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 confidently 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, 2000*a*, 2000*b*; 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-

Key words: heterogeneous GC content, unequal base composition, maximum-likelihood phylogeny, nonhomogeneous phylogeny models, molecular phylogeny, Drosophilidae.

Address for correspondence and reprints: Francisco Rodríguez-Trelles % Francisco J. Ayala, Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, California 92697-2525. E-mail: ftrelles@iiag.cesga.es.

Mol. Biol. Evol. 18(8):1464–1473. 2001
 ISSN: 0737-4038

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 busckii* were newly obtained for this study. The strategies for amplification, cloning, and sequencing are described in Tarrío, Rodríguez-Trelles, and Ayala (1998) and Rodríguez-Trelles, Tarrío, 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 coin-

Table 1
The 14 Species Studied, with the GenBank Accession Numbers for the Gene Sequences

Family	Genus	Subgenus	Group	Species	Accession no.				
					Adh	*Ddc*	*Gpdh*	*Sod*	*Xdh*
Drosophilidae	*Drosophila*	*Sophophora*	*melanogaster*	*melanogaster*	X78384	AF091328	X14179	X13780	Y00307
			obscura	*pseudoobscura*	X62181	—	L41249	U47871	M33977
				bogotana	—	*	—	—	—
			willistoni	*willistoni*	L08648	*	L37038	L13281	AF093206
		Drosophila	*virilis*	*virilis*	U26846	*	D10697	X13831	AF093215
			repleta	*hydei*	X58694	*	L41650	U37714	AF226974
			Hawaiian	*mimica*	M60792	*	—	AF022218	AF395402
		Hirtodrosophila		*pictiventris*	AF021832	*	L41649	AF21824	AF93214
		Dorsilopha		*busckii*	—	*	*	U39445	AF395403
		Zaprionus		*tuberculatus*	X63955	*	L37039	AF021823	AF093216
	Scaptodrosophila			*lebanonensis*	M7637	AF091329	*	AF021822	AF058984
	Chymomyza			*amoena*	—	*	L36961	X61687	AF093217
				procnemis	AF021831	—	—	—	—
Tephritidae.....	*Ceratitis*			*capitata*	Z30194	*	L36960	M76975	AF093218

Note.—Two *Xdh* sequences newly obtained by us are underlined. *Scaptodrosophila* was classified by Wheeler (1981) as a subgenus of *Drosophila* but has been raised to genus by Grimaldi (1990). *Zaprionus* is classified as a genus by Wheeler (1981); in this paper we refer to it, as well as to *Hirtodrosophila* and *Dorsilopha*, as subgenera within the genus *Drosophila*. *Ddc* and *Gpdh* sequences represented by asterisks are from Tatarenkov et al. (1999) and Kwiatowski and Ayala (1999), respectively.

cides with that shown in figure 3*a*. *Drosophila mimica* and Dorsilopha (*busckii*), not shown in the figure, are positioned according to the *Adh* + *Sod* + *Xdh* data as the sister clades to *virilis-repleta* and Zaprionus, 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 Dorsilopha 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 (Tavaré 1986; Yang 1994), which allows for unequal nucleotide frequencies at equilibrium ($A \neq C \neq G \neq T$), 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 $\neq$ 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 neither 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. 3*a*), a model (H_1) with p parameters and log likelihood L_1 fits the data significantly better than a nested submodel (H_0) with $q = p - n$ restrictions and likelihood L_0 if the deviance $2\delta = 2 \ln(L_1/L_0) = -2(\ln L_1 - \log L_0)$ falls in the rejection region of a χ^2_n (where n represents degrees of freedom). Specifically for the test of rate constancy among sites, where the H_0 ($\alpha = \infty$) is equivalent to fixing α at the boundary of the parameter space of the H_1 ($\alpha < \infty$), 2δ follows a 50:50 mixture of χ^2_{n-1} and χ^2_n distributions (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 (Cunninghan, 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-

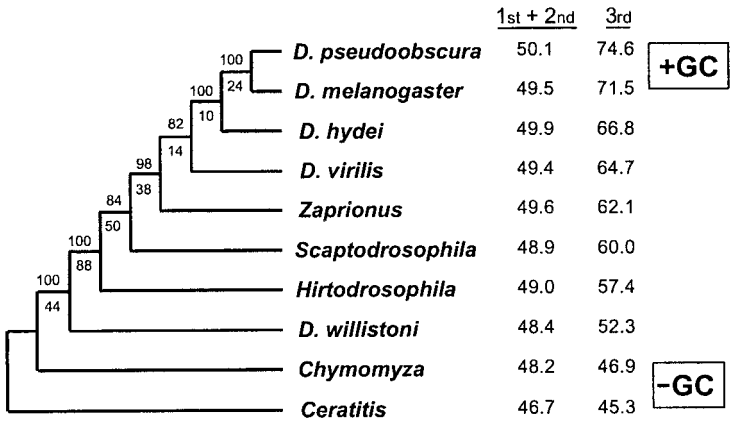

FIG. 1.—Relationships as inferred from the nucleotide composition of the *Adh*, *Ddc*, *Gpdh*, *Sod*, and *Xdh* gene regions pooled together. The cladogram was obtained with the neighbor-joining algorithm based on the Euclidean distances between nucleotide frequencies for each pair of taxa using all three codon positions. Percentages of GC content in first-plus-second and in third codon positions, respectively, are given. Numbers at the nodes are percentage bootstrap values (based on 100 pseudoreplications) considering all sites (above) or solely the first-plus-second positions of codons (below), respectively. GC content in third positions decreases monotonically from top (+GC) to bottom (−GC); a mild trend in the same direction can be seen for first-plus-second positions.

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. Scaptodrosophila, 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*, 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
Model Fitting for the Sequence Data Sets Considered in this Study

		$2[\ln L_1 - \ln L_0]$					
$H_0 : H_1$	df	*Adh* (11; 513)	*Ddc* (12; 963)	*Gpdh* (11; 747)	*Sod* (12; 342)	*Xdh* (12; 2,085)	Total (10; 4,650)
JC69 : K80	1	141.68	208.38	337.78	141.06	700.20	1,652.38
K80 : T92	1	NS	NS	NS	NS	NS	NS
K80 : T92+dG	2	416.38	1,593.46	776.62	460.74	2,672.86	4,152.46
K80 : T92+GC	18, 20, 22	96.38	172.44	106.54	89.92	447.14	314.50
T92+dG : T92+dG+GC	17, 19, 21	114.36	157.38	90.36	94.50	469.72	664.46

NOTE.—In each row, the null model (H_0) is compared with the next full model (H_1), assuming that the likelihood ratio statistic ($2[\ln L_1 - \ln L_0]$) follows a χ^2 distribution, with degrees of freedom (df) indicated. Log likelihood scores were obtained assuming the topology shown in figure 3*a* (see *Materials and Methods*). All tests are significant ($P < 10^{-6}$), 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.

Table 3
Estimates of the Shape Parameter (α) of the Discrete Gamma Distribution for the Data Sets of this Study

Model	*Adh*	*Ddc*	*Gpdh*	*Sod*	*Xdh*	Total
T92+dG.....	0.368	0.173	0.139	0.310	0.319	0.285
T92+dG+GC...	0.371	0.179	0.192	0.311	0.331	0.296
GTR+dG.....	0.379	0.179	0.168	0.313	0.310	0.278

NOTE.—α values were obtained under the topology shown in figure 3*a*. See table 2 footnote for model definitions.

nonnested models; therefore, their relative fit cannot be evaluated by chi-square. The T92+dG+GC model yields greater likelihood scores than the GTR+dG model in all cases (−3,785.8 vs. −3,809.7, −6,536.9 vs. −6,572.8, −3,960.1 vs. −3,971.9, −2,635.5 vs. −2,662.7, −15,942.9 vs. −16,102.3, and −29,730.7 vs. −29,894.9; log likelihood scores produced by T92+dG+GC vs. GTR+dG for *Adh*, *Ddc*, *Gpdh*, *Sod*, *Xdh*, and the combined data set), evidencing the importance of accounting for nucleotide composition differences among lineages. The results do not change when model fitting is conducted assuming the topology proposed by Throckmorton (1975; first topology in table 4; results not shown), which strengthens the conclusions from other studies (e.g., Yang 1994; Yang, Goldman, and Friday 1994), indicating that, in general, tree topology differences have only a minor effect on model selection. At any rate, it should be kept in mind that the log likelihood ratio test values shown in table 2 were obtained under a topology resulting from the use of phylogenetic methods that are stationary (see *Materials and Methods*); therefore, it would be expected that any bias that might exist owing to the adoption of this topology for model fitting should occur in a direction favoring the GTR+dG model.

Table 3 shows the estimates of the shape parameter α of the discrete gamma distribution for the different data sets obtained with the T92+dG, the T92+dG+GC, and the GTR+dG models assuming the first working topology. The three models yield basically the same estimates, with the T92+dG+GC model producing slightly larger α values than the other models. Overall, the value of α appears to be quite insensitive to the number of substitution types and nucleotide frequency parameters included in the model. Substitution rate differences from site to site are largest for *Gpdh* and *Ddc* (lowest α values), suggesting that these loci are subject to stronger functional constraints than *Adh*, *Sod*, and *Xdh*.

Phylogenetic Relationships of the Drosophilidae

Several simple methods for tree reconstruction were used to generate a topology with the five gene regions of this study all pooled together (see *Materials and Methods*). We used this topology and that proposed by Throckmorton (1975; topology 1 in table 4) as working hypotheses for modeling the molecular evolution of the sequence data by means of the likelihood ratio test. Using this approach, we found that the T92+dG+GC model gave a reasonable representation of the evolution of the sequences. Next, we used this description to gen-

Table 4
Kishino, Miyata, and Hasegawa's (1990) Resampling Estimated Log Likelihood (RELL) Test Applied to Seven Different Phylogenetic Hypotheses of the Drosophilidae

		Maximum-Likelihood Model			
		GTR+dG		T92+dG+GC	
Hypothesis	Reference	ln(*L*)	RELL	ln(*L*)	RELL
1. (Le, (Ch, (So, (VR, (Hi, Za)))))	Throckmorton (1975)	−29,918.88	0.10	−29,739.64	0.16
2. (Le, Ch, (Hi, (Za, (So, VR))))	Grimaldi (1990)	−29,944.57	0.00	−29,774.07	0.00
3. (Le, (Hi, (Ch, (Za, (So, VR)))))	DeSalle (1992)	−29,982.21	0.00	−29,806.49	0.00
4. (Le, Ch, (Hi, (So, (Za, VR))))	Powell and DeSalle (1995)	−29,931.38	0.00	−29,758.14	0.02
5. (Le, Ch, (So, (Za, (Hi, VR))))	Remsen and DeSalle (1998); Tatarenkov et al. (1999)	−29,895.01	1.30	−29,720.07	0.68
6. (Ch, (Le, (Wi, (So, (Za, (Hi, VR))))))	This study: NJ on GTR+dG distance	−29,894.86	25.00	−29,730.71	3.40
7. (Le, (Ch, (So, (Za, (Hi, VR)))))	This study: NJ on T92+dG+GC distance	−29,886.92	73.60	−29,708.15	95.74

NOTE.—Le = *Scaptodrosophila lebanonensis*; Ch = Chymomyza; Za = Zaprionus; Hi = Hirtodrosophila; VR = *virilis-repleta*; So = Sophophora subgenus; Wi = *Drosophila willistoni*. See table 2 footnote for model definitions.

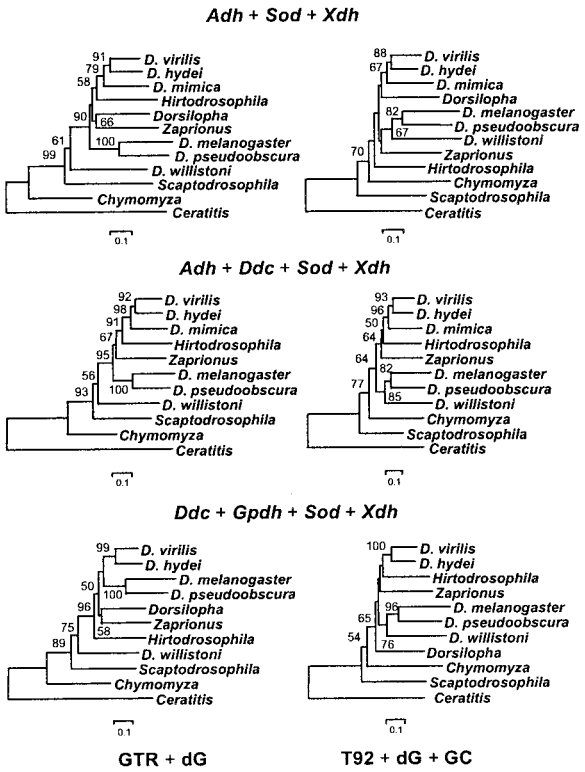

FIG. 2.—Neighbor-joining trees based on the general time-reversible distance allowing rate variation among sites (GTR+dG) and the Tamura (1992) distance allowing rate variation among sites and GC content variation among lineages (T92+dG+GC). α values used in distance computations are 0.285, 0.251, and 0.321 from top to bottom respectively, obtained with the T92+dG+GC model using the topology in figure 3*a*. Branch lengths are proportional to the scale, given in substitutions per nucleotide. Bootstrap values, based on 1,000 replications, are given on the nodes.

erate 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 *D. melanogaster* + *D. pseudoobscura* (supported by *Ddc*, *Sod*, and *Xdh* data when they are analyzed with the GTR+dG model, and by *Gpdh* and *Xdh* under the T92+dG+GC model) and *D. hydei* + *D. virilis* (supported by *Adh* under GTR+dG, and by *Xdh* using either model). *Adh* analyzed under the GTR+dG model also supported the connection of *D. mimica* to the cluster consisting of *D. hydei* and *D. virilis.* All of these are well-established relationships from other studies. In addition, analysis of the *Xdh* data with the GTR+dG model supported the cluster consisting of Ceratitis, Chymomyza, *D. willistoni*, and Scaptodrosophila, and also the association of Zaprionus with *D. busckii.*

Figure 2 shows the NJ trees derived from the GTR+dG and T92+dG+GC distance matrices using the *Adh* + *Sod* + *Xdh* data set (for which all species are available), and this data set combined separately with *Ddc* (*D. busckii* unavailable) and *Gpdh* (*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 Scaptodrosophila as more closely related to Drosophila than Chymomyza, and it places *D. willistoni* outside all other species of the Drosophila genus. In contrast, the T92+dG+GC model identifies Scaptodrosophila as the first derived lineage after Ceratitis (followed by Chymomyza) and places *D. willistoni* within the subgenus 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 *D. mimica* as the sister lineage to the clade consisting of *D. hydei* + *D. virilis* and the association of *D. melanogaster* with *D. pseudoobscura.* In addition, the *Adh* + *Ddc* + *Sod* + *Xdh* data set supports inclusion within the subgenus Drosophila of Zaprionus, which derives first, followed successively by Hirtodrosophila, *D. mimica*, and the clade consisting of *D. hydei* + *D. virilis.*

Figure 3*a* and *b* presents the NJ trees derived from the GTR+dG and T92+dG+GC models after combining all the information. *Drosophila mimica* and *D. busckii* are not included because they are unavailable for *Ddc* and *Gpdh*, respectively. 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 Chymomyza as the first derived after Ceratitis, followed by Scaptodrosophila and *D. willistoni*, while the T92+dG+GC model places Scaptodrosophila as derived before Chymomyza and places *D. willistoni* within the subgenus Sophophora. The two models agree, however, in that Hirtodrosophila splits after Zaprionus, followed by the clade consisting of *D. hydei* + *D. virilis*, all four pertaining to the subgenus 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 Chymomyza and/or *D. willistoni* in different positions with respect to each other and to the AT-rich outgroup *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 *Adh* and *Sod* regions, and hypothesis 5 was arrived at by Tatarenkov et al. (1999) after a sta-

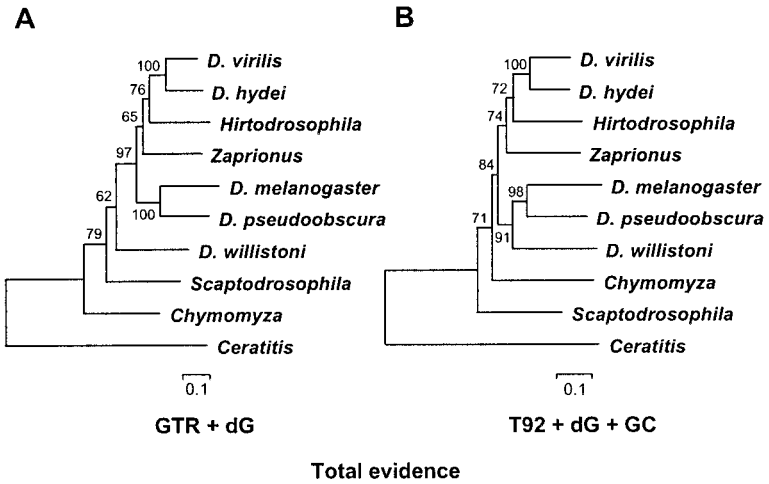

FIG. 3.—Neighbor-joining trees based on the general time-reversible distance allowing rate variation among sites (GTR+dG) and the Tamura (1992) distance allowing rate variation among sites and GC content variation among lineages (T92+dG+GC) for the total-evidence data set. Branch lengths are proportional to the scale, given in substitutions per nucleotide. Bootstrap values, based on 1,000 replications, are given on the nodes.

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 3*a* 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. 3*b*) 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 3*b* 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. 3*a*) 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, Tarrío, and Ayala 2000*c*), and also for inferring nucleotide composition of ribosomal RNA in the ''cenancestor'' (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

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. 3*a*; 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 Drosophila genus (likewise the GTR+dG model; see fig. 3*a*); 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. 3*a* 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. 3*a*). Our results strongly suggest that the GC-poor *D. willistoni* sequence is artifactually 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 on a more solid basis previous conclusions about the branching order of Zaprionus 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 can be problematic (see Whitfield and Cameron 1998; Steel, Huson, and Lockhart 2000). It has been shown that in situations like the one tackled in our study, in which the substitution process is nonstationary and substitution rates are unequal across sites, additive pairwise distance methods lose the ability to recognize the parametric topology (see Baake 1998). Despite these caveats, the results from the T92+dG+GC distance may be preferred, on the one hand, because no other pairwise distance measure exists, apart from the LogDet transformation, that could be applied to our data on a better-grounded theoretical basis. Moreover, it recovers a topology which is fully congruent with the topology achieved by explicit ML methods through the joint comparison of all sequences, although in this regard, it can be argued that we did not perform an exhaustive search (we just limited the ML analysis to a few hypotheses of interest; see table 4) or that the assumed gamma distribution does not appropriately accommodate the true among-sites rate variation present in the sequences (which, given the observed absence of stationarity, 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

BAAKE, H. 1998. What can and what cannot be inferred from pairwise sequence comparisons? Math. Biosci. **154**:1–21.

CUNNINGHAN, C. W., H. ZHU, and D. M. HILLIS. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution **52**:978–987.

DESALLE, R. 1992. The phylogenetic relationships of the flies in the family Drosophilidae deduced from mtDNA sequences. Mol. Phylogenet. Evol. **1**:31–40.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

———. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FOSTER, P. G., and D. A. HICKEY. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. **48**:284–290.

GALTIER, N., and M. GOUY. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA **92**:11317–11321.

———. 1998. Inferring the pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. **15**:871–879.

GALTIER, N., and D. MOUCHIROUD. 1998. Isochore evolution in mammals: a human-like ancestral structure. Genetics **150**:1577–1584.

GALTIER, N., N. TOURASSE, and M. GOUY. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science **283**:220–221.

GOLDMAN, N., J. P. ANDERSON, and A. G. RODRIGO. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49**:652–670.

GOLDMAN, N., and S. WHELAN. 2000. Statistical tests of gamma distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol. Biol. Evol. **17**:975–978

GRIMALDI, D. A. 1990. A phylogenetic revised classification of genera in the Drosophilidae (Diptera). Bull. Am. Mus. Nat. Hist. **197**:1–139.

HILLIS, D. M., and J. J. BULL. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42**:182–192.

HUELSENBECK, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. **44**:17–48.

HUELSENBECK, J. P., and K. A. CRANDAL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. **28**:437–466.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from the DNA sequence data, and the branching order in *Hominoidea*. J. Mol. Evol. **29**:170–179.

KISHINO, H., T. MIYATA, and M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

KWIATOWSKI, J., and F. J. AYALA. 1999. Phylogeny of Drosophila and related genera: conflict between molecular and anatomical analyses. Mol. Phylogenet. Evol. **13**:319–328.

KWIATOWSKI, J., D. SKARECKY, K. BAILEY, and J. A. AYALA. 1994. Phylogeny of Drosophila and related genera inferred from the nucleotide sequence of the Cu,Zn *Sod* gene. J. Mol. Evol. **38**:443–454.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, and D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

PÉLANDAKIS, M., and M. SOLIGNAC. 1993. Molecular phylogeny of Drosophila based on ribosomal RNA sequences. J. Mol. Evol. **37**:525–543.

POWELL, J. R. 1997. Progress and prospects in evolutionary biology: the Drosophila model. Oxford University Press, New York.

POWELL, J. R., and R. DESALLE. 1995. Drosophila molecular phylogenies and their uses. Evol. Biol. **28**:87–138.

REMSEN, J., and R. DESALLE. 1998. Character congruence of multiple data partitions and the origin of the Hawaiian Drosophilidae. Mol. Phylogenet. Evol. **9**:225–235.

RODRÍGUEZ-TRELLES, F., R. TARRÍO, and F. J. AYALA. 1999. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. Genetics **153**:339–350.

———. 2000*a*. Fluctuating mutation bias and the evolution of the base composition in Drosophila. J. Mol. Evol. **50**:1–10.

———. 2000*b*. Disparate evolution of paralogous introns in the *Xdh* gene of Drosophila. J. Mol. Evol. **50**:123–130.

———. 2000*c*. Evidence for a high ancestral GC content in Drosophila. Mol. Biol. Evol. **17**:1710–1717.

RUSSO, C. A., N. TAKEZAKI, and M. NEI. 1995. Molecular phylogeny and divergence times of *Drosophilid* species. Mol. Biol. Evol. **12**:391–404.

RZHETSKY, A., and M. NEI. 1995. Tests of the applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. **12**:131–151.

STEEL, M. A., D. HUSON, and P. J. LOCKHART. 2000. Invariable sites models and their use in phylogenetic reconstruction. Syst. Biol. **49**:225–232.

STEEL, M. A., P. J. LOCKHART, and D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. Nature **364**:440–442.

STEEL, M. A., L. A. SZIKELY, and M. D. HENDY. 1994. Reconstructing trees when sequence sites evolve at variable rates. J. Comput. Biol. **1**:153–163.

SWOFFORD, D. L. 1999. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0b2. Sinauer, Sunderland, Mass.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. Sinauer, Sunderland, Mass.

TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Mol. Biol. Evol. **9**:678–687.

TARRIO, R., F. RODRIGUEZ-TRELLES, and F. J. AYALA. 1998. New Drosophila introns originate by duplication. Proc. Natl. Acad. Sci. USA **95**:1658–1662.

———. 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. Phylogenet. Evol. **16**:344–349.

TATARENKOV, A., J. KWIATOWSKI, D. SKARECKY, E. BARRIO, and F. J. AYALA. 1999. On the evolution of *Dopa decarboxylase* (*Ddc*) and Drosophila systematics. J. Mol. Evol. **48**:445–462.

TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Pp. 57–86 *in* R. M. MIURA, ed. Some mathematical questions in biology—DNA sequence analysis. Lec. Math. Life Sci. Vol. 17, Providence, R.I.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

THROCKMORTON, L. H. 1975. The phylogeny ecology and geography of Drosophila. Pp. 421–436 *in* R. C. KING, ed. Handbook of genetics. Vol. 3. Plenum Press, New York.

TOURASSE, N. J., and W. H. LI. 1999. Performance of the relative-rate test under nonstationary models of nucleotide substitution. Mol. Biol. Evol. **16**:1068–1078.

WHEELER, M. R. 1981. The Drosophilidae: a taxonomic overview. Pp. 1–97 *in* M. ASHBURNER, H. L. CARSON, and J. N. THOMPSON JR., eds. The genetics and biology of Drosophila. Academic Press, New York.

WHELAN, S., and N. GOLDMAN. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol. Biol. Evol. **16**:1292–1299.

WHITFIELD, J. B., and S. CAMERON. 1998. Hierarchical analysis of variation in the mitochondrial 16S rRNA gene among Hymenoptera. Mol. Biol. Evol. **15**:1728–1743.

WOJTAS, K. M., L. VON KALM, J. R. WEAVER, and D. T. SULLIVAN. 1992. The evolution of duplicate glyceraldehyde-3-phosphate dehydrogenase genes in Drosophila. Genetics **132**:789–797.

YANG, Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

———. 1996. The among-site rate variation and its impact on phylogenetic analyses. TREE **11**:367–372.

———. 1999. PAML: phylogenetic analysis by maximum likelihood. Version 2.0g. Distributed by the author, Department of Biology, Galton Laboratory, University College London.

YANG, Z., N. GOLDMAN, and N. E. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol. Biol. Evol. **11**: 316–324.

YANG, Z., and D. ROBERTS. 1995. On the use of nucleic acid sequences to infer branchings in the tree of life. Mol. Biol. Evol. **12**:451–458.

MANOLO GOUY, reviewing editor

Accepted April 10, 2001