J Mol Evol (2003) 57:S214–S225
DOI: 10.1007/s00239-003-0030-3

JOURNAL OF MOLECULAR EVOLUTION

Analysis of a Shift in Codon Usage in *Drosophila*

Jeffrey R. Powell,[1] Erminia Sezzi,[1,*] Etsuko N. Moriyama,[2] Jennifer M. Gleason,[1,] Adalgisa Caccone[1,3]**

[1] Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520-8106, USA
[2] School of Biological Sciences, University of Nebraska, Lincoln, NE 68588-0660, USA
[3] Yale Institute for Biospherics Studies, Yale University, New Haven, CT 06520-8106, USA

Received: 21 August 2002 / Accepted: 11 March 2003

Abstract. In order to gain further insight into a shift in codon usage first observed in *Drosophila willistoni* we have analyzed seven genes in six species in the lineage leading to *D. willistoni*. This lineage contains the *willistoni* and *saltans* species groups. Sequences were obtained from GenBank or newly sequenced for this study. All species studied showed significant difference in codon usage compared to *D. melanogaster* for about one third of all amino acids. Within the *willistoni/saltans* lineage, codon usage is homogeneous, indicating that the shift in codon usage occurred prior to the diversification of extant species in this lineage which we estimate to date to about 20 million years ago. Thus the shift is old and has been stable. We also examined introns from these genes and the G/C composition at four-fold degenerate sites in an effort to detect a change in mutation bias. There is little or no evidence for a difference in mutation bias compared to *D. melanogaster*. We also considered whether relaxed selection (possibly due to reduced population sizes) or reduced recombination (due to numerous naturally occurring inversions) could account for the shift and concluded these factors alone are insufficient to explain the patterns observed. A change in the relative abundance of isoaccepting tRNAs is one of the few explanations that can account for the observations. Particularly intriguing is the fact that the greatest changes in codon usage have occurred for amino acids with two-fold C/T ending codons for which it is known that posttranscriptional modification occurs in tRNAs from a G in the wobble position to Queuosine that changes optimal binding from C to a slight preference for U. However, we do not argue that this shift was adaptive in nature, rather it may be an example of a "frozen accident."

Key words: Codon usage bias — *Drosophila willistoni* — *Drosophila saltans* — tRNA

Introduction

Codon usage bias is almost certainly the result of interactions among at least three evolutionary forces: mutation bias, genetic drift, and selection. The relative influence of these forces is determined by several factors, e.g., mutation mechanism, mutation rate, population size, relative abundance of isoaccepting tRNAs. Further, it is well-established that there is no single pattern of codon usage bias evident in all organisms. For example, *Drosophila melanogaster* favors C and G in the wobble position, while yeast (*S. cerevisiae*) favors A and T (Codon Usage Database, http://www.kazusa.or.jp/codon/). Thus, it is difficult to envision a physical chemical reason for certain codon preferences that would apply to all organisms sharing the (nearly) universal genetic code and very similar protein synthesis systems. However, it was generally thought that a particular pattern of codon usage was common to organisms within broad

**Present address:* Dipartimento di Scienze Ambientali, Universitá della Tuscia, Viterbo, Italy
***Present address:* Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS 66045-7534, USA
Correspondence to: J.R. Powell; *email:* jeffrey.powell@yale.edu

Table 1. Sequences used in this study with GenBank accession numbers and number of codons analyzed

	GenBank accession numbers (codons)[a]								
Species	*RpL32*	*Adh*	*Hsp83*	*Sod*	*Gpdh*	*Amyrel*	*Ddc*	*per*	*ry*
D. melanogaster	X00848 (97)	Z00030 (222)	X03810 (317)	Z19591 (126)	J04567 (255)	U69607 (494)	NM78876 (377)	M30114 (432/590)[c]	Y00308 (695)
D. yakuba/ D. teissieri[b]	NA	X54120 (222)	NA	AF127159 (126)	U47809 (255)	AF03956 (494)	NA	X61127 (418)	AF169401 (695)
D. pseudoobscura	S59382 (97)	Y00602 (222)	X03812 (317)	U47871 (107)	U59682 (255)	U82556 (495)	AF293746 (377)	X13878 (507/651)[c]	M33977 (695)
D. virilis	AJ245566 (81)	U26846 (222)	X03813 (316)	X13831 (126)	D10697 (255)	NA	NA	X13877 (330)	AF093215 (695)
D. willistoni	AY335233 (88)	L08648 (222)	AY335220 (317)	L13281 (126)	L41248 (255)	AF039560 (493)	AF293750 (377)	U51055 (402/555)[c]	AF093206 (695)
D. nebulosa	AY335234 (97)	U95275 (222)	AY334221 (319)	AF021830 (126)	L41250 (255)	NA	AF293742 (377)	U51090 (385)	AF093213 (695)
D. capricorni	AY335235 (97)	AY335196 (222)	AY335222 (321)	AY335226 (126)	AY335214 (255)	NA	NA	U51092 (375)	AF093212 (695)
D. sucinea	AY335236 (96)	AY335197 (222)	AY335223 (317)	AY335229 (126)	AY335215 (255)	NA	NA	U51091 (375)	AF093211 (695)
D. saltans	AY335237 (97)	AY335198 (222)	AY335224 (321)	U37590 (126)	AY335216 (255)	NA	NA	AY335218 (355/508)[c]	AF058978 (695)
D. sturtevanti	AY335238 (96)	AY335199 (210)	AY335225 (320)	AY335230 (126)	AY335217 (255)	NA	NA	AY335219 (360)	AF058983 (695)

[a] The number of codons used in this study.

[b] U47809 and AF169401 are *D. teissieri* sequences, and others are *D. yakuba* sequence.

[c] An additional 153 codons were determined from *D. saltans per* gene. The numbers of codons are given both excluding and including this region for *D. melanogaster*, *D. pseudoobscura*, *D. willistoni*, and *D. saltans*.

NA: Sequences not available.

groups, such as *Drosophila*. Thus, it was unexpected to find that within the single genus *Drosophila* one lineage, *D. willistoni*, evidently has undergone a significant shift in codon usage pattern (Anderson et al. 1993). Almost certainly the pattern of codon usage bias in the *willistoni* lineage is a derived condition as all other species of *Drosophila* so far examined have a very similar codon usage pattern which is different from that in *D. willistoni* (Moriyama and Powell 1997a; Rodriguez et al. 2000a). Here we expand our analysis of this shift by including more genes and species to both confirm the shift in codon usage as well as achieve a phylogenetic perspective of the shift. This expanded data set allows us to evaluate the relative roles of mutation and selection in patterning codon usage.

The lineage leading to *D. willistoni* diverged from *D. melanogaster* about 35 to 45 mya (Powell and DeSalle 1995) and consists of two major groups, the *willistoni* and *saltans* groups. Within each group, subgroups are recognized. We present data here for four subgroups: *D. saltans* and *D. sturtevanti* belong to different subgroups of the *saltans* group; *D. willistoni* belongs to the *willistoni* subgroup; while *D. sucinea* and *D. capricorni* are members of the *bocainensis* subgroup of the *willistoni* group. We also studied *D. nebulosa*, a member of the *willistoni* group, but with ambiguous affinity to the subgroups (Val et al. 1981; Tarrio et al. 2000), or has even been aligned with the *saltans* group (O'Grady and Kidwell 2002).

Materials and Methods

Species

D. saltans, *D. sturtevanti*, *D. sucinea*, *D. capricorni*, and *D. nebulosa*, were obtained from the National Drosophila Species Resource Center, Bowling Green State University, Ohio (now at the University of Arizona, Tucson). *D. willistoni* data are from strains maintained in our laboratory.

Sequences

We chose seven genes based on their codon usage bias in *D. melanogaster*: three genes with high bias (*RpL32*, *Adh*, and *Hsp*83), two with low bias (*per* and *ry*), and two with intermediate bias (*Sod* and *Gpdh*). The sequences of their homologs in four *willistoni* and two *saltans* group species were either obtained from the GenBank database, if it was available, or newly sequenced for this study. A total of 25 new sequences were produced for this study using standard PCR methods and automated sequencing on an ABI model 377; in all cases, sequencing was performed in both directions. Two or three flies per strain were sequenced. When attempting to root phylogenetic trees, two additional genes (*Amyrel* and *Ddc*) that were available for the outgroups and at least one

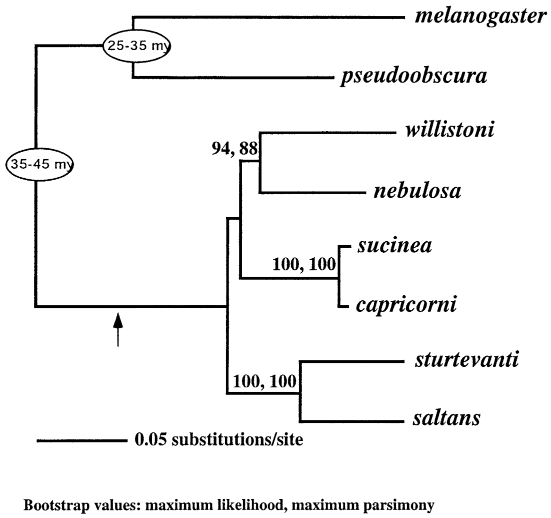

Fig. 1. Resulting phylogenetic hypothesis from a maximum likelihood analysis. Branch lengths are proportional to the number of substitutions per site. Bootstrap values are given from 1000 replicates of the ingroup taxa only and correspond to the maximum likelihood analysis and a maximum parsimony analysis using the first and second coding positions only. Dates on nodes are from Powell and DeSalle (1995) and arrow indicates branch on which codon usage likely changed (see text).

species in the ingroup were included in the analysis. All of the sequences for *D. melanogaster*, *D. yakuba*, *D. teissieri*, and *D. pseudoobscura* were obtained from GenBank. Table 1 lists gene sizes and the accession numbers. For the phylogenetic analysis, we also included COI and COII mtDNA sequences, both those previously reported (Gleason et al. 1998) and some newly added for this study (GenBank accession numbers AY335202–AY335210).

Phylogenetic Analyses

Phylogenetic analyses were carried out using maximum parsimony (MP; Farris 1970), maximum likelihood (ML; Felsenstein 1981), and neighbor joining (NJ; Saitou and Nei 1987) using PAUP* (4.0b10) (Swofford 2001). Because we sequenced more than one fly per species we used consensus sequences for ML and NJ analyses; for MP, variable sites within a species were coded as polymorphic. Gaps were treated as missing data. Branch and bound searches were run using the ACCTRAN character state optimization. Searches were performed using all substitutions unweighted, excluding third codon position, or excluding only Ti (transitions) from third codon positions. ML analyses were carried out using empirically determined Ti/Tv ratios. Rates were assumed to be variable following a gamma distribution with an empirically determined shape parameter (alpha). For analyses that included the outgroups *D. melanogaster* and *D. pseudoobsura*, alpha was first estimated without the outgroups. For the NJ analyses Tamura and Nei (1993) distances were calculated using the same empirically derived gamma parameter. To determine whether datasets are compatible, the ILD test (Farris et al. 1995; also called the Partition Homogeneity Test) was performed in PAUP*. The robustness of the phylogenetic hypothesis was tested by bootstrap over 1000 replicates for the MP and NJ analyses and 100 for the ML analyses. Alternative MP tree topologies were tested using the Templeton (1983) two-tailed Wilcoxon rank test (Larson 1994).

The dataset was examined with the following partitions: each gene separately, nuclear genes together, mitochondrial genes together, all genes. This was done for coding regions (all genes) and introns (five genes). For the analyses with introns only, the outgroups (*D. melanogaster* and *D. pseudoobscura*) were not included because these are too phylogenetically diverged to be aligned.

Codon Usage Bias

Two methods were used to estimate the degree of bias in synonymous codon usage. The codon adaptive index (CAI; Sharp and Li 1987a) measures the degree of departure from the use of a set of "optimal" codons determined from highly expressed *D. melanogaster* genes. It ranges between 0 (no bias) and 1.0 (completely biased). Optimal codons for *D. melanogaster* were taken from Shields et al. (1988). Therefore, high estimates of CAI indicate the codon bias toward "optimal" codons found in *D. melanogaster* genes. Low estimates of CAI, on the other hand, imply a departure from it, which does not necessarily mean weak codon usage bias, simply a different codon preference than found in *D. melanogaster*. The second method, effective number of codons (ENC; Wright 1990), measures the degree of departure from the equal use of synonymous codons. It ranges from 20 (completely biased) to 61 (no bias). While ENC detects any deviation from completely random codon usage, unlike CAI it does not have directionality. We should also note that low or no bias (ENC ≈ 61) does not necessarily imply that there is no selection on synonymous codon use. Because of mutation bias, ENC becomes lower than 61 even with no selection. For example, the expected ENC based on 40% G + C (assuming 30% T, 20% C, 30% A, and 20% G) is 57.8. ENC becomes 52.0 for 30% G + C (30% T, 15% C, 35% A, and 15% G).

Numbers of Synonymous Substitutions

Numbers of synonymous substitutions per site were estimated by methods developed by Moriyama and Powell (1997b). These methods consider the base composition bias in synonymous substitution estimation. We used the method incorporating the four base frequencies for all of the distance calculation, except for those involving *D. melanogaster* and *D. pseudoobscura*. For eight of nine genes for these two species comparisons, we used the method incorporating only GC content bias (arithmetic exceptions prevented the use of the first method for some genes). For the *Sod* gene, we could use only the Li (1993) method, which does not consider base composition bias.

Correspondence Analysis

Correspondence analysis on codon usage among nine *Drosophila* species (including *D. virilis*) was done by using CodonW ver. 1.4.2 written by John Peden (ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z). Five datasets were created based on codon usage bias (ENC values): Higher (*RpL32*, *Adh*, and *Hsp83*), High (*RpL32*, *Adh*, *Hsp83*, and *Sod*), Low (*Gpdh*, *Amyrel*, *Ddc*, *per*, and *ry*), Lower (*per* and *ry*), and All, containing all nine genes. We expect that the Higher and High datasets better represent the optimum pattern of synonymous codon use. Sequences were concatenated for each dataset, and RSCU (Relative Synonymous Codon Usage; Sharp and Li 1987a) values were used for the correspondence analysis to remove the effect of amino acid composition bias.

Results

Phylogeny

The coding sequences comprised 7629 nucleotides, of which 1482 were parsimony informative. The introns

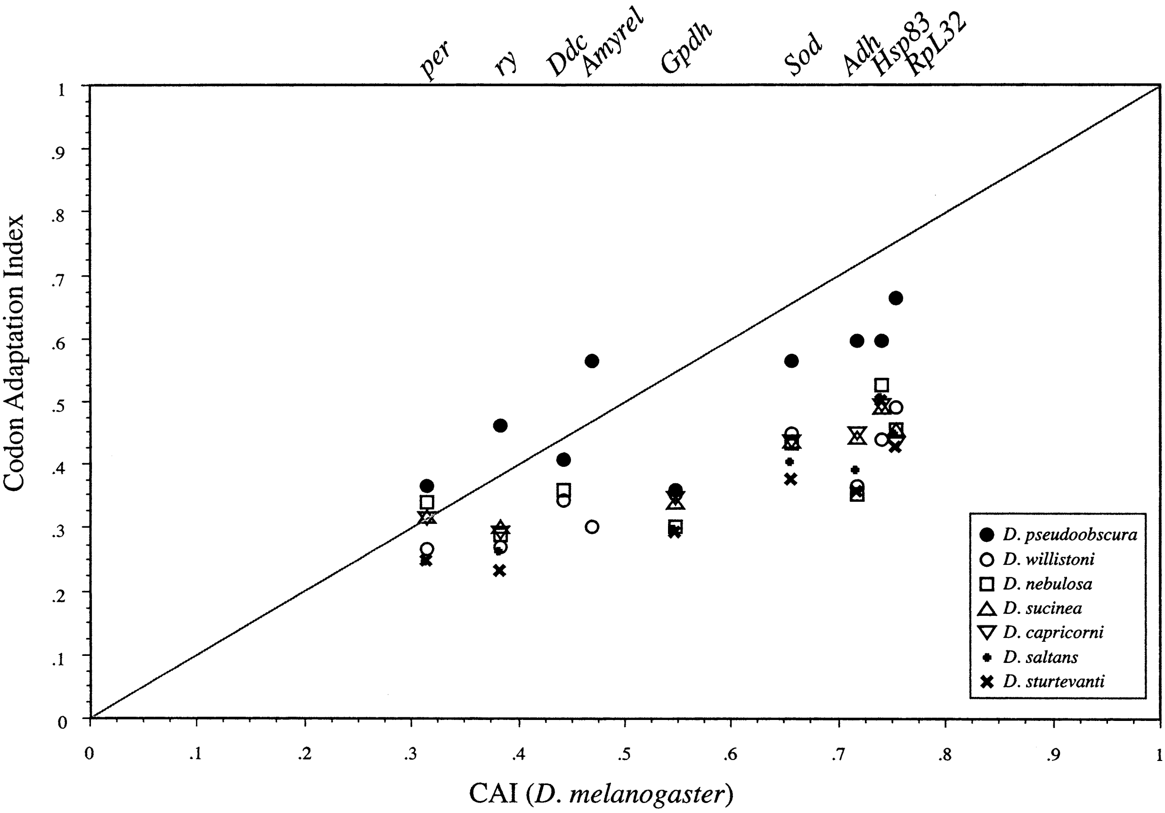

Fig. 2. Comparison of Codon Adaptation Index (CAI) usage based on preferred codons in *D. melanogaster* for the nine genes listed above the graph. Diagonal line represents a slope of one. All *willistoni/saltans* species deviate from *melanogaster*, whereas *pseudoobscura* does not.

comprised 1631 nucleotides, of which 261 were parsimony informative. For coding sequences, there were no significant partitions found in the dataset using the partition homogeneity test (as implemented in PAUP*); thus they were combined in all analyses. The partition homogeneity test was significant (p = 0.0016) for the coding versus intron partition. Therefore, all analyses on introns were done separately before also being combined with the coding sequences.

Because of the great phylogenetic distance between the outgroups (*D. melanogaster* and *D. pseudoobscura*) and the ingroup, a ML analysis was first conducted on only the ingroup species (i.e., members of the *willistoni/saltans* lineage). Rooting was performed by doing a subsequent analyses constraining the tree to the ingroup topology. The resulting topology for all coding data (Fig. 1) is well supported by both ML and MP; the identical topology was obtained when excluding third positions. In this and all other trees, *D. sturtevanti* and *D. saltans* are sister taxa, as are *D. sucinea* with *D. capricorni*. In Fig. 1, *D. nebulosa* and *D. willistoni* are each other's closest relatives and together they are the sister clade to the *D. sucinea/D. capricorni* clade. However, the relative position of *D. willistoni* and *D. nebulosa* did vary in trees derived from other analyses. Adding intron data to the coding sequences in ML or MP resulted in the same topology as the coding sequences alone, but the bootstrap values of the *D. willistoni/D. nebulosa* node are reduced to less than 75%. NJ analysis produced another topology supporting the competing hypothesis that *D. nebulosa* is more closely related to the *D. sucinea/D. capricorni* clade than to *D. willistoni.* Using only intron data, all three analyses (MP, ML, NJ) produced a third hypothesis: *D. willistoni* is closest to the *D. sucinea/D. capricorni* clade with *D. nebulosa* basal.

All the differences among the various topologies are in the relative positions of *D. willistoni* and *D. nebulosa*, which are either placed in the same clade or placed within the *D. sucinea/D. capricorni* clade, with either *D. willistoni* or *D. nebulosa* being the basal taxon. All these alternative tree topologies are not statistically significantly different from the topology in Fig. 1, as determined by the Templeton (1983) test. Although the amount of data employed to address this phylogenetic question is large, including both nuclear and mitochondrial genes, coding and noncoding regions, phylogenetic analyses of these datasets were unable to produced an unambiguous phylogenetic hypothesis for all species. Previous attempts to reconstruct molecular phylogenies for species in the *willistoni/saltans* lineage have reached

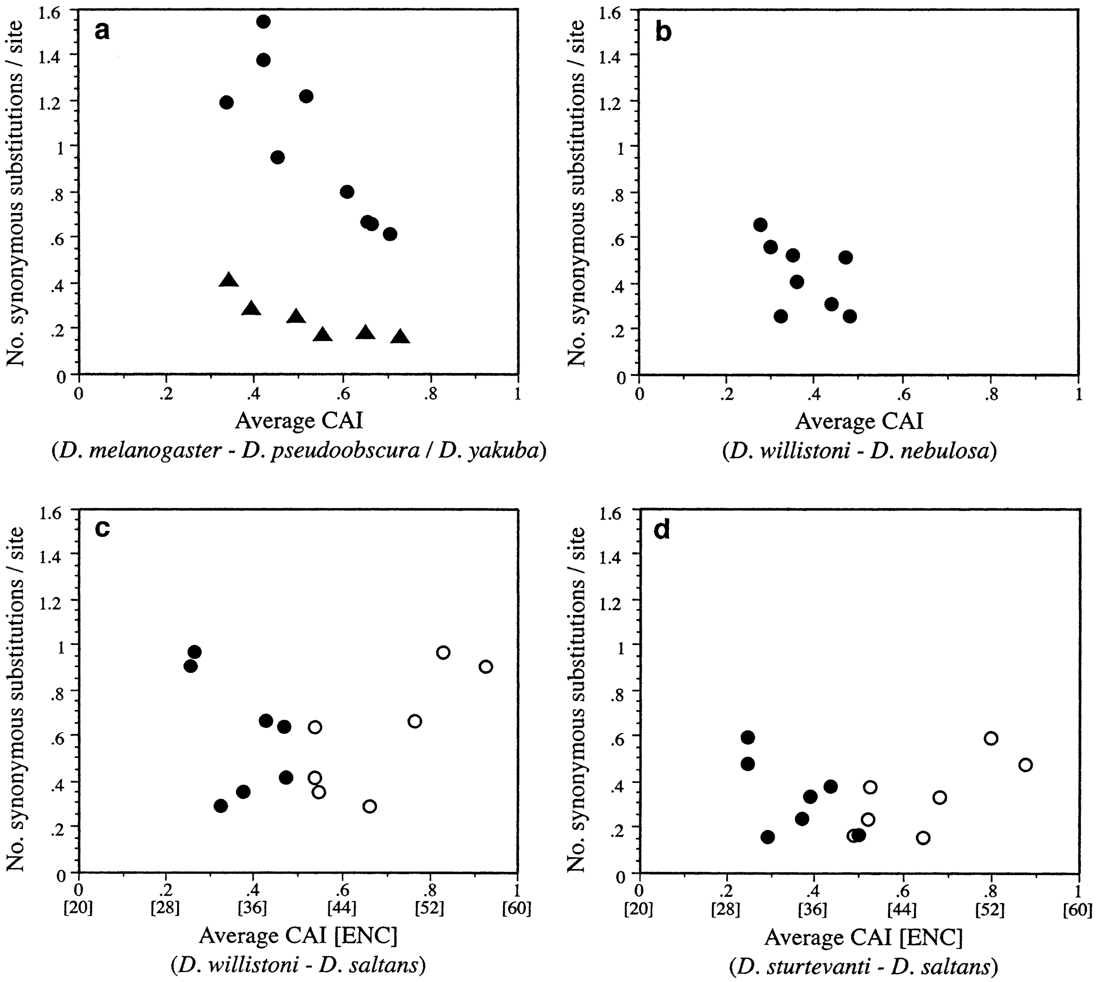

Fig. 3. Effect of codon usage bias on rates of synonymous substitutions. In (**a**), circles are for *melanogaster–pseudoobscura* and triangles are for *melanogaster–yakuba* or *teissieri*. In (**c**) and (**d**), closed circles are for CAI and open for ENC, with ENC values shown in brackets on abscissa.

similar conclusions (O'Grady et al. 1998; Rodriguez-Trelles et al. 1999; Tarrio et al. 2000; O'Grady and Kidwell 2002).

Overall Codon Usage Bias and Effects

Codon usage bias in the *willistoni/saltans* lineage deviates significantly from that in *D. melanogaster*. Figure 2 plots CAI (based on "optimal" codons in *melanogaster*) for six species of the *willistoni/saltans* lineage for nine genes. In all cases, the *willistoni/saltans* species have lower CAI than does *D. melanogaster (p* < 0.02 by Wilcoxon signed rank test). For comparison, *D. pseudoobscura* is also plotted for these same genes showing no significant deviation from *D. melanogaster*. Similarly, the nondirectional measure of codon bias, ENC, is also on average greater (indicating less bias) in the *willistoni/saltans* species compared to *D. melanogaster* (p < 0.02, Wilcoxon signed rank test). Thus we can conclude that, not only do species in the *willistoni/saltans* lineage differ in codon usage from *D. melanogaster* (as indicated by lower CAI), but they are also on average less biased (indicated by ENC).

The level of codon usage bias of a gene has been shown to correlate with the rate of synonymous substitution. A negative correlation between level of bias and synonymous substitution rates have been documented for bacteria (e.g., Sharp and Li 1987b) and *Drosophila* (e.g., Shields et al. 1988; Sharp and Li 1989; Powell and Moriyama 1997; but also see Dunn et al. 2001). This negative correlation has been considered evidence that selection affects codon usage bias, i.e., highly biased genes are more constrained in codon usage and thus evolve more slowly for silent substitutions. Figure 3 shows five species comparisons for the genes we studied. As examples of this pattern outside the *willistoni/saltans* lineage, Fig. 3a shows a strong negative correlation between *D. melanogaster* and *D. pseudoobscura* for the particular genes studied here (closed circles) and between the more closely related species *D. melanogaster* and *D. yakuba* (closed

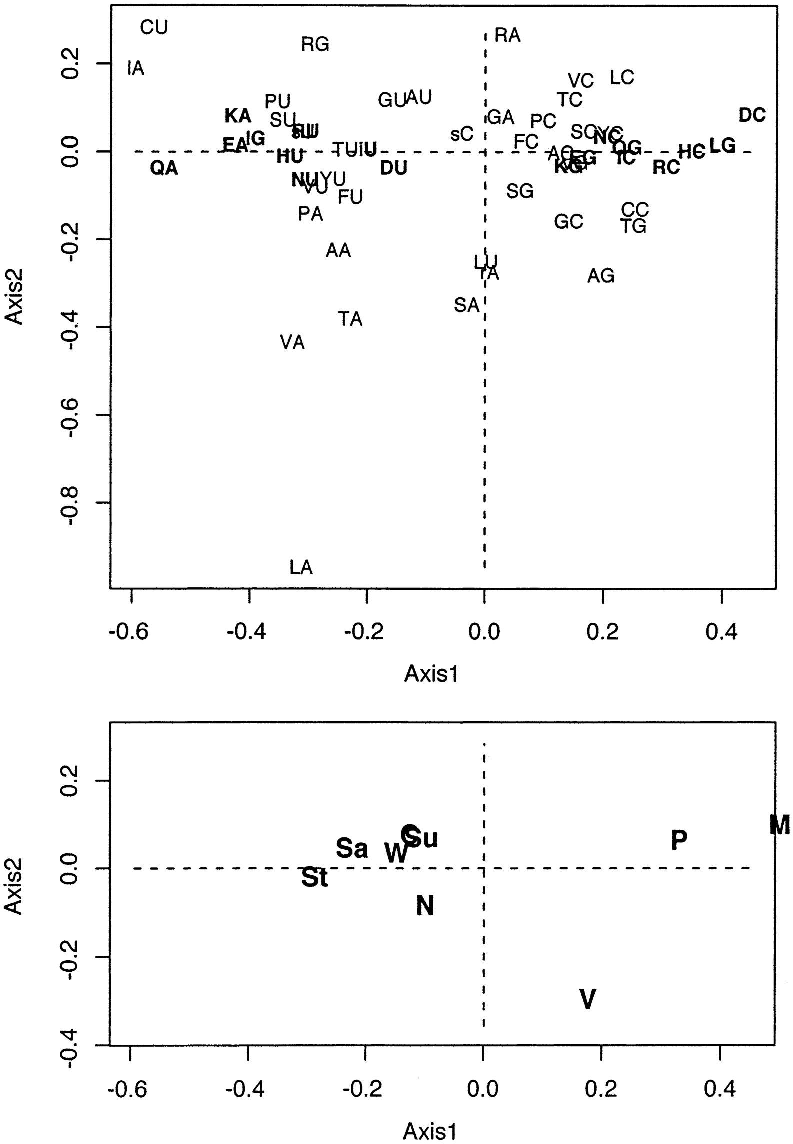

Fig. 4. Plots of principal coordinate analysis for codons and species profiles from High bias data-set. In upper plot, each codon is represented by the amino acid in one letter code followed by the nucleotide type at the third codon position (e.g., 'PC' for proline, CCC). For six-fold degenerate amino acids (Leu, Ser, and Arg), the upper case is used for the four-fold degenerate group, and the lower case for the two-fold degenerate group (e.g., IA for UUA and LA for CUA). Codons with more than 90% of inertia explained by Axis 1 are **bold**. The lower plot shows each species abbreviated: M for *D. melanogaster*, P for *D. pseudoobscura*, V for *D. virilis*, W for *D. willistoni*, N for *D. nebulosa*, Su for *D. sucinea*, C for *D. capricorni*, Sa for *D. saltans*, and St for *D. sturtevanti*.

triangles; *D. teissieri*, a close relative of *D. yakuba*, was used for two genes; see Table 1). Although we could use only a small number of genes in this study, these correlations are significant at the 5% level both for CAI and ENC. (The correlation coefficient, $R = 0.76$, between ENC and synonymous substitution numbers for the *D. melanogaster–D. yakuba* comparison was not significant, but a nonparametric Spearman rank correlation was significant for all of the four other comparisons). This is consistent with the previous studies involving analysis of the *melanogaster/pseudoobscura* lineage (Shields et al. 1988; Sharp and Li 1989; Powell and Moriyama 1997).

The relationship of level of bias and rates of synonymous substitutions within the *willistoni/saltans* species is not as strong as found in the *melanogaster/pseudoobscura* studies (Fig. 3b, c, d). Correlations for *D. willistoni–D. capricorni*, *D. willistoni–D. sucinea*, *D. willistoni–D. nebulosa* are all very similar and nonsignificant (for both CAI and ENC); the last is shown in Fig. 3b. However, interestingly, when *D. saltans* or *D. sturtevanti* are used in the comparison (e.g., Fig. 3c, d), correlations were slightly stronger when ENC (open circles in the figure), rather than CAI (closed circles in the figure), was used. The correlation coefficients ($R \approx 0.73$–0.78) are, or are very close to, significant at the 5% level. As described in Materials and Methods, ENC measures bias as a departure from the equal use of synonymous codons, whereas CAI estimates bias toward (or departure from) the optimum codon usage pattern found in *D. melanogaster* genes. Therefore, ENC may represent codon bias better if the codon bias pattern in question is repulsive from that found in *D. melanogaster*. So while these species generally have weaker codon usage bias than *D. melanogaster*, there still appears to be a negative correlation with synonymous substitution rates within this lineage.

Table 2. Difference in codon usage among *D. melanogaster*, *D. willistoni*, and *D. sturtevanti*

	No. codons (RSCU)				No. codons (RSCU)		
Codon	M	W	St	Codon	M	W	St
UUA:L	0 (0.00)	1 (0.10)	1 (0.10)	GCU:A	15 (1.20)	15 (1.25)	19 (1.58)
UUG:L	5 (0.48)	**34 (3.34)**	**29 (3.00)**	GCC:A	32 (2.56)	24 (2.00)	20 (1.67)
CUU:L	2 (0.19)	1 (0.10)	4 (0.41)	GCA:A	1 (0.08)	7 (0.58)	7 (0.58)
CUC:L	8 (0.77)	5 (0.49)	1 (0.10)	GCG:A	2 (0.16)	2 (0.17)	2 (0.17)
CUA:L	0 (0.00)	1 (0.10)	4 (0.41)				
CUG:L	**47 (4.55)**	19 (1.87)	19 (1.97)	GGU:G	22 (1.47)	24 (1.57)	32 (2.25)
				GGC:G	24 (1.60)	26 (1.70)	14 (0.98)
UCU:S	4 (0.60)	6 (0.95)	10 (1.46)	GGA:G	14 (0.93)	11 (0.72)	11 (0.77)
UCC:S	21 (3.15)	15 (2.37)	12 (1.76)	GGG:G	0 (0.00)	0 (0.00)	0 (0.00)
UCA:S	1 (0.15)	1 (0.16)	2 (0.29)				
UCG:S	6 (0.90)	8 (1.26)	7 (1.02)	CAA:Q	0 (0.00)	8 (0.94)	10 (1.11)
AGU:S	0 (0.00)	2 (0.32)	2 (0.29)	CAG:Q	**20 (2.00)**	9 (1.06)	8 (0.89)
AGC:S	8 (1.20)	6 (0.95)	8 (1.17)				
				AAA:K	5 (0.14)	25 (0.68)	24 (0.65)
CGU:R	5 (1.03)	**15 (3.33)**	**19 (3.93)**	AAG:K	67 (1.86)	48 (1.32)	50 (1.35)
CGC:R	**22 (4.55)**	9 (2.00)	8 (1.66)				
CGA:R	1 (0.21)	0 (0.00)	1 (0.21)	GAA:E	3 (0.10)	21 (0.67)	22 (0.75)
CGG:R	0 (0.00)	2 (0.44)	0 (0.00)	GAG:E	56 (1.90)	42 (1.33)	37 (1.25)
AGA:R	1 (0.21)	1 (0.22)	1 (0.21)				
AGG:R	0 (0.00)	0 (0.00)	0 (0.00)	UUU:F	1 (0.07)	6 (0.41)	9 (0.60)
				UUC:F	28 (1.93)	23 (1.59)	21 (1.40)
AUU:I	16 (0.92)	**30 (1.80)**	**30 (1.76)**				
AUC:I	**36 (2.08)**	20 (1.20)	21 (1.24)	UAU:Y	6 (0.55)	11 (1.16)	11 (1.16)
AUA:I	0 (0.00)	0 (0.00)	0 (0.00)	UAC:Y	16 (1.45)	8 (0.84)	8 (0.84)
GUU:V	7 (0.55)	17 (1.31)	**21 (1.58)**	CAU:H	3 (0.32)	**13 (1.37)**	**13 (1.24)**
GUC:V	**23 (1.08)**	17 (1.31)	13 (0.98)	CAC:H	**16 (1.68)**	6 (0.63)	8 (0.76)
GUA:V	1 (0.08)	2 (0.15)	7 (0.53)				
GUG:V	**20 (1.57)**	16 (1.23)	12 (0.91)	AAU:N	6 (0.30)	20 (0.98)	20 (0.98)
				AAC:N	**34 (1.70)**	21 (1.02)	21 (1.02)
CCU:P	1 (0.14)	6 (0.92)	3 (0.48)				
CCC:P	22 (3.14)	15 (2.31)	16 (2.56)	GAU:D	27 (1.02)	**39 (1.59)**	**44 (1.76)**
CCA:P	2 (0.29)	5 (0.77)	6 (0.96)	GAC:D	26 (0.98)	10 (0.41)	6 (0.24)
CCG:P	3 (0.43)	0 (0.00)	0 (0.00)				
				UGU:C	0 (0.00)	1 (0.50)	3 (1.20)
ACU:T	7 (0.52)	12 (0.79)	21 (1.42)	UGC:C	6 (2.00)	3 (1.50)	2 (0.80)
ACC:T	38 (2.81)	33 (2.16)	24 (1.63)				
ACA:T	1 (0.07)	9 (0.59)	12 (0.81)				
ACG:T	8 (0.59)	7 (0.46)	2 (0.14)				

Total numbers of codons from High gene dataset. Synonymous codons with RSCU >1.5 and $p < 0.01$ by χ^2-test (or Fisher's exact test if available) for the synonymous group between *D. melanogaster* and *D. willistoni*/*D. saltans* are **bold**. M, *D. melanogaster*; W, *D. willistoni*; and St, *D. sturtevanti*.

Individual Amino Acids

The above has concerned overall patterns and we now present more details of codon usage for individual amino acids. We used Correspondence Analysis (CA) as a means to discern statistically distinct codon usage patterns. For all nine species considered (three outgroup and six ingroup) both the numbers of synonymous codons and the RSCU values were used in the analysis. Consistent results were obtained, however, from both analyses. In the High and Higher datasets, there were four codons that were seldom used (AUA, AGG, CCG, and GGG). These four codons were removed from the analysis on these two datasets to avoid artifacts caused by their erroneously large relative contributions.

The first and second principal axes (Axes 1 and 2 in Fig. 4) accounted for axis approximately 63% and 13% of the total variance (inertia), respectively. Axis 1 explains the difference between the three outgroup species (*D. melanogaster*, *D. pseudoobscura*, and *D. virilis*) and the *willistoni*/*saltans* group species. The difference between the two groups is seen not only in High and Higher datasets, but also in All, Low, and Lower datasets. All of the six *willistoni*/*saltans* group species use relatively more A/T-ending synonymous codons compared to the three outgroup species. *D. sturtevanti* has the largest difference in codon usage pattern from the outgroup species as shown in the

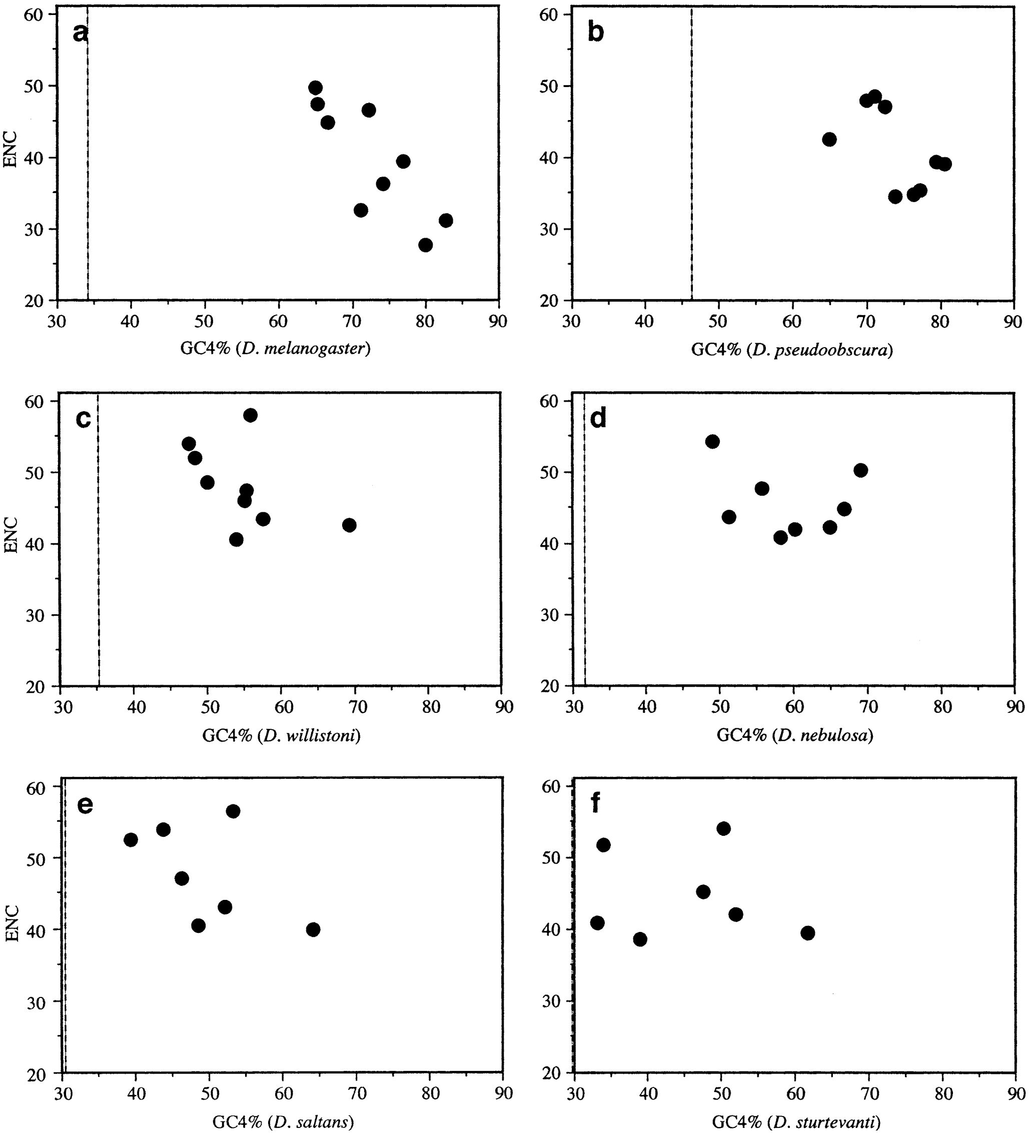

Fig. 5. Relationship between ENC and G + C content of four-fold degenerate sites (GC4%). Vertical lines in graphs are G + C content of introns (> 90 bp) for each species (Table 3).

leftmost position on Axis 1. Synonymous codons that had more than 90% of their inertia explained by Axis 1 are **bold**. Table 2 shows that those synonymous codons are used significantly differently between the two groups. In general, A-ending codons are not used in any of the *Drosophila* species, with a slight increase in use among the *willistoni*/*saltans* group. On the other hand, U-ending (or U-starting for Leu) codons are used relatively more frequently in the *willistoni*/*saltans* group. We should note that, however, the increase in A/U-ending synonymous codon use among the *willistoni*/*saltans* group species did not cause completely different codon usage patterns between the two *Drosophila* groups.

Table 2 shows the codon usage for each amino acid. First, we note that in comparing the six species of the *D. willistoni*/*saltans* lineage, chi-squared homogeneity tests for each amino acid individually showed no significant differences among these six species. Because of effects of phylogenetic correlation, we only show *D. willistoni* and *D. sturtevanti* in Table 2; species connected through the deepest node and thus less affected by phylogenetic correlations. Two things stand out in this table. First, only some amino acids display a change in codon preference between *D. melanogaster* and the *willistoni*/*saltans* species. Six (Leu, Ile, Val, His, Asn, Arg) out of 18 (discounting Met and Trp with single codons) have a different favored codon, while twelve amino acids appear not to have changed favored codon. The nineteenth amino acid, Asp, is generally quite unbiased in *D. melanogaster* (nearly 50% use of its two codons; see also Moriyama and Powell 1997a) but is quite highly biased in the *willistoni*/*saltans* lineage favoring

Table 3. Base composition of introns in the seven *Drosophila* genes

	Average base frequencies in introns (%)							
	Length ≤90 bp				Length >90 bp			
Species	T	C	A	G	G + C	T	C	A
D. melanogaster	29.3	22.3	27.3	21.1	43.4	33.7	16.1	32.5
D. pseudoobscura	33.1	21.0	30.2	15.8	36.8	22.2	25.3	31.4
D. willistoni	40.1	17.4	27.1	15.5	32.9	31.0	16.6	33.5
D. nebulosa	41.0	12.7	30.4	16.0	28.7	30.5	15.3	37.7
D. capricorni	40.9	13.9	28.8	16.3	30.2	34.8	17.7	31.2
D. sucinea	40.7	13.6	28.7	17.0	30.6	34.8	17.7	31.0
D. saltans	38.0	13.1	34.9	14.0	27.1	36.5	16.4	32.9
D. sturtevanti	41.1	12.7	34.3	11.9	24.6	36.1	14.6	34.2

the U-ending codon (Table 2). Second, in all cases when a shift has occurred it is from C preference to U preference; this is true for both third position change as well as in the single case of a first position change (Leu).

Introns

One possible explanation for the change in pattern and level of codon usage bias in the *willistoni/saltans* lineage is that there has been a change in the level of mutation bias. If so, this should be reflected in introns. Seven of the nine genes studied had introns. Table 3 summarizes the average base composition of introns in the eight *Drosophila* species. Short introns are presumably under selective constraints for both length and mutations, e.g., splicing mechanisms (Moriyama et al. 1998) and thus we expect longer introns to reflect better the mutation process. Therefore, Table 3 lists short (≤90 bp) and long (>90 bp) introns separately. The average base compositions are based on six to nine short and two or three long introns depending on the species. Although the number of long introns (>90 bp) is small, the average base composition for *D. melanogaster* is similar to that obtained previously from a larger number of introns longer than 500 bp (31% T, 19% C, 30% A, and 19% G; Moriyama and Hartl 1993). Especially for longer introns, there are no statistically significant differences in base composition between *D. melanogaster* and species in the *willistoni/saltans* lineage ($p > 0.05$ for all species pairs by Wilcoxon signed rank test). We have no obvious explanation why shorter introns in seven out of eight of the species have a lower G + C content than longer introns ($p < 0.05$ Spearman rank correlation test).

Figure 5 displays another way to assess the effect of mutation bias on codon usage, namely to compare the G + C content of introns to that found at fourfold degenerate sites (what we designate GC4%). As can be seen in Fig. 5, in general, species of the *willistoni/saltans* lineage have a lower GC4% than that of *D. melanogaster* and *D. pseudoobscura*, but not nearly as low as that found in introns.

Discussion

Phylogenetic Perspective

We had hoped that by sequencing more species in the lineage leading to *D. willistoni* that we would be able to obtain a phylogenetic perspective on the change in codon usage first observed in *D. willistoni*. We used members of all extant groups in this lineage and found that all had very similar codon usage. Thus evidently the shift occurred before the extant species diversified, i.e., before the *willistoni* and *saltans* groups diverged (see also Rodriguez et al. 2000a, 2000b). The genetic distance through the node connecting all the *willistoni/saltans* species is about half that between this group and *melanogaster/pseudoobscura*, this latter node being dated to about 40 my. So we can estimate that the extant species of the *willistoni/saltans* lineage began diversifying about 20 mya. This leaves on the order of 20 my on the branch leading from this node to the node leading to *melanogaster/pseudoobscura* (arrow in Fig. 1), so it is not possible to determine if the shift was gradual, punctuated, occurred for all amino acids simultaneously, etc. Unfortunately, there are no other known members of this lineage that could be used to break this long branch to examine the shift in more detail. However, this phylogenetic perspective does allow us to conclude that the shift is relatively old and has been stable for a long time.

Causes of the Shift in Codon Usage

Several previous studies have documented the shift in codon usage in the *willistoni/saltans* lineages (Anderson et al. 1993; Carew and Powell 1997;

Gleason and Powell 1997; Rodriguez-Trelles et al. 1999, 2000a; O'Grady and Kidwell 2002). Various tentative explanations for this shift have been presented, with no definitive conclusion. With this expanded dataset, both in terms of numbers of genes and species, we are in a better position to evaluate possible causes of the shift in codon usage in this lineage.

We can identify at least four possible explanations for the observations presented here: (1) Relaxation of selection possibly due to small population sizes and/or bottlenecks. This could account for the relatively low bias of genes in the *willistoni*/*saltans* lineage compared to *D. melanogaster* and *D. pseudoobscura* (Fig. 2). Also, relaxation of selection is consistent with a relatively high degree of replacement polymorphism at the *Adh* locus in species of the *willistoni* group (Griffith and Powell 1997), although Rodriguez et al. (2000b) found no evidence for increased replacement substitutions in the *willistoni*/*saltans* lineage. Furthermore, if the relaxation of selection was due to small population size, then it must have persisted for a very long time (at least 20 my) despite the fact that these species are very widespread and have large contemporary population sizes. Complete relaxation of selection is also inconsistent with the observation that codon usage bias affects rates of synonymous substitutions (Fig. 3). Finally, relaxation of selection should affect coding for all amino acids whereas the evidence is that only about one-third of amino acids have changed their preferred codon (Table 2). One could argue around this last point by speculating that not all amino acids are subject to the same strength of selection on codon usage and therefore those with the weakest selection would be selectively more subject to relaxed selection due to such factors as population size fluctuation. However, this *ad hoc* hypothesis has no good basis and, in fact, the observation that all amino acids (except Asp) contribute strongly and about equally to overall bias (Moriyama and Powell 1997a) would argue against there being much variance in selective constraint on codon usage for most amino acids. Below we speculate as to why the particular set of amino acids has changed in codon usage.

(2) High inversion polymorphism reduces recombination, reducing the effectiveness of selection. It is well-established that *D. willistoni* is one of the most polymorphic species for naturally occurring inversions (Dobzhansky and Powell 1975) and that regions of genomes with reduced recombination tend to have reduced codon usage bias (Kliman and Hey 1993; Moriyama and Powell 1996). However, it is not known if all members of the *willistoni*/*saltans* group have such high inversion polymorphism and, again, one would expect the effect of reduced recombination to affect all amino acids equally, a predication at odds with the pattern observed.

(3) Change in mutation bias and/or rate. We addressed this issue above by examining both the base composition of introns as well as comparing introns to the base composition of fourfold degenerate sites in coding regions. There is little or no evidence of change in base composition of reasonably long introns between *D. melanogaster* and species in the *willistoni*/*saltans* lineage (Table 3). While there has been some lowering of G + C content at fourfold degenerate sites in the *willistoni*/*saltans* lineage, it is not as low as in introns (Fig. 5). One could still argue that the level of mutation bias has changed in the *willistoni*/*saltans* lineage, but the change has been too recent to have reached an equilibrium, i.e., the small change in GC4% is the beginning of the effect of mutation bias. Because the shift in pattern is virtually identical in all species of the *willistoni*/*saltans* lineage, the most parsimonious explanation is that whatever caused the shift occurred sometime between the split of this lineage from the *melanogaster*/*pseudoobscura* lineage and the split between the *saltans* and *willistoni* groups (arrow in Fig. 1). As noted above, this means 20 or more mya. Even conservatively estimating five generations a year for these tropical species, this means that at least 10^8 generations have elapsed, presumably time to reach equilibrium for mutation bias. And finally, like the two previous explanations, we would expect a change in mutation rate/bias to affect all amino acids equally.

(4) Shift in relative abundance of isoaccepting tRNAs. It is well-established in unicellular organisms such as yeast and *E. coil* that the preferred codons are optimally decoded by the most abundant isoaccepting tRNAs, especially for highly expressed genes (reviewed in Ikemura 1992). There is also evidence for this in *D. melanogaster* (Shields et al. 1988; Moriyama and Powell 1997a; Powell and Moriyama 1997). Thus, there may have been a shift in the relative abundance of isoaccepting tRNAs in the lineage leading to the present day *willistoni*/*saltans* species. This is one of the few hypotheses that would be consistent with the observation that only some amino acids have shifted in codon preferences, while others have not. This makes the clear prediction that tRNA pools for some amino acids have changed and others not. We are presently pursuing this hypothesis by measuring the relative abundance of isoaccepting tRNAs in *D. willistoni.*

Two-Fold, C/U-Ending Codons

Of particular interest is that the most significant shifts in codon usage are for two-fold degenerate amino acids coded by C/U ending codons (Table 2). This has also been observed when comparing *D. virilis* to *D. melanogaster* (Moriyama and Powell 1997a). For some of the two-fold degenerate C/U amino acids, it is known that base modification occurs in tRNAs at the first position of the anticodon (the third for mRNA). The modification is from G to Queuosine (Q; White et al. 1973; Owenby et al. 1979). This modification results in a change from optimally decoding C-ending codons to a slight preference for U-ending codons (Meier et al. 1985). The shift from G to Q occurs for tRNAs decoding Asp, Tyr, His, and Asn, precisely those identified as undergoing the greatest shift in codon preference in the *willistoni/saltans* lineage (Table 2); it does not occur for tRNA^{Cys} and tRNA^{Phe} for which no significant shift has been identified. Precisely the same is seen in comparing *D. virilis* to *D. melanogaster*: Asp, Tyr, His, and Asn change in codon usage, Cys and Phe do not (Moriyama and Powell 1997a). The relative level of G versus Q tRNAs varies as a function of age, nutrition, and genotype in *D. melanogaster* (Owenby et al. 1997; Hosbach and Kubli 1979). Thus it is a fairly labile process. It is conceivable that this tRNA modification selectively accounts for more interspecific differences in *Drosophila* in two-fold C/T amino acids compared to all other amino acids.

Adaptation or Chance?

It is certainly possible that changes in mutation rates and/or mutation bias as well as reduced selection in the *willistoni/saltans* lineage accounts for some of the shift in codon usage. However, as we argued above, when considering details of the shift, especially considering each amino acid separately, it is difficult to see how mutation and drift alone can account for the observed data. Thus we are led to the conclusion that selection is likely to be at least partially responsible for the codon usage pattern observed in the *willistoni/saltans* lineage. The negative correlation between rate of synonymous substitution and degree of codon usage bias (Fig. 3) is further confirmation of natural selection affecting codon usage in this lineage. Selection based on the relative abundance of isoaccepting tRNAs is one of the better-documented routes for selection on codon usage and is the only explanation that appears to be consistent with all the observations. However, it is important to point out that the tRNA selection hypothesis can account only for the maintenance of codon usage bias but begs the question of the origin of a particular pattern of bias (Powell and Moriyama 1997). It is very difficult to conceive of an ecological/adaptationist explanation for the change in codon usage in the *willistoni/saltans* lineage compared to *D. melanogaster*. Both groups are primarily tropical, breeding on rotting tropical fruits. Rather we hypothesize that the shift may have originated solely by chance. For example, in the lineage leading to *willistoni/saltans* some random event (e.g., strong drift due to a bottleneck) may have changed the relative use of some codons in highly expressed genes leading to selection to change the relative abundance of isoaccepting tRNAs which would then have a feedback on selection for codons, etc. Alternatively, the random perturbation could have initially affected tRNA pools leading to the selection for change in codon usage in highly expressed genes, leading again to feedback selection. However, once codon usage and tRNA pools become adjusted to one another, it is difficult to break out of the particular pattern being maintained primarily by selection. In this view, codon usage bias would be similar to what Crick (1968) called the genetic code, a "frozen accident."

Acknowledgments. Research was supported by the National Science Foundation grant DEB9726236 (to JRP).

References

Anderson CL, Carew EA, Powell JR (1993) Evolution of the *Adh* locus in the *Drosophila willistoni* group: The loss of an intron, and shift in codon usage. Mol Biol Evol 10:605–618

Dobzhansky TH, Powell JR (1975) The *willistoni* group of sibling species. In: King R (ed) Handbook of genetics, volume 3. Plenum Press, New York, pp 589–622

Dunn KA, Bielawski JP, Yang Z (2001) Substitution rates in *Drosophila* nuclear genes: Implications for translational selection. Genetics 157:295–305

Farris JS (1970) Methods for computing Wagner trees. Syst Zool 18:374–402

Farris JS, Kallersjo M, Kluge AG, Bult C (1995) Testing significance of incongruence. Cladistics 10:315–319

Felsestein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368–376

Gleason JM, Powell JR (1997) Interspecific and intraspecific comparisons of the *period* locus in the *Drosophila willistoni* sibling species. Mol Biol Evol 14:741–753

Griffith EC, Powell JR (1997) *Adh* nucleotide variation in *Drosophila willistoni*: High replacement polymorphism in an electrophoretically monomorphic protein. J Mol Evol 45:232–237

Hosbach HA, Kubli E (1979) Transfer RNA in aging *Drosophila*: II. Isoacceptor patterns. Mech Ageing Dev 10:141–149

Ikemura T (1992) Correlation between codon usage and tRNA content in microorganisms. In: Hatfield DL, Lee BJ, Pirtle RM (eds) Transfer RNA in protein synthesis. CRC Press, Boca Raton, pp 87–111

Kliman RM, Hey J (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol Biol Evol 10:1239–1258

Larson A (1994) The comparison of morphological and molecular data in phylogenetic systematics. In: Schierwater B, Streit B, Wagner GP, DeSalle R (eds) Molecular ecology and evolution: Approaches and applications. Birkhäuser, Basil, pp 371–390

Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Meier F, Suter B, Grosjean H, Keith G, Kubli E (1985) Queuosine modification of the wobble base in tRNA^{His} influences in vivo decoding properties. EMBO J 4:823–836

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

Moriyama EN, Petrov DA, Hartl DL (1998) Genome size and intron size in *Drosophila*. Mol Biol Evol 15:770–773

Moriyama EN, Powell JR (1996) Intraspecific nuclear DNA variation in *Drosophila*. Mol Biol Evol 13:261–277

Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in *Drosophila*. J Mol Evol 45:514–523

Moriyama EN, Powell JR (1997) Synonymous substitution rates of *Drosophila:* Mitochondrial versus nuclear genes. J Mol Evol 45:378–391

O'Grady PM, Kidwell MG (2002) Phylogeny of the subgenus *Sophophora* (Diptera: Drosophilidae) based on combined analysis of nuclear and mitocondrial sequences. Mol Phylogenet Evol 22:442–453

O'Grady PM, Clark JB, Kidwell MG (1998) Phylogeny of the *Drosophila saltans* species group based on combined analysis of nuclear and mitochondrial DNA sequences. Mol Biol Evol 15:656–664

Owenby RK, Stulberg MP, Jacobson KB (1979) Alteration of the Q family of transfer RNAs in adult *Drosophila melanogaster* as a function of age, nutrition, and genotype. Mech Ageing Dev 11:91–103

Powell JR, DeSalle R (1995) *Drosophila* molecular phylogenies and their uses. Evol Biol 28:87–138

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

Rodriguez-Trelles F, Tarrio R, Ayala FJ (2000a) Evidence for a high ancestral GC content in *Drosophila*. Mol Biol Evol 17:1710–1717

Rodriguez-Trelles F, Tarrio R, Ayala FJ (2000b) Fluctuating mutation bias and the evolution of base composition in *Drosophila*. J Mol Evol 50:1–10

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sharp PM, Li W-H (1987) The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Sharp PM, Li W-H (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Sharp PM, Li W-H (1989) On the rate of DNA sequence evolution in *Drosophila*. J Mol Evol 28:398–402

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

Swofford D (2001) PAUP*, phylogenetic analysis using parimony (* and other methods). Version 4.0b. Sinauer Associates, Sunderland, MA

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tarrío R, Rodríguez-Trelles F, Ayala FJ (2000) Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The *Drosophila saltans* and *willistoni* groups, a case study. Mol Phylog Evol 16:344–349

Templeton A (1983) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. Evolution 37:221–244

Val FC, Vilela CR, Marques MD (1981) Drosophilidae of the neotropical region. In: Ashburner M, Carson HL, Thompson JN (eds) The genetics and biology of *Drosophila*, volume 3a. Academic Press, New York, pp 124–168

White BN, Tener GM, Holden J, Suzuki DT (1973) Analysis of tRNAs during the development of *Drosophila*. Develop Biol 33:185–195

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