# RESEARCH ARTICLE

# Nucleotide Substitution Bias within the Genus *Drosophila* Affects the Pattern of Proteome Evolution

*Mihai Albu,*[1] Xiang Jia Min,†[1] G. Brian Golding,‡ and Donal Hickey**

*Department of Biology, Concordia University, Montréal, Québec, Canada; †Department of Biological Sciences, Youngstown State University; and ‡Department of Biology, McMaster University, Hamilton, Ontario, Canada

The availability of complete genome sequences for 12 Drosophila species provides an unprecedented resource for large-scale studies of genome evolution. In this study, we looked for correlated shifts in the patterns of genome and proteome evolution within the genus *Drosophila*. Specifically, we asked if the nucleotide composition of the *Drosophila willistoni* genome—which is significantly less GC rich than the other 11 sequenced Drosophila genomes—is reflected in an altered pattern of amino acid substitutions in the encoded proteins. Our results show that this is indeed the case: There are large and highly significant asymmetries in the patterns of amino acid substitution between *D. willistoni* and *Drosophila melanogaster*, and they are in the direction predicted by the nucleotide biases. The implication of this result, combined with previous studies on long-term proteome evolution, is that substitutional biases at the DNA level can be a major factor in determining both the long-term and the short-term directions of proteome evolution.

## Introduction

Molecular sequence data have provided major insights into the process of biological evolution. Essentially, the positive correlation between levels of sequence divergence and the time since the existence of a common ancestor allows us to use sequence divergence as a "molecular clock" (King and Jukes 1969; Zuckerkandl 1972). Many studies have demonstrated, however, that this is not a simple clock; there are many factors in addition to the age of the common ancestor that affect the rate of sequence divergence (Roger and Hug 2006). One obvious complicating factor is natural selection, which can act to either constrain sequence change in order to conserve biological function or can accelerate change in cases of positive selection (Yang 1998; Yang and Nielsen 2002). In addition to natural selection, variations in the rate and direction of both mutation and DNA repair can also have a major impact on the patterns of sequence divergence. For example, it has been shown that molecular phylogenies of eukaryotes based on ribosomal RNA sequences are affected by biases in the nucleotide composition of those sequences (Hasegawa and Hashimoto 1993). Subsequently, many studies have shown that nucleotide bias—usually expressed as GC content—is a pervasive phenomenon (Sueoka 1992), and a host of sophisticated statistical techniques have been developed to minimize its effects on phylogenetic reconstruction (Lockhart et al. 1994; Van Den Bussche et al. 1998; Wang et al. 2008).

One approach to avoid the problem of biased nucleotide content has been to construct phylogenies based on the encoded protein sequences rather than on the DNA sequences themselves (Hashimoto et al. 1994). This reduces the problem because the most extreme compositional bias is observed at synonymous sites that do not affect the amino acid sequence. Nevertheless, the problem still persists for protein-based phylogenies because compositionally biased DNA sequences encode biased amino acid sequences (Lobry 1997; Foster et al. 1997; Singer and Hickey 2000; Knight et al. 2001; Wang et al. 2004). The problem is especially troublesome in the case of genome-wide compositional biases because, in these cases, adding more data simply compounds the problem (Foster and Hickey 1999; Leigh et al. 2008; Wang et al. 2008).

Previous studies of compositional bias have involved comparisons of widely diverged lineages. This is because more closely related organisms tend to have, on average, more similar nucleotide and amino acid compositions. However, the extensive genomic data that are now available for the genus Drosophila (Ashburner 2007; Drosophila 12 Genomes Consortium 2007) provide us with the possibility of looking at broadscale patterns of nucleotide and protein evolution over a relatively short evolutionary period. Specifically, in this case, there are different species within the same genus that show distinctly different nucleotide compositions. This allows us to look at the shorter term evolutionary effects of substitution biases, both at the DNA and protein levels. In other words, we have focused particularly on the minority of sites where evolutionary change has happened between related sequences, that is, the nonconserved sites.

## Materials and Methods

### Data Collection

We downloaded the complete set of aligned protein-coding DNA sequences from all 12 Drosophila species from Fly Base Genome project (ftp://ftp.flybase.net). From this set, we extracted only the aligned sequences for *Drosophila melanogaster* and *Drosophila willistoni*. Out of the 9,850 files with paired sequences from the two species, we generated a nonredundant gene set by removing genes that have two or more copies encoding identical (100%) protein sequences (only one was chosen). We also tried to avoid possible alignment errors by removing gene pairs that showed a large number of multiple consecutive changes.

After this filtering of the data, we obtained 7,780 gene pairs. The aligned sequences were then scanned for gaps, and we removed codons from gapped regions in either

[1] These authors contributed equally to this work.

Key words: nucleotide content, amino acid composition, GC content.

E-mail: donal.hickey@concordia.ca.

*Genome. Biol. Evol.* 1:288–293.
doi:10.1093/gbe/evp028
Advance Access publication August 4, 2009

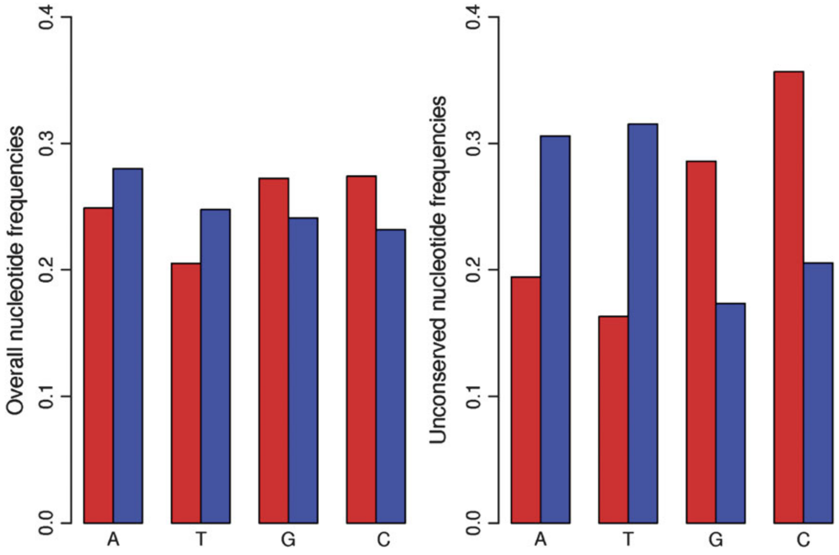

FIG. 1.—Differences in nucleotide content between the coding sequences of *D. melanogaster* and *D. willistoni*. Panel (*A*) shows a comparison of the frequency of each nucleotide in both species based on all aligned nucleotide sites (conserved and nonconserved). The results for *D. melanogaster* are shown in red and those for *D. willistoni* are shown in blue. Panel (*B*) shows the same comparison as in Panel but limited to the nonconserved, that is, variable, sites only. These frequency differences between the two species were highly significant ($P \ll 0.00001$).

species. This resulted in a total gap-free alignment of 11,290,860 bases. The aligned sequences were then compared site by site for both nucleotide and amino acid substitutions.

### Data Analysis

All statistical tests were performed using the R statistical package (http://www.R-project.org). Kolmogorov–Smirnov tests (KS tests, Marsaglia et al. 2003) were used to detect significant differences between the aligned *D. melanogaster* and *D. willistoni* DNA and protein sequences. These tests were applied to both the concatenated genome sequences and the collection of individual gene sequences.

A computer program (available upon request) was implemented to allow analysis of conserved and variable sites. The software scans the reading frames of all gene sequence files and counts conserved and nonconserved nucleotide sites. The program then produces a 64-by-64 codon substitution matrix, with each row corresponding to the occurrences in *D. willistoni* and each column to the occurrences in *D. melanogaster*, respectively (supplementary table S1, Supplementary Material online). In silico translation of this codon matrix was used to produce a 20-by-20 amino acid substitution matrix (supplementary table S2, Supplementary Material online). These matrices were then used to extract information about the overall patterns of nucleotide and amino acid substitutions (see Results).

The nonconserved sites in the DNA alignments were subdivided into synonymous and nonsynonymous substitutions. The latter—which result in amino acid substitutions—were further subdivided into those that alter the number of amino acids encoded by GC-rich or AT-rich codons. Codons were classified as being GC rich, GC neutral, or GC poor according to the classification used previously (Foster et al. 1997; Singer and Hickey 2000).

## Results

First, we compared the nucleotide composition of the *D. willistoni* genome with that of the extensively studied *D. melanogaster*. Our final DNA sequence data set contains approximately 11.3 million bp from each of these two species. When the sequences are aligned, there are a total of 3,157,787 nonconserved nucleotides (27.9%). It is already known that the genome of *D. willistoni* has a lower GC content than that of other Drosophila species such as *D. melanogaster* (Drosophila 12 Genomes Consortium 2007; Vicario et al. 2007). But since there is greater than 70% sequence identity between homologous coding sequences from these two species and since the identical sites necessarily have identical GC contents, the global difference in GC content underestimates the differences at those sites where nucleotide divergence has occurred. This effect is shown in figure 1. Although we see a marked difference in nucleotide content between the two species (fig. 1*A*), this difference becomes much greater when we consider the variable sites only (fig. 1*B*). From this figure, we also see that the reduction in GC content in the *D. willistoni* genome involves both G and C nucleotides, with a concomitant increase in both A and T nucleotides. These differences in GC content at the variable sites are statistically highly significant (KS test; Marsaglia et al. 2003; $D = 0.8913$, $P \ll 0.00001$). The fact that the GC content of the *D. willistoni* genome is significantly lower than the average of the other 11 species strongly suggests that there has been a reduction in the GC content of the *D. willistoni* genome rather than an increase in the other 11 genomes. We confirmed the direction of the change by comparing with the GC content at 4-fold degenerate sites in *D. willistoni* with both the other eight species within the subgenus Sophophora and with the three outgroup species that fall within the subgenus *Drosophila* (*Drosophila virilis*, *Drosophila grimshawi*, and

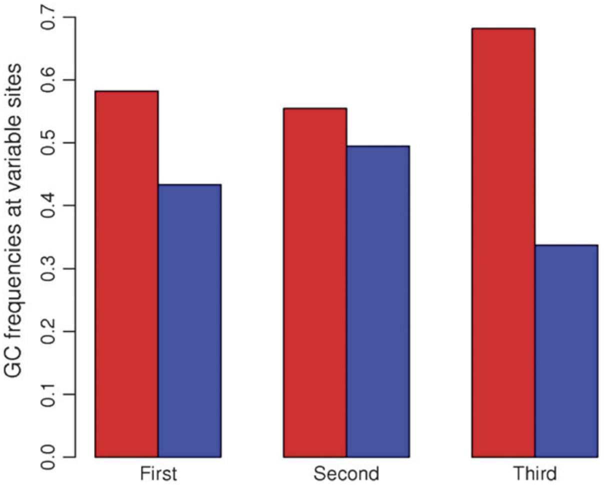

FIG. 2.—GC content at each of the three codon positions. The results for *D. melanogaster* are shown in red and those for *D. willistoni* are shown in blue. These data are based on the nucleotide frequencies at variable sites (see fig. 1*B*). As can be seen from this figure, *D. melanogaster* has a higher GC content at each of the three codon positions than does *D. willistoni*. The absolute numbers of GC nucleotide pairs at each position are shown in supplementary figure S1 (Supplementary Material online).

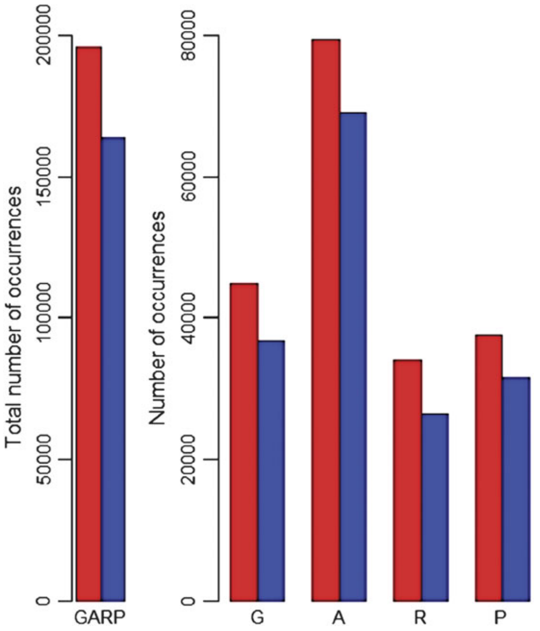

FIG. 3.—Interspecific differences in amino acid content of homologous protein sequences at nonconserved sites. Panel (*A*): number of amino acids encoded by GC-rich codons in each of the two species. In this Panel, all four of the amino acids that encoded by GC-rich codons (G, A, R, P) are grouped together. The number in *D. melanogaster* is shown by a red bar and the number in *D. willistoni* by a blue bar. Panel (*B*): this panel shows the data for each of the four amino acids—Gly, Ala, Arg, and Pro—separately. The color coding is the same as in Panel (*A*).

*Drosophila mojavensis*). The GC content of *D. willistoni* at these sites (51%) is significantly lower ($P < 0.0001$) than the average of the other Sophophora species (68%; see Vicario et al. 2007). It is also significantly lower ($P < 0.01$) than the average value for the three outgroup species (64.4%). Thus, it is reasonable to conclude that there has been a reduction in GC content within the *D. willistoni* genome rather than an increase in the other 11 genomes. Because the results reported by Vicario et al. (2007) are based on all coding sequences—and not just on aligned sequences as we used here—we double checked that this did not bias the results. Specifically, we aligned the sequences of the outgroup species, *D. virilis*, with *D. melanogaster* and *D. willistoni*, and we then calculated the GC content at the third codon position of the aligned sequences. The result, 64% GC, is entirely consistent with the value of 64.4% reported by Vicario et al. (2007) for the average of the three outgroup species. In addition to using the outgroup comparison, there is a more direct method for inferring the GC content of the common ancestor of *D. melanogaster* and *D. willistoni*; that is to calculate the GC content of the conserved sites, that is, those sites which have remained unchanged since the time of species divergence. The GC content of the third codon position at conserved sites is 65%, which is close to the value of 68% GC at the variable sites in *D. melanogaster*; more important, it is much higher than the value of 34% GC at the variable sites in *D. willistoni*. This provides further confirmation that the trend has been toward a reduction in GC content in the *D. willistoni* lineage since its divergence from *D. melanogaster*.

We then investigated the distribution of the interspecific nucleotide changes among the three codon positions (see fig. 2). The results are again highly significant for each of the three codon positions (first codon position $D = 0.7049$, second codon position $D = 0.2582$, third codon position $D = 0.9613$, and $P \ll 0.00001$ in all three cases). As expected, the majority of the changes occur at the largely synonymous third codon position (supplementary fig. S1, Supplementary Material online), and greatest degree of nucleotide bias is also seen at this position (fig. 2; supplementary fig. S1, Supplementary Material online). If we focus on 4-fold degenerate codons only, we see that the trend is highly consistent among the five codon groups (see supplementary fig. S2, Supplementary Material online); A and T ending codons are generally more frequent in *D. willistoni*, whereas G and C ending codons are more in *D. melanogaster*. A less expected finding was that a significant difference in GC content occurs at the second codon position (fig. 2). Because changes at the second codon position lead to changes in the amino acid sequence, this led us to predict that the differences in GC content would be reflected in differences in the amino acid contents of the encoded proteins, especially at the nonconserved sites.

In order to assess the effect of nucleotide bias on amino acid substitutions, the amino acids were categorized into three groups: 1) those encoded by GC-rich codons—Glycine, Alanine, Arginine, and Proline; 2) those encoded by GC-poor codons—Phenylalanine, Tyrosine, Methionine, Isoleucine, Asparagine, and Lysine; and 3) those encoded by GC-neutral codons—Serine, Aspartate, Glutamate, Valine, Threonine, Leucine, Histidine, Cysteine, Tryptophan, and Glutamine. Figure 3*A* shows a comparison of the numbers of the first category (G, A, R, P) at the nonconserved sites between the two proteomes. The *D. willistoni* proteome has 31,959 fewer of these amino acids than the homologous sequences from *D. melanogaster* (see table 1). Not only are there differences between the two species when we group these four amino acids into

**Table 1**
**Amino Acid Substitution Matrix between *D. melanogaster* and *D. willistoni* Homologous Sequences**

| | GARP | CDEHLQSTVW | FYMINK | Totals (*D. willistoni*) |
|---|---|---|---|---|
| GARP | 41,338 | 96,162 | 26,376 | 163,876 |
| CDEHLQSTVW | 117,200 | 207,916 | 98,625 | 423,741 |
| FYMINK | 37,297 | 117,613 | 36,847 | 191,757 |
| Totals (*D. melanogaster*) | 195,835 | 421,691 | 161,848 | 779,374 |

Note.—This summary table contains the amino acids at variable sites only. The amino acids are grouped into those encoded by GC-rich codons (G, A, R, and P), those encoded by GC-neutral codons (C, D, E, H, L, Q, S, T, V, and W), and those encoded by GC-poor codons (F, Y, M, I, N, and K).

a single category but also the same trend is seen for each of the four amino acids separately (fig. 3*B*). A similar, but opposite trend is seen for the amino acids encoded by GC-poor codons (see supplementary fig. S3, Supplementary Material online). The asymmetries in the amino acid substitution matrix are summarized qualitatively in figure 4. From this figure, it is clear that there is pervasive tendency for the *D. willistoni* proteome to lose amino acids encoded by GC-rich codons and to gain amino acids encoded by GC-poor codons. Out of the 780,000 (approximately) amino acid substitutions between the aligned proteome sequences, there are 272,000 substitutions in the direction predicted by the nucleotide bias and 221,000 in the opposite direction—a difference of more than 50,000 amino acid substitutions (see table 1). This difference is highly significant ($P \ll 0.00001$).

## Discussion

Our results show that the difference in GC content between the *D. willistoni* and *D. melanogaster* genomes is reflected in a bias in the amino acid substitution pattern of their proteomes. Of course, one could ask if this correlation between nucleotide bias and amino acid bias was due to selection for certain amino acids at the protein level, rather than a substitution bias at the DNA level. If we look at nucleotide changes at 4-fold degenerate synonymous sites, we can resolve this question because selection at the protein level would not affect these sites. At these sites, the nucleotide difference is even more marked—68% GC in *D. melanogaster* and 51% GC in *D. willistoni*. Moreover, the nucleotide bias affects all five 4-fold synonymous groups (see supplementary fig. S2, Supplementary Material online). This is also true for the 6-fold degenerate codons, for example, Arginine codons (see supplementary fig. 4, Supplementary Material online). Moreover, the same nucleotide bias is also seen in noncoding regions. This can be illustrated by comparing the average GC content of introns within the *D. willistoni* genome (35% GC) with the average for the other eight species within the Sophophora subgenus (42% GC); this difference in the nucleotide content of introns is also statistically significant ($P < 0.0001$). Thus, there is an underlying and pervasive DNA substitution bias that affects all nucleotide sites; the effect at synonymous sites is dramatic, whereas, at nonsynonymous sites, the effect is less dramatic but it is still highly significant.

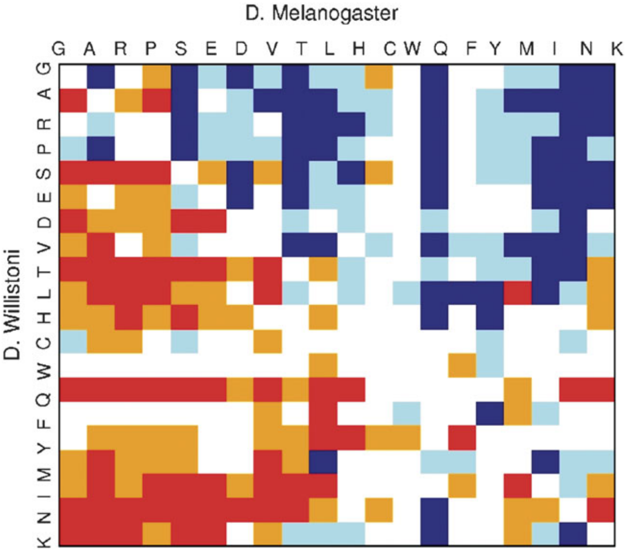

Fig. 4.—Biased patterns of amino acid substitution between *D. melanogaster* and *D. willistoni* protein sequences. We constructed an amino acid substitution matrix between the two species (see supplementary table S2, Supplementary Material online). Differences between the upper and lower diagonals were then color coded as follows to illustrate the asymmetry in the matrix. Differences of 250 or greater are shown in red; differences between 50 and 250 are shown in orange; and differences less than 50 are uncolored. Similarly, large negative values are shown in dark blue and intermediate negative values in light blue.

Our study focused on the evolutionary effects of nucleotide bias rather than on the molecular causes of these biases. It is generally agreed that the nucleotide bias is the result of an interplay between AT-biased mutation and GC-biased DNA repair (Brown and Jiricny 1988). Gene conversion, which involves heteroduplex repair, has been shown to result in increased GC content, both in Drosophila (Hickey et al. 1991) and in mammals (Galtier 2003). Over the course of evolution, there is a shifting balance between mutation and repair, resulting in fluctuating GC content that can be modeled as a Brownian motion process (Haywood-Farmer and Otto 2003). In the case of the *D. willistoni* genome, the decreased GC content could be explained by some combination of increased AT-biased mutation and/or decreased levels of GC-biased repair.

DNA substitution biases, if they persist for a long periods of evolutionary time, can have profound effects on the overall composition of both genomes and proteomes (Lobry 1997; Foster et al. 1997; Singer and Hickey 2000; Knight et al. 2001; Wang et al. 2004). At the early stages of the process, however, the cumulative effect is not so obvious because the majority of sites have not yet undergone a substitution. Thus, a simple calculation of overall GC content and amino acid composition does not reflect the amount of bias that is actually occurring at the sites undergoing substitution. A more accurate estimate is obtained if one calculates the nucleotide contents at the sites that have undergone substitution, as we have done in this study. Then it becomes clear that the effect is very pronounced, even in the short term.

Although the *D. willistoni* genome has been losing GC-rich codons and gaining AT-rich codons, this has not occurred through a direct substitution of GC-rich codons with AT-rich codons. Instead, it occurs by a two-step process whereby GC-rich codons become GC-neutral and GC-neutral codons become GC-poor (i.e., AT-rich). For example, if we look at the amino acid substitutions involving the abundant GC-neutral Serine codons (see supplementary table S2, Supplementary Material online), we see that *D. willistoni* gains 37,514 Serine codons from the GC-rich codons (encoding amino acids G, A, R, and P) of *D. melanogaster*, whereas it loses only 31,156 Serine codons to the same class—a difference of more than 6,000 codons. In other words, GC-neutral codons such as those encoding Serine act as an intermediate, "flow through" step in the biased transformation of the amino acid composition of the *D. willistoni* proteome.

Although the nucleotide bias affects all codons equally, the countervailing selective pressure at the protein level varies depending on the encoded amino acid (see Urbina et al. 2006). We can see evidence for these differential selective constraints in our data also. For example, there are relatively few substitutions involving the highly conserved amino acids Cysteine and Tryptophan (see supplementary table S1, Supplementary Material online). On the other hand, there are many substitutions involving biochemically similar amino acid pairs such as Lysine and Arginine, and these substitutions are asymmetric, consistent with the nucleotide bias. For example, the *D. willistoni* genome has gained approximately 500 more Lysines (encoded by AT-rich codons) from Arginine codons than it has lost. As expected, there are also many substitutions between the biochemically similar Isoleucine and Valine residues, but there is an approximately 7,000 excess Valine-to-Isoleucine changes from the *D. melanogaster* sequences to the *D. willistoni* sequences (see supplementary table S1, Supplementary Material online). This excess is expected because Isoleucine is encoded by more AT-rich codons than Valine.

All the four amino acids that are encoded by GC-rich codons follow the predicted trend (see fig. 3*B*). Although the AT-rich group as a whole follows the predicted trend, this does not apply to all six of the amino acids when scored individually (see supplementary fig. 3, Supplementary Material online). For example, Methionine (M) is not enriched at the variable sites in *D. willistoni* and Phenylalanine (F) appears to counter the prediction. This counterintuitive result can be explained, however, by a more detailed look at the codon substitution table (supplementary table S1, Supplementary Material online). We see that there is a tendency for the ATG Methionine codons to be converted into even more AT-rich Isoleucine codon, ATA. Likewise, the deficiency of Phenylalanine codons can be explained by the fact that Phenylalanine is also converted into even more AT-rich codons. For example, there are only 565 substitutions of the TAT codon (encoding Tyrosine) by TTC (encoding Phenylalanine), but there are 2,587 substitutions in the opposite direction. In other words, the deficiency in Phenylalanine codons in *D. willistoni* is not because they have mutated to more GC-rich codons (which would be against the prediction) but because the TTC codons been substituted by even more AT-rich codons such as TAT.

In summary, our results show that substitution biases can affect protein evolution and that the direction of such biases can change relatively rapidly over the course of evolution (within the genus Drosophila in this case). An important practical implication of our work is that substitution bias between related sequences may not be evident when one simply compares the nucleotide or amino acid composition of the entire sequences. This is because the majority of the sites, which are by definition invariant in closely related sequences, tend to mask the differences at the variant sites. It is necessary to look specifically at the variant sites in order to get an accurate estimate of the amount of bias.

## Funding

This work was supported by Discovery Grant from the Natural Sciences and Engineering Research Council of Canada [8516 to D.A.H.].

## Supplementary Material

Supplementary figures S1–S3, 3, and 4 and tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

## Literature Cited

Ashburner M. 2007. Drosophila Genomes by the Baker's Dozen. Preface. Genetics. 177:1263–1268.

Brown TC, Jiricny J. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell. 54:705–711.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450: 203–218.

Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol. 44:282–288.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol. 48:284–290.

Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. 19:65–68.

Hasegawa M, Hashimoto T. 1993. Ribosomal RNA trees misleading? Nature. 361:23.

Hashimoto T, et al. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. Mol Biol Evol. 11:65–71.

Haywood-Farmer E, Otto SP. 2003. The evolution of genomic base composition in bacteria. Evolution. 57:1783–1792.

Hickey DA, Bally-Cuif L, Abukashawa S, Payant V, Benkel BF. 1991. Concerted evolution of duplicated protein-coding genes in Drosophila. Proc Natl Acad Sci USA. 88:1611–1615.

King JL, Jukes TH. 1969. Non-Darwinian evolution. Science. 164:788–798.

Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2:research 0010.1–research 0010.13.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol. 11:605–612.

Leigh JW, Susko E, Baumgartner M, Roger AJ. 2008. Testing congruence in phylogenomic analysis. Syst Biol. 57: 104–115.

Lobry JR. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene. 205:309–316.

Marsaglia G, Tsang WW, Wang J. 2003. Evaluating Kolmogorov's distribution. J Stat Softw. 8(18):1–4.

Roger AJ, Hug LA. 2006. The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. Philos Trans R Soc Lond B Biol Sci. 361:1039–1054.

Singer GA, Hickey DA. 2000. Nucleotide bias causes a genome-wide bias in the amino acid composition of proteins. Mol Biol Evol. 17:1581–1588.

Sueoka N. 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol. 34:95–114.

Urbina D, Tang B, Higgs PG. 2006. The response of amino acid frequencies to directional mutation pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and to the structure of the genetic code. J Mol Evol. 62:340–61.

Van Den Bussche RA, Baker RJ, Huelsenbeck JP, Hillis DM. 1998. Base compositional bias and phylogenetic analyses: a test of the "flying DNA" hypothesis. Mol Phylogenet Evol. 10:408–16.

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

Wang HC, Singer GA, Hickey DA. 2004. Mutational bias affects protein evolution in flowering plants. Mol Biol Evol. 21:90–6.

Wang HC, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol. 8:331.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.

Zuckerkandl E. 1972. Some aspects of protein evolution. Biochimie. 54:1095–1102.

William Martin, Associate Editor

Accepted July 30, 2009