J Mol Evol (2003) 56:691–701
DOI: 10.1007/s00239-002-2437-7

JOURNAL OF MOLECULAR EVOLUTION

# Codon Usage Patterns in Cytochrome Oxidase I Across Multiple Insect Orders

**Joshua T. Herbeck,[1] John Novembre[2]**

[1] Division of Insect Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[2] Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA

Received: 13 December 2002 / Accepted: 16 December 2002

**Abstract.** Synonymous codon usage bias is determined by a combination of mutational biases, selection at the level of translation, and genetic drift. In a study of mtDNA in insects, we analyzed patterns of codon usage across a phylogeny of 88 insect species spanning 12 orders. We employed a likelihood-based method for estimating levels of codon bias and determining major codon preference that removes the possible effects of genome nucleotide composition bias. Three questions are addressed: (1) How variable are codon bias levels across the phylogeny? (2) How variable are major codon preferences? and (3) Are there phylogenetic constraints on codon bias or preference? There is high variation in the level of codon bias values among the 88 taxa, but few readily apparent phylogenetic patterns. Bias level shifts within the lepidopteran genus *Papilio* are most likely a result of population size effects. Shifts in major codon preference occur across the tree in all of the amino acids in which there was bias of some level. The vast majority of changes involves double-preference models, however, and shifts between single preferred codons within orders occur only 11 times. These shifts among codons in double-preference models are phylogenetically conservative.

**Key words:** Cytochrome oxidase I — Codon bias — Phylogeny

*Correspondence to:* Joshua Herbeck, JBP Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA; *email:* herbeck@mbl.edu

## Introduction

Codon bias is the unequal usage of synonymous codons within amino acid families. The pattern was first described by Grantham et al. (1981) and has been ascribed to several potential causes. First, codon usage can be influenced by genome compositional constraints and mutational biases, as seen in mammalian (D'Onofrio et al. 1991; Karlin and Mrazek 1996; but see Smith and Eyre-Walker 2001), protozoan (Musto et al. 1999), and endosymbiotic bacterial (Wernegreen and Moran 1999) genomes. Second, selection among synonymous codons for either translational efficiency or translational accuracy can result in bias, as seen in bacterial, fungal, and insect genes (Ikemura 1982; Sharpe and Cowe 1991; Powell and Moriyama 1997). This selection is a result of the relationship between local tRNA abundance and major codon preference, in which a particular codon of an amino acid family pairs most optimally with the most abundant tRNA (Bulmer 1988; Ikemura 1992). Such optimal pairing of codons and tRNAs will increase translation speed (translational efficiency) and decrease amino acid misincorporation (translational accuracy) (Akashi 1994). Third, codon usage can be determined by a combination of mutational biases, selection, and genetic drift (Bulmer 1991; Sharp and Li 1986), in what is known as the mutation–selection–drift theory of codon bias.

Comparing codon usage patterns among species and among genes, including bias levels and major codon preferences, can help in clarifying the causes underlying observed codon bias. While bias levels are believed to be generally conserved across closely related taxa (Powell and Moriyama 1997), exceptions have been noted in *Drosophila* and related to population size differences among species (Akashi 1995). As the mutation–selection–drift theory of bias posits that genetic drift and selection pressure on major codons will be determinants of bias levels (Akashi 1997, Bulmer 1991), population size differences among species may result in variation in bias. As well, differences in mutational biases among species may result in variation in levels of bias. In comparing bias levels among taxa, then, care must be taken not to accredit all noticeable variation to selective differences. Major codon preferences are seen to be conserved across related taxa (Kreitman and Antezana 1999) but do shift across distantly related lineages (e.g., Sharp 1989). Whether variation in codon preferences across taxa is due primarily to selection or mutational biases is also unclear. This question can be resolved with polymorphism and divergence data for preferred and unpreferred mutations (Akashi 1995), but such an analysis is not undertaken here with cytochrome oxidase I (COI).

While studies of codon bias are numerous, most studies examine multiple genes or genomes with limited phylogenetic sampling, involving instead only pairwise or several taxon comparisons (e.g., Akashi 1995; Duret and Mouchiroud 1999; Morton 1998). One aim of the present work is to examine the codon usage patterns in a single gene across a broad phylogeny of 88 insect species from 12 orders. This approach allows us to ask the following questions, among others: (1) How variable are codon bias levels across a phylogeny? (2) How variable are major codon preferences across a phylogeny? and (3) Are there phylogenetic constraints on codon bias or codon preference? Based on predictions from the mutation–selection–drift theory of bias, single-gene codon bias levels may vary little across all taxa, assuming relatively limited effects of drift and similar selection pressures on the homologous genes, while codon preferences may vary among distantly related orders but should be conserved in closely related species (Powell and Moriyama 1997).

COI is a mitochondrial gene approximately 1.5 kb in length, encoding a polypeptide subunit of cytochrome *c* oxidase, the terminal enzyme in the respiratory chain. COI is widely used in insect phylogenetics, as its rate of nucleotide evolution allows it to resolve evolutionary histories at the family, genus, and species levels (Caterino et al. 2000). Popular though COI may be for specific phylogenetic questions, however, its nucleotide characteristics and evolutionary dynamics have been only cursorily examined across broader evolutionary levels (Lunt et al. 1996). A study of particular aspects of COI's evolution, such as codon usage, may prove beneficial to the phylogenetic community, if increased knowledge of COI leads to more informed use of COI sequences in phylogenetic analyses. As well, the large existing database of COI sequences spanning a wide evolutionary range of taxa is an ideal resource for studies of molecular evolution.

The present study of codon usage employs a measure of codon bias that removes the effects of genome nucleotide composition on estimated values of codon bias. Removing the effects of composition bias is an attempt to isolate the variation among taxa to that which must necessarily have either drift- or selection-based explanations. Variations in mutational biases among species, beyond those reflected in the background compositional biases, may exist, however, and this possibility is also considered.

## Methods

### *Insect Phylogeny*

The phylogeny used in this study includes 88 taxa from 12 insect orders (Table 1). We chose taxa for inclusion in the phylogeny and subsequent sequence analysis based on two criteria: (1) they increased the breadth of taxonomic sampling, and (2) there was a COI sequence of at least 700 bp (the full COI sequence is 1.53 kb) available in GenBank. Of the 88 sequences included in this study, three are under 1 kb, seven are between 1 and 1.5 kb, and 78 are the full 1.53 kb. The phylogeny is a composite tree based on hypothesized insect evolutionary relationships described in the following publications: Diptera, McAlpine and Wood (1989); Lepidoptera, Caterino and Sperling (1999) and Kristensen (1997); Hymenoptera, Dowton et al. (1994); Coleoptera, Kukalova-Peck and Lawrence (1993); and insect orders, Whiting et al. (1997).

The phylogeny contains two disparate levels of evolutionary divergence. The greater level of divergence is represented by the broad scope of the full phylogeny, including orders from Odonata to Diptera. The lesser level of divergence is represented by extensive sampling in the Lepidoptera, including 23 taxa from one species group, *Papilio*. This allows us to study relative variation in codon bias and preference at different divergence levels.

### *Codon Usage*

When measuring bias in synonymous codon usage, it is important to account for variation in background nucleotide composition among taxa. None of the commonly used codon bias measures, such as "effective number of codons" (ENC), CAI, and FOP, control for nucleotide composition. There is the scaled $\chi^2$ (Shields et al. 1988), as well as its modifications (Akashi and Schaeffer 1997; Wernegreen and Moran 1999), and our method is a likelihood-based extension of those methods that also allows for the determination of separate codon preference models. Two methods, MCB (Urrutia and Hurst 2001) and ENC′ (Novembre 2002), do control for background nucleotide composition but do not include a likelihood-based identification of codon preference.

To accomplish this we generated a null distribution of codon usage for each taxon that is based on nucleotide composition. Background nucleotide composition is at best a rough estimate of the diverse set of mutational biases ongoing across a genome and within a gene. While our method explicitly accounts for only background nucleotide content, in our analyses and interpretation we use nucleotide composition as an approximation for mutational biases. Given the phylogenetic scale of this study it was unfeasible to include, for each taxon, parameters of mutational biases such as

**Table 1.** Taxa included in this study, with corresponding total bias, ENC, GC content, and GenBank accession number

| Taxon | Total bias | ENC | GC% | Accession No. |
|---|---|---|---|---|
| Odonata | | | | |
| *Lybella cyanea* | 2.132 | 32.102 | 0.33878 | AF195739 |
| Orthoptera | | | | |
| *Gryllus ovisopis* | 3.119 | 29.950 | 0.31096 | U88333 |
| *Gryllus pennsylvanicus* | 3.069 | 29.918 | 0.31011 | U88332 |
| *Locusta migratoria* | 4.015 | 29.675 | 0.30915 | NC_001712 |
| *Chorthippus parallelus* | 3.726 | 31.140 | 0.30458 | X95574 |
| Dictyoptera | | | | |
| *Blatella germanica* | 3.132 | 35.714 | 0.31238 | S72627 |
| Phasmatodea | | | | |
| *Anisomorpha buprestoidea* | 2.050 | 26.949 | 0.28550 | AF005347 |
| *Timema bartmani* | 1.910 | 39.430 | 0.34091 | AF005331 |
| Hemiptera | | | | |
| *Triatoma sordida* | 2.668 | 39.714 | 0.39987 | AF021213 |
| *Panstrongylus megistus* | 2.279 | 34.000 | 0.33333 | AF021182 |
| *Limnoporus rufoscutellatus* | 3.726 | 29.874 | 0.29304 | LRU83338 |
| Phthiraptera | | | | |
| *Heterodoxus macropus* | 2.088 | 33.931 | 0.26900 | NC_002651 |
| Coleoptera | | | | |
| *Tetraopes sublaevis* | 3.362 | 30.815 | 0.27495 | AF267482 |
| *Phaea mirabilis* | 3.572 | 29.418 | 0.28346 | AF267468 |
| *Xylosandrus mancus* | 2.879 | 35.477 | 0.34749 | AF187143 |
| *Neochlamisus platani* | 1.365 | 45.783 | 0.38365 | AF093367 |
| *Aleochara heeri* | 2.825 | 32.774 | 0.30433 | 13169439 |
| *Drusilla canaliculata* | 3.059 | 31.834 | 0.29809 | 13161778 |
| Hymenoptera | | | | |
| *Apis mellifera* | 2.214 | 28.183 | 0.24575 | NC_001566 |
| *Wiebesia punctatae* | 2.242 | 30.054 | 0.24587 | AF200414 |
| *Formica exsecta* | 1.242 | 32.589 | 0.28337 | AB010927 |
| *Formica fusca* | 1.432 | 32.506 | 0.28029 | AB010925 |
| *Apanteles nephoptericis* | 1.219 | 31.503 | 0.24606 | AF102720 |
| *Pholotesor bedellidae* | 1.913 | 28.255 | 0.23696 | AF102715 |
| Trichoptera | | | | |
| *Hesperophylax designatus* | 1.387 | 41.483 | 0.33333 | AF375614 |
| *Lepidostoma ojanum* | 2.316 | 35.403 | 0.31336 | AF375612 |
| Lepidoptera | | | | |
| *Feltia herilis* | 2.636 | 27.714 | 0.27071 | U60991 |
| *Lambdina fiscellaria* | 2.808 | 31.857 | 0.28366 | AF064521 |
| *Choristoneura fumiferana* | 3.280 | 28.851 | 0.28463 | LI9098 |
| *Choristoneura rosaceana* | 3.355 | 30.632 | 0.27291 | LI9099 |
| *Ostrinia nubilalis* | 3.064 | 28.941 | 0.28432 | AF170853 |
| *Bombyx mori* | 3.779 | 26.954 | 0.27240 | AF149768 |
| *Hemileuca electra* | 2.385 | 32.029 | 0.29117 | AF170856 |
| *Macrosoma* sp. | 2.745 | 30.399 | 0.27273 | AF170854 |
| *Coenonympha tullia* | 2.552 | 33.197 | 0.29739 | AF170860 |
| *Boloria epithone* | 2.629 | 30.446 | 0.27898 | AF170862 |
| *Colias eurytheme* | 2.428 | 32.837 | 0.27974 | AF044024 |
| *Pieris napi* | 2.315 | 30.394 | 0.28431 | AF170861 |
| *Pyrgus communis* | 3.428 | 28.228 | 0.26536 | AF170857 |
| *Erynnis tristic* | 3.131 | 29.810 | 0.28308 | AF170858 |
| *Euphilotes bernardino* | 2.669 | 29.723 | 0.27321 | AF170864 |
| *Apodemia mormo* | 3.060 | 29.327 | 0.26732 | AF170863 |
| *Baronia brevicornis* | 3.832 | 26.649 | 0.26993 | AF170865 |
| *Eurytides marcellus* | 2.671 | 30.371 | 0.26667 | AF044022 |
| *Troides helena* | 3.257 | 30.833 | 0.28431 | AF170878 |
| *Iphiclides podalirius* | 3.258 | 29.301 | 0.26144 | AF170873 |
| *Allancastria cerisvi* | 3.591 | 28.471 | 0.27974 | AF170869 |
| *Parnassius clodius* | 3.334 | 28.584 | 0.26470 | AF170871 |
| *Battus philenor* | 2.976 | 29.639 | 0.28693 | AF170875 |
| *Sericinus montela* | 3.641 | 28.352 | 0.27255 | AF170868 |
| *Pachliopta neptunus* | 3.287 | 29.587 | 0.27516 | AF044023 |
| *Parides alcinous* | 3.406 | 29.922 | 0.28040 | AF170876 |
| *Papilio constantinus* | 3.115 | 31.008 | 0.28366 | AF044002 |
| *P. cresphontes* | 3.282 | 30.365 | 0.28628 | AF044004 |

**Table 1.** Continued

| Taxon | Total bias | ENC | GC% | Accession No. |
|---|---|---|---|---|
| *P. glaucus* | 2.875 | 29.123 | 0.27516 | AF044013 |
| *P. hospiton* | 3.440 | 29.393 | 0.26340 | AF044009 |
| *P. indra* | 2.895 | 30.091 | 0.26862 | AF044011 |
| *P. xuthus* | 1.964 | 32.542 | 0.28300 | AF043999 |
| *P. zelicaon* | 2.615 | 29.778 | 0.27059 | AF044008 |
| *P. polyxenes* | 2.726 | 30.231 | 0.27516 | AF044010 |
| *P. canadensis* | 2.281 | 28.569 | 0.27190 | AF044014 |
| *P. garamus* | 2.234 | 32.559 | 0.29281 | AF044021 |
| *P. oregonius* | 2.638 | 30.255 | 0.27320 | AF044007 |
| *P. multicaudatus* | 3.500 | 28.564 | 0.27124 | AF044016 |
| *P. rutulus* | 2.513 | 32.557 | 0.27648 | AF044015 |
| *P. dardanus* | 2.279 | 32.747 | 0.29608 | AF044003 |
| *P. palamedes* | 3.462 | 31.116 | 0.29608 | AF044018 |
| *P. scamander* | 2.760 | 31.116 | 0.29020 | AF044020 |
| *P. anchisiades* | 2.867 | 28.161 | 0.27909 | AF044005 |
| *P. demoleus* | 3.460 | 29.583 | 0.27517 | AF044000 |
| *P. troilus* | 2.570 | 30.096 | 0.29216 | AF044017 |
| *P. phorcas* | 2.665 | 33.216 | 0.30000 | AF044001 |
| *P. alexanor* | 2.173 | 32.985 | 0.28627 | AF044012 |
| *P. machaon* | 2.626 | 30.341 | 0.27320 | AF044007 |
| *P. pilumnus* | 2.742 | 32.518 | 0.30065 | AF044019 |
| Mecoptera | | | | |
| *Panorpa vulgaris* | 1.768 | 28.656 | 0.26826 | AF180105 |
| *Panorpedes paradoxus* | 2.616 | 26.997 | 0.27932 | AF180100 |
| Diptera | | | | |
| *Chrysomya chlorpyga* | 5.384 | 28.876 | 0.31111 | AF295554 |
| *Cochliomyia macellaria* | 5.497 | 29.379 | 0.31438 | AF295555 |
| *Anopheles gambiae* | 4.195 | 31.350 | 0.31360 | NC_002084 |
| *Drosophila melanogaster* | 4.142 | 28.351 | 0.29100 | NC_001709 |
| *Drosophila vakuba* | 3.819 | 29.672 | 0.30131 | X03240 |
| *Drosophila simulans* | 3.181 | 29.121 | 0.30248 | AF200854 |
| *Apocephalus paraponerae* | 2.607 | 28.893 | 0.28482 | AF217481 |
| *Ceratitis capitata* | 4.416 | 31.312 | 0.30653 | AJ242872 |
| *Scathophaga stercoraria* | 4.570 | 27.719 | 0.28954 | AF104625 |
| *Oestrus ovis* | 2.182 | 32.607 | 0.30648 | AF257118 |
| *Gastrophilus intestinalis* | 2.018 | 40.960 | 0.35782 | AF257117 |

transition/transversion bias or dinucleotide biases, alongside the contigent background nucleotide content, in the likelihood-based estimation of codon bias levels. The removal of background nucleotide content to isolate codon bias possibly due to selection has been attempted (Urrutia and Hurst 2001), but it is important to state that codon bias values estimated using these types of methods cannot automatically be interpreted as adaptive bias. There may be alternate explanations, including complex mutational dynamics and genetic drift, that can explain codon bias variation. Yet these methods are substantial steps in the study of codon bias variation, and, in this study particularly, the study of codon bias variation in a single gene across a broad phylogeny. The popular measure of codon bias ENC (Wright 1990) does not account for mutational bias and, therefore, is not suitable for comparing bias levels among species.

A null distribution based on nucleotide composition can be used as the basis for $\chi^2$ tests. In addition, extensions of the null model can be made that incorporate parameters describing the amount of bias for each codon. Maximum likelihood estimates (MLEs) of the bias parameters can be obtained for each model, and the models can be compared using likelihood-ratio tests. A similar approach that uses GC content to generate the null distribution is described by Slatkin and Novembre, (2003). The method in this analysis is a slight modification, as it includes relative percentages of all four nucleotides rather than solely GC and AT.

The null distribution is generated by considering what nucleotides make up the codon sequences of a synonymous codon family. For a $k$-fold ($k = 2, 3, 4, 6, 8$) degenerate amino acid, we have an expected proportion $e_i$ ($i = 1,\ldots,k$) for each codon in a synonymous codon family. We label the four nucleotide compositions (expressed as a proportion) $f_A$, $f_C$, $f_G$, and $f_T$. For a codon $i$ that has the sequence XYZ, the corresponding expected frequency $e_i = f_X f_Y f_Z / C$, where $C$ is a renormalization constant that ensures that the sum of the $e_i$ for an amino acid equals one.

Expected numbers for each codon are obtained by multiplying each $e_i$ by the total number of observed codons $n$. These can be compared to the observed numbers of each codon, $n_i$, using a standard $\chi^2$ test with $k - 1$ degrees of freedom and a 5% significance level. If the null model is not rejected, the number of overrepresented codons for the amino acid in question is set to zero. If it is rejected, more elaborate models are explored to estimate the number of overrepresented codons.

The first model, *the single-preference model*, augments the expected frequency of the overrepresented codon by a parameter $b$ and subtracts $b/(k - 1)$ from the remaining $e_i$. This corresponds to having one codon being overrepresented while all others are under represented relative to their expected proportions. The overrepresented codon is chosen to be the one with the highest value of $(n_i/n) - e_i$ (i.e., the one that is most overrepresented). By assuming that the data are drawn from a multinomial sample of size $n$ with expected proportions $e_i$, the likelihood of the data can be obtained for all of the models considered here. Under the single-preference model, the likeli-

**Table 2.** Tests for equal variances and means, for total bias and ENC between the 49 lepidopteran taxa and the 39 other insect taxa

| | Lepidoptera | All other taxa |
|---|---|---|
| Total bias | | |
| Mean | 2.922 | 2.855 |
| Variance | 0.212 | 1.216 |
| *N* | 49 | 39 |
| *F* test for variances | | |
| *F* | 5.740 | |
| $p$ $(F \leq f)$, one-tail | 1.643 E-08 | |
| *F* critical, one-tail | 1.652 | |
| *Reject null of equal variances for total bias* | | |
| *t* test assuming unequal variances | | |
| *t*-stat | −0.359 | |
| $p$ $(T \leq t)$, one-tail | 0.361 | |
| *t* critical, one-tail | 1.677 | |
| $p$ $(T \leq t)$, two-tail | 0.722 | |
| *t* critical, two-tail | 2.010 | |
| *Do not reject null of equal means for total bias* | | |
| | ENC | |
| Mean | 30.203 | 32.112 |
| Variance | 2.723 | 18.825 |
| *N* | 49 | 39 |
| *F* test for variances | | |
| *F* | 6.913 | |
| $p$ $(F \leq f)$, one-tail | 6.788 E-10 | |
| *F* critical, one-tail | 1.652 | |
| *Reject null of equal variances for ENC* | | |
| *t* test assuming unequal variances | | |
| *t*-stat | −0.359 | |
| $p$ $(T \leq t)$, one-tail | 0.361 | |
| *t* critical, one-tail | 1.677 | |
| $p$ $(T \leq t)$, two-tail | 0.722 | |
| *t* critical, two-tail | 2.010 | |
| *Reject null of equal means for ENC* | | |

hood of the data under this model is a function of $b$ and a MLE of $b$ is obtained by maximizing the likelihood function with a grid search.

The second model, *the double-preference model*, is applied only to amino acids that are greater than twofold redundant ($k > 2$). The two codons with the largest value of $(n_i/n) - e_i$ are identified and given the expected proportions $e_i + b_1$ and $e_j + b_2$, while the remaining codons are assigned proportions $e_i - ((b_1 + b_2)/2)$. MLE estimates of $b_1$ and $b_2$ are found using the same methods as for the single-preference model.

Likelihood-ratio tests are then used to determine which model fits the data best. The likelihood of the null $L_0$ is tested against the maximum likelihood of the single-preference model, $L_1$, and the double-preference model, $L_2$. The likelihood-ratio test statistic $R = -2In(L_m/L_n)$ was assumed to be $\chi^2$ distributed, with degrees of freedom equal to the difference in the number of free parameters between model $m$ and model $n$ ($m = 1, 2$; $n = 0, 1$). If $L_2$ was found to be significantly greater than $L_1$ or $L_0$, then the double-preference model was accepted. If $L_2$ was not significantly greater than $L_1$, and $L_1$ was significantly greater than $L_0$, then the single-preference model was accepted. If neither the single- nor the double-preference model had a significantly higher likelihood than the null model, the null model was accepted. Finally, in cases where the double-preference model was accepted, but the $b_1$ value was 10 times greater than the $b_2$ value, we chose the single-preference model to describe the data.

For each amino acid, this analysis yields a description of the preference model, overrepresented codons, and bias parameters. A summary statistic describing the overall amount of codon bias for each taxon (hereafter labeled "*total bias*" [TB] can be constructed by summing the bias parameters for each amino acid's chosen preference model over all amino acids.

In our use of the method we have assumed that the nucleotide composition of each COI sequence in question is closely related to the overall nucleotide composition of each respective mitochondrial genome. Until more fully sequenced mitochondrial genomes are available, this method is forced to use the nucleotide composition of single genes as a proxy for the genome composition (which, in turn, is a best approximation for mutational biases, as described above). We believe that this use of COI nucleotide frequencies as an approximation of genome nucleotide frequencies is reliable. For 11 taxa (*Hetrodoxus*, *Locusta*, *Bombyx*, *Apis*, *Drosophila melanogaster*, *Drosophila simulans*, *Ceratitis*, *Cochliomyia*, *Chrysomya*, *Triatoma*, and *Anopheles*), fully sequenced mitochondrial genomes were available in GenBank. For these taxa we compared genomic GC frequencies to COI GC frequencies and found a high correlation ($\rho = 0.85$, $p = 0.0013$) between the two estimates.

The methods described here are implemented in a program entitled *biasml* that is available upon request from the authors.

## *Comparative Analyses*

Qualitative patterns of codon bias and major codon preference were studied in the insect phylogeny using MacClade version 3.06 (Maddison and Maddison 1996). To distinguish objectively among low, medium, and high levels of TB across all taxa, we used the following criteria: *low bias*, < (mean TB − 1 SD); *high bias*, > (mean TB + 1 SD); and *medium bias*, all remaining values. This

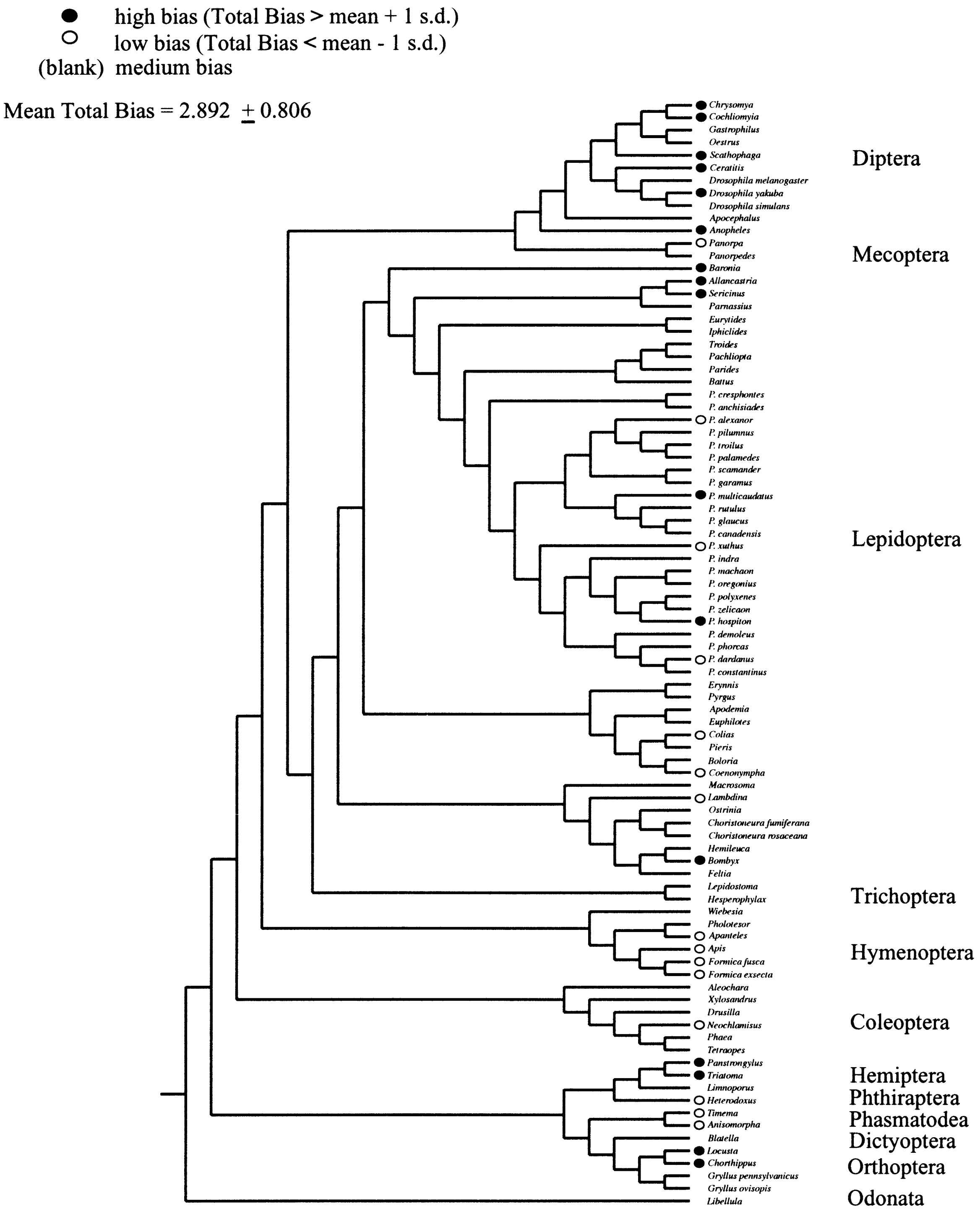

**Fig. 1.** Insect composite phylogeny, with relative total bias levels mapped on. The phylogeny is based on hypothesized insect relationships culled from the following papers: for Diptera, McAlpine and Wood (1989); for Lepidoptera, Caterino and Sperling (1999) and Kristensen (1997); for Hymenoptera, Dowton and Austin (1994); for Coleoptera, Kukalova-Peck and Lawrence (1993); and for the insect orders, Whiting et al. (1997).

method of defining bias values was necessary, as a predicted distribution of TB values among gene sequences of varying bias levels is unknown. As well as TB, we estimated levels of codon bias in COI using ENC (Wright 1990) and ENC′ (Novembre 2002).

Lineage-specific synonymous substitution rates, $d_S$, were calculated across the phylogeny using the program *codeml* in the PAML software package (Yang 2000). Correlations among TB, ENC, ENC′, and $d_S$ were calculated.

## Results

### Codon Bias

There is high variation in the level of codon bias values among the 88 taxa. The mean TB value is 2.892, with a standard deviation of 0.806. The values

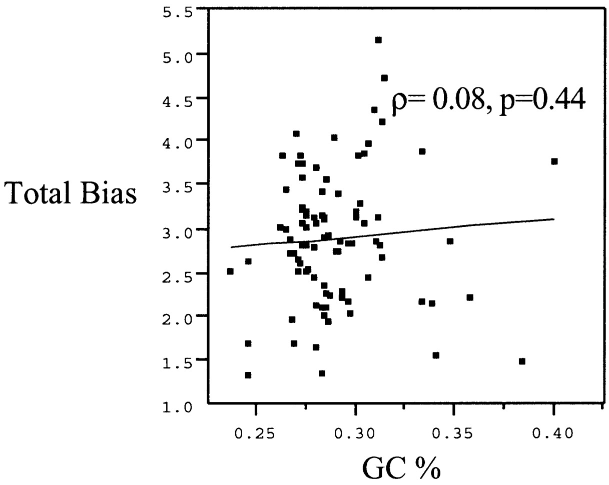

**Fig. 2.** Correlation of GC% with total bias.

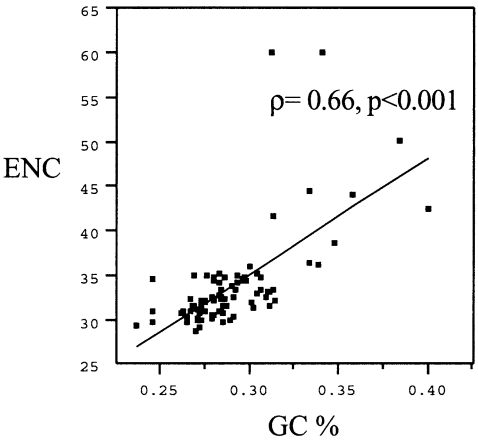

**Fig. 3.** Correlation of GC% with ENC (Wright 1990).

range from a low of 1.219, in the hymenopteran species *Apanteles*, to a high of 5.497, seen in the dipteran species *Cochliomyia.* For comparison, ENC values range from 26.65 in the lepidopteran *Baronia* to 45.78 in the coleopteran *Neochlamisus*, with a mean of 31.05 and a standard deviation of 3.26. ENC′ is less variable than ENC, ranging from 37.61 in the dipteran *Chrysomya* to 40.96 in *Gastrophilus*, with a mean of 45.22.

The mean TB in the 49 lepidopteran taxa is 2.922, with a standard deviation of 0.460. The values range from 1.964 *in Papilio xuthus* to 3.832 in *Baronia.* ENC values range from the aforementioned 26.65 in *Baronia* to 33.22 in *Phorcas*, with a mean of 30.20 and a standard deviation of 1.65. Variances of TB and ENC are significantly different between the 49 Lepidoptera and the 39 other taxa; mean TB values are not significantly different between the Lepidoptera and the other taxa, while means of ENC are different (Table 2).

Mapping the three levels of codon bias onto the composite insect phylogeny reveals few apparent large-scale patterns (Fig. 1). However, two orders contain *low* bias values: the Phasmatidae, *Anisomorpha and Timemaj* and four of the six Hymenoptera. Three clades contain *high* bias values: the reduviid Hemiptera, *Triatoma* and *Panstrongylus*; the acridid Orthoptera, *Chorthippus* and *Locusta*; and six of the nine sampled Diptera.

The TB values show no correlation with the GC frequencies in the respective COI sequences ($\rho = 0.08$, $p = 0.44$) (Fig. 2). There is a significant correlation between ENC and GC content ($\rho = 0.66$, $p = 0.001$) (Fig. 3), supporting our decision to use our developed likelihood-based estimation, as comparisons among taxa using ENC will be affected by variations in contigent nucleotide composition.

Across the phylogeny, TB was not correlated with $d_S$ ($\rho = 0.05$, $p = 0.05$). ENC was correlated only slightly with $d_S$ ($\rho = 0.17$, $p = 0.0002$).

### *Major Codon Preference*

On average, only 3% (0.03 ± 0.005) of a species' overrepresented codons in this data set are G/C ending. The vast majority of overrepresented codons present, then, is A/T ending. In the three *Drosophila* species included, *melanogaster*, *yakuba*, and *simulans*, all overrepresented codons are A/T ending (Table 3). The A/T-ending overrepresented codons in COI of *Drosophila* correspond exactly to the tRNA anticodons for the respective amino acids, as described for the mitochondrial genome of *D. yakuba* (Clary and Wolstenholme 1985), in only 10 of the 19 amino acids with measurable bias in the *Drosophila* taxa. However, when wobble is considered, 17 of 19 amino acids correspond to the anticodon. The two amino acids not corresponding are methionine, the start codon, and lysine.

Shifts in major codon preference occur across the tree in all of the amino acids in which there was bias of some level. However, when looking only at preference changes *within* orders rather than *among* orders, the vast majority of changes involves a double-preference model. These include single- to double-preference, double-preference to double-preference, and double- to single-preference. All shifts are conservative; shifts from one to two codons always retain the previous single codon, shifts between double-preference models always retain one of the previous codons, and shifts from two to one major codon always retain one of the original codon pair.

Shifts between single overrepresented codons, within single orders, occur only 11 times across the phylogeny. We are defining shifts in this case as any difference between single-preference models within a single order. Sampling within orders is admittedly limited, including sister-species relationships only in the *Papilio*, and determining the exact point of major codons *shifts* is impossible; hence for

**Table 3.** Selected codon usage for the three *Drosophila* species used in this study, with corresponding total bias and ENC[a]

| | *D. melanogaster* | *D. simulans* | *D. yakuba* | *yakuba* mtDNA tRNA anticodon[b] |
|---|---|---|---|---|
| Total bias | 4.14 | 3.18 | 3.82 | |
| ENC | 28.35 | 29.12 | 29.67 | |
| Glycine | | | | |
| (GGA) | 40 | 33 | 35 | UUC |
| GGC | 0 | 0 | 0 | |
| GGG | 1 | 6 | 3 | |
| GGT | 6 | 8 | 9 | |
| Leucine | | | | |
| (CTA) | 7 | 3 | 2 | UAG |
| CTC | 0 | 0 | 0 | |
| CTG | 0 | 0 | 0 | |
| CTT | 4 | 3 | 7 | |
| (TTA) | 53 | 56 | 54 | UAA |
| TTG | 1 | 0 | 2 | |
| Tryptophan | | | | |
| (TGA) | 15 | 15 | 13 | UCA |
| TGG | 0 | 0 | 2 | |
| Proline | | | | |
| CCA | 9 | 8 | 9 | |
| CCC | 1 | 1 | 2 | |
| CCG | 1 | 1 | 1 | |
| (CCT) | 14 | 14 | 13 | UGG |
| Phenylalanine | | | | |
| TTC | 3 | 5 | 7 | GAA |
| (TTT) | 36 | 33 | 31 | |
| Histidine | | | | |
| CAC | 2 | 4 | 3 | GUG |
| (CAT) | 16 | 12 | 14 | |

[a] Major codons, as determined by the method described in this paper, are in parentheses.

[b] tRNA anticodons present in the mitochondrial genome of *D. yakuba*, as described by Clary and Wolstenholme (1985).

our purposes simple *differences* within orders are treated as equivalent to shifts. These shifts are often accompanied by extremes of bias, low or high, in the lineage in which the preference shift has occurred.

The majority of preference/model changes is homoplastic, although preference changes seen in select amino acids are correlated with the phylogeny. An example is the fourfold amino acid proline (Fig. 4). Codon usage in proline is not biased in any of the taxa examined. When bias does exist, the type of codon preference model (one or two codons overrepresented) varies across all levels of the phylogeny—as both models, as well as *no* bias, are seen within the Lepidoptera and the genus *Papilio*. Comparing preference *models* to major codon *preference* in proline reveals the following phylogenetic pattern: when *one* codon is overrepresented, that codon is CCT in Lepidoptera and Mecoptera and CCA in Coleoptera, Diptera, and Orthoptera; when *two* codons are overrepresented, they are CCC and CCT in Lepidoptera and CCA and CCT in Diptera and Dictyoptera. In the Hymenoptera there is a shift between single-preference models, and the genera *Apis* and *Wiebesia* have extremely low bias levels.

## Conclusions

We conducted comparative analyses of codon usage across 12 insect orders, with both broad and focused sampling. The maximum likelihood bias estimation procedure we used removes effects of background nucleotide bias on estimates of codon bias and appears to be a useful measure. Because this procedure is new, we are still exploring its properties. Results here are suggestive of patterns in codon usage evolution that deserve more exploration.

### *Codon Bias*

The extensive variation in codon bias levels seen across the insect orders is expected, given the known relative rates of COI evolution and probable distribution of variation in effects of genetic drift across such a broad expanse of insect taxa. Mapped onto the phylogeny, TB is not conserved or phylogenetically informative. The only obvious patterns of shared bias levels include the Diptera, with high bias, and the Hymenoptera, with low bias. Variation in bias is not as extensive within the Lepidoptera or within the genus *Papilio*. This is not

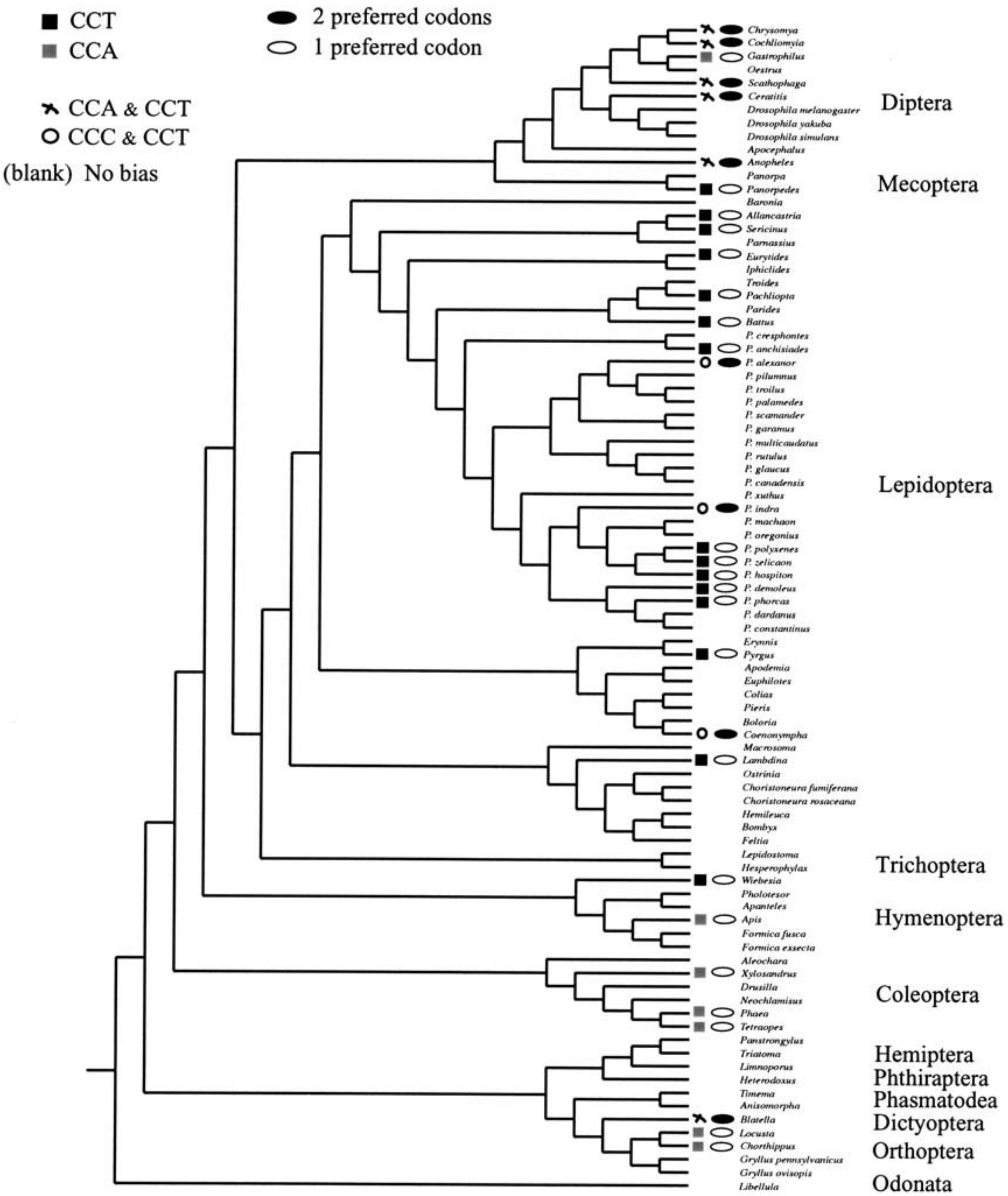

**Fig. 4.** Example of codon preference model evolution, using the fourfold amino acid proline.

surprising, as one would expect bias levels to be fairly uniform at this divergence level (Powell and Moriyama 1997).

Following predictions of the mutation–selection–drift theory of bias, then, we predict that the variation seen within Lepidoptera, and *Papilio* specifically, will be due to variation in effects of either genetic drift or strength of selection on codon bias. We assume that selection intensity on COI should not vary significantly at the genus level. While population sizes of *Papilio* butterflies are not believed to vary drastically (F. Sperling, personal communication), genetic drift may be initiated by any factors changing the effective population size in recent history. Lacking a clear understanding of *Papilio* populations, and following predictions of the mutation–selection–drift theory, we propose that the variation in codon bias within the genus is most likely a result of historical changes in effective population size rather than variation in selection on COI among *Papilio* species, while substantial changes in taxon-specific mutational biases may also play a role.

### *Codon Preference*

As well as variation in bias levels, there are shifts in codon overrepresentation across the insect phylogeny and within the Lepidoptera. The preference shifts within Lepidoptera are unexpected, given the supposedly constrained nature of overrepresented codons (Kreitman and Antezana 1999). Designating the *type* of codon preference shift is important, however, and preference shifts within lineages appear less dramatic when these are considered. Most shifts involve the double-preference model, and these shifts may be likened to general conservation of particular single major codons, with toggling of second major codons. Shifts among single-preference models within orders are extremely rare, occurring only 11 times across the 88-taxon phylogeny.

Accompanying these shifts among single-preference models are extremes of bias. As predicted by the mutation–selection–drift theory, genetic drift can impair the ability of selection to maintain a particular codon bias level; comparatively low codon bias may be explained by the action of genetic drift. Also predicted, and detailed by Kreitman and Antezana (1999), are the possible causes for codon preference shifts: preference shifts are either *allowed* by the inability of selection to maintain codon preference in the face of genetic drift or *caused* by high levels of selection for the novel preferred codon.

Looking at the preference shifts among insects, all shifts accompanied by *high* levels of bias can be attributed to selection. But preference shifts accompanied by *low* levels of bias cannot be automatically attributed to genetic drift (Kreitman and Antezana 1999). However, if the low bias is a result of genetic drift, the probability of major codon shifts within that lineage should increase, since drift should affect all amino acids. Within this insect phylogeny, the only preference shifts accompanied by high bias are seen in the Diptera, where *Chocliomyia* and *Chrysomya* have high bias and show preference shifts in three amino acids, histidine, phenylalanine, and threonine. The coleopteran *Neochlamisus* has very low bias and shows preference shifts in alanine, histidine, and phenylalanine. These three taxa are the only taxa with more than one preference shift. We may conclude that *Chocliomyia* and *Chrysomya* have selection-induced shifts, while *Neochlamisus* has drift-induced shifts. Other taxa have only single shifts, and with only one observed shift our ability to differentiate between selection and drift is limited. At this point we can only corroborate the expected rarity of preference shifts within (relatively) closely related species (Kreitman and Antezana 1999) and present the tendency for shifts to be accompanied by extreme—high or low—bias levels.

Preferred codons in the three *Drosophila* taxa are A/T ending, in direct contrast to the greater abundance of G/C-ending preferred codons in *Drosophila* nuclear genes (Moriyama and Hartl 1993; Shields et al. 1988). The overall correspondence of the *Drosophila* preferred codons to the tRNA anticodons, when considering wobble, supports the selection–mutation–drift theory of codon bias, and in fact the level of correspondence seen in COI is greater than that seen *in Drosophila* nuclear genes (Powell and Moriyama 1997), as would be predicted given mtDNA's higher level of conservation.

We present a method for estimating codon bias that controls for background nucleotide composition, as well as identifying codon families with multiple overrepresented codons. Our hope was to limit plausible explanations of bias differences among taxa to selection or genetic drift and to gain insight into possible mutational dynamics or drift effects by studying codon preference. The method is used to analyze codon usage patterns in COI across a broad insect phylogeny. Given the extensive variation in codon bias levels and codon preference seen in COI, future studies attempting to isolate causes for shifts in bias and preference across closely related taxa may be feasible.

*Acknowledgments.* We thank George Roderick, Montgomery Slatkin, Felix Sperling, and Dennis Wall for discussions and assistance. J.H. supported by NSF PEET Grant 9521835 to J. Powell and F. Sperling. J.N. was supported in part by NIH Grant GM40282 to M. Slatkin. We greatly appreciate the helpful comments of Martin Kreitman and two anonymous reviewers.

## References

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

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. Genetics 139:1067–1076

Akashi H (1997) Codon bias evolution in *Drosophila:* Population genetics of mutation-selection-drift. Gene 205:269–278

Akashi H, Schaeffer S (1997) Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics 146:295–307

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by selection-mutation balance. J Evol Biol 1:15–26

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

Caterino MS, Sperling FAH (1999) Papilio phylogeny based on mitochondrial cytochrome oxidase I and II genes. Mol Phylogenet Evol 11:122–137

Caterino MS, Cho S, Sperling FAH (2000) The current state of insect molecular systematics: A thriving Tower of Babel. Annu Rev Entomol 45:1–54

Clary DO, Wolstenholme DR (1985) The mitochondrial DNA molecule of *Drosophila yakuba:* Nucleotide sequence, gene organization and genetic code. J Mol Evol 22:252–271

D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol 32:504–510

Dowton M, Austin AD, Dillon N, Bartowsky E (1997) Molecular phylogeny of the apocritan wasps: The Proctotrupomorpha and Evaniomorpha. Syst Entomol 22:245–255

Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and Arabidopsis. Proc Natl Acad Sci USA 96: 4482–4487

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 10:7055–7074

Ikemura (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573–597

Karlin S, Mrazek J (1996) What drives codon choices in human genes? J Mol Biol 262:459–472

Kreitman M, Antezana M (1999) The population and evolutionary genetics of codon bias. In: Singh RS, Krimbas CB (eds) Evolutionary genetics: From molecules to morphology. Cambridge University Press, Cambridge, Vol 1, pp 83–101

Kristensen NP (1997) Early evolution of the Lepidoptera plus Trichoptera lineage: Phylogeny and the ecological scenario. Mem Mus Natl Hist Nat 173:253–271

Kukalova-Peck J, Lawrence JF (1993) Evolution of the hind wing in Coleoptera. Can Entomol 125:181–258

Lunt DH, Zhang DX, Szymura JM, Hewitt GM (1996) The insect cytochrome oxidase I gene: Evolutionary patterns and conserved primers for phylogenetic studies. Insect Mol Biol 5:153–165

Maddison WP, Maddison DR (1996) MacClade version 3.06. Sinauer Associates, Sunderland, MA

McAlpine JF (1989) Phylogeny and classification of the Muscomorpha. In: McAlpine JF, Wood DM (eds) Manual of nearctic Diptera 3. Research Branch, Agriculture Canada, Monograph, 32 pp 1397–1518

Moriyama EN, Hartl DL (1993) Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics 134:847–858

Morton B (1998) Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol 46:449–459

Musto H, Romero H, Zavala A, Jabbari K, Bernardi G (1999) Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum:* Compositional constraints and translational selection. J Mol Evol 49:27–35

Novembre J (2002) Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19:1390–1394

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

Sharp PM, Cowe E (1991) Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast 7:657–678

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

Sharp PM, Shields DC, Wolfe KH, Li WH (1989) Chromosomal location and evolutionary rate variation in enterobacterial genes. Science 246:808–810

Shields DC, Sharp PM, Higgins DG, Wright F (1988) 'Silent' sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Slatkin M, Novembre J (2003) Appendix to paper by Wall and Herbeck. J Mol Evol 56:689–690

Smith NGC, Eyre-Walker A (2001) Synonymous codon bias is not caused by mutation bias in G + C-rich genes in humans. Mol Biol Evol 18:982–986

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

Wernegreen JJ, Moran NA (1999) Evidence for genetic drift in endosymbionts (*Buchnera*): Analysis of protein-coding genes. Mol Biol Evol 16:83–97

Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC (1997) The strepsiptera problem: Phylogeny of the holometabolous insect order inferred from 18S and 28S ribosomal DNA sequences and morphology. Syst Biol 46:1–68

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

Yang Z (2000) Phylogenetic Analysis by Maximum likelihood (PAML), version 2.0. University College London, London