Genet. Res., Camb. (2010), **92**, *pp.* 25–38.
doi:10.1017/S001667231000008X

A supermatrix-based molecular phylogeny of the family Drosophilidae

KIM VAN DER LINDE[1]*, DAVID HOULE[1], GREG S. SPICER[2] AND SCOTT J. STEPPAN[1]

[1] *Department of Biological Science, Florida State University, Tallahassee, FL 32306-4295, USA*
[2] *Department of Biology, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132-1722, USA*

(*Received 30 July 2009 and in revised form 1 March 2010*)

Summary

The genus *Drosophila* is diverse and heterogeneous and contains a large number of easy-to-rear species, so it is an attractive subject for comparative studies. The ability to perform such studies is currently compromised by the lack of a comprehensive phylogeny for *Drosophila* and related genera. The genus *Drosophila* as currently defined is known to be paraphyletic with respect to several other genera, but considerable uncertainty remains about other aspects of the phylogeny. Here, we estimate a phylogeny for 176 drosophilid (12 genera) and four non-drosophilid species, using gene sequences for up to 13 different genes per species (average: 4333 bp, five genes per species). This is the most extensive set of molecular data on drosophilids yet analysed. Phylogenetic analyses were conducted with maximum-likelihood (ML) and Bayesian approaches. Our analysis confirms that the genus *Drosophila* is paraphyletic with 100% support in the Bayesian analysis and 90% bootstrap support in the ML analysis. The subgenus *Sophophora*, which includes *Drosophila melanogaster*, is the sister clade of all the other subgenera as well as of most species of six other genera. This sister clade contains two large, well-supported subclades. The first subclade contains the Hawaiian *Drosophila*, the genus *Scaptomyza*, and the *virilis-repleta* radiation. The second contains the *immigrans-tripunctata* radiation as well as the genera *Hirtodrosophila* (except *Hirtodrosophila duncani*), *Mycodrosophila*, *Zaprionus* and *Liodrosophila*. We argue that these results support a taxonomic revision of the genus *Drosophila*.

1. Introduction

Model organisms offer us our deepest understanding of many biological phenomena. Scientists are now capitalizing on the knowledge of these model systems to perform comparative studies of the phylogenetic relatives of many model organisms, such as *Arabidopsis thaliana* (Mitchell-Olds, 2001), *Caenorhabditis elegans* (Harris *et al*., 2004), *Danio rerio* (Quigley *et al*., 2005) and *Drosophila melanogaster* (Singh *et al*., 2009). Such studies promise to unravel the genetic basis of phenotypic evolution and strongly implicate the evolutionary forces responsible for species divergence. Clearly, comparative studies of phenotypic differences between species must be based on a good understanding of the phylogenetic relationships among the taxa involved (Felsenstein, 1985).

If the phylogeny is poorly known, the quality of the conclusions from comparative analyses will be poor. Unfortunately, the research traditions of those working on model organisms have not emphasized a phylogenetic framework, leaving us with an inadequate understanding of the relationships of model organisms to some of their close relatives (Al-Shehbaz & Kane, 2002; Kiontke *et al*., 2004; Quigley *et al*., 2004). The unfortunate result is that our phylogenetic information on clades containing model organisms is often fairly weak, even though these are precisely the clades best suited to comparative studies.

A prime example of this paradox is the genus *Drosophila* and closely related genera. Over the past 20 years, many studies dealing with parts of the

* Corresponding author. Department of Biological Science, Florida State University, Tallahassee, FL 32306-4295. Tel: ++1-850-645-8521. Fax: ++1-850-644-9829. e-mail: kim@kimvdlinde.com

Table 1. *ML estimates of parameter values for each data partition under the GTR+Γ model*

	Parameter										
Partition	A	C	G	T	A↔C	A↔G	A↔T	C↔G	C↔T	G↔T	α
nuc 1st	0·268	0·207	0·347	0·177	1·401	1·634	1·254	0·665	4·033	1	0·364
nuc 2nd	0·283	0·244	0·205	0·268	1·766	2·917	0·973	2·284	3·542	1	0·299
nuc 3rd	0·132	0·368	0·275	0·225	4·979	13·403	7·225	2·077	12·609	1	0·964
mt 1st	0·272	0·164	0·256	0·309	0·416	3·232	1·539	0·493	77·898	1	0·173
mt 2nd	0·227	0·223	0·151	0·399	1·661	2·587	0·819	1·527	1·029	1	0·125
mt 3rd	0·436	0·068	0·029	0·466	1·638	277·252	6·260	29·226	344·918	1	0·388
16S	0·363	0·161	0·126	0·351	0·153	0·929	1·880	0·657	1·835	1	0·195
28S	0·317	0·164	0·204	0·315	0·854	2·448	2·726	0·315	2·401	1	0·192
tRNA	0·314	0·188	0·180	0·319	0·914	5·416	3·429	0·602	2·087	1	0·494

Parameter values include the base frequencies and the instantaneous substitution rates between nucleotides. The G↔T is set to 1·0 by default. α is the value of the shape parameter for the Γ distribution describing among-site rate variation. Data partitions include the codon positions for protein coding regions of each genome (nuclear, 'nuc'; mitochondrial, 'mt'). Values were calculated by RAxML as described in the text. Particular values of note include the very strong AT bias in mt 3rd positions in contrast to the GC bias in nuc 3rd positions, much less even substitution rates at 3rd positions of both genomes, and the weak transition/transversion ratio in 16S.

Drosophila phylogeny have been published (for an overview, see van der Linde & Houle, 2008), but surprisingly few have adequately addressed the phylogenetic relationships among the subgenera of *Drosophila* in the context of the various closely related genera. Consequently, even a cursory examination of the literature reveals that many aspects of drosophilid phylogeny are controversial or poorly studied (Ashburner *et al*., 2005; Markow & O'Grady, 2006).

Grimaldi's (1990) phylogeny based on morphological characters is the most recent comprehensive family wide treatment. An important competing phylogenetic hypothesis is that of Throckmorton (1975), which differs from it in many respects. Throckmorton's work was clearly based on many sources of evidence (e.g. Throckmorton, 1962, 1965, 1966); unfortunately, he did not use explicit and reproducible methods. More recently, many phylogenetic hypotheses based on molecular data have been published (see Table 1, van der Linde & Houle, 2008, for the most important studies). The best of these have emphasized clades well below the genus level or have been based on small numbers of genes. Figure 1 shows the combination of gene numbers and species numbers in other phylogenetic studies using molecular data. Some aspects of the phylogeny, such as relationships within the *melanogaster* species subgroup (see Coyne *et al*., 2004), now seem robustly supported by analysis of large sets of molecular data. At the same time, the results from other studies show that various clades within the phylogeny differ in many key respects.

The underlying source of this lack of consensus is that the available data are fragmentary. Taxon sampling of this very large group has been haphazard, and a few studies sequence more than a small number of genes. The result is that the sequence available for a pair of closely related species is likely to come from

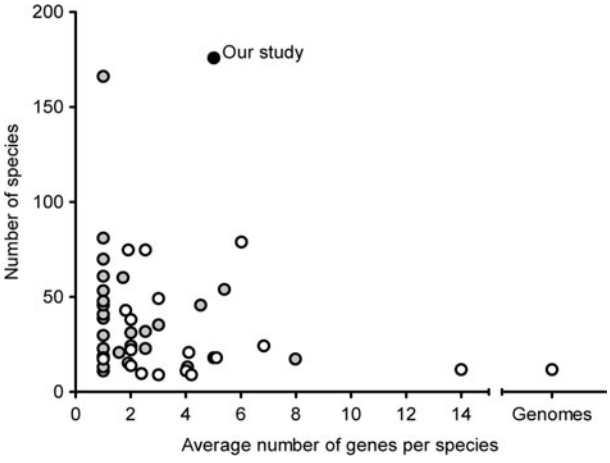

Fig. 1. Overview of average number of genes per species and number of species included in previous studies compared with those in our study. Black dot: our study; grey dots: studies limited to a single genus or *Drosophila* subgenus or major clade within the subgenus *Drosophila* (e.g. *immigrans-tripunctata* radiation, *virilis-repleta* radiation and Hawaiian *Drosophila*); white dots: studies covering at least two genera and/or *Drosophila* subgenera and/or major clades of the subgenus *Drosophila*.

different genes, making meaningful phylogenetic analyses difficult. We refer to this situation as a lack of overlap.

In this situation, phylogenetic hypotheses can be generated through a supertree analysis of published results (see, e.g. Bininda-Emonds *et al*., 2002; Bininda-Emonds, 2004). Our supertree analysis and review covering *Drosophila* and its closely related genera (van der Linde & Houle, 2008) resulted in a generally well-resolved tree and clearly showed that the genus *Drosophila* as currently defined is paraphyletic with respect to various genera (e.g. *Scaptomyza*, *Hirtodrosophila*, *Samoaia* and *Zaprionus*) placed within it

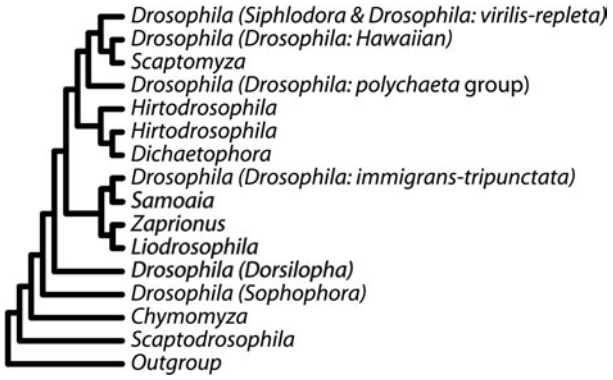

Fig. 2. The phylogenetic relationships as understood on the basis of the available literature before the study reported here.

(Fig. 2). Supertree methods have been criticized on various grounds, including that they give too much weight to weakly supported or erroneous nodes, difficulty in dealing with biased data, failure to use the original data, inapplicability of model-based methods of analysis and failure to use all the data efficiently (Kluge, 1989; Gatesy *et al*., 2004; but see also Bininda-Emonds *et al*., 2003; Bininda-Emonds, 2004). Some of these issues are apparent in our own supertree analysis (van der Linde & Houle, 2008).

Here, we report the results of a supermatrix analysis in which we used publicly available sequence data, plus a limited amount of new sequence chosen to increase overlap. Our data stand out in the number of species included (180; Fig. 1) and in the variety of taxa included. We note that, although ours is the most comprehensive study to date, overlap of the sequence data for many species remains limited. The focus of the study is to resolve the basal nodes within the genus *Drosophila sensu lato*. The major issue is the topology of the three main clades of the subgenus *Drosophila* and the various genera that are positioned among them (e.g. *Hirtodrosophila*, *Zaprionus* and *Scaptomyza*). In addition, the order of the two genera placed sister to the genus *D. sensu lato* (*Scaptodrosophila* and *Chymomyza*) is still insufficiently resolved (Tarrio *et al*., 2001). The phylogenetic relationships between the various species groups in both the *immigrans-tripunctata* clade and the *virilis-repleta* clade are not yet fully resolved. Finally, the monophyletic nature of many groups is questioned.

2. Materials and methods

(i) *Species and data*

We compiled data for 541 species: 535 species in family Drosophilidae and six outgroup species in the families Tephritidae (one species), Ephydridae (four) and Lauxaniidae (one). We screened loci in GenBank for which multiple drosophilids had been sequenced and selected 13 to maximize taxonomic overlap among gene sequences. These include nine nuclear loci (*Adh*, *AmyRel*, *per*, *Ddc*, *Sod*, *yp1*, *28Sd1*, *28Sd2* and *28Sd8*) and four mitochondrial loci (*COI*, *COII*, *COIII* and *16S*). Most sequences were obtained from GenBank, supplemented with new sequences for *COI*, *COII*, *COIII*, *28Sd1* and *28Sd8*. Only protein-coding regions of the non-ribosomal genes were used. The aligned sequences for *AmyRel* were kindly provided by J.-L. Da Lage. After generating the full set of data, we selected species on the basis of number of genes available, number of base pairs and our need for a distance matrix covering a large number of species for which we have wing-shape data. On the basis of these criteria, we selected 180 species (see Supplementary Table 1). The total number of base pairs per species ranged from 339 to 13 539 bp (mean: 4333 bp); only seven taxa had fewer than 1000 bp. Accession codes can be found in Supplementary Table 1; new sequences are available from GenBank under accession codes GU597372 to GU597535. Collection locations of the 39 species for which we collected new sequences can be found in Supplementary Table 2.

(ii) *Alignment*

Nucleotide sequences were aligned with Clustal X (Thompson *et al*., 1997) and manually inspected in MacClade (Maddison & Maddison, 2005) for resolution of regions of ambiguity or disagreement and consolidated indels. Alignment of all protein-coding regions was trivial because amino acid indels were rare and readily interpreted. Sequences for the genes were concatenated for each taxon. The alignment and trees [maximum parsimony (MP), partitioned maximum-likelihood (ML) and Bayesian] are available from TreeBase under accession code SN4940.

(iii) *Phylogenetic analysis*

Heterogeneity of nucleotide composition among informative sites was estimated with PAUP* version 4.0b10 (Swofford, 2002). Only two of the 13 genes, AmyRel and Adh, showed heterogeneity. Phylogenetic analyses were conducted under MP, neighbour-joining (NJ), ML and Bayesian approaches. Because of the number of species without overlapping data, we did not test for congruence using, for example, a partition-homogeneity test (Farris *et al*., 1994, 1995). All MP analyses were conducted with PAUP* version 4.0b10 (Swofford, 2002) with heuristic searches with tree bisection-reconnection (TBR) branch swapping and 20 random-addition starting trees; the first 2000 trees found were retained. All substitutions were weighted equally; gaps were treated as missing data. To determine whether changes in base composition in AmyRel and Adh might mislead phylogenetic reconstruction, we analysed these genes separately using NJ with LogDet distances in PAUP* and

compared the results to the ML and Bayesian trees for well-supported conflicts. LogDet is less sensitive to base-composition heterogeneity (Swofford *et al.*, 1996). Although the LogDet NJ tree and the ML/Bayesian trees differed, the differences were not well supported, and the ML gene tree was not appreciably more similar to the concatenated tree than was the LogDet NJ tree, suggesting that convergence on nucleotide frequencies was not misleading the combined-data analyses.

We conducted ML analyses in several ways to account for the complexity of the data. First, a series of analyses on unpartitioned data was conducted with PAUP*. ML parameter values were estimated under a nested array of substitution models for a random MP tree as implemented in Modeltest 3.04 (Posada & Crandall, 1998); likelihood-ratio tests (Yang *et al.*, 1995) and the Akaike Information Criterion (Akaike, 1973) were used to identify the simplest models of sequence evolution that adequately fit the data and phylogeny. The most complex model was selected by both criteria: GTR + I + Γ. Parameters were fixed for the values estimated on the initial MP tree. Heuristic searches were then conducted with two alternatives for generating starting trees. In the first search, the first 20 trees were saved from each of 10 random-addition replicates from MP analyses (200 maxtrees total), and these 200 trees became the starting trees for the ML search, providing some initial sampling of tree space. MP trees were used rather than a single NJ tree because non-overlapping data among sets of taxa resulted in undefined distances and anomalous placement of some taxa in the NJ tree. Parsimony appeared less affected by non-overlapping data, in that recovered trees were consistent with published studies. In addition, five parallel random-sequence-addition runs were conducted, each of which required approximately 200 h to build the starting trees. The search starting with the MP trees yielded a more likely tree ($\ln L = -178603$ as opposed to -178606), and that tree is reported here.

Our data include protein-coding and RNA genes from mitochondrial and nuclear genomes, so a single model may be insufficient to account for the data. We therefore conducted partitioned ML analyses using RAxML (Stamatakis *et al.*, 2008) using the Cipres Portal 1.15 to access the San Diego Supercomputer Center. Conditions followed the default settings, including a GTR + Γ model, and no per-gene branch length optimization (branch lengths were proportional across partitions). We partitioned the data by codon position separately for the three nuclear codon positions and the three mitochondrial codon positions, as well as by genome for the ribosomal genes and for the tRNAs, producing nine partitions (1st, 2nd and 3rd nuclear; 1st, 2nd and 3rd mitochondrial; mt rRNA, nu rRNA and tRNA). We repeated the analyses unpartitioned for comparison with the PAUP* results to distinguish the effects of partitioning from software-specific tree-search strategies. Parameter values as estimated by RAxML for each partition are reported in Table 1.

Non-parametric bootstrapping (Felsenstein, 1985) was performed under ML with two approaches, one using the genetic algorithm approach in GARLI version 0.951 (Zwickl, 2006) with the data unpartitioned and the other using RAxML to allow partitioned data. Garli analyses used 200 replicates and the GTR + I + Γ model with parameters estimated by GARLI. The search was conducted with a random starting tree and an automatic run termination after a minimum of 5000 generations that did not improve topology, a $\ln L$ improvement of less than 0·02 due to topological changes, a 0·05 score improvement threshold, and default genetic algorithm settings. The second bootstrapping approach used RAxML with both unpartitioned and partitioned data sets. Bootstrapping was run for 250 replicates (RAxML selected 150 as sufficient), representing the combined output of two independent runs (100 and 150 replicates) with different starting seeds, with default settings. Partitioning was the same as in the ML searches. MP bootstrapping was conducted with 500 replicates, and 50 trees were saved in each random-addition replicate.

Bayesian analyses used the mpi (multiple processors) version of MrBayes 3.2 (Ronquist & Huelsenbeck, 2003) distributed over eight processors. We partitioned the data as with RAxML. Two independent analyses of four heated chains each were run for 80 million generations. Parameters were estimated for each partition separately ('unlinked'); trees and parameters were recorded every 1000 generations. Convergence was estimated by means of cumulative and sliding plots from "Are We There Yet?" (AWTY) (Wilgenbusch *et al.*, 2004) as well as by examination of likelihood plots and posterior probabilities of individual clades for subsets of the runs. The chains converged slowly on the basis of the AWTY and likelihood plot diagnostics, possibly because large blocks of data were missing and overlap among several taxa was limited, yielding a burn-in of 50%. Split-frequency standard deviations never indicated full convergence, plateauing around 0·1 by 20 million generations. We calculated a majority-rule consensus tree of the post-burn-in trees from both runs to summarize posterior probabilities.

3. Results and discussion

At the level of the previously recognized genera, the results of the ML and Bayesian analyses were strongly concordant with each other (Fig. 3). The main differences between the analyses are in the placement of the *immigrans-tripuncata*, *Zaprionus* and

Fig. 3. Partitioned ML phylogram of the Drosophilidae. Taxon names along branches are the traditional classification, whereas names within the coloured boxes are the proposed genus names for species currently assigned to *Drosophila*. Branches coloured by current genus; *Drosophila* appears in black. Support values above branches are unpartitioned ML bootstrap (bs; PAUP), partitioned ML bs (RAxML) and Bayesian posterior probabilities (pp), respectively, expressed as percentages. Only bs >70 or pp >85 are shown. Symbols: '*' indicates 100; '—' indicates the node is below 50% in the bootstrap majority tree and not present in the Bayesian majority rule tree.

Hirtodrosophila/*Mycodrosophila* clades and among the poorly supported nodes near the root of *immigrans-tripuncata* clade. The concordance between the ML and Bayesian analyses within genera differed dramatically for different genera. The MP results were also largely concordant, lacking any significant conflict (MP bs $>80\%$); the strict-consensus MP tree and the ML tree had 140 nodes in common. All but three differences (all in subgenus *Drosophila*, discussed individually below) were confined to regions poorly supported in all analyses and with $<42\%$ MP bootstrap support. MP bootstrap values were strongly correlated with the other support values and were nearly always the lowest of the four values. Because MP bootstrap values do not provide a strong independent signal, only the model-based values are reported on Fig. 3.

(i) *Steganinae and Drosophilinae*

Traditionally, the family Drosophilidae is split into two subfamilies, the Drosophilinae and the Steganinae (Hendel, 1917; Duda, 1924; Throckmorton, 1962, 1965, 1975; Okada, 1989; Grimaldi, 1990; Sidorenko, 2002). Our small sample of subfamily Steganinae (four taxa) suggests that it may be paraphyletic with

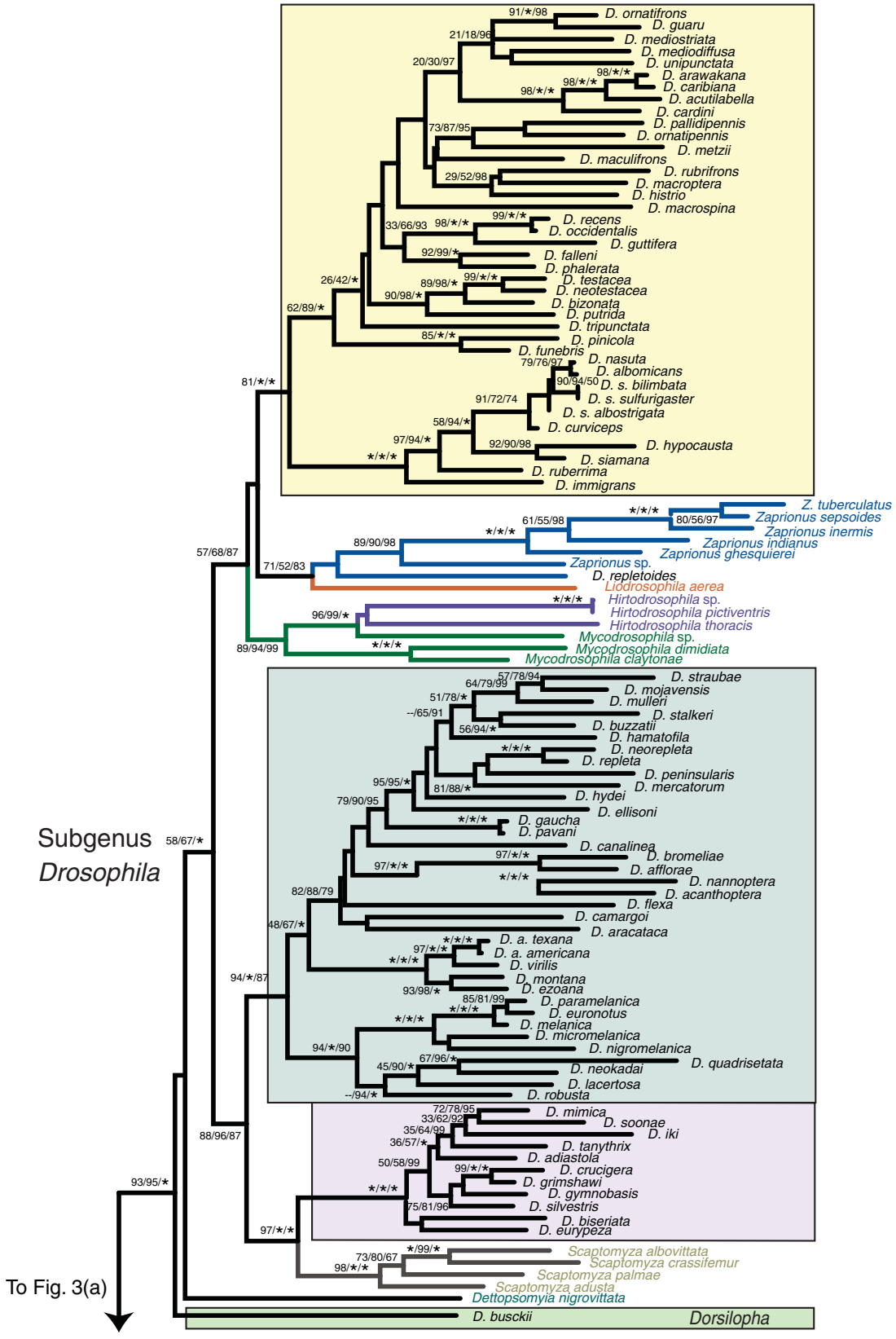

Fig. 3. (*Cont.*)

respect to the subfamily Drosophilinae. Although the Steganinae are monophyletic in the partitioned ML tree (Fig. 3*a*), the single species of the genus *Leucophenga* is the sister taxon to the Drosophilinae in the Bayesian and the unpartitioned ML analyses. The node supporting paraphyly was poorly supported

[ML unpartitioned bootstrap (MLu bs): 42; Bayesian posterior probability (pp), reported as a percentage: 81] relative to the nodes for the family Drosophilidae (MLu bs: 95; pp: 100) and subfamily Drosophilinae (MLu bs: 58; pp: 100).

Previous analyses of these subfamilies suggest that no single character distinguishes the two subfamilies (see Ashburner *et al*., 2005, for discussion). The only molecular study that has included species of both subfamilies (Remsen & O'Grady, 2002) confirmed this basal subdivision, although one of the unrooted molecular trees (16S) also suggests that this subfamily is paraphyletic. Our results stress the need for additional sampling of species from both subfamilies before any conclusions regarding the basal nodes within the family Drosophilidae can be drawn.

(ii) Scaptodrosophila *and* Chymomyza

Our analyses are equivocal with respect to the positions of *Chymomyza* and *Scaptodrosophila*, and the two alternative arrangements are both poorly supported. *Chymomyza* appears monophyletic [two species, MLu bs, ML partitioned bootstrap (MLp bs) and pp support all 100], whereas *Scaptodrosophila* is paraphyletic with respect to two moderately to well-supported clades (*Scaptodrosophila deflexa* plus *Scaptodrosophila lebanonensis* group, 99–100 for all analyses; *Scaptodrosophila dorsocentralis*/*latifasciaformis* only moderate). Examination of the underlying data reveals a very limited overlap in sequences between species of the two *Scaptodrosophila* clades, which could explain the lack of support for monophyly of the genus.

Scaptodrosophila and *Chymomyza* appear to be the closest relatives to *Drosophila* for which sequences are available. Most morphological (Okada, 1963; Hu & Toda, 2001) and molecular (DeSalle, 1992*a*; Kwiatowski *et al*., 1994, 1997) studies do not contradict Tarrio *et al*. (2001), who suggested that *Scaptodrosophila* diverged from other drosophilids before *Chymomyza* did, on the basis of almost 5000 bp of sequence from five nuclear genes. Remsen & O'Grady (2002) found that *Scaptodrosophila* and *Chymomyza* formed a sister clade to the *Sophophora* but with low support. Other studies have been unable to resolve this node (Throckmorton, 1975; Grimaldi, 1990; Remsen & DeSalle, 1998; Kwiatowski & Ayala, 1999; Tatarenkov *et al*., 1999; Da Lage *et al*., 2007). Our results do not provide clear support for any of the topologies and raise questions about the monophyly of the genus *Scaptodrosophila*. These issues can only be resolved with further data.

(iii) *Genus* Drosophila *and included genera*

The genus *Drosophila*, together with the included genera, forms a moderately supported clade (MLu bs: 58; MLp bs: 79; pp: 100), and most species are placed in two major clades. One is the monophyletic subgenus *Sophophora* (MLu bs: 58; MLp bs: 79; pp: 100; Fig. 3*a*). The remaining genera and subgenera form a separate well-supported clade (MLu bs: 93; MLp bs: 95; pp: 100; Fig. 3*b*), within which most species are distributed over two major clades. The first major clade contains the Hawaiian drosophilids – Hawaiian *Drosophila* clade and *Scaptomyza* – and the *virilis-repleta* radiation and is well supported (MLu bs: 88; MLp bs: 96; pp: 87). The two Hawaiian clades are each monophyletic and sister to each other (98–100 for all analyses). The sister taxon to the Hawaiian drosophilids is the *virilis-repleta* radiation, which is monophyletic in all analyses with strong support (MLu bs: 94; MLp bs: 100; pp: 87).

The second major clade contains three groups (Fig. 3*b*), the *immigrans-tripunctata* radiation, *Zaprionus* and *Hirtodrosophila*/*Mycodrosophila*. Together with some of the smaller genera, they form a weakly supported clade (MLu bs: 51; MLp bs: 68; pp: 87). Unfortunately, the analyses failed to resolve the topology among the three basal lineages consistently. The *immigrans-tripunctata* radiation was monophyletic in all analyses (MLu bs: 81; MLp bs: 100; pp: 100), as was the genus *Zaprionus* (MLu bs: 89; MLp bs: 90; pp: 98). The *Hirtodrosophila* and *Mycodrosophila* species, except *H. duncani*, form a clade (MLu bs: 89; MLp bs: 94; pp: 99). The genus *Liodrosophila* was placed in the sister clade of the genus *Zaprionus* in all analyses, although with weak support.

The subgenus *Dorsilopha* and the genus *Dettopsomyia* were placed basal to these two major clades. In the unpartitioned ML analysis, the two genera form a clade with weak bootstrap support (61%), whereas in the Bayesian analysis, they are successive outgroups to the major lineage, but the pp for the relevant node is only 62%.

The largest *Drosophila* phylogeny to date with respect to number of genes used is based on the genomes of the 12 sequenced species (*Drosophila* 12 Genomes Consortium, 2007). The topology of the 12-genome study is identical with our topology for the same species, underlining the robustness of our analysis. The subgenus *Sophophora* is the sister clade of the remaining subgenera, as well as of the regularly included genera (Beverley & Wilson, 1984; DeSalle, 1992*b*; Wojtas *et al*., 1992; Pélandakis & Solignac, 1993; Thomas & Hunt, 1993; Kwiatowski *et al*., 1994, 1997; Russo *et al*., 1995; Remsen & DeSalle, 1998; Kwiatowski & Ayala, 1999; Tatarenkov *et al*., 1999; Tarrio *et al*., 2001; Remsen & O'Grady, 2002; Robe *et al*., 2005; Da Lage *et al*., 2007). The position of the major clades in the subgenus *Drosophila* combined with the included genera is not recovered consistently, but our results support the grouping of the Hawaiian

Drosophila clade with *Scaptomyza* (Throckmorton, 1975; DeSalle, 1992*a*; Thomas & Hunt, 1993; Kambysellis *et al*., 1995; Russo *et al*., 1995; Remsen & DeSalle, 1998; Kwiatowski & Ayala, 1999; Tatarenkov *et al*., 1999, 2001; Davis, 2000; Davis *et al*., 2000*b*; Remsen & O'Grady, 2002; Da Lage *et al*., 2007; O'Grady & DeSalle, 2008), which together form the sister clade of the *virilis-repleta* radiation (Kambysellis *et al*., 1995; Russo *et al*., 1995; Remsen & DeSalle, 1998; Kwiatowski & Ayala, 1999; Tatarenkov *et al*., 1999, 2001; Gailey *et al*., 2000; Tarrio *et al*., 2001; Tatarenkov & Ayala, 2001; Remsen & O'Grady, 2002; Da Lage *et al*., 2007). Most studies recover the remaining three major clades – *Zaprionus*, the *immigrans-tripunctata* radiation and *Hirtodrosophila*/*Mycodrosophila* – as sister taxa to each other, but no consensus has emerged about their branching order (Kwiatowski & Ayala, 1999; Tatarenkov *et al*., 1999; Davis *et al*., 2000*a*; Gailey *et al*., 2000; Robe *et al*., 2005; Da Lage *et al*., 2007).

(iv) *Subgenus* Sophophora

Our analyses recover the previously identified Neotropical and Old World clades within the subgenus *Sophophora*. The Neotropical clade contains the *willistoni* and *saltans* species groups (all support values: 100), whereas the 'Old World' clade contains the *obscura*, *ananassae*, *montium* and *melanogaster* species groups (MLu bs: 71; MLp bs: 94; pp: 100). The *obscura* and *ananassae* species groups form sequential sister groups to the *montium* and *melanogaster* species groups (cf. Da Lage *et al*., 2007). Each of the six species groups was monophyletic (MLbs: 96–100; pp: 100), and the topology was well supported in both the ML and Bayesian analyses.

The subgenus *Sophophora* has been traditionally split into eight species groups, of which the four largest – *melanogaster*, *obscura*, *saltans* and *willistoni* – are generally included in phylogenetic studies (Pitnick *et al*., 1999; Tatarenkov *et al*., 1999; Bächli, 1999–2009; O'Grady & Kidwell, 2002; Remsen & O'Grady, 2002; Ashburner *et al*., 2005; Da Lage *et al*., 2007). Recently, Da Lage *et al*. (2007) proposed to elevate the *montium* and *ananassae* species subgroups to the level of species groups, bringing the total number to 10. The 10 groups are distributed among the two major clades, one containing the *melanogaster*, *montium*, *ananassae* and *obscura* species groups and the other containing the *willistoni* and *saltans* species groups (Pélandakis *et al*., 1991; Pélandakis & Solignac, 1993; Russo *et al*., 1995; Kwiatowski & Ayala, 1999; Pitnick *et al*., 1999; O'Grady & Kidwell, 2002; Remsen & O'Grady, 2002; Da Lage *et al*., 2007).

Our study confirms the generally accepted topology of the subgenus *Sophophora*, as well as the validity of the proposal by Da Lage *et al*. (2007) to split the *melanogaster* species group into three: *ananassae*, *montium* and *melanogaster*. Among these three species groups, the *montium* and *melanogaster* groups are sister to each other (MLu bs: 83; MLp bs: 97; pp: 100). This topology has been observed in many studies (cf. Hsu, 1949; Inomata *et al*., 1997; Goto & Kimura, 2001; O'Grady & Kidwell, 2002; Kastanis *et al*., 2003; Lewis *et al*., 2005*b*; Kopp, 2006; Prud'homme *et al*., 2006; Da Lage *et al*., 2007). In this context, the placement of the *fima* species group as the sister clade of the *ananassae* species subgroup (Pélandakis & Solignac, 1993) provides an additional argument for accepting the proposed split of the *melanogaster* species group.

The subgenus *Sophophora* as currently defined is already recognized as paraphyletic; the genus *Lordiphosa* (not included in our study) is the sister clade of the *willistoni-saltans* clade (Katoh *et al*., 2000; Hu & Toda, 2001). In our study, the only member of the *duncani* species group, *H. duncani* (Wheeler, 1949), was also placed in this subgenus as the sister to the *willistoni-saltans* clade.

(v) Virilis-repleta *radiation*

Most studies suggest division of the *virilis-repleta* radiation into *repleta* and *virilis* clades (Pitnick *et al*., 1999; Katoh *et al*., 2000; Carrasco *et al*., 2003; Robe *et al*., 2005; Wang *et al*., 2006). In our analyses, the *virilis* species group was placed sister to the *repleta* clade with high support in the Bayesian analysis (pp = 100) but not in the ML analyses (MLu bs: 48; MLp bs: 67). The *repleta* clade, which includes one of the two species of the subgenus *Siphlodora* (*Drosophila flexa*), was reasonably well supported (MLbs: 82; MLp bs: 88; pp: 79), but the topology of the species groups within the clade differed slightly in the MP, ML and Bayesian analyses, primarily in the position of the *dreyfusi* species group (*Drosophila camargoi*). Generally, the support for the nodes dealing with the placement of the various species groups is low. Monophyly for most of the species groups is strongly supported, except for the *robusta* species group, which is paraphyletic, in line with a previous study including this group (Wang *et al*., 2006).

The topology within the *virilis-repleta* radiation is poorly resolved, although some consensus on aspects of the tree has emerged. The *robusta*, *melanica*, *angor* and *quadrisetata* species groups generally form a clade (MLu bs: 94; MLp bs: 100; pp: 90) (Watabe & Peng, 1991; Pitnick *et al*., 1999; Remsen & O'Grady, 2002; Wang *et al*., 2006), whereas the *repleta* clade consists of the *repleta*, *mesophragmatica*, *bromeliae*, *dreyfusi*, *annulimana*, *flavopilosa* and *canalinea* species groups (Pitnick *et al*., 1999; Tatarenkov & Ayala, 2001; Remsen & O'Grady, 2002; Carrasco *et al*., 2003;

Robe *et al*., 2005; Wang *et al*., 2006; Da Lage *et al*., 2007) with moderate support (MLu bs: 82; MLp bs: 88; pp: 79). The studies differ in the position of the *virilis* species group; Pitnick *et al*. (1999) and Wang *et al*. (2006) place it sister to the *robusta-melanica* clade, whereas others place it sister to the *repleta* clade (Pélandakis & Solignac, 1993; Tatarenkov & Ayala, 2001; Remsen & O'Grady, 2002). Our results favour its placement sister to the *repleta* clade (MLu bs: 48; MLp bs: 67; pp: 100), but further study is clearly needed to resolve this issue. The *nannoptera* species group is generally placed within the *repleta* clade (Pitnick *et al*., 1999; Tatarenkov & Ayala, 2001; Carrasco *et al*., 2003; Wang *et al*., 2006), whereas Robe *et al*. (2005) place it outside the whole genus on the basis of a Bayesian analysis of the *amd* gene, while Remsen & O'Grady (2002) place it sister to the *melanica-robusta* clade.

D. flexa belongs to the small subgenus *Siphlodora*, which has only been included in a single analysis before ours (Remsen & O'Grady, 2002), where it was placed as the sister species of the *nannoptera* species group. Our results show it to be a basal member of the *repleta* lineage, which includes the *nannoptera* species group.

Parsimony differs from the model-based results, showing moderate conflict in two respects. MP places *Drosophila hydei* sister to *Drosophila hamatofila* (MP bs: 63%) and *Drosophila buzzatii* sister to the *Drosophila mulleri*/*mojaviensis* clade (MP bs: 54%).

(vi) Immigrans-tripunctata *radiation*

Our results place the *immigrans* clade (bs and pp: 100) as the sister group to the *tripunctata*/*funebris* groups (MLu bs: 62; MLp bs: 89; pp: 100), but relationships within the *tripunctata*/*funebris* clade are unstable, and many nodes are poorly supported. In agreement with previous studies, the *tripunctata* and *guarani* clades are not monophyletic (Frota-Pessoa, 1954; Throckmorton, 1975; Carrasco *et al*., 2003; Yotoko *et al*., 2003; Robe *et al*., 2005; Da Lage *et al*., 2007; Hatadani *et al*., 2009). The broad coverage of species in this study also suggests that the *testacea* and *funebris* species groups are not monophyletic. *Drosophila bizonata* (*bizonata* species group) is placed within the *testacea* species group, whereas the two species of the *funebris* species group are placed at different locations in the tree. Additional work is needed in this group.

Previous studies are generally consistent regarding the major splits in the *immigrans-tripunctata* radiation. All but two (Da Lage *et al*., 2007; Katoh *et al*., 2007) have concluded that the *immigrans-tripunctata* group is monophyletic. Most authors place the *immigrans* species group sister to the remaining members of the *immigrans-tripunctata* radiation (Pélandakis & Solignac, 1993; Remsen & O'Grady, 2002; Carrasco *et al*., 2003; Perlman *et al*., 2003; Yotoko *et al*., 2003; Robe *et al*., 2005), but mirroring our results, previous studies are equivocal about relationships within the *tripunctata*/*funebris* clade (Pélandakis & Solignac, 1993; Remsen & O'Grady, 2002; Carrasco *et al*., 2003; Yotoko *et al*., 2003; Robe *et al*., 2005; Da Lage *et al*., 2007; Hatadani *et al*., 2009). Additional studies with a better coverage of species will be required before the relationships between the various groups can be resolved.

(vii) Drosophila repletoides

Yassin and co-workers (Yassin, 2007; Yassin *et al*., 2008, 2010) have shown that the *tumiditarsus* species group, which contains *D. repletoides*, is positioned close to the genus *Zaprionus* and that several species of the *Zaprionus* subgenus *Anaprionus* actually belong to the *tumiditarsus* species group. Our only representative of the subgenus *Anaprionus* has not been affected by this taxonomic change, and is closely related to the remaining species of the genus (cf. Yassin *et al*., 2010; Amir Yassin, personal communication). Our model-based analyses indicate that *D. repletoides* is most probably in a clade with the genera *Zaprionus* and *Liodrosophila* (MLu bs: 71; MLp bs: 52; pp: 83), together sister to the *immigrans-tripunctata* radiation. In contrast, the MP analysis weakly places it sister to the Hawaiian *Drosophila*/*Scaptomyza*/*Dettopsomyia* clade, with no intervening node with greater than 27% MP bs. The ML and Bayesian results are more decisive; several intervening clades are well supported (e.g. the *virilis-repleta* radiation/Hawaiian *Drosophila*/*Scaptomyza* clade).

(viii) *Paraphyly of taxa*

Our results confirmed that several species groups, e.g. *tripunctata*, *guarani*, *testacea*, *robusta* and *funebris*, as well as genera, e.g. *Drosophila* and *Hirtodrosophila*, were paraphyletic or even polyphyletic. Our results indicated that *Hirtodrosophila* is not monophyletic, as suggested by Bächli *et al*. (2004), whereas the monophyly of *Scaptodrosophila* is unclear. For *Scaptodrosophila*, the paraphyly might just be an artefact of the underlying data (see above), in that the overlap in sequences between the two clades is relatively small. The situation for *Hirtodrosophila* is clearly different. In our analysis, the overlap in the sequences between *H. duncani* and *Hirtodrosophila thoracis* spans four different genes. *H. duncani* is placed within the subgenus *Sophophora* sister to the *willistoni* and *saltans* species groups, whereas *H. thoracis* is grouped with the other *Hirtodrosophila* species, nested within *Mycodrosophila*. Previous work

has suggested this placement of *H. duncani*. It was placed in its own unique species group on the basis of its unique male genitalia (Wheeler, 1949). Nater (1950, 1953) concluded that the male genitalia are most similar to those of the *obscura* subgroup. Burla (1956) noted that, among *Hirtodrosophila*, this species is closest to *Drosophila* in several internal and external morphological characteristics, whereas Throckmorton (1962) placed it close to or in *Sophophora*. Finally, Grimaldi (1990) concluded that apart from 'the presence of the long sensilla trichodea on the first flagellomere … *Drosophila duncani* has no other *Hirtodrosophila* features'.

Several species groups were paraphyletic or polyphyletic as well. The most striking is the *tripunctata* species group, whose members were scattered within the *immigrans-tripunctata* radiation. This result is in agreement with previous studies (Frota-Pessoa, 1954; Throckmorton, 1975; Carrasco *et al*., 2003; Yotoko *et al*., 2003; Robe *et al*., 2005; Da Lage *et al*., 2007; Hatadani *et al*., 2009). Additional studies are needed on this group.

The *guarani* species group is paraphyletic with two major species subgroups – *guarani* (including *Drosophila ornatifrons* and *guaru*) and *guaramun* (*Drosophila maculifrons*) – positioned in different locations of the *immigrans-tripunctata* radiation. The support for splitting this group was especially strong in the Bayesian analysis (two intervening nodes with pp >95), but weak in the ML analyses (largest intervening MLp bs $=50\%$). This result agrees with previous studies (Kastritsis, 1969; Clayton & Wheeler, 1975; Throckmorton, 1975; Yotoko *et al*., 2003; Robe *et al*., 2005; Hatadani *et al*., 2009); treating both subgroups as species groups seems to be fully justified.

A third species group suspected to be polyphyletic is the *robusta* group (Wang *et al*., 2006), and this conclusion was corroborated in our analysis as well. An unexpected find in the *immigrans-tripunctata* radiation was the placement of the single included species of the *bizonata* species group (*D. bizonata*) within the *testacea* species group. Before our study, no analysis has included members of the *bizonata* group together with multiple species of the *testacea* group. The support in both the ML (MLu bs: 90; MLp bs: 98) and the Bayesian analyses is strong (100). Finally, the *funebris* species group was not monophyletic. The two subgroups – *funebris* and *macrospina* – were positioned at different locations within the phylogeny, although the ML and Bayesian analyses differed in the exact position. Support was strong for a closer relationship of *Drosophila funebris* with *Drosophila pinicola* than with *Drosophila macrospina* (MLu bs: 85; MLp bs: 100; pp: 100).

The paraphyletic nature of the genus *Drosophila* has been reported before by various authors (Throckmorton, 1962, 1965, 1975; Beverley & Wilson, 1984; Grimaldi, 1990; DeSalle, 1992*a*, *b*; Pélandakis & Solignac, 1993; Thomas & Hunt, 1993; Kwiatowski *et al*., 1994, 1997; Kambysellis *et al*., 1995; Russo *et al*., 1995; Remsen & DeSalle, 1998; Tatarenkov *et al*., 1999, 2001; Davis *et al*., 2000*b*; Gailey *et al*., 2000; Katoh *et al*., 2000; Hu & Toda, 2001; Tarrio *et al*., 2001; Remsen & O'Grady, 2002; Da Lage *et al*., 2007; O'Grady & DeSalle, 2008; van der Linde & Houle, 2008). Our results confirm that several genera – *Hirtodrosophila*, *Mycodrosophila*, *Zaprionus*, *Scaptomyza* and *Liodrosophila* – are placed between the three major clades of the subgenus *Drosophila*. The ML and Bayesian analyses differed slightly in the exact placement of the *Hirtodrosophila*/*Mycodrosophila*, *Zaprionus* and *immigrans-tripunctata* clades, but agreed on all other nodes.

A resolution to the paraphyletic nature of the genus *Drosophila* will be addressed separately. Three solutions are available: (1) do nothing (O'Grady & Markow, 2009); (2) sink the included genera into *Drosophila* and (3) split the genus along the major clades. Splitting the genus is clearly the most desirable from a purely taxonomic point of view but has the major practical disadvantage that the type specimen for the genus is *D. funebris* (Fabricius) which is not in the same clade as *D. melanogaster*. To avoid the widespread confusion that would result from renaming *D. melanogaster*, an application to preserve the name *D. melanogaster* has been submitted to the International Commission on Zoological Nomenclature and is currently under consideration (van der Linde *et al*., 2007). If the commission rules in favour of this application, proposals to split the genus are more likely to be entertained by the enormous *Drosophila* community.

(ix) *Impact of analytical approach*

Despite the limited character overlap for some species and uneven sampling, our results are remarkably robust across the methods of analysis. Results obtained with different methods show no strong conflict. Many clades are well supported in all of the analyses. In addition, partitioning under ML had little effect on the topology. The partitioned and unpartitioned results differ in only a small number of nodes, and none of these differences received strong support under either partitioning scheme. Partitioned bootstrap values were generally greater than unpartitioned. For example, 45 nodes had greater bootstrap values under partitioning (of four percentage points or more) than without, compared to 14 with the reverse, for those values reported on Fig. 3. Sixty-three (35·6%) nodes differed by no more than three points. The most noticeable differences were between Bayesian and ML support values. As has often been reported, posterior

probabilities can be much higher than bootstrap values, sometimes artefactually so (Lewis *et al*., 2005*a*). The most striking examples in our analyses were in the *immigrans-tripunctata* radiation, where bootstrap values of less than 40, some as low as 18, were associated with near 100% posterior probabilities. We therefore interpret some of these high pp values with caution.

Considerable debate has surrounded the relative merits of the supermatrix approach as used here and the supertree approach, the latter synthesizing topological information across studies (Bininda-Emonds *et al*., 2003; Bininda-Emonds, 2004; Gatesy *et al*., 2004; de Queiroz & Gatesy, 2007). A particular concern about the supermatrix approach is the potential bias created by large numbers of missing data, an issue side-stepped by supertree approaches because the latter do not directly analyse characters. Most studies exploring missing data in large sets have concluded that the supermatrix approach is relatively robust to this problem and that the real question is how many informative data is present, not how many might be missing (Wiens, 2003; de Queiroz & Gatesy, 2007; Wiens & Moen, 2008). Reassuringly, the primary results we report here are well supported and consistent with those revealed by a supertree approach (van der Linde & Houle, 2008), a pattern that Baker *et al*. (2009) argued was evidence against a misleading bias in either method. We note, however, that the simulation studies have concentrated on the number of missing data, not their distribution within the matrix – in other words, on the behaviour of phylogenetic methods when overlap between some taxa is limited. This focus certainly reduces the effective phylogenetic information below the total number of characters for each taxon. In addition, Lemmon *et al*. (2009) have shown that ML and Bayesian methods can be positively misleading or provide inflated support values as a result of ambiguous (missing) data. We therefore look forward to future studies that are more complete.

4. Conclusions

Our study includes more drosophilid taxa than any previous molecular phylogenetic study. We obtained better taxon sampling of subfamily Drosophilinae than previous studies by focusing attention on species that are not traditionally assigned to the genus *Drosophila*, which have been omitted from most previous studies. We obtained this coverage by assembling a matrix of data with a great number of missing data. Despite the potential pitfalls of analyses of such data, results obtained by different methods produced similar results, adequately resolving many aspects of the overall phylogeny, including several long-standing issues. Our study confirms the general observation that the genus *Drosophila* is paraphyletic and points toward issues that still need attention.

We thank Dr Jean-Luc Da Lage for providing us with the aligned AmyRel data, Clemens Lakner for his help with the parallel version of MrBayes, Dr Jean David for providing us with stocks of several species, Jeff Birdsley for his fly collections and contributions and Dr Anne B. Thistle for her editorial assistance. We also thank the two anonymous reviewers for their constructive comments. This work was supported by National Science Foundation grants DEB-0129219 and the NIH Roadmap for Medical Research, Grant U54 RR021813 to DH and DEB-0454673 and 0841447 to SJS.

References

Akaike, H. (1973). 2nd International Symposium on Information Theory, Akademiai Kiado, Budapest, p. 267–281.

Al-Shehbaz, I. A. & Kane, S. L. (2002). Taxonomy and phylogeny of Arabidopsis (Brassicaceae). The Arabidopsis Book, pp. 1–22.

Ashburner, M., Golic, K. G. & Hawley, R. S. (2005). *Drosophila: A Laboratory Handbook*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Bächli, G. (1999–2009). *TaxoDros: The Database on Taxonomy of Drosophilidae*. Available at http://taxodros.unizh.ch/

Bächli, G., Vilela, C. R., Escher, S. A. & Saura, A. (2004). *The Drosophilidae* (*Diptera*) *of Fennoscandia and Denmark*. Leiden, The Netherlands, and New York: Brill.

Baker, W. J., Savolainen, V., Asmussen-Lange, C. B., Chase, M. W., Dransfield, J., Forest, F., Harley, M. M., Uhl, N. W. & Wilkinson, M. (2009). Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. *Systematic Biology* **58**, 240–256.

Beverley, S. M. & Wilson, A. C. (1984). Molecular evolution in *Drosophila* and the higher Diptera. II. A time scale for fly evolution. *Journal of Molecular Evolution* **21**, 1–13.

Bininda-Emonds, O. R. P. (2004). The evolution of supertrees. *Trends in Ecology and Evolution* **19**, 315–322.

Bininda-Emonds, O. R. P., Gittleman, J. L. & Steel, M. A. (2002). The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* **33**, 265–289.

Bininda-Emonds, O. R. P., Jones, K. E., Price, S. A., Grenyer, R., Cardillo, M., Habib, M., Purvis, A. & Gittleman, J. L. (2003). Supertrees are a necessary not-so-evil: a comment on Gatesy *et al*. *Systematic Biology* **52**, 724–729.

Burla, H. (1956). Die Drosophilidengattung *Zygothrica* und ihre Beziehung zur *Drosophila*-Untergattung *Hirtodrosophila*. *Mittheilungen aus dem Zoologischen Museum zu Berlin* **32**, 189–321.

Carrasco, S. F., Prado, L. F. & Godoy-Herrera, R. (2003). Molecular phylogeny of the *mesophragmatica* species group inferred from cytochrome oxidase II sequence. *Drosophila Information Service* **86**, 72–75.

Clayton, F. E. & Wheeler, M. R. (1975). A catalog of *Drosophila* metaphase chromosome configurations. In *Handbook of Genetics* (ed. R. C. King), pp. 471–512. New York: Plenum Press.

Coyne, J. A., Elwyn, S., Kim, S. Y. & Llopart, A. (2004). Genetic studies of two sister species in the *Drosophila melanogaster* subgroup, *D. yakuba* and *D. santomea*. *Genetical Research* **84**, 11–26.

Da Lage, J.-L., Kergoat, G. J., Maczkowiak, F., Silvain, J.-F., Cariou, M.-L. & Lachaise, D. (2007). A phylogeny of Drosophilidae using the Amyrel gene: questioning the *Drosophila melanogaster* species group boundaries. *Journal of Zoological Systematics and Evolutionary Research* **45**, 47–63.

Davis, T. (2000). On the relationship between the *Scaptomyza* and the Hawaiian *Drosophila*. *Hereditas* **132**, 257–259.

Davis, T., Kurihara, J. & Yamamoto, D. (2000*a*). Genomic organisation and characterisation of the neural sex-determination gene fruitless (fru) in the Hawaiian species *Drosophila heteroneura*. *Gene* **246**, 143–149.

Davis, T., Kurihara, J., Yoshino, E. & Yamamoto, D. (2000*b*). Genomic organisation of the neural sex determination gene fruitless (fru) in the Hawaiian species *Drosophila silvestris* and the conservation of the fru BTB protein-protein-binding domain throughout evolution. *Hereditas* **132**, 67–78.

de Queiroz, A. & Gatesy, J. (2007). The supermatrix approach to systematics. *Trends in Ecology and Evolution* **22**, 34–41.

DeSalle, R. (1992*a*). The origin and possible time of divergence of the Hawaiian Drosophilidae: evidence from DNA sequences. *Molecular Biology and Evolution* **9**, 905–916.

DeSalle, R. (1992*b*). The phylogenetic relationships of flies in the family Drosophilidae deduced from mtDNA sequences. *Molecular Phylogenetics and Evolution* **1**, 31–40.

Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218.

Duda, O. (1924). Beitrag zur Systematik der Drosophiliden unter besonderer Berücksichtigung der palaarktischen u. orientalischen Arten (Dipteren) (in German). *Archiv für Naturgeschichte* **90**, 172–234.

Farris, J. S., Kallersjo, M., Kluge, A. G. & Bult, C. (1994). Testing significance of incongruence. *Cladistics – The International Journal of the Willi Hennig Society* **10**, 315–319.

Farris, J. S., Kallersjo, M., Kluge, A. G. & Bult, C. (1995). Constructing a significance test for incongruence. *Systematic Biology* **44**, 570–572.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist* **125**, 1–15.

Frota-Pessoa, O. (1954). Revision of the *tripunctata* group of *Drosophila* with description of fifteen new species (Drosophilidae, Diptera). *Arquivos do Museu Paranaense* **10**, 253–330.

Gailey, D. A., Ho, S. K., Ohshima, S., Liu, J. H., Eyassu, M., Washington, M. A., Yamamoto, D. & Davis, T. (2000). A phylogeny of the Drosophilidae using the sex-behaviour gene fruitless. *Hereditas* **133**, 81–83.

Gatesy, J., Baker, R. H. & Hayashi, C. (2004). Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. *Systematic Biology* **53**, 342–355.

Goto, S. G. & Kimura, M. T. (2001). Phylogenetic utility of mitochondrial COI and nuclear Gpdh genes in *Drosophila*. *Molecular Phylogenetics and Evolution* **18**, 404–422.

Grimaldi, D. A. (1990). A phylogenetic, revised classification of genera in the Drosophilidae (Diptera). *Bulletin of the American Museum of Natural History* **197**, 1–128.

Harris, T. W., Chen, N. S., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Chan, J., Chen, C. K., Chen, W. J., Davis, P., Kenny, E., Kishore, R., Lawson, D., Lee, R., Muller, H. M., Nakamura, C., Ozersky, P., Petcherski, A., Rogers, A., Sabo, A., Schwarz, E. M., Van Auken, K., Wang, Q. H., Durbin, R., Spieth, J., Sternberg, P. W. & Stein, L. D. (2004). WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Research* **32**, D411–D417.

Hatadani, L. M., McInerney, J. O., de Medeiros, H. F., Martins Junqueira, A. C., de Azeredo-Espin, A. M. & Klaczko, L. B. (2009). Molecular phylogeny of the *Drosophila tripunctata* and closely related species groups (Diptera: Drosophilidae). *Molecular Phylogenetics and Evolution* **51**, 595–600.

Hendel, F. (1917). Beiträge zur Kenntnis der acalyptraten Musciden (in German). *Deutsche entomologische Zeitschrift* (*Berliner entomologische Zeitschrift*) **1917**, 33–47.

Hsu, T. C. (1949). The external genital apparatus of male Drosophilidae in relation to systematics. *University of Texas Publication* **4920**, 80–142.

Hu, Y.-G. & Toda, M. J. (2001). Polyphyly of *Lordiphosa* and its relationships in Drosophilinae (Diptera: Drosophilidae). *Systematic Entomology* **26**, 15–31.

Inomata, N., Tachida, H. & Yamazaki, T. (1997). Molecular evolution of the Amy multigenes in the subgenus *Sophophora* of *Drosophila*. *Molecular Biology and Evolution* **14**, 942–950.

Kambysellis, M. P., Ho, K. F., Craddock, E. M., Piano, F., Parisi, M. & Cohen, J. (1995). Pattern of ecological shifts in the diversification of Hawaiian *Drosophila* inferred from a molecular phylogeny. *Current Biology* **5**, 1129–1139.

Kastanis, P., Eliopoulos, E., Goulielmos, G. N., Tsakas, S. & Loukas, M. (2003). Macroevolutionary relationships of species of *Drosophila melanogaster* group based on mtDNA sequences. *Molecular Phylogenetics and Evolution* **28**, 518–528.

Kastritsis, C. D. (1969). The chromosomes of some species of the *guarani* group of *Drosophila*. *Journal of Heredity* **60**, 50–57.

Katoh, T., Tamura, K. & Aotsuka, T. (2000). Phylogenetic position of the subgenus *Lordiphosa* of the genus *Drosophila* (Diptera: Drosophilidae) inferred from alcohol dehydrogenase (Adh) gene sequences. *Journal of Molecular Evolution* **51**, 122–130.

Katoh, T., Nakaya, D., Tamura, K. & Aotsuka, T. (2007). Phylogeny of the *Drosophila immigrans* species group (Diptera: Drosophilidae) based on Adh and Gpdh sequences. *Zoological Science* **24**, 913–921.

Kiontke, K., Gavin, N. P., Raynes, Y., Roehrig, C., Piano, F. & Fitch, D. H. A. (2004). *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proceedings of the National Academy of Sciences of the USA* **101**, 9003–9008.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicates* (Boidae, Serpentes). *Systematic Zoology* **38**, 7–25.

Kopp, A. (2006). Basal relationships in the *Drosophila melanogaster* species group. *Molecular Phylogenetics and Evolution* **39**, 787–798.

Kwiatowski, J. & Ayala, F. J. (1999). Phylogeny of *Drosophila* and related genera: conflict between molecular and anatomical analyses. *Molecular Phylogenetics and Evolution* **13**, 319–328.

Kwiatowski, J., Skarecky, D., Bailey, K. & Ayala, F. J. (1994). Phylogeny of *Drosophila* and related genera inferred from the nucleotide sequence of the Cu,Zn Sod gene. *Journal of Molecular Evolution* **38**, 443–454.

Kwiatowski, J., Krawczyk, M., Jaworski, M., Skarecky, D. & Ayala, F. J. (1997). Erratic evolution of glycerol-3-phosphate dehydrogenase in *Drosophila*, *Chymomyza*, and *Ceratitis*. *Journal of Molecular Evolution* **44**, 9–22.

Lemmon, A. R., Brown, J. M., Stanger-Hall, K. & Lemmon, E. M. (2009). The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology* **58**, 130–145.

Lewis, P. O., Holder, M. T. & Holsinger, K. E. (2005*a*). Polytomies and Bayesian phylogenetic inference. *Systematic Biology* **54**, 241–253.

Lewis, R. L., Beckenbach, A. T. & Mooers, A. Ø. (2005*b*). The phylogeny of the subgroups within the *melanogaster* species group: likelihood tests on COI and COII sequences and a Bayesian estimate of phylogeny. *Molecular Phylogenetics and Evolution* **37**, 15–24.

Maddison, D. R. & Maddison, W. P. (2005). *MacClade*. v. 4.08. Sunderland, MA: Sinauer Associates.

Markow, T. A. & O'Grady, P. M. (2006). *Drosophila: A Guide to Species Identification and Use*. London: Elsevier.

Mitchell-Olds, T. (2001). *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends in Ecology and Evolution* **16**, 693–700.

Nater, H. (1950). Der Samenpumpen-Sklerit von *Drosophila* als taxonomisches Merkmal. *Archiv der Julius Klaus-Stiftung für Verebungsforschung, Sozialanthropologie und Rassenhygiene* **25**, 623–625.

Nater, H. (1953). Vergleichend-morphologische Untersuchung des ausseren Geschlechtsapparates innerhalb der Gattung *Drosophila*. *Zoologische Jahrbucher* **81**, 437–486.

O'Grady, P. & DeSalle, R. (2008). Out of Hawaii: the origin and biogeography of the genus *Scaptomyza* (Diptera: Drosophilidae). *Biology Letters* **4**, 195–199.

O'Grady, P. M. & Kidwell, M. G. (2002). Phylogeny of the subgenus *Sophophora* (Diptera: Drosophilidae) based on combined analysis of nuclear and mitochondrial sequences. *Molecular Phylogenetics and Evolution* **22**, 442–453.

O'Grady, P. M. & Markow, T. A. (2009). Phylogenetic taxonomy in *Drosophila*. Problems and prospects. *Fly* **3**, 10–14.

Okada, T. (1963). Law of unspecialized applied to the family Drosophilidae. *Drosophila Information Service* **38**, 39.

Okada, T. (1989). A proposal of establishing tribes for the family Drosophilidae with key to tribes and genera (Diptera). *Zoological Science* **6**, 391–399.

Pélandakis, M. & Solignac, M. (1993). Molecular phylogeny of *Drosophila* based on ribosomal RNA sequences. *Journal of Molecular Evolution* **37**, 525–543.

Pélandakis, M., Higgins, D. G. & Solignac, M. (1991). Molecular phylogeny of the subgenus *Sophophora* of *Drosophila* derived from large subunit of ribosomal RNA sequences. *Genetica* **84**, 87–94.

Perlman, S. J., Spicer, G. S., Shoemaker, D. D. & Jaenike, J. (2003). Associations between mycophagous *Drosophila* and their *Howardula* nematode parasites: a worldwide phylogenetic shuffle. *Molecular Ecology* **12**, 237–249.

Pitnick, S., Markow, T. & Spicer, G. S. (1999). Evolution of multiple kinds of female sperm-storage organs in *Drosophila*. *Evolution* **53**, 1804–1822.

Posada, D. & Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V. A., Williams, T. M., Yeh, S. D., True, J. R. & Carroll, S. B. (2006). Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* **440**, 1050–1053.

Quigley, A. K., Turner, J. M., Nuckels, R. J., Manuel, J. L., Budi, E. H., MacDonald, E. L. & Parichy, D. M. (2004). Pigment pattern evolution by differential deployment of neural crest and post-embryonic melanophore lineages in *Danio* fishes. *Development* **131**, 6053–6069.

Quigley, I. K., Manuel, J. L., Roberts, R. A., Nuckels, R. J., Herrington, E. R., MacDonald, E. L. & Parichy, D. M. (2005). Evolutionary diversification of pigment pattern in *Danio* fishes: differential fms dependence and stripe loss in *D. albolineatus*. *Development* **132**, 89–104.

Remsen, J. & DeSalle, R. (1998). Character congruence of multiple data partitions and the origin of the Hawaiian Drosophilidae. *Molecular Phylogenetics and Evolution* **9**, 225–235.

Remsen, J. & O'Grady, P. (2002). Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support. *Molecular Phylogenetics and Evolution* **24**, 249–264.

Robe, L. J., Valente, V. L. S., Budnik, M. & Loreto, E. L. S. (2005). Molecular phylogeny of the subgenus *Drosophila* (Diptera, Drosophilidae) with an emphasis on Neotropical species and groups: a nuclear versus mitochondrial gene approach. *Molecular Phylogenetics and Evolution* **36**, 623–640.

Ronquist, F. & Huelsenbeck, J. P. (2003). MRBAYES3: Bayesian phylogenetic interference under mixed models. *Bioinformatics* **19**, 1572–1574.

Russo, C. A. M., Takezaki, N. & Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Molecular Biology and Evolution* **12**, 391–404.

Sidorenko, V. S. (2002). Phylogeny of the tribe Steganini Hendel and some related taxa (Diptera, Drosophilidae). *Far Eastern Entomologist* **111**, 1–20.

Singh, N. D., Larracuente, A. M., Sackton, T. B. & Clark, A. G. (2009). Comparative genomics on the *Drosophila* phylogenetic tree. *Annual Review of Ecology Evolution and Systematics* **40**, 459–480.

Stamatakis, A., Hoover, P. & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* **57**, 758–771.

Swofford, D. L. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony* (**and Other Methods*). v. 4.0b10. Sunderland, MA: Sinauer Associates.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996). Phylogenetic inference. In *Molecular Systematics* (ed. D. M. Hillis, C. Moriz & B. K. Mable), pp. 407–514. Sunderland, MA: Sinauer Associates.

Tarrio, R., Rodriguez-Trelles, F. & Ayala, F. J. (2001). Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. *Molecular Biology and Evolution* **18**, 1464–1473.

Tatarenkov, A. & Ayala, F. J. (2001). Phylogenetic relationships among species groups of the *virilis-repleta* radiation of *Drosophila*. *Molecular Phylogenetics and Evolution* **21**, 327–331.

Tatarenkov, A., Kwiatowski, J., Skarecky, D., Barrio, E. & Ayala, F. J. (1999). On the evolution of Dopa decarboxylase (Ddc) and *Drosophila* systematics. *Journal of Molecular Evolution* **48**, 445–462.

Tatarenkov, A., Zurovcova, M. & Ayala, F. J. (2001). Ddc and amd sequences resolve phylogenetic relationships of *Drosophila*. *Molecular Phylogenetics and Evolution* **20**, 321–325.

Thomas, R. H. & Hunt, J. A. (1993). Phylogenetic relationships in *Drosophila*: a conflict between molecular and morphological data. *Molecular Biology and Evolution* **10**, 362–374.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**, 4876–4882.

Throckmorton, L. H. (1962). The problem of phylogeny in the genus *Drosophila*. *University of Texas Publication* **6205**, 207–343.

Throckmorton, L. H. (1965). Similarity versus relationship in *Drosophila*. *Systematic Zoology* **14**, 221–236.

Throckmorton, L. H. (1966). The relationships of the endemic Hawaiian Drosophilidae. *University of Texas Publication* **6615**, 335–396.

Throckmorton, L. H. (1975). The phylogeny, ecology, and geography of *Drosophila*. In *Invertebrates of Genetic Interest* (ed. R. C. King), pp. 421–469. New York: Plenum Press.

van der Linde, K. & Houle, D. (2008). A supertree analysis and literature review of the genus *Drosophila* and closely related genera. *Insect Systematics and Evolution* **39**, 241–267.

van der Linde, K., Bächli, G., Toda, M. J., Zhang, W.-X., Katoh, T., Hu, Y.-G. & Spicer, G. S. (2007). Case 3407: *Drosophila* Fallén, 1823 (Insecta, Diptera): proposed conservation of usage. *Bulletin of Zoological Nomenclature* **64**, 238–242.

Wang, B.-C., Park, J., Watabe, H.-A., Gao, J.-J., Xiangyu, J.-G., Aotsuka, T., Chen, H.-W. & Zhang, Y.-P. (2006). Molecular phylogeny of the *Drosophila virilis* section (Diptera: Drosophilidae) based on mitochondrial and nuclear sequences. *Molecular Phylogenetics and Evolution* **40**, 484–500.

Watabe, H. & Peng, T. X. (1991). The *Drosophila virilis* section (Diptera: Drosophilidae) from Guangdong Province, Southern China. *Zoological Science* **8**, 147–156.

Wheeler, M. R. (1949). Taxonomic studies on the Drosophilidae. *University of Texas Publication* **4920**, 157–195.

Wiens, J. J. (2003). Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology* **52**, 528–538.

Wiens, J. J. & Moen, D. S. (2008). Missing data and the accuracy of Bayesian phylogenetics. *Journal of Systematics and Evolution* **46**, 307–314.

Wilgenbusch, J. C., Warren, D. L. & Swofford, D. L. (2004). AWTY: a system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference. Available at http://ceb.csit.fsu.edu/awty

Wojtas, K. M., Vonkalm, L., Weaver, J. R. & Sullivan, D. T. (1992). The evolution of duplicate glyceraldehyde-3-phosphate dehydrogenase genes in *Drosophila*. *Genetics* **132**, 789–797.

Yang, Z., Goldman, N. & Friday, A. (1995). Maximum-likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Systematic Biology* **44**, 384–399.

Yassin, A. (2007). A revision of the *tumiditarsus* group of the subgenus *Drosophila* and its relation to the genus *Zaprionus*. *Drosophila Information Service* **90**, 20–22.

Yassin, A., Araripe, L. O., Capy, P., Da Lage, J.-L., Klaczko, L. B., Maisonhaute, C., Ogereau, D. & David, J. R. (2008). Grafting the molecular phylogenetic tree with morphological branches to reconstruct the evolutionary history of the genus *Zaprionus* (Diptera: Drosophilidae). *Molecular Phylogenetics and Evolution* **47**, 903–915.

Yassin, A., Da Lage, J.-L., David, J. R., Kondo, M., Madi-Ravazzi, L., Prigent, S. R. & Toda, M. J. (2010). Polyphyly of the *Zaprionus* genus group (Diptera: Drosophilidae). *Molecular Phylogenetics and Evolution* **55**, 335–339.

Yotoko, K. S. C., Medeiros, H. F., Solferini, V. N. & Klaczko, L. B. (2003). A molecular study of the systematics of the *Drosophila tripunctata* group and the *tripunctata* radiation. *Molecular Phylogenetics and Evolution* **28**, 614–619.

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. Ph.D. thesis, University of Texas at Austin, Austin.