# The Codon Statistics Database: A Database of Codon Usage Bias

Krishnamurthy Subramanian,[†,‡] Bryan Payne,[†] Felix Feyertag, and David Alvarez-Ponce *

Biology Department, University of Nevada, Reno, Reno, NV 89557, USA

***Corresponding author:** E-mail: dap@unr.edu.

[†]These authors contributed equally to this work.

[‡]Present address: Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

**Associate editor:** Naruya Saitou

## Abstract

**We present the Codon Statistics Database, an online database that contains codon usage statistics for all the species with reference or representative genomes in RefSeq (over 15,000). The user can search for any species and access two sets of tables. One set lists, for each codon, the frequency, the Relative Synonymous Codon Usage, and whether the codon is preferred. Another set of tables lists, for each gene, its GC content, Effective Number of Codons, Codon Adaptation Index, and frequency of optimal codons. Equivalent tables can be accessed for (1) all nuclear genes, (2) nuclear genes encoding ribosomal proteins, (3) mitochondrial genes, and (4) chloroplast genes (if available in the relevant assembly). The user can also search for any taxonomic group (e.g., "primates") and obtain a table comparing all the species in the group. The database is free to access without registration at http://codonstatsdb.unr.edu.**

***Key words:*** **codon bias, codon usage, database, synonymous codons.**

## Introduction

Most amino acids are encoded by multiple synonymous codons. Despite encoding for the same amino acid, some synonymous codons are used significantly more often than others, a phenomenon known as codon usage bias. Species significantly differ in their codon preferences—for instance, glutamic acid is preferentially encoded by GAG in human, whereas the same amino acid is preferentially encoded by GAA in *Escherichia coli* (Ikemura 1982; Sharp et al. 2010). In addition, genes within any given genome differ in their patterns of codon usage. In particular, gene expression levels significantly correlate with gene-specific metrics of codon usage such as the Effective Number of Codons (ENC; Wright 1990), the Codon Adaptation Index (CAI; Sharp and Li 1987), or the frequency of optimal codons ($F_{op}$; Ikemura 1985) (e.g., Gouy and Gautier 1982).

Codon preferences can be affected by a number of factors, including the genome's nucleotide composition (e.g., AT-rich genomes tend to use codons ending in A or T) and translational selection (codons that are translated by highly abundant tRNAs are translated faster and with fewer errors; e.g., Ikemura 1985; Hershberg and Petrov 2008).

Understanding codon preferences across the different species and genes is important not only to understanding genome evolution, but can also inform tasks such as heterologous expression, gene prediction, or phylogenetic inference (e.g., Gustafsson et al. 2004; Christianson 2005; Rota-Stabelli et al. 2013). In addition, the patterns of codon usage of viruses tend to be similar to those of their host species (e.g., Shackelton et al. 2006). Despite the relevance of maintaining species- and gene-specific codon usage information, existing databases have not been updated in a long time, focus on specific taxa, and/or do not provide gene-specific metrics (Nakamura et al. 2000; Hilterbrand et al. 2012; Athey et al. 2017).

## Implementation

For each of the species with reference or representative genomes in the RefSeq database (release 207), we chose one full assembly (in order of preference, the one labeled as "reference," the one with the highest assembly level, or the most recent one) and retrieved the corresponding coding sequences (CDSs) file. Using that file as input, a number of tables were pre-computed using an R pipeline. To avoid codon redundancy, only one CDS per gene was used (if multiple were available, the longest one was chosen). The web interface was created using PERL CGI.

For each species, we computed the total frequency of each codon, and used the information to compute the Relative Synonymous Codon Usage (RSCU) of each codon. For each gene, we computed the GC content for the entire CDS (GC), the GC content at third codon positions (GC3), the ENC, and the RSCU for each codon.

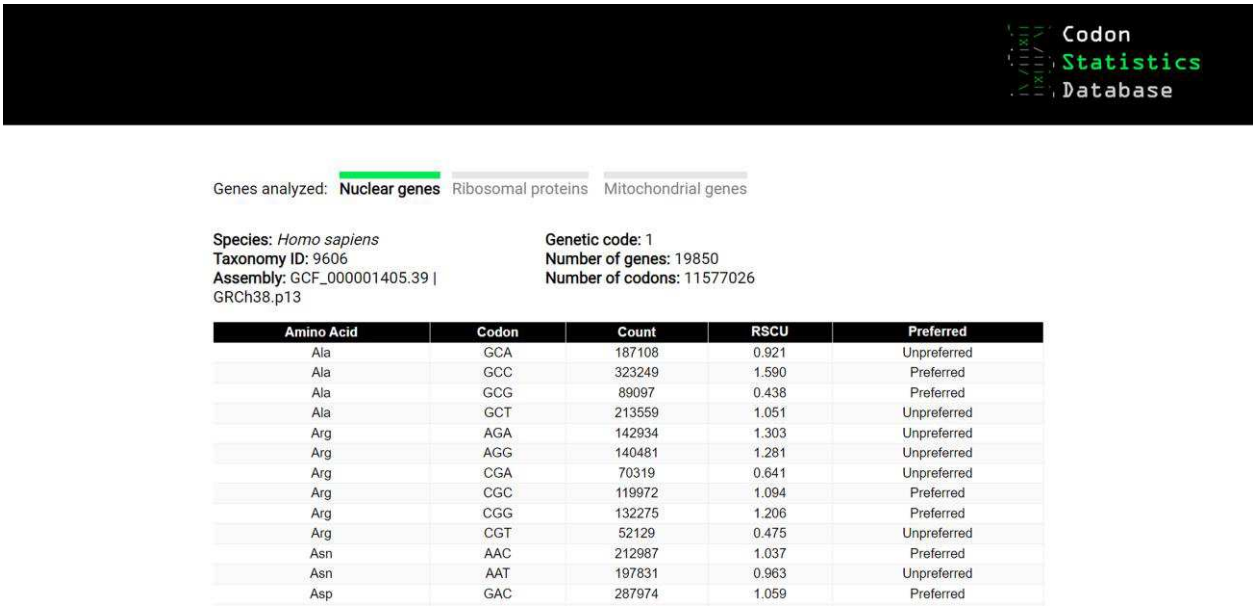
For species with over 1,000 genes, we also compared genes inferred to be highly expressed (bottom 10% ENC values) with genes inferred to be lowly expressed (top 10% ENC values). Codons with significantly higher RSCU values in the highly expressed gene set (according to a







Brief Communication




| Amino Acid | Codon | Count | RSCU | Preferred |
|---|---|---|---|---|
| Ala | GCA | 187108 | 0.921 | Unpreferred |
| Ala | GCC | 323249 | 1.590 | Preferred |
| Ala | GCG | 89097 | 0.438 | Preferred |
| Ala | GCT | 213559 | 1.051 | Unpreferred |
| Arg | AGA | 142934 | 1.303 | Unpreferred |
| Arg | AGG | 140481 | 1.281 | Unpreferred |
| Arg | CGA | 70319 | 0.641 | Unpreferred |
| Arg | CGC | 119972 | 1.094 | Preferred |
| Arg | CGG | 132275 | 1.206 | Preferred |
| Arg | CGT | 52129 | 0.475 | Unpreferred |
| Asn | AAC | 212987 | 1.037 | Preferred |
| Asn | AAT | 197831 | 0.963 | Unpreferred |
| Asp | GAC | 287974 | 1.059 | Preferred |

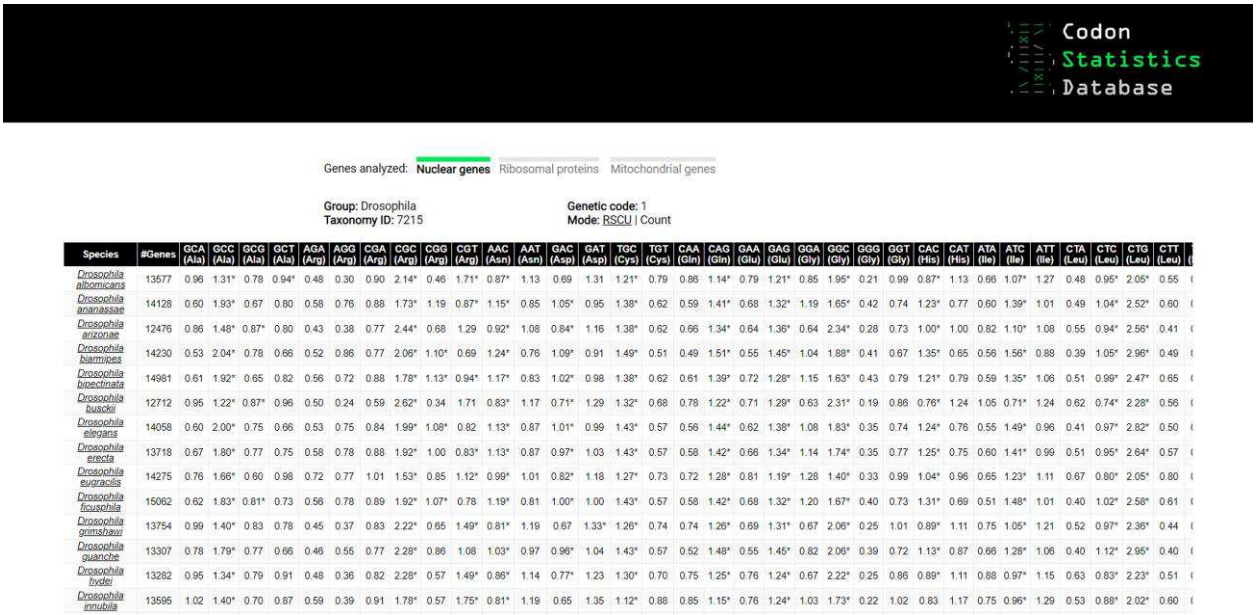
**Fig. 1.** Species summary. Codon statistics corresponding to all human nuclear genes are shown.

Mann–Whitney *U* test) were considered preferred/optimal. We then computed the $F_{op}$ for each gene. The highly expressed gene set was also used as reference to compute the CAI of each gene.

## The Codon Statistics Database

We present the Codon Statistics Database, an online database that contains codon usage information for all species with reference or representative genomes in RefSeq (over 15,000). The user can search for any species or taxonomic group by taxonomic ID (e.g., "9606"), scientific name (e.g., "*Homo sapiens*"), or common name (e.g., "human"), and select an option from a drop-down menu.

If a species is selected, the user is directed to a table that lists, for each codon, the encoded amino acid, the total count in the genome, the RSCU, and whether the codon is preferred or unpreferred (fig. 1). The user can visualize and download equivalent tables for (1) all nuclear genes (default option), (2) nuclear genes encoding ribosomal proteins (this subset is included since such proteins are often highly expressed and thus subjected to strong codon bias), (3) mitochondrial genes, and (4) chloroplast genes (if such gene sets are available in the relevant genome assembly). For viruses, only one option including all genes is available. Additionally, for each gene set, the user can download a tab-delimited file (.tsv) listing the following statistics for each gene: GC, GC3, ENC, CAI, and $F_{op}$.




| Species | #Genes | GCA (Ala) | GCC (Ala) | GCG (Ala) | GCT (Ala) | AGA (Arg) | AGG (Arg) | CGA (Arg) | CGC (Arg) | CGG (Arg) | CGT (Arg) | AAC (Asn) | AAT (Asn) | GAC (Asp) | GAT (Asp) | TGC (Cys) | TGT (Cys) | CAA (Gln) | CAG (Gln) | GAA (Glu) | GAG (Glu) | GGA (Gly) | GGC (Gly) | GGG (Gly) | GGT (Gly) | CAC (His) | CAT (His) | ATA (Ile) | ATC (Ile) | ATT (Ile) | CTA (Leu) | CTC (Leu) | CTG (Leu) | CTT (Leu) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Drosophila albomicans* | 13577 | 0.96 | 1.31* | 0.78 | 0.94* | 0.48 | 0.30 | 0.90 | 2.14* | 0.46 | 1.71* | 0.87* | 1.13 | 0.69 | 1.31 | 1.21* | 0.79 | 0.86 | 1.14* | 0.79 | 1.21* | 0.85 | 1.95* | 0.21 | 0.99 | 0.87* | 1.13 | 0.66 | 1.07* | 1.27 | 0.48 | 0.95* | 2.05* | 0.55 |
| *Drosophila ananassae* | 14128 | 0.60 | 1.93* | 0.67 | 0.80 | 0.58 | 0.76 | 0.88 | 1.73* | 1.19 | 0.87* | 1.15* | 0.85 | 1.05* | 0.95 | 1.38* | 0.62 | 0.59 | 1.41* | 0.68 | 1.32* | 1.19 | 1.65* | 0.42 | 0.74 | 1.23* | 0.77 | 0.60 | 1.39* | 1.01 | 0.49 | 1.04* | 2.52* | 0.60 |
| *Drosophila arizonae* | 12476 | 0.86 | 1.48* | 0.87* | 0.80 | 0.43 | 0.38 | 0.77 | 2.44* | 0.68 | 1.29 | 0.92* | 1.08 | 0.84* | 1.16 | 1.38* | 0.62 | 0.66 | 1.34* | 0.64 | 1.36* | 0.64 | 2.34* | 0.28 | 0.73 | 1.00* | 1.00 | 0.82 | 1.10* | 1.08 | 0.55 | 0.94* | 2.56* | 0.41 |
| *Drosophila biarmipes* | 14230 | 0.53 | 2.04* | 0.78 | 0.66 | 0.52 | 0.86 | 0.77 | 2.06* | 1.10* | 0.69 | 1.24* | 0.76 | 1.09* | 0.91 | 1.49* | 0.51 | 0.49 | 1.51* | 0.55 | 1.45* | 1.04 | 1.88* | 0.41 | 0.67 | 1.35* | 0.65 | 0.56 | 1.56* | 0.88 | 0.39 | 1.05* | 2.96* | 0.49 |
| *Drosophila bipectinata* | 14961 | 0.61 | 1.92* | 0.65 | 0.82 | 0.56 | 0.72 | 0.88 | 1.78* | 1.13* | 0.94* | 1.17* | 0.83 | 1.02* | 0.98 | 1.38* | 0.62 | 0.61 | 1.39* | 0.72 | 1.28* | 1.15 | 1.63* | 0.43 | 0.79 | 1.21* | 0.79 | 0.59 | 1.35* | 1.06 | 0.51 | 0.99* | 2.47* | 0.65 |
| *Drosophila busckii* | 12712 | 0.95 | 1.22* | 0.87* | 0.96 | 0.50 | 0.24 | 0.59 | 2.62* | 0.34 | 1.71 | 0.83* | 1.17 | 0.71* | 1.29 | 1.32* | 0.68 | 0.78 | 1.22* | 0.71 | 1.29* | 0.63 | 2.31* | 0.19 | 0.86 | 0.76* | 1.24 | 1.05 | 0.71* | 1.24 | 0.62 | 0.74* | 2.28* | 0.56 |
| *Drosophila elegans* | 14058 | 0.60 | 2.00* | 0.75 | 0.66 | 0.53 | 0.75 | 0.84 | 1.99* | 1.08* | 0.82 | 1.13* | 0.87 | 1.01* | 0.99 | 1.43* | 0.57 | 0.56 | 1.44* | 0.62 | 1.38* | 1.08 | 1.83* | 0.35 | 0.74 | 1.24* | 0.76 | 0.55 | 1.49* | 0.96 | 0.41 | 0.97* | 2.82* | 0.50 |
| *Drosophila erecta* | 13718 | 0.67 | 1.80* | 0.77 | 0.75 | 0.58 | 0.78 | 0.88 | 1.92* | 1.00 | 0.83* | 1.13* | 0.87 | 0.97* | 1.03 | 1.43* | 0.57 | 0.58 | 1.42* | 0.66 | 1.34* | 1.14 | 1.74* | 0.35 | 0.77 | 1.25* | 0.75 | 0.60 | 1.41* | 0.99 | 0.51 | 0.95* | 2.64* | 0.57 |
| *Drosophila eugracilis* | 14275 | 0.76 | 1.66* | 0.60 | 0.98 | 0.72 | 0.77 | 1.01 | 1.53* | 0.85 | 1.12* | 0.99* | 1.01 | 0.82* | 1.18 | 1.27* | 0.73 | 0.72 | 1.28* | 0.81 | 1.19* | 1.28 | 1.40* | 0.33 | 0.99 | 1.04* | 0.96 | 0.65 | 1.23* | 1.11 | 0.67 | 0.80* | 2.05* | 0.80 |
| *Drosophila ficusphila* | 15062 | 0.62 | 1.83* | 0.81* | 0.73 | 0.56 | 0.78 | 0.89 | 1.92* | 1.07* | 0.78 | 1.19* | 0.81 | 1.00* | 1.00 | 1.43* | 0.57 | 0.58 | 1.42* | 0.68 | 1.32* | 1.20 | 1.67* | 0.40 | 0.73 | 1.31* | 0.69 | 0.51 | 1.48* | 1.01 | 0.40 | 1.02* | 2.58* | 0.61 |
| *Drosophila grimshawi* | 13754 | 0.99 | 1.40* | 0.83 | 0.78 | 0.45 | 0.37 | 0.83 | 2.22* | 0.65 | 1.49* | 0.81* | 1.19 | 0.67 | 1.33* | 1.26* | 0.74 | 0.74 | 1.26* | 0.69 | 1.31* | 0.67 | 2.06* | 0.25 | 1.01 | 0.89* | 1.11 | 0.75 | 1.05* | 1.21 | 0.52 | 0.97* | 2.36* | 0.44 |
| *Drosophila guanche* | 13307 | 0.78 | 1.79* | 0.77 | 0.66 | 0.46 | 0.55 | 0.77 | 2.28* | 0.86 | 1.08 | 1.03* | 0.97 | 0.96* | 1.04 | 1.43* | 0.57 | 0.52 | 1.48* | 0.55 | 1.45* | 0.82 | 2.06* | 0.39 | 0.72 | 1.13* | 0.87 | 0.66 | 1.28* | 1.06 | 0.40 | 1.12* | 2.95* | 0.40 |
| *Drosophila hydei* | 13282 | 0.95 | 1.34* | 0.79 | 0.91 | 0.48 | 0.36 | 0.82 | 2.28* | 0.57 | 1.49* | 0.86* | 1.14 | 0.77* | 1.23 | 1.30* | 0.70 | 0.75 | 1.25* | 0.76 | 1.24* | 0.67 | 2.22* | 0.25 | 0.86 | 0.89* | 1.11 | 0.88 | 0.97* | 1.15 | 0.63 | 0.83* | 2.23* | 0.51 |
| *Drosophila innubila* | 13595 | 1.02 | 1.40* | 0.70 | 0.87 | 0.59 | 0.39 | 0.91 | 1.78* | 0.57 | 1.75* | 0.81* | 1.19 | 0.65 | 1.35 | 1.12* | 0.88 | 0.85 | 1.15* | 0.76 | 1.24* | 1.03 | 1.73* | 0.22 | 1.02 | 0.83 | 1.17 | 0.75 | 0.96* | 1.29 | 0.53 | 0.88* | 2.02* | 0.60 |

**Fig. 2.** Taxonomic group summary. Codon preferences for species in the genus *Drosophila* are shown.

If a taxonomic group with multiple species is selected (e.g., "7215," "*Drosophila*," or "fruit flies"), the user is presented with a table comparing all the species in the group (fig. 2). The user has the option to visualize either codon counts or RSCU values. Preferred codons in each species are marked with asterisks.

## Acknowledgments

We are grateful to Alejandra Nores for assistance with web design, and to the Office of Information Technology of the University of Nevada, Reno for computational resources. This work was supported by Nevada INBRE (funded by grant P20GM103440, National Institute of General Medical Sciences, National Institutes of Health) and by the National Science Foundation (grant MCB 1818288).

## Data Availability

All data used in this work are derived from the RefSeq database.

## References

Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, Simonyan V, Kimchi-Sarfaty C. 2017. A new and updated resource for codon usage tables. *BMC Bioinform.* **18**: 1–10.

Christianson ML. 2005. Codon usage patterns distort phylogenies from or of DNA sequences. *Am J Bot.* **92**:1221–1233.

Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10**:7055–7074.

Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. *Trends Biotech.* **22**(7):346–353.

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

Hilterbrand A, Saelens J, Putonti C. 2012. CBDB: the codon bias database. *BMC Bioinform.* **13**:1–7.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* **2**:13–34.

Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* **28**:292.

Rota-Stabelli O, Lartillot N, Philippe H, Pisani D. 2013. Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. *Syst Biol.* **62**(1):121–133.

Shackelton LA, Parrish CR, Holmes EC. 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J Mol Evol.* **62**:551–563.

Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci.* **365**: 1203–1212.

Sharp PM, Li WH. 1987. The codon adaptation index: a measure of directional synonymous codon usage, and its potential applications. *Nucleic Acids Res.* **15**:1281–1295.

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