Codon Statistics Database User Manual

Availability: The database is freely available, with no registration, at http://codonstatsdb.unr.edu/.

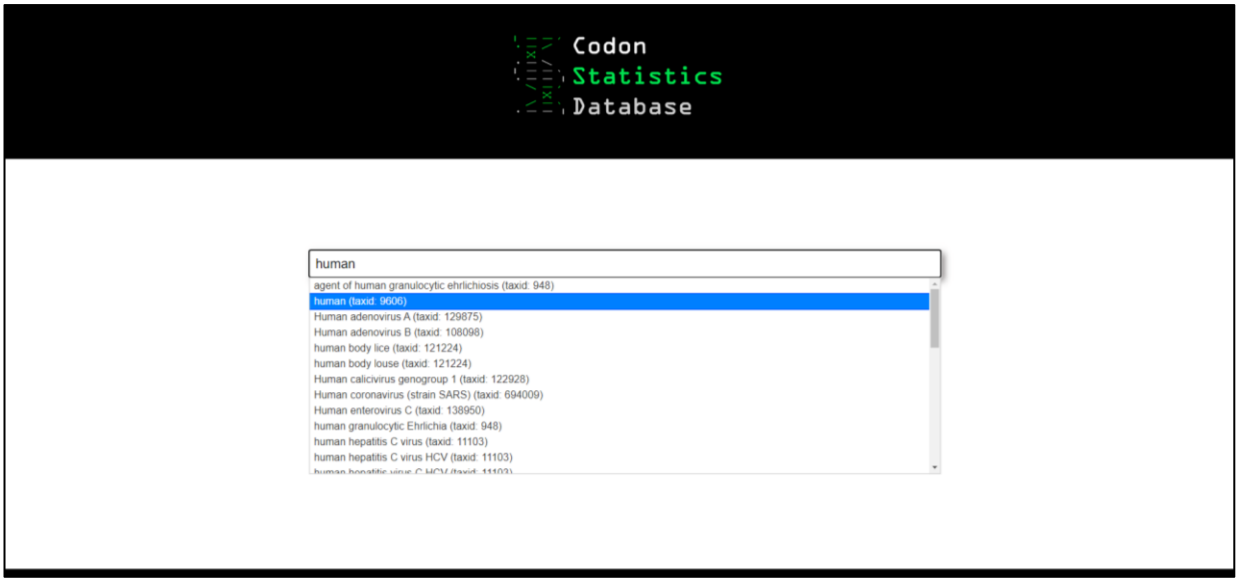
Citation: The database is described in the following article:

> Subramanian K, Payne B, Feyertag F, Alvarez-Ponce D. 2022. The Codon Statistics Database: a Database of Codon Usage Bias. *Mol Biol Evol*. https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msac157/6647594

Last updated: This manual was last updated on July 21, 2022.

1. Input

The user can search for any species or a taxonomic group represented in the RefSeq database (version 207). Searches can be done by taxonomic ID (e.g., "9606"), scientific name (e.g., "*Homo sapiens*") or common name (e.g., "human" or "primates"). The user can then select one option from a dropdown menu (Figure 1).


human

Codon
Statistics
Database
agent of human granulocytic ehrlichiosis (taxid: 948)
human (taxid: 9606)
Human adenovirus A (taxid: 129875)
Human adenovirus B (taxid: 108098)
human body lice (taxid: 121224)
human body louse (taxid: 121224)
Human calicivirus genogroup 1 (taxid: 122928)
Human coronavirus (strain SARS) (taxid: 694009)
Human enterovirus C (taxid: 138950)
human granulocytic Ehrlichia (taxid: 948)
human hepatitis C virus (taxid: 11103)
human hepatitis C virus HCV (taxid: 11103)
human hepatitis virus C HCV (taxid: 11103)

Figure 1. Front page of the Codon Statistics Database.

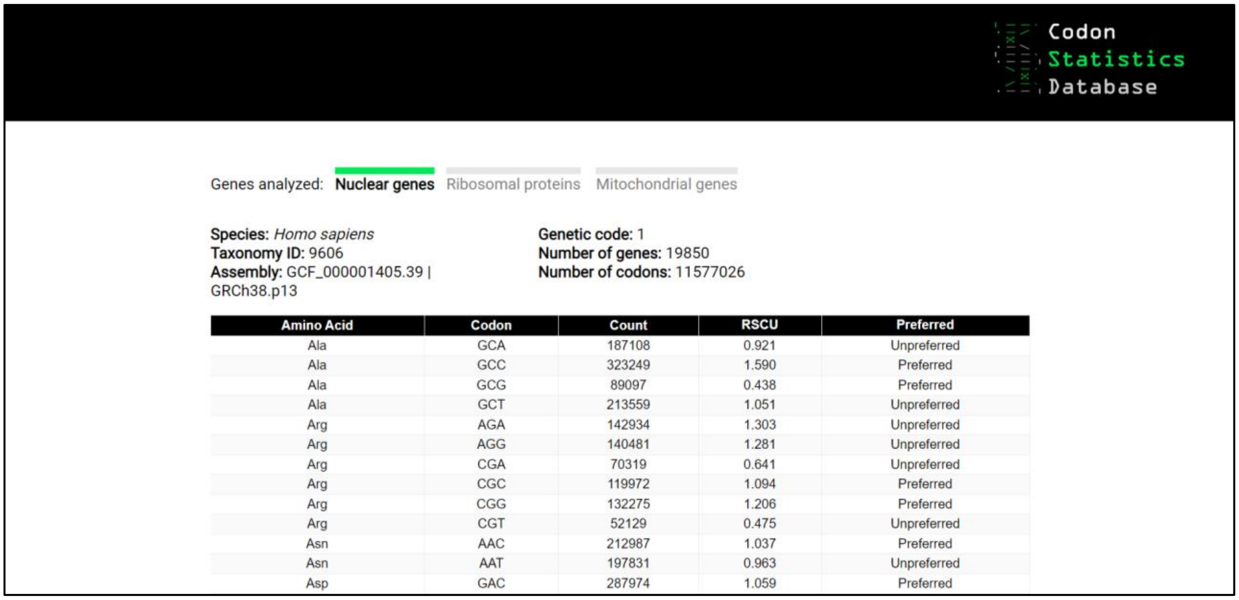
2. Species output

If a species has been selected, the user will access a table presenting the following information (Figure 2):

- Species name.
- Species' taxonomy ID.
- Genome assembly.
- Genetic code used, following NCBI standards (e.g., 1 = standard genetic code).
- Number of protein-coding genes analyzed.
- Number of codons analyzed.

For each codon, the following information is shown (Figure 2):

- Encoded amino acid.
- Count: the number of times that the codon is used in the entire genome. N-terminal methionines encoded by codons other than ATG were excluded from the analysis.
- Relative Synonymous Codon Usage (RSCU): The observed codon count divided by the count that would be expected if all synonymous codons were used at the same frequency. For each amino acid, the average of the RSCUs of the codons encoding the amino acid is 1.
- Whether the codon is preferred or unpreferred. This information is only available for genomes with at least 1000 genes. Preferred codons are those that exhibit a significantly higher RSCU in highly expressed genes than in lowly expressed genes. We used as set of highly expressed genes those with the lowest ENC values (bottom decile), and as set of lowly expressed genes those with the highest ENC values (top decile).



Genes analyzed: **Nuclear genes** Ribosomal proteins Mitochondrial genes

Species: *Homo sapiens*
Taxonomy ID: 9606
Assembly: GCF_000001405.39 | GRCh38.p13

Genetic code: 1
Number of genes: 19850
Number of codons: 11577026

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

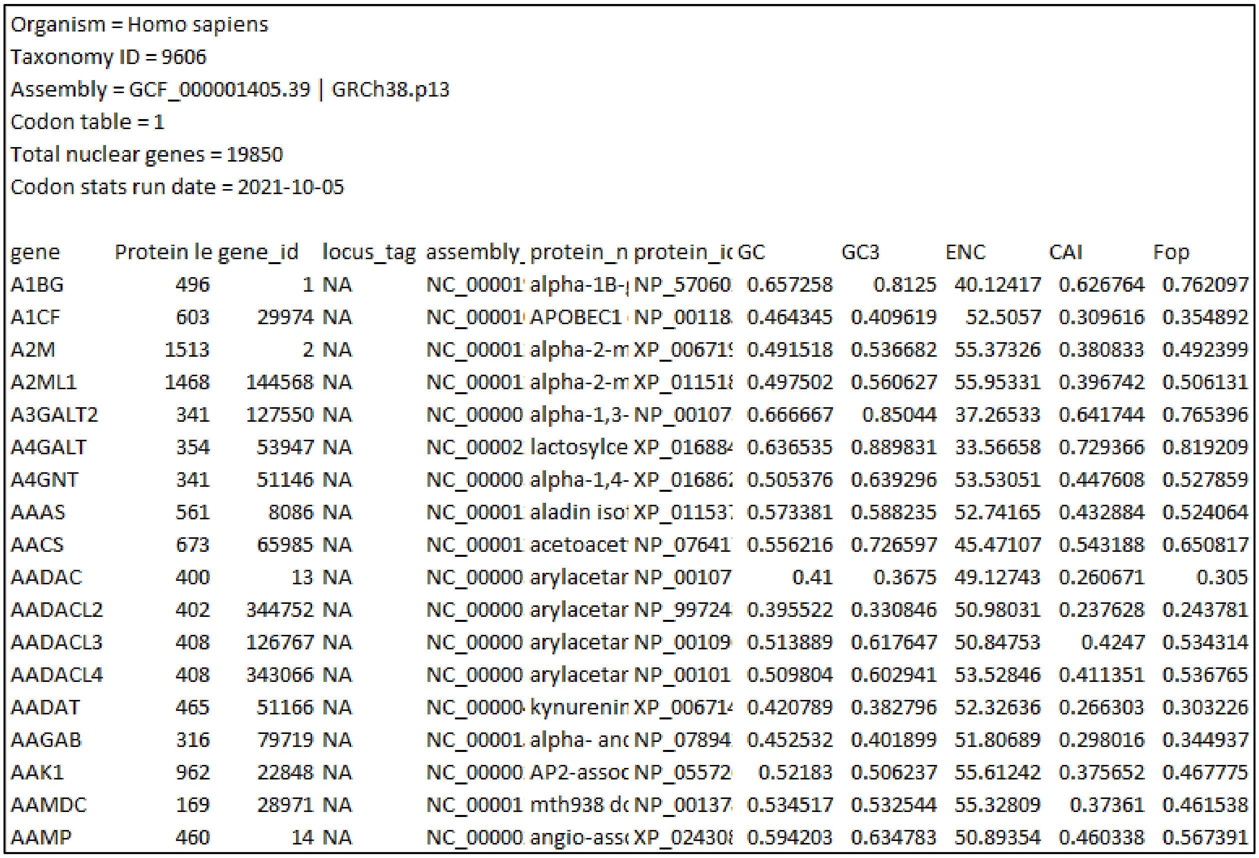
Figure 2. Summary codon statistics for human. This visualization corresponds to all nuclear genes. Only the first lines are shown.

For each species, the user can access codon statistics for the following gene sets (if such gene sets are available in the corresponding genome assembly):

- All nuclear protein-coding genes. In the case of prokaryotes and viruses, this category is substituted by an "All genes" category.
- Nuclear genes encoding ribosomal proteins. Nuclear genes whose descriptions included the substrings "Ribosom" or "ribosom" were included in this category. These genes are particularly interesting since they're highly expressed on average, and thus expected to exhibit a high level of codon bias. The preferred/unpreferred codons listed in this table are the same as the ones for all nuclear genes.
- Mitochondrial genes. Since mitochondria contain less than 1000 genes, information on preferred/unpreferred codons is not included.
- Chloroplast genes. Since chloroplasts contain less than 1000 genes, information on preferred/unpreferred codons is not included.

The tables can be visualized online (Fig. 2) or downloaded as a tab-delimited (.tsv) file by using the "Download codon stats" button. For each dataset, the user can also push the "Download gene stats" button to download a tab-delimited file with the following statistics for each gene:

- Gene symbol.
- Protein length (number of amino acids).
- NCBI gene ID.
- Locus tag.
- Assembly unit.
- Protein name.
- Protein ID.
- GC content for the entire CDS.
- GC content at third codon positions (GC3).
- Effective number of codons (ENC), as described by Wright (1990). This statistic is low for genes with strong codon bias, and thus negatively correlates with expression levels.
- Codon Adaptation Index (CAI), as described by Sharp and Li (1987). This statistic is high for genes with strong codon bias, and thus positively correlates with expression levels. Calculation of this statistic requires a set of highly expressed genes. For that purpose, we used genes with a low ENC (bottom decile).
- Frequency of optimal codons (F_{op}): the fraction of codons that are preferred (as described above).



Organism = Homo sapiens
Taxonomy ID = 9606
Assembly = GCF_000001405.39 | GRCh38.p13
Codon table = 1
Total nuclear genes = 19850
Codon stats run date = 2021-10-05

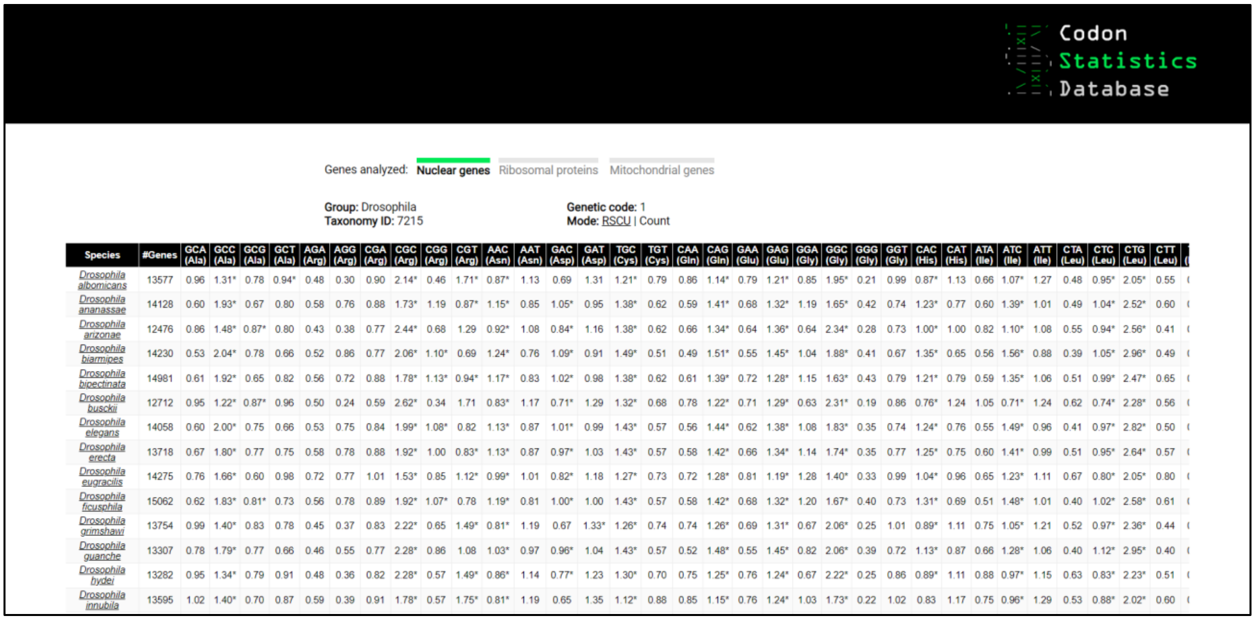
gene	Protein le	gene_id	locus_tag	assembly_	protein_n	protein_i	GC	GC3	ENC	CAI	Fop
A1BG	496	1	NA	NC_00001	alpha-1B-	NP_57060	0.657258	0.8125	40.12417	0.626764	0.762097
A1CF	603	29974	NA	NC_00001	APOBEC1	NP_00118	0.464345	0.409619	52.5057	0.309616	0.354892
A2M	1513	2	NA	NC_00001	alpha-2-m	XP_00671	0.491518	0.536682	55.37326	0.380833	0.492399
A2ML1	1468	144568	NA	NC_00001	alpha-2-m	XP_01151	0.497502	0.560627	55.95331	0.396742	0.506131
A3GALT2	341	127550	NA	NC_00000	alpha-1,3-	NP_00107	0.666667	0.85044	37.26533	0.641744	0.765396
A4GALT	354	53947	NA	NC_00002	lactosylce	XP_01688	0.636535	0.889831	33.56658	0.729366	0.819209
A4GNT	341	51146	NA	NC_00000	alpha-1,4-	XP_01686	0.505376	0.639296	53.53051	0.447608	0.527859
AAAS	561	8086	NA	NC_00001	aladin iso	XP_01153	0.573381	0.588235	52.74165	0.432884	0.524064
AACS	673	65985	NA	NC_00001	acetoacet	NP_07641	0.556216	0.726597	45.47107	0.543188	0.650817
AADAC	400	13	NA	NC_00000	arylacetar	NP_00107	0.41	0.3675	49.12743	0.260671	0.305
AADACL2	402	344752	NA	NC_00000	arylacetar	NP_99724	0.395522	0.330846	50.98031	0.237628	0.243781
AADACL3	408	126767	NA	NC_00000	arylacetar	NP_00109	0.513889	0.617647	50.84753	0.4247	0.534314
AADACL4	408	343066	NA	NC_00000	arylacetar	NP_00101	0.509804	0.602941	53.52846	0.411351	0.536765
AADAT	465	51166	NA	NC_00000	kynurenin	XP_00671	0.420789	0.382796	52.32636	0.266303	0.303226
AAGAB	316	79719	NA	NC_00001	alpha- an	NP_07894	0.452532	0.401899	51.80689	0.298016	0.344937
AAK1	962	22848	NA	NC_00000	AP2-asso	NP_05572	0.52183	0.506237	55.61242	0.375652	0.467775
AAMDC	169	28971	NA	NC_00001	mth938 d	NP_00137	0.534517	0.532544	55.32809	0.37361	0.461538
AAMP	460	14	NA	NC_00000	angio-ass	XP_02430	0.594203	0.634783	50.89354	0.460338	0.567391

Figure 3. Gene codon statistics for human nuclear genes. Only the first lines are shown.

3. Taxonomic group output

If a taxonomic group with multiple species has been selected (e.g., the genus "*Drosophila*" or the order "Primates"), a comparative table for all species is shown. For each species and codon, the count or RSCU is shown. Preferred codons (as defined above) are marked with an asterisk.

The user can select among: (1) all nuclear genes (or all genes in the case of prokaryotes and viruses), (2) nuclear genes encoding ribosomal proteins, (3) mitochondrial genes, and (4) chloroplast genes (if such gene sets are available).



Group: Drosophila
Taxonomy ID: 7215

Genetic code: 1
Mode: RSCU | Count

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	14981	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

Figure 4. Summary gene statistics for all species in the genus Drosophila. RSCU values for nuclear genes are shown.

4. References

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

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