'/'AC' dinucleotide23, b. Two-fold degenerate amino acids with 'AA'/'AG' dinucleotide23, c. Four-fold degenerate amino acids, and d. Proline and threonine.

Table 5 Pagel's λ estimate for the mean sENC-X values of 18 amino acids

Amino acid	Pagel's λ	*P-value*
Alanine	6.86E-05	1
Glycine	0.000139	0.999227
Leucine	0.123656	0.454539
Arginine	0.405609	0.164516
Serine	0.595586	0.589475
Tyrosine	0.8541	1.50E-08
Histidine	0.881316	5.29E-07
Asparagine	0.889795	0.001935
Phenylalanine	0.925138	1.14E-06
Cysteine	0.953992	1.91E-05
Valine	0.96357	0.028574
Isoleucine	0.970706	1.03E-11
Proline	0.979507	1.57E-09
Glutamic acid	0.980207	4.76E-06
Glutamine	0.982469	2.21E-08
Lysine	0.991687	3.11E-09
Aspartic acid	1.019727	3.25E-08
Threonine	1.023205	1.21E-06

Among the fourfold degenerate amino acids, the pattern was not as evident when all the five amino acids were compared to each other (Figure 5C). However, two amino acids, proline and threonine, showed a very close association in their codon usage pattern in all species (Figure 5D). Notably, these two amino acids are biochemically dissimilar. The commonality between these amino acids is that their respective synonymous codons have C at the 2nd position, and the preferred codons for both the amino acids end with CC. However, this is not a feature unique to proline and threonine. Alanine also is a fourfold degenerate amino acid with C in the 2nd codon position. It is evident from the mean RSCU values (Supplementary Figure S2, B, D, and E) that proline and threonine have similar codon usage patterns as compared to alanine, and this may attribute to the unique similarities reflected in the mean sENC-X values of proline and threonine (Figure 5D). These findings strengthen our hypothesis that CUB is noticeably influenced by the nucleotides at the dinucleotide$_{23}$ position.

Codon preferences correlate with dinucleotide$_{23}$ preferences

As indicated by the RSCU and sENC-X analysis, the peculiar codon preferences suggest that the dinucleotide$_{23}$ patterns may play a significant role in CUB. To investigate this observation further, we calculated the SDU values at the dinucleotide$_{23}$ position for the SCO genes in all 29 species. As shown in Figure 6, the dinucleotide preferences were in concordance with the codon preferences. Species of the subgenus Sophophora showed an overrepresentation (SDU > 1.0) for the CC dinucleotide$_{23}$ and an underrepresentation (SDU < 1.0) of the CA dinucleotide$_{23}$. Similarly, AC at the dinucleotide$_{23}$ position was underrepresented, whereas AU was overrepresented in species of the subgenus *Drosophila*. When the SDU values were compared between the two subgenera, species of the subgenus *Drosophila* showed an overrepresentation of GC dinucleotide$_{23}$. As deduced from the RSCU analysis, species of the subgenus *Drosophila* show a reduced preference for C-ending codons compared to those of the subgenus *Sophophora*, except in the case of GC dinucleotide$_{23}$. These

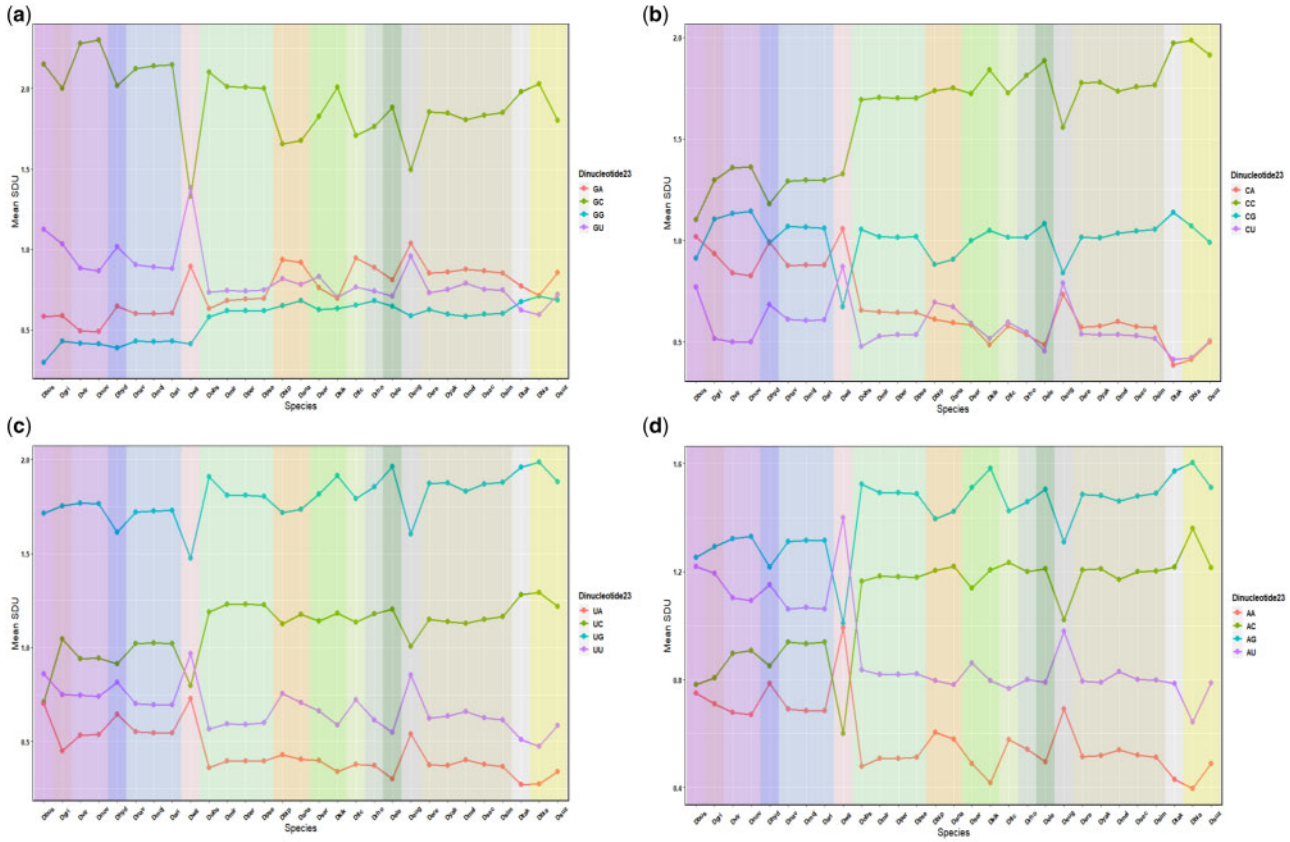

Figure 6 Synonymous dinucleotide usage in SCO genes from 29 *Drosophila* species. The background shading indicates species from the same species sub-group. a. G in the 2nd nucleotide position of the codon, b. C in the 2nd nucleotide position of the codon, c. U in the 2nd nucleotide position of the codon, and d. A in the 2nd nucleotide position of the codon.

findings clearly illustrate how codon preferences correspond to the dinucleotide$_{23}$ preferences among species.

The irregularity in the codon usage pattern in *D. willistoni* was reflected in the dinucleotide$_{23}$ preference as well. The dinucleotide$_{23}$ representation in *D. willistoni* was unlike either subgenus. The AU dinucleotide$_{23}$ was overrepresented, and AC was underrepresented. These results coincided with the mean RSCU values of aspartic acid, asparagine, histidine, and tyrosine (Figure 3, A–D). The AG and AA dinucleotides$_{23}$ had SDU values of 1.0, indicating no codon usage preference for codons ending with these dinucleotides. This observation agrees with the mean RSCU values of glutamine, glutamic acid, and lysine (Figure 3, G–I). The results from *D. willistoni* fortify our observed connection between dinucleotide$_{23}$ and codon preferences.

The mean SDU values for all 16 dinucleotides$_{23}$ exhibited a strong phylogenetic signal (Pagel's $\lambda > 0.97$, $P < 0.001$, Table 6), establishing a strong correlation between dinucleotide$_{23}$ preferences and speciation.

Table 6 Pagel's λ estimate for the mean SDU values of 16 dinucleotides$_{23}$

Dinucleotide$_{23}$	Pagel's λ	*P-value*
GG	1.030876	4.05E-15
CC	1.015944	5.69E-14
CA	1.00912	1.39E-11
AC	1.005113	1.34E-10
AU	1.005108	1.35E-10
GA	0.97812	1.26E-09
UC	1.024857	2.09E-09
AG	1.007625	8.97E-09
AA	1.007593	9.04E-09
UA	1.009185	5.65E-08
GU	1.003112	2.81E-07
GC	0.975313	8.84E-07
UU	0.9934	2.76E-05
CG	1.014305	0.000129
UG	0.98143	0.000217
CU	1.007935	0.0006

Discussion

In this study, we investigated CUB in the genus *Drosophila* by analyzing 668 SCO genes in 29 *Drosophila* species. We (1) show a difference in CUB within the genus *Drosophila*, (2) show a strong correlation between phylogenetic distance and CUB, (3) identify the specific differences in CUB among species, (4) describe an association between codon preference and dinucleotide$_{23}$ preference, and (5) show a connection between codon preference, dinucleotide$_{23}$ preference, and speciation.

Our data indicate that distantly related species show greater differences in CUB, while closely related species have similar CUB. The codon usage patterns showed the most substantial differences between the two subgenera, but it was also evident, although to a lesser degree, down to the species subgroup level. Hence, the observed patterns across the phylogenetic tree established from our chosen 29 *Drosophila* species support the previously held hypothesis that each species appears to prefer a different set of codons (Sharp and Li 1986; Hershberg and Petrov 2008; Plotkin and Kudla 2011).

We note that our findings are partially contradicting a previous study. Vicario *et al.* (2007) were unable to detect any substantial differences in codon preference among 11 of the 12 species (except *D. willistoni*) that they chose for their investigation. The discrepancy between our and their data may be explained by the different data set sizes and species composition. Our study uses 29 species with a better representation of species from both subgenera and species groups than the 12 species represented in the Vicario *et al.* (2007) study. Another possible explanation for this discrepancy could be that in the previous study, the authors used 10% most highly biased genes (determined by a low ENC) from each species for the CUB evaluation. It is now well established that genes from the same genome show variation in the codon usage pattern and that gene length and expression levels impact CUB (Behura and Severson 2012; Paul *et al.* 2018). On the other hand, our study analyzes only 668 SCO genes present in all 29 species, irrespective of their ENC values. We conclude that selecting genes with low ENC (≤35) as a criterion for CUB studies may introduce a bias that could potentially affect the optimal evaluation of CUB, especially across species.

The codon usage patterns between the two subgenera, *Drosophila* and Sophophora, differ in the preferred codons and the extent of codon preference, where species of the subgenus *Drosophila* show a reduced preference for C-ending codons, except for codons with GC in their dinucleotide$_{23}$ position. Our study is the first report describing this unusual trait that distinguishes the two subgenera. Further tests are required to identify the selective pressures acting at this evolutionary juncture.

D. willistoni was described as an "outlier" in the previous publication examining CUB in *Drosophila* (Vicario *et al.* 2007), and our findings corroborate their results. Coding sequences of *D. willistoni* had the lowest GC content, which fits with the mean RSCU values of the preferred codons that are predominantly U-ending rather than C-ending. The relatively high ENC value for the orthologous genes also correlates well with the codon usage pattern, as *D. willistoni* showed reduced CUB for all amino acids, except for aspartic acid, asparagine, histidine, and tyrosine. The codons preferred for these four amino acids were also precisely opposite to the codons preferred by the other species of the subgenus Sophophora. Further studies to understand the codon usage patterns in *D. willistoni* and other closely related species will be necessary to explain how speciation of the willistoni species group has affected CUB.

Behura and Severson (2012) observed a potential association between CUB and amino acid composition. They found that codons with A in the 2nd codon position were more abundant than codons with G, U, or C in the 2nd position in the insect genomes they studied. We propose that the preference for codons with A in the 2nd codon position may have a simpler explanation. Codons with A in the 2nd codon position encode seven amino acids, whereas codons with G, U, and C in the 2nd position code for four amino acids each. Furthermore, Behura and Severson (2012) associated a preference for codons with A in the 2nd position with the hydrophilic nature of the amino acid encoded, suggesting that CUB may be related to the biochemical properties of the amino acid. We have studied 668 SCO genes that have a similar amino acid composition (Supplementary Figure S4 and Supplementary Table S3) and still found changes in CUB that correlated with phylogenetic distance. Therefore, we propose that the amino acid composition may not have a substantial role to play in CUB.

Rather than the role of encoded amino acids in CUB, we observed that codon preferences are associated with the nucleotide composition at the dinucleotides$_{23}$ position of the codon. In twofold degenerate amino acids, the three amino acids that have codons with AG or AA as the dinucleotide$_{23}$ showed almost identical mean RSCU values and comparable mean sENC-X values in all the 29 species. These three amino acids (lysine, glutamine, and glutamic acid) are polar, hydrophilic but differently charged. Lysine is positively charged, glutamine is uncharged, and glutamic acid is negatively charged. Similarly, the mean sENC-X values for amino acids with AU or AC at the dinucleotide$_{23}$ position of their respective synonymous codons showed a convincing pattern for the four amino acids (aspartic acid, asparagine, histidine, and threonine), again with biochemically distinct properties. In the fourfold degenerate amino acids, the association between the two amino acids, proline, and threonine, was particularly robust. Their preferred codons had CC dinucleotide$_{23,}$ followed by CG. These instances, in which biochemically distinct amino acids that share the same dinucleotide$_{23}$ show a similar codon usage pattern, suggest that the nucleotides at the dinucleotide$_{23}$ position may have a significant contribution toward CUB. Further, the codon preferences and dinucleotide$_{23}$ preferences coincided with each other in all 29 *Drosophila* species, confirming this observation.

Although the effect of dinucleotide preference on CUB has been reported in viruses (Castells *et al.* 2017; Gu *et al.* 2019), very few reports describe this effect in prokaryotes and eukaryotes (Paul *et al.* 2018; Roy and van Staden 2019; Wang *et al.* 2019). Roy and van Staden (2019) studied five species of the fungal genus *Puccinia*. While they found an overrepresentation of certain dinucleotides, they have not described a correlation between the preferred codons and the preferred dinucleotides. Another group of researchers studied three dicot species and found a correlation between dinucleotide preferences and codon preferences (Paul *et al.* 2018). However, they have not established a connection between codon preferences and speciation. We found a strong phylogenetic signal for codon preferences as well as dinucleotides$_{23}$ preferences. Our research reports, for the first time, an association between speciation, CUB, and dinucleotide preferences in a large dataset of 29 *Drosophila* species.

In conclusion, CUB is strongly correlated with phylogenetic distance. Our study in 29 *Drosophila* species demonstrates that CUB may be influenced by dinucleotide$_{23}$ preferences. Further studies are necessary to identify the causes and the consequences of the selection acting at the dinucleotide$_{23}$ positions that, in turn, are related to codon preferences.

Limitations

Each *Drosophila* species genome contains approximately 14,000 genes. Our study is based on 668 SCO genes. We understand that the number of genes studied is low and raises the possibility that the entire genome may not necessarily show the pattern reflected in these orthologs. However, CUB studies throughout the genome in a holistic manner may have certain drawbacks. The presence of pseudogenes, paralogues, and different codon usage patterns between genes of the same genome can produce confounding results in the CUB analysis. Furthermore, the length and expression levels of genes are known to influence CUB. The 668 SCO genes used in our study essentially have the same length and a similar amino acid composition. Thus, we are confident that the study of single-copy orthologs is an appropriate method to identify evolving codon usage patterns. Also, gene prediction and gene annotation of eukaryotic genomes involve various technical challenges (Yandell and Ence 2012). Therefore, the OrthoFinder (Emms and Kelly 2019) indirectly ensures better curation of the genes. We recommend the use of SCO genes for the comparison of CUB among species.

Acknowledgments

The authors would like to thank the following: Paresh Kokate for his valuable suggestions that expedited the data analyses for the manuscript, two anonymous reviewers, and the journal editor for their valuable comments that have improved the quality of the manuscript.

Funding

This work was supported by the National Science Foundation (DOB/DEB-1737877) to T.W.

Conflicts of interest

None declared.

Literarutre cited

Alberts BJA, Johnson A, Lewis J, Raff M, Roberts K, *et al.* 2002. *Drosophila* and the molecular genetics of pattern formation: genesis of the body plan. In: Molecular Biology of the Cell. 4th ed. New York: Garland Science. pp. 1177-1190.

Anwar AM, Soudy M, Mohamed R. 2019. vhcub: virus-host codon usage co-adaptation analysis. F1000Res. 8:2137.

Behura SK, Severson DW. 2012. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS One. 7:e43111.

Blomberg SP, Garland T. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. J Evol Biol. 15:899–910.

Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, *et al.* 2011. A synonymous variant in *IRGM* alters a binding site for miR-196 and causes deregulation of *IRGM*-dependent xenophagy in Crohn's disease. Nat Genet. 43:242–245.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12:59–60.

Castells M, Victoria M, Colina R, Musto H, Cristina J. 2017. Genome-wide analysis of codon usage bias in Bovine coronavirus. Virol J. 14:115–121.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: U Bastolla, M Porto, HE Roman, M Vendruscolo, editors. Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 207–232.

Elek A, Kuzman M, Vlahoviček K. 2020. coRdon: codon usage analysis and prediction of gene expressivity. R package version 1.8.0. https://github.com/BioinfoHR/coRdon (Accessed: 2021 March 22).

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Galili T, O'Callaghan A, Sidi J, Sievert C. 2018. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics. 34:1600–1602.

Guan DL, Ma LB, Khan MS, Zhang XX, Xu SQ, *et al.* 2018. Analysis of codon usage patterns in Hirudinaria manillensis reveals a preference for GC-ending codons caused by dominant selection constraints. BMC Genomics. 19:542.

Gu H, Fan RLY, Wang D, Poon LLM. 2019. Dinucleotide evolutionary dynamics in influenza A virus. Virus Evol. 5:vez038.

Han F, Peng Y, Xu L, Xiao P. 2014. Identification, characterization, and utilization of single copy genes in 29 angiosperm genomes. BMC Genomics. 15:504.

Heger A, Ponting CP. 2007. Variable strength of translational selection among 12 *Drosophila* species. Genetics. 177:1337–1348.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet. 42:287–299.

Ismail SNFB, Baharum SN, Fazry S, Low CF. 2019. Comparative genome analysis reveals a distinct influence of nucleotide composition on virus-host species-specific interaction of prawn-infecting nodavirus. J Fish Dis. 42:1761–1772.

James TD, Salguero-Gómez R, Jones OR, Childs DZ, Beckerman AP. 2020. Bridging gaps in demographic analysis with phylogenetic imputation. Conserv Biol. 0:1–12.

LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2019. Variation and selection on codon usage bias across an entire subphylum. PLoS Genet. 15:e1008304.

Lamolle G, Fontenla S, Rijo G, Tort JF, Smircich P. 2019. Compositional analysis of flatworm genomes shows strong codon usage biases across all classes. Front Genet. 10:771.

Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32:11–16.

Lytras S, Hughes J. 2020. Synonymous dinucleotide usage: a codon-aware metric for quantifying dinucleotide representation in viruses. Viruses. 12:462.

Molina-Venegas R, Rodríguez M. 2017. Revisiting phylogenetic signal; strong or negligible impacts of polytomies and branch length information? BMC Evol Biol. 17:53–62.

Novoa EM, Jungreis I, Jaillon O, Kellis M. 2019. Elucidation of codon usage signatures across the domains of life. Mol Biol E. 36: 2328–2339.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature. 401:877–884.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 20: 289–290.

Paul P, Malakar AK, Chakraborty S. 2018. Codon usage vis-a-vis start and stop codon context analysis of three dicot species. J Genet. 97:97–107.

Payne BL, Alvarez-Ponce D. 2019. Codon usage differences among genes expressed in different tissues of *Drosophila* melanogaster. Genome Biol E. 11:1054–1065.

Pinheiro J, Bates D, DebRoy S, Sarkar D. and R core team 2021. {nlme}: Linear and Nonlinear Mixed Effects Models}. R package version 3.1-152. https://CRAN.R-project.org/package=nlme (Accessed: 2021 March 22).

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 12:32–42.

Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in *Drosophila*. Proc Natl Acad Sci USA. 94:7784–7790.

Quax TEF, Claassens NJ, Söll D, van der Oost J. 2015. Codon bias as a means to fine-tune Gene expression. Mol Cell. 59:149–161.

Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology. Methods Ecol E. 3:217–223.

Roy A, van Staden J. 2019. Insights into the riddles of codon usage patterns and codon context signatures in fungal genus Puccinia, a persistent threat to global agriculture. J Cell Biochem. 120: 19555–19566.

Sahyoun AH, Hölzer M, Jühling F, Höner zu Siederdissen C, Al-Arab M, *et al.* 2015. Towards a comprehensive picture of alloacceptor tRNA remolding in metazoan mitochondrial genomes. Nucleic Acids Res. 43:8044–8056.

Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 41:2073–2094.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 24:28–38.

Subramanian A, Sarkar RR. 2015. Data in support of large scale comparative codon usage analysis in Leishmania and Trypanosomatids. Data Brief. 4:269–272.

Van Rossum G, Drake FL. 2009. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. BMC Evol Biol. 7:226.

Wang Y, Zeng Z, Liu TL, Sun L, Yao Q, *et al.* 2019. TA, GT and AC are significantly under-represented in open reading frames of prokaryotic and eukaryotic protein-coding genes. Mol Genet Genomics. 294:637–647.

Wright F. 1990. The 'effective number of codons' used in a gene. Gene. 87:23–29.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:329–342.

Communicating editor: J. Comeron