Shared Nucleotide Composition Biases Among Species and Their Impact on Phylogenetic Reconstructions of the Drosophilidae

Rosa Tarrío,*† Francisco Rodríguez-Trelles,*‡ and Francisco J. Ayala*†

*Department of Ecology and Evolutionary Biology, University of California at Irvine; †Misión Biológica de Galicia (CSIC), Pontevedra, Spain; and ‡Instituto de Investigaciones Agrobiológicas de Galicia (CSIC), Santiago de Compostela, Spain

Compositional changes are a major feature of genome evolution. Overlooking nucleotide composition differences among sequences can seriously mislead phylogenetic reconstructions. Large compositional variation exists among the members of the family Drosophilidae. Until now, however, base composition differences have been largely neglected in the formulations of the nucleotide substitution process used to reconstruct the phylogeny of this important group of species. The present study adopts a maximum-likelihood framework of phylogenetic inference in order to analyze five nuclear gene regions and shows that (1) the pattern of compositional variation in the Drosophilidae does not match the phylogeny of the species; (2) accounting for the heterogeneous GC content with Galtier and Gouy’s nucleotide substitution model leads to a tree that differs in significant aspects from the tree inferred when the nucleotide composition differences are ignored, even though both phylogenetic hypotheses attain strong nodal support in the bootstrap analyses; and (3) the LogDet distance correction cannot completely overcome the distorting effects of the compositional variation that exists among the species of the Drosophilidae. Our analyses clearly place the Chymomyza genus as an outgroup closer than the genus Scaptodrosophila to the Drosophila genus and conclusively support the monophyly of the Sophophora subgenus.

Introduction

Homologous DNA sequences from different organisms frequently differ in nucleotide base composition. Failure to account for nucleotide base composition variation among sequences can lead to incorrectly reconstructed tree topologies (sequences of similar base compositions may become erroneously clustered; Steel, Lockhart, and Penny 1993; Lockhart et al. 1994; Galtier and Gouy 1995) and to branch lengths that reflect changes in nucleotide composition rather than changes in substitution rate (Tourasse and Li 1999). Accounting for nucleotide composition differences among sequences is critical for correct phylogenetic assessment.

The family Drosophilidae exhibits extensive nucleotide composition variation (Rodríguez-Trelles, Tarrío, and Ayala 1999, 2000a, 2000b; Tarrío, Rodríguez-Trelles, and Ayala 2000 ). Despite its relevance as a model for evolutionary studies, significant aspects of the phylogeny of this family remain unresolved. Two unsettled cases involve taxa with extremely low GC contents: (1) the position of the genus Chymomyza relative to the genera Scaptodrosophila and Drosophila, and (2) the monophyly of the Sophophora subgenus of the genus Drosophila. Morphological (Throckmorton 1975; Grimaldi 1990) and molecular (Kwiatowski et al. 1994; Powell and DeSalle 1995; Tatarenkov et al. 1999) surveys agree that Chymomyza and Scaptodrosophila are distantly related to the rest of drosophilids, but the question of which one derived earlier remains uncertain. Because they are well known and easily available, these two lineages are often used as outgroups to Drosophila (Powell 1997); therefore, knowing which of them originated first is important for correct assessment of the plesiomorphy in this genus. The monophyly of Sophophora is well established based on morphology (Throckmorton 1975) and on the evolution of structural features of several genes (Wojtas et al. 1992; Tatarenkov et al. 1999). However, determination of the monophyletic status of this subgenus from the substitution process of the sequences has proven elusive. Molecular studies characteristically achieve weak bootstrap support for the critical node (Kwiatowski et al. 1994; Russo, Takezaki, and Nei 1995; Remsen and DeSalle 1998; Kwiatowski and Ayala 1999; Tatarenkov et al. 1999), with some studies placing the willistoni (and its sister clade saltans) species group outside the Drosophila genus (Pélandakis and Solignac 1993). The uncertainties remain despite an increasing number of nucleotide regions included in the analyses.

Current knowledge of the molecular systematics of the Drosophilidae is based on the strength of bootstrap support for nodes, but virtually no attention has been paid to the substitution models employed for the reconstruction of the trees (although the topic is discussed by Whitfield and Cameron [1998] and Steel, Huson, and Lockhart [2000] in connection with the evolution of mitochondrial rDNA genes in insects). The extensive nucleotide composition differences that occur among representatives of the family have been neglected in formulations of the substitution processes. The situation is aggravated because Ceratitis capitata, a member of the sister family Tephritidae, which is frequently used for rooting the tree of the Drosophilidae, exhibits a highly biased AT content. Additional potentially relevant parameters, such as the variation among nucleotide sites in their rates of substitution, have also been neglected.

In the present study, we address the systematics of the Drosophilidae with a focus on the substitution processes governing the evolution of the sequences. We adopted a maximum-likelihood (ML) framework of phy-
logenetic inference in order to investigate 4,650 nucleotide characters pertaining to five nuclear loci: alcohol dehydrogenase (Adh), dopa-decarboxilase (Ddc), glycerophosphate dehydrogenase (Gpdh), superoxide dismutase (Sod), and xanthine dehydrogenase (Xdh). We demonstrate that accounting for the large nucleotide composition differences among sequences yields a phylogeny that significantly differs from the relationships obtained when the heterogeneous GC content is omitted from the substitution model. Yet, the topologies obtained under the two different sets of assumptions were statistically highly supported in the bootstrap analyses. Our study (1) favors Chymomyza as the sister genus to Drosophila, with Scaptodrosophila derived earlier, and (2) confidently places the willistoni group within the Sophophora subgenus.

Materials and Methods
Species and Sequences
We investigated 13 Drosophilidae species, plus C. capitata as an outgroup (table 1). We listed Zaprionus, classified as a genus by Wheeler (1981), as a Drosophila subgenus following Tatarenkov et al. (1999), but we listed Scaptodrosophila as a genus following Grimaldi (1990), Kwiatowski et al. (1994), and Tatarenkov et al. (1999; see also Remsen and DeSalle 1998).

Table 1 gives the GenBank accession numbers for the sequences. The Xdh sequences from Drosophila mimica and Drosophila buscki were newly obtained for this study. The strategies for amplification, cloning, and sequencing are described in Tarrô, Rodríguez-Trelles, and Ayala (1998) and Rodríguez-Trelles, Tarrô, and Ayala (1999). We replaced one species with another in two cases because of unavailability: we used Drosophila bogotana (rather than Drosophila pseudoobscura) for Ddc, and Chymomyza procnemis (rather than Chymomyza amoena) for Adh. These exchanges are not expected to bias the conclusions of our analyses (see Tatarenkov et al. 1999).

Sequences were aligned using the default option of CLUSTAL W, version 1.5 (Thompson, Higgins, and Gibson 1994). After the removal of gaps and incompletely determined columns, the alignment of the five gene coding regions spanned 4,650 nucleotide positions: 513 from Adh, 963 from Ddc, 747 from Gpdh, 342 from Sod, and 2,085 from Xdh. To our knowledge, this is the largest number of regions and nucleotide characters jointly employed to investigate the phylogeny of the Drosophilidae.

Statistical Analyses
In order to control possible errors due to imperfect knowledge of the phylogeny, we considered two working tree topologies for model fitting. The first topology (hereinafter referred to as the first working topology) was the strict consensus of the topologies that resulted after applying the computer programs DNADIS, DNAML, and DNAPARS from the PHYLIP package (Felsenstein 1993), using the default options with the five gene regions pooled together. This topology coi-
cides with that shown in figure 3a. Drosophila mimica and Dorsilopa (busckii), not shown in the figure, are positioned according to the Adh + Sod + Xdh data as the sister clades to virilis-repleta and Zapronius, respectively. The second topology represents the relationships proposed by Throckmorton (1975; see hypothesis 1 in table 4) on the basis of morphological data. In Throckmorton’s (1975) scheme, D. mimica and Dorsilopa form a trichotomy together with Hirtodrosophila. These two topologies are substantially different; use of other reasonable tree topologies for model fitting is not expected to change the best-fit models identified in this study (see Yang 1994; Yang, Goldman, and Friday 1994).

We considered two sets of nested models. Models in one set were all special forms of the general time-reversible (GTR) Markov process model (Tavare 1986; Yang 1994), which allows for unequal nucleotide frequencies at equilibrium (A ≠ C ≠ G ≠ T), and six substitution classes (two transition and four transversion types). The GTR model assumes that (1) the substitution pattern has remained constant over the tree (i.e., the uniformity premise), and (2) all lineages exhibit the same nucleotide composition (i.e., the stationarity premise). Models in the second set are nested versions of the model of Galtier and Gouy (1998) (hereinafter denoted T92+GC). This model is based on Tamura’s (1992) (T92) representation of the substitution process, which allows unequal transition and transversion rates, and GC ≠ AT (with G = C and A = T) at equilibrium. Galtier and Gouy’s (1998) implementation of the T92 model allows the nucleotide composition to change from branch to branch by assigning a different equilibrium GC content parameter to each branch. The model is not homogeneous nor stationary, since equilibrium GC content can vary among lineages. Because the model lacks reversibility, trees are rooted.

Among-sites rate variation was accommodated into the models by treating rate differences among sites as a random effect using the discrete gamma distribution (eight equal-probability categories of rates, represented by the mean) with shape parameter α (denoted as dG models). The value of α is inversely related to the extent of rate variation (Yang 1996). Analyses were conducted with the BASEML program of PAML, version 2.0g (Yang 1999), and the EVAL_NH and EVAL_NHG programs from the NHML package (Galtier and Gouy 1998; Galtier, Tourasse, and Gouy 1999).

The relevance of specific parameters for describing the evolution of the sequences was evaluated by means of the likelihood ratio test (Yang 1994; Huelsenbeck and Crandall 1997). For a given tree topology (e.g., fig. 3a), a model (H_i) with p parameters and log likelihood L_i fits the data significantly better than a nested submodel (H_0) with q = p - n restrictions and likelihood L_0 if the deviance 2δ = 2 \ln L_i / L_0 = -2(\ln L_i - \log L_0) falls in the rejection region of a \chi^2_n distribution (Whelan and Goldman 1999). For this test, we used the critical values for the rejection of the H_0 provided by Goldman and Whelan (2000).

Varying the parameter addition sequence can affect best-fit model selection (Cunningham, Zhu, and Hillis 1998). We took into account this potential source of bias by assaying different parameter addition sequences. Identified best models remained the same (results not shown).

The model found to satisfactorily describe the substitution process was used for generating candidate tree topologies by the distance-based neighbor-joining (NJ) criterion. Estimates of the shape parameter α used in distance computation were those obtained simultaneously by the joint likelihood comparison of all sequences in the first stage, which can be considered the most reliable (Yang 1996). NJ trees were generated using the best-fit model identified by the likelihood ratio test in the ML analysis. Statistical support for nodes of the NJ trees was assessed with the bootstrap method (retaining nodes representing >50% of 1,000 bootstrap replications; Felsenstein 1985). Galtier and Gouy’s (1995) gamma distances and the NJ trees built from them were obtained with the GGG95 and SK programs, kindly provided by Dr. Nicolas Tourasse.

Phylogenetic hypotheses derived from the analyses were compared by the resampling estimated log likelihood (RELL) method of Kishino, Miyata, and Hasegawa (1990) (as implemented in PAML 2.0g; Yang 1999). For a given model of evolution, this test provides an estimate of the significance of a difference between the log likelihood scores of several candidate tree topologies.

Results

Variation of Nucleotide Composition in the Drosophilidae

In order to evaluate the extent to which base composition varies among the sequences under scrutiny, we tested the stationarity of base composition with the method of Rzhetsky and Nei (1995); unlike alternative approaches, such as chi-square, this method takes into account possible phylogenetic correlations, so it seems more appropriate. Stationarity of nucleotide base composition was clearly rejected. Separately, all five regions deviated from stationarity (P < 10^-6) both when the complete sequences were included and when only third codon positions were included in the analysis; in addition, all but the Gpdh and Sod regions were nonstationary (P < 10^-4) in first and first-plus-second codon positions. When the five genes were combined, stationarity was rejected (P < 10^-6) for first, first-plus-second, third, and all three codon positions pooled together.

We have shown that base composition in the Drosophilidae is nonstationary. Now we are interested in the pattern of compositional differences across taxa, because it can help to identify potential biases in reconstructed topologies elicited by the heterogeneous composition of the sequences. Figure 1 depicts the relationships inferred from the nucleotide composition of the Adh, Ddc, Gpdh, Sod, and Xdh sequences pooled together, taking into ac-
count the three codon positions. Similar cladograms were obtained for the five gene regions analyzed separately (results not shown). Because of its low GC content, *Drosophila willistoni* is repelled from its subgenus (i.e., Sophophora; herein represented by the GC-rich species *Drosophila melanogaster* and *Drosophila pseudoobscura*) and becomes associated with the cluster of GC-poor taxa Ceratitis, Chymomyza, and Hirtodrosophila. Scapeodrosophila, currently viewed as representing a different genus (Grimaldi 1990; Kwiatowski et al. 1994; Tatarenkov et al. 1999), clusters with species (including the Drosophila subgenus) that exhibit intermediate GC contents. The GC contents of *D. busckii* and *D. mimica* are not shown in the figure, are also intermediate (50.3% and 51.0%, and 61.8% and 61.9%, in first-plus-second and in third codon positions, respectively, for *Ddc, Sod*, and *Xdh* combined). The relationships in figure 1 are strongly supported statistically (bootstrap values above the nodes are all at or near 100), reflecting the extensive GC content differences among taxa. The topology remains the same after excluding third codon positions, but the bootstrap support (values below the nodes) decreases, surely because fewer sites showing biased multiple substitution are included in the analysis.

The Process of Nucleotide Substitution

Table 2 shows the log likelihood ratio statistic values for models obtained assuming the first working topology (see Materials and Methods), separately for each gene region, and for the five gene regions pooled together. Except for the comparison of Kimura’s (1980) two-parameter model versus Tamura’s (1992) model, nested models are always rejected when contrasted against the next full model. All data sets (including the *Ddc + Sod + Xdh* data set; not shown in the table) are best described with the nonhomogeneous nonstationary *T92+dG+GC* model, which allows two substitution types (transitions and transversions), discrete gamma-distributed rates across sites (dG component), and variable GC content among lineages (GC component). The best homogeneous stationary representation of the substitution process is attained with the GTR+dG model (results not shown). *T92+dG+GC* and GTR+dG are

Table 2

<table>
<thead>
<tr>
<th>Model Fitting for the Sequence Data Sets Considered in this Study</th>
<th>2[ln L₁−ln L₀]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H₀ : H₁</strong></td>
<td><strong>Adh</strong> (11; 513)</td>
</tr>
<tr>
<td>JC69 : K80</td>
<td>1</td>
</tr>
<tr>
<td>K80 : T92</td>
<td>1</td>
</tr>
<tr>
<td>K80 : T92+dG</td>
<td>2</td>
</tr>
<tr>
<td>K80 : T92+GC</td>
<td>18, 20, 22</td>
</tr>
<tr>
<td>T92+dG : T92+dG+GC</td>
<td>17, 19, 21</td>
</tr>
</tbody>
</table>

**NOTE.**—In each row, the null model (H₀) is compared with the next full model (H₁), assuming that the likelihood ratio statistic (2[ln L₁−ln L₀]) follows a χ² distribution, with degrees of freedom (df) indicated. Log likelihood scores were obtained assuming the topology shown in figure 3a (see Materials and Methods). All tests are significant (P < 10⁻⁴), except for the comparison of K80 versus T92, which is nonsignificant (NS) for all data sets. The numbers of taxa and the lengths of the sequences are given in parentheses. JC69 = Jukes and Cantor (1969); K80 = Kimura (1980); T92 = Tamura (1992); T92+dG = T92 assuming discrete gamma-distributed rates at sites; T92+GC = T92 as implemented by Galtier and Gouy (1998) to account for heterogeneous GC content among lineages; T92+dG+GC = T92 with discrete gamma-distributed rates at sites and variable GC content among lineages.
Several simple methods for tree reconstruction were used to test for a model topology with the first gene regions of this study. Although we used the topology from the T92+GC, we also used the T92+GC topology in all cases, as the T92+GC topology was previously used by several authors. We used the topology from the T92+GC, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We also used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We also used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology. We then used the T92+GC topology, and then used the T92+GC topology to model the data sets obtained with the T92+GC topology.

### Table 4

Kishino, Miyata, and Hasegawa's (1990) Resampling Estimated Log Likelihood (RELL) Test Applied to Seven Different Phylogenetic Hypotheses of the Drosophilidae

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Reference</th>
<th>GTR+dG RELL</th>
<th>T92+dG+GC RELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (Le, (Ch, (So, (VR, (Hi, Za)))))) .....................................</td>
<td>Throckmorton (1975)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. (Le, (Ch, (Hi, (Za, (So, VR)))))) .....................................</td>
<td>Grimaldi (1990)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. (Le, (Hi, (Ch, (Za, (So, VR)))))) .....................................</td>
<td>DeSalle (1992)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. (Le, (Hi, (So, (Za, VR)))) ................................................</td>
<td>Powell and DeSalle (1995)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. (Le, (Ch, (Za, (Za, VR)))) ................................................</td>
<td>Renssen and DeSalle (1998); Tatarenkov et al. (1999)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. (Ch, (Le, (Wi, (So, (Za, (Hi, VR)))))) ................................</td>
<td>This study: NJ on GTR+dG distance</td>
<td>-29,918.88</td>
<td>0.10</td>
</tr>
<tr>
<td>7. (Le, (Ch, (So, (Za, (Hi, VR)))))) ....................................</td>
<td>This study: NJ on T92+dG+GC distance</td>
<td>-29,918.88</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Note.—Le = Scaptodrosophila lebanonensis; Ch = Chymomyza; Za = Zapironus; Hi = Hirtodrosophila; VR = virilis-repleta; So = Sophophora subgenus; Wi = Drosophila williamsi. See table 2 footnote for model definitions.
generate a hypothesis for the phylogenetic relationships of the Drosophilidae using the distance-based NJ criterion. The GTR+dG model was also considered for comparison.

Gene regions were first considered separately. The results of these analysis (not shown) indicated that each gene alone lacked sufficient information to resolve most relationships (the majority of the bootstrap values are below 70%: Hillis and Bull 1993). The only reasonably well defined clades were \textit{D. melanogaster} + \textit{D. pseudoobscura} (supported by \textit{Ddc}, \textit{Sod}, and \textit{Xdh} data when they are analyzed with the GTR+dG model, and by \textit{Gpdh} and \textit{Xdh} under the T92+dG+GC model) and \textit{D. hydei} + \textit{D. virilis} (supported by \textit{Adh} under GTR+dG, and by \textit{Xdh} using either model). \textit{Adh} analyzed under the GTR+dG model also supported the connection of \textit{D. mimica} to the cluster consisting of \textit{D. hydei} and \textit{D. virilis}. All of these are well-established relationships from other studies. In addition, analysis of the \textit{Xdh} data with the GTR+dG model supported the cluster consisting of \textit{Ceratitis}, \textit{Chymomyza}, \textit{D. willistoni}, and \textit{Scaptodrosophila}, and also the association of \textit{Zaprionus} with \textit{D. busckii}.

Figure 2 shows the NJ trees derived from the GTR+dG and T92+dG+GC distance matrices using the \textit{Adh} + \textit{Sod} + \textit{Xdh} data set (for which all species are available), and this data set combined separately with \textit{Ddc} (\textit{D. busckii} unavailable) and \textit{Gpdh} (\textit{D. mimica} unavailable). Combining the data sets results in increased resolution of the phylogeny and reveals conflicts between the GTR+dG and T92+dG+GC models in the resulting branching pattern of the topologies. The GTR+dG model always places \textit{Scaptodrosophila} as more closely related to \textit{Drosophila} than \textit{Chymomyza}, and it places \textit{D. willistoni} outside all other species of the \textit{Drosophila} genus. In contrast, the T92+dG+GC model identifies \textit{Scaptodrosophila} as the first derived lineage after \textit{Ceratitis} (followed by \textit{Chymomyza}) and places \textit{D. willistoni} within the subgenus \textit{Sophophora}. These two alternative branching patterns receive strong bootstrap support from their respective models. With regard to the remaining relationships, the two models are congruent across data sets in the well-resolved nodes. Both models support \textit{D. mimica} as the sister lineage to the clade consisting of \textit{D. hydei} + \textit{D. virilis} and the association of \textit{D. melanogaster} with \textit{D. pseudoobscura}. In addition, the \textit{Adh} + \textit{Ddc} + \textit{Sod} + \textit{Xdh} data set supports inclusion within the subgenus \textit{Drosophila} of \textit{Zaprionus}, which derives first, followed successively by \textit{Hirtodrosophila}, \textit{D. melanogaster}, and \textit{D. virilis}. The two trees are fairly well resolved, with statistical support (somewhat greater for the T92+dG+GC model), but depict conflicting phylogenetic relationships. The GTR+dG model places \textit{Chymomyza} as the first derived after \textit{Ceratitis}, followed by \textit{Scaptodrosophila} and \textit{D. willistoni}, while the T92+dG+GC model places \textit{Scaptodrosophila} as derived before \textit{Chymomyza} and places \textit{D. willistoni} within the subgenus \textit{Sophophora}. The two models agree, however, in that \textit{Hirtodrosophila} splits after \textit{Zaprionus}, followed by the clade consisting of \textit{D. hydei} + \textit{D. virilis}, all four pertaining to the subgenus \textit{Drosophila}.

Table 4 shows the results of Kishino, Miyata, and Hasegawa’s (1990) RELL test for seven different phylogenetic hypotheses of interest. In particular, we are interested in the effect of placing the AT-rich taxa \textit{Chymomyza} and/or \textit{D. willistoni} in different positions with respect to each other and to the AT-rich outgroup \textit{C. capitata}. The hypotheses considered in table 4 are based on data and analyses as follows: Throckmorton (1975; hypothesis 1) and Grimaldi (1990; hypothesis 2) used morphological data; DeSalle (1992; hypothesis 3) carried out a parsimony analysis of the mitochondrial 16S rDNA region; these same sequence data together with the nuclear 28S rDNA region and several morphological and behavioral characters were combined in a parsimony analysis by Powell and DeSalle (1995; hypothesis 4); also adopting a parsimony framework, Remsen and DeSalle (1998; hypothesis 5) added to these data sequences from the \textit{Adh} and \textit{Sod} regions, and hypothesis 5 was arrived at by Tatarenkov et al. (1999) after a sta-
tionary, constant rate from site-to-site distance model-based NJ and parsimony analysis of the Adh, Ddc, Gpdh, and Sod regions. Hypotheses 6 and 7 were generated in our study and correspond to the topologies shown in figure 3a and b, respectively. RELL tests were conducted using the GTR+dG and T92+dG+GC models with the Adh, Ddc, Gpdh, Sod, and Xdh data sets combined. Under the T92+dG+GC model, which accounts for the observed large nucleotide composition differences among the sequences, hypothesis 7 (see also fig. 3b) is statistically superior to all of the alternative phylogenetic hypotheses considered (yields the best log likelihood score out of all hypotheses in 95.7% of 10,000 resampled likelihood scores). Hypothesis 7 also produces a better ML score than the alternatives under the stationary GTR+dG representation, although this model does not allow conclusive discrimination between this hypothesis and hypothesis 6 (RELL support 73.6 vs. 25.0 for hypotheses 7 and 6, respectively; table 4). If we assume that the topology shown in figure 3b reflects the correct biological tree, the fact that this topology achieves higher support with the GTR+dG model applied in an ML framework than when used in a distance formulation (see fig. 3a) would be expected because of the greater robustness of the former approach (see Felsenstein 1988; Huelsenbeck 1995).

Discussion

Molecular approaches to the systematics of the Drosophilidae have focused on reconstructed tree topologies. Virtually no attention has been paid to the intricacies of the substitution processes governing the evolution of the sequences. Substitution models employed for tree building have characteristically been arbitrarily chosen. In no case has the extensive nucleotide composition variation across members of the family been taken into consideration in formulations of the substitution process. We adopted an ML framework of phylogenetic inference because it provides a rationale for choosing between increasingly realistic descriptions of the evolution of the sequences by means of the likelihood ratio test. We demonstrate that (1) the pattern of nucleotide composition biases across the Drosophilidae does not match the phylogeny of the species (see fig. 1), and (2) there is clearly an effect of nucleotide composition, since the same tree selection procedures give different trees, depending on the model used to account for multiple changes. We conclude that accommodation of compositional biases (together with the among-sites rate variation) into the substitution model is critical for a minimally realistic assessment of the phylogeny (see fig. 3). This conclusion is worth emphasis, owing to the elevated number of nucleotide characters and regions included in the study, to our knowledge, the largest so far employed to address the evolutionary relationships of the Drosophilidae; apparently, an increase in the size of the data set is less relevant to the phylogeny than the use of an appropriate model of substitution. Notice that although the homogeneous-stationary GTR+dG model is more realistic than all previously used representations, it is not robust enough given the observed variation in base composition.

Several ML approaches have been devised to deal with the problem of varying compositional biases between lineages. Galtier and Gouy’s (1998) implementation of the Tamura (1992) model is faster than other approximations (e.g., Yang and Roberts 1995) as a tool for describing the substitution process (Galtier and Gouy 1998). The method has proven useful for the study of GC content evolution in mammals (Galtier and Mouchiroud 1998), as well as Drosophila (Rodríguez-Trelles, Tarrio, and Ayala 2000c), and also for inferring nucleotide composition of ribosomal RNA in the “cenaces-tor” (i.e., the most recent common ancestor of all extant life forms) (Galtier, Tourasse, and Gouy 1999). However, the algorithm is computationally too time-demanding for tree reconstruction from data sets as large as ours (see Galtier and Gouy 1998). We circumvented this

Fig. 3.—Neighbor-joining trees based on the general time-reversible distance allowing rate variation among sites (GTR+dG) and the Tamura (1992) distance allowing rate variation among sites and GC content variation among lineages (T92+dG+GC) for the total-evidence data set. Branch lengths are proportional to the scale, given in substitutions per nucleotide. Bootstrap values, based on 1,000 replications, are given on the nodes.
drawback by using the distance-based NJ implementation of the TN92+GC+dG model (Galtier and Gouy 1995; Galtier, Tourasse, and Gouy 1999) to infer the tree. This method outperforms ML and distance-based tree-making methods that assume homogeneous and stationary conditions, as well as maximum-parsimony methods in cases of heterogeneous base composition (Galtier and Gouy 1995).

Probably the most popular distance correction for coping with the problem of heterogeneous base composition is the LogDet transformation (Lockhart et al. 1994). Compared with Galtier and Gouy’s (1995) distance model, LogDet has the disadvantage that it generally does not yield the amount of change along branches and that it assumes that substitution rates are equal across sites (Lockhart et al. 1994). Unlike Galtier and Gouy’s (1995) distance measure, LogDet distances cannot be directly modified to take account of a specific distribution of rates, such as the gamma distribution (see Swofford et al. 1996). Inclusion of invariant sites in the LogDet calculation tends to underestimate the amount of change, and sites that vary greatly are problematic because of saturation (Lockhart et al. 1994). It has been shown to be useful to exclude both these extremes by using only parsimony-informative sites (Lockhart et al. 1994). Substitution rate varies widely from site to site in our data set (see table 3). Therefore, we calculated LogDet distances using only parsimony-informative sites, considering first-plus-second (420 sites) or third (1,233 sites) codon positions. When parsimony sites from first-plus-second codon positions were used, 2 out of the 40 pairwise comparisons (i.e., D. pseudoobscura, and Hirtodrosophila vs. C. capitata) had negative determinants, for which the logarithm (and thus the distance) is undefined. In other words, there is such a large divergence between these two pairs of taxa that their sequences are effectively random with respect to each other (see Foster and Hickey 1999). In order to build the NJ tree from the distance matrix, the program PAUP*, version 4.0 (Swofford 1999), arbitrarily sets the values of these undefined distances at twice the distance of the largest defined distance in the distance matrix (i.e., 2 × 2.5332, the distance between D. hydei and C. capitata).

To guard against the effects of choosing these distances on the topology, we additionally tried factors of 1.1×, and 5×. When undefined distances were set to 1.1 times the largest defined distance in the matrix, the resulting NJ topology was identical to the T92+dG+GC topology except that it placed Chymomyza closer than Scaptodrosophila to C. capitata (likewise the GTR+dG model; see fig. 3a; note that Chymomyza is still compositionally more biased than D. willistoni toward C. capitata; see fig. 1). When the factor was set to 2× (i.e., PAUP* choice), Chymomyza remained closer than Scaptodrosophila to C. capitata, and D. willistoni appeared displaced to an external position to the Drosophilidae genus (likewise the GTR+dG model; see fig. 3a); when the factor was set to 5×, the resulting NJ topology exhibited disparate relationships. Similar analyses conducted using the parsimony sites of third codon positions (13 out of the 40 pairwise comparisons yielded undefined distances) also produced inconsistent configurations. Therefore, it seems that by limiting the analysis to parsimony sites from first-plus-second codon positions and arbitrarily adjusting undefined distances in the LogDet transformation, it is possible to cope with some (i.e., the compositional bias of D. willistoni), but not all (i.e., the even larger compositional bias of Chymomyza), of the nucleotide composition variation present in our data set. In this respect, our study corroborates the results of other authors who point out that the LogDet correction can fail when there are large nucleotide composition differences among sequences (Foster and Hickey 1999).

Failure to account for nucleotide substitution differences among sites when they exist can dramatically affect phylogenetic inferences (Yang 1996). Phylogenetic studies of the Drosophilidae have faced this problem by arbitrarily dropping fast-changing third codon positions from the analysis, thus dismissing any phylogenetic signal they may contain (e.g., Kwiatowski et al. 1994; Tatarenkov et al. 1999). paradoxically, because first and second codon positions are usually under stronger functional constraints and can greatly vary along the sequences, they generally exhibit more extensive among-sites rate variation than when they are analyzed in conjunction with third codon positions. Here we show that among-sites rate variation is a significant feature of the data (see table 3). The ML methods that we used to account for it made use of the full length of the sequences, such that sites were given a phylogenetic weight inversely related to their rate of change in an objective manner (see Yang 1996).

Our study shows that two different representations of the substitution process generate two different tree topologies, each attaining high nodal bootstrap support. Our study illustrates a well-known property of bootstrapping: high nodal bootstrap support indicates that the optimal tree would be unlikely to change as sequence length increases, but it gives absolutely no indication as to whether the results are converging to the right tree (see Swofford et al. 1996). Previous discussions about bootstrap support values for nodes of trees of the Drosophilidae from unrealistic models of substitution should therefore be taken cautiously. Similarly, caution should be exercised in adopting topological congruency among phylogenetic algorithms as a criterion to use in choosing among candidate trees: the GTR+dG and the LogDet (to an extent that can depend on arbitrary choices) distance methods both agree in supporting a wrong topology.

The Kishino, Miyata, and Hasegawa (1990) RELL test is a popular means to test competing evolutionary hypotheses in an ML framework. Strictly speaking, the RELL test is only valid for comparison of tree topologies that have been specified a priori. (Kishino and Hasegawa 1989; Kishino, Miyata, and Hasegawa 1990; Swofford et al. 1996). Several authors have warned about the risks of including one or more a posteriori specified trees in the comparison, specifically the ML tree resulting from the data used to conduct the test (Goldman, Anderson, and Rodrigo 2000). Our applica-
tion of the Kishino, Miyata, and Hasegawa (1990) RELL test is correct because the phylogenetic hypotheses generated by our analyses (see fig. 3a and b) were obtained using distance-based methods (i.e., we cannot assume that they are ML trees), while the other competing hypotheses were derived from other sources.

The monophyly of the Sophophora subgenus has been determined from anatomical and biogeographical evidence (Throckmorton 1975) and is in agreement with the evolution of structural properties of several coding regions: the absence of an intron in the Gpdh gene (Wojtas et al. 1992) and the deletion of three coding nucleotides in the Ddc gene (Tatarenkov et al. 1999) are features specific to the four major species groups of Sophophora (i.e., melanogaster, obscura, saltans, and willistoni). So far, however, attempts at confirming this positioning by tree-making methods based on conventional descriptions of the nucleotide substitution process have tended to place the saltans and willistoni groups outside the genus Drosophila (see fig. 3a). Our results strongly suggest that the GC-poor D. willistoni sequence is artifactualy attracted by the relatively GC-poor C. capitata outgroup sequence when the heterogeneous base composition is not accounted for by the substitution model (see figs. 1 and 3). A similar effect impacts the GC-poor Chymomyza sequence. Its position as a closer outgroup to Drosophila than the Scaptodrosophila genus obtained in our study is consistent with the hypothesis of Throckmorton (1975) based on the evolution of morphological characters. Our results corroborate a more solid basis previous conclusions about the branching order of Zapironus and Hirtodrosophila and their position closer to the subgenus Drosophila than the Sophophora subgenus.

The fact that the phylogenetic hypothesis produced by our study is based on a more realistic approach than previous assessments by no means guarantees that we have arrived at the correct tree. Dealing with different causes of tree-building inconsistency at the same time would lead ML to the problem of loss of identifiability, mentioned above for distances; see Steel, Székely, and Hendy 1994; Baake 1998). However, there is now better agreement between different classes of data, including morphological and molecular evidence.

**Acknowledgments**

We are indebted to Nicolas Galtier and Nicolas Tourasse for making their software available and to two anonymous reviewers for valuable comments. F.R.-T. received support from the Spanish Council for Scientific Research (Contrato Temporal de Investigación) and grant AGL2000-1073 from the Ministerio de Ciencia y Tecnología to A. Ballester. Research was supported by NIH grant GM42397 to F.J.A.

**LITERATURE CITED**


