DOI: 10.1111/j.1365-3113.2005.00305.x

How do insect nuclear ribosomal genes compare to protein-coding genes in phylogenetic utility and nucleotide substitution patterns?

BRYAN N. DANFORTH, CHUNG-PING LIN[1] and JENNIFER FANG
Department of Entomology, Cornell University, Ithaca, NY, U.S.A.

Abstract. The expanding data set on insect molecular systematics allows examination of phylogenetic performance and molecular evolution of different types of gene. Studies combining more than one gene in the same analysis allow examination of the relative contribution and performance of each gene partition and can help inform gene choice for resolving deep and/or problematic divergences. We compared results obtained from analyses of twelve insect data sets in which authors combined one or more nuclear ribosomal genes (28S and/or 18S) with one or more protein-coding genes [elongation factor-1α (EF-1α), histone H3, carbamoylphosphate synthetase domain (CPS domain of CAD, or rudimentary), long-wavelength rhodopsin (LW opsin), glucose-6-phosphate dehydrogenase (G_6pd), phosphoenolpyruvate carboxykinase (PEPCK), arginine kinase, and *white*]. Data sets examined spanned eight orders of insects (Odonata, Ephemeroptera, Hemiptera, Coleoptera, Trichoptera, Lepidoptera, Diptera and Hymenoptera), providing a broad range of divergence times and taxonomic levels. We estimated the phylogenetic utility of the individual genes (using parsimony methods) and characterized the nucleotide substitution patterns (using Bayesian methods) to ask which type of data is preferable for phylogenetic analysis in insects. Nuclear ribosomal and protein coding genes differed little in our measures of phylogenetic performance and patterns of nucleotide substitution. We recommend combining nuclear ribosomal gene data with nuclear protein-coding gene data because each data set has distinct advantages. We do not recommend using mitochondrial genes for higher-level studies of insect phylogeny because reviewed studies demonstrate substitution patterns that lead to high levels of homoplasy.

Introduction

Insect molecular systematists interested in reconstructing deeper (i.e. Mesozoic and older) divergences focus their efforts on either nuclear ribosomal genes or nuclear protein-coding genes. Mitochondrial genes are considered to be too rapidly evolving for these deep divergences and show substitution patterns that are problematic for reconstructing ancient divergences (Lin & Danforth, 2004). Nuclear ribosomal genes are by far the most commonly used genes for higher-level insects phylogenetics. A cursory search on Biosis in April 2004 revealed nearly 100 papers on insect phylogeny using one or more nuclear ribosomal genes. Reviews of the utility of ribsomal gene data in phylogenetic analysis include Hillis & Dixon (1991), Simon *et al.* (1994) and Caterino *et al.* (2000). While nuclear ribosomal genes (i.e. 5.8S, 18S and 28S) have been the most common choice for deep divergences in insects (e.g. Whiting *et al.*, 1997; Wiegmann *et al.*, 2000; Dietrich *et al.*,

Correspondence: Bryan N. Danforth, Department of Entomology, Comstock Hall, Cornell University, Ithaca, NY 14853-0901, U.S.A. Tel.: +1 607 255 3563; fax: +1 607 255 0939; e-mail: bnd1@cornell.edu

[1] Present address. Department of Life Science, Tunghai University, No.1 81, Sec. 3, Taichung-Kan Road, Taichung, Taiwan 40704.

2001; Giribet *et al*., 2001; Kjer *et al*., 2001; Wheeler *et al*., 2001; Belshaw & Quicke, 2002; Hovmöller *et al*., 2002; Schulmeister *et al*., 2002; Kjer, 2004; but see Ogden & Whiting, 2003), these genes sometimes present major alignment problems (see Hickson *et al*., 2000, for a comparison of various methods). Kjer (2004) recently criticized previous studies of insect ordinal relationships based on 'direct optimization' of ribosomal sequences (Whiting *et al*., 1997; Wheeler *et al*., 2001).

Nuclear protein-coding genes may be a better choice for phylogenetic analysis of insects because they are easily alignable and exhibit variable rates of substitution both within and among genes (Friedlander *et al*., 1992, 1994). Such genes include EF-1α (Cho *et al*., 1995; Mitchell *et al*., 1997; Danforth & Ji, 1998; Reed & Sperling, 1999; Clark *et al*., 2000; Regier *et al*., 2000; Caterino *et al*., 2001; Cognato & Vogler, 2001; Kjer *et al*., 2001; Sipes & Wolf, 2001; Buckley *et al*., 2002; Danforth, 2002), PEPCK (Friedlander *et al*., 1996; Wiegmann *et al*., 2000; Sota & Vogler, 2001; Leys *et al*., 2002), dopa-decarboxylase (DDC; Fang *et al*., 1997, 2000; Friedlander *et al*., 1998, 2000; Tatarenkov *et al*., 1999), *wingless* (Brower & Egan, 1997; Brower & DeSalle, 1998; Brower, 2000; Campbell *et al*., 2000; Morris *et al*., 2001; Brady, 2003), *white* (Baker *et al*., 2001), LW opsin (Mardulyn & Cameron, 1999; Ascher *et al*., 2001; Cameron & Mardulyn, 2001; Hsu *et al*., 2001; Danforth *et al*., 2003; Kawakita *et al*., 2003, 2004; Ortiz-Rivas *et al*., 2004), *hunchback* (Baker & DeSalle, 1997), *period* (Regier *et al*., 1998), arginine kinase (Kawakita *et al*., 2003, 2004) and others (see Brower & DeSalle, 1994 and Caterino *et al*., 2000 for complete lists of nuclear protein-coding genes thus far used in insects). The coding regions of these genes are easily and unambiguously alignable. Based on the two complete insect genomes so far analysed (*Drosophila melanogaster* and *Bombyx mori*), the insect genome consists of between 13 379 (Adams *et al*., 2000) and 18 510 (Xia *et al*., 2004) protein-coding genes.

However, we know little about how nuclear ribosomal and nuclear protein-coding genes compare in phylogenetic utility or in the nature of their substitution patterns. Are nuclear protein-coding and ribosomal genes of equal phylogenetic utility? In the last three years, several published studies have combined both ribosomal and protein-coding gene data sets, allowing direct comparisons of both the phylogenetic utility of these genes and their nucleotide substitution patterns.

Here we present a comparison of the phylogenetic utility and nucleotide substitution patterns of nuclear ribosomal and nuclear protein-coding genes using twelve recently published, combined insect data sets. We use parsimony methods to assess the relative contribution of the ribosomal and protein-coding genes to the overall analysis and Bayesian methods to understand how the substitution patterns compare among ribosomal and nuclear protein-coding genes. Bayesian methods provide an ideal framework for investigating and characterizing substitution patterns in molecular data sets (Huelsenbeck *et al*., 2001, 2002). Models in Bayesian analyses can be complex, incorporating many aspects of the nucleotide substitution process, including variation in base composition, rate variation among sites (either through site-specific rates models, gamma models, or gamma + invariant sites models; see Swofford *et al*., 1996), and variation in rates of transformation among bases. Furthermore, within the Bayesian framework, the parameters are estimated over many plausible tree topologies so that substitution parameters are independent of any *particular* tree topology (Huelsenbeck *et al*., 2001).

Materials and methods

To compare ribosomal and protein-coding genes within the same analysis, we obtained twelve data sets that combine the two types of data. Such data sets were rare because insect molecular systematists seem to fall into two catagories: those that use nuclear ribosomal genes and those that use nuclear, protein-coding genes. The broad range of pterygote orders studied included examples from eight orders: Odonata, Ephemeroptera, Hemiptera, Coleoptera, Trichoptera, Lepidoptera, Diptera and Hymenoptera (Table 1), with a slight bias toward the Diptera. Most studies employed a combination of 28S and EF-1α ($n = 8$), other studies utilized PEPCK, LW opsin, histone, *white* and CAD, and two studies combined 18S with protein-coding nuclear genes. In all cases (except two), the data sets were obtained directly from the authors or were downloaded from author or journal websites. In two cases (Cameron & Mardulyn, 2001; Moulton & Wiegmann, 2004) we downloaded additional data from GenBank that were unavailable at the time of publication. Appendices 1 and 2 list the GenBank accession numbers used for these two studies.

Data sets varied from eighteen taxa to over 100 taxa (Table 1) and individual gene regions varied in size from 339 bp to over 3800 bp. Maximum likelihood (and presumably Bayesian) parameter estimates are sensitive to taxon sampling (Sullivan *et al*., 1999; Yang & Yoder, 1999) and presumably also to data set size. In all cases we relied only on the alignment provided by the authors.

Parsimony analyses

Initially we performed an equal weights parsimony analysis on the combined ribosomal and protein-coding data sets with gaps coded either as missing data or as a fifth state, or additional gap-coded characters were included (according to the authors' preferences). We excluded regions considered to be unalignable by the authors. For calculating the parsimony parameters described below we coded gaps as missing data to apply a single standard across all data sets. The effect of coding gaps as missing data probably will reduce the phylogenetic utility of the ribosomal genes to the extent that the gaps provide useful phylogenetic information. However, this conservative

Table 1. Summary of parsimony and Bayesian results. 'A + T' is proportion A and T in each data partition. 'Base comp. hetero.' lists the *P*-value of the chi-square test for the similarity in base composition across taxa. Significant *P*-values indicate significant deviation among taxa in base composition. 'Total sites' lists the total number of sites in each data partition. 'PI sites' lists the number of parsimony informative sites in each data partition. 'CI' is the consistency index (excluding the invariant sites). 'DD' lists the data decisiveness of each data partition. 'PBS/min steps' is the partitioned Bremer support divided by the minimum number of steps. 'Alpha' is the shape parameter of the gamma distribution

(n taxa)	Data partitions	A + T	base comp. hetero.	Total sites	PI sites	CI	DD	PBS/ min steps	alpha	Reference
Palaeoptera	18S	0.48	1.000	2224	375	0.504	0.664	0.264	0.252	Ogden & Whiting, 2003
(*n* = 34 taxa)	histone (H3)	0.41	< 0.0001	339	130	0.254	0.274	0.484	0.099	
	total	0.47	0.997	2563	505	0.376	0.487	0.314		
Treehoppers	EF-1α	0.50	0.999	958	330	0.225	0.517	0.788	0.194	Cryan *et al.*, 2000
(*n* = 79 taxa)	28S	0.42	< 0.0001	2363	405	0.370	0.505	– 0.121	0.238	
	total	0.45	0.012	3321	735	0.276	0.497	0.233		
Chrysomelidae	EF-1α	0.56	0.503	420	137	0.343	0.259	0.116	0.159	Kim *et al.*, 2003
(*n* = 27 taxa)	28S	0.44	0.291	448	138	0.438	0.388	0.121	0.383	
	total	0.50	< 0.0001	868	275	0.377	0.324	0.124		
Therevid flies	EF-1α	0.58	0.083	993	299	0.290	0.430	0.086	0.156	Yang *et al.*, 2000
(*n* = 39 taxa)	28S	0.60	1.000	1162	138	0.445	0.515	0.142	0.164	
	total	0.59	0.879	2155	437	0.321	0.441	0.106		
Empidoid flies	EF-1α	0.60	0.337	986	360	0.329	0.298	0.212	0.177	Collins & Wiegmann, 2002
(*n* = 28 taxa)	28S	0.58	1.000	1341	99	0.546	0.683	0.417	0.111	
	total	0.59	1.000	2327	459	0.347	0.341	0.260		
Eremoneura	CAD	0.59	0.000	3888	1845	0.385	0.223	0.121	0.322	Moulton & Wiegmann, 2004
(*n* = 18 taxa)	28S	0.60	0.996	2075	251	0.509	0.468	0.158	0.160	
	total	0.60	0.000	5963	2096	0.393	0.240	0.126		
Mosquitoes	white	0.46	0.000	801	338	0.397	0.357	0.314	0.273	Krzywinski *et al.*, 2001 a,b
(*n* = 24 taxa)	G6PD	0.47	0.869	462	197	0.396	0.307	0.078	0.317	
	28S	0.37	0.421	577	285	0.545	0.484	0.082	1.015	
	total	0.44	0.000	1607	710	0.413	0.354	0.207		
Caddis flies	EF-1α	0.44	0.000	1099	447	0.183	0.282	na	0.274	Kjer *et al.*, 2001
(*n* = 101 taxa)	28S	0.46	0.799	1078	468	0.380	0.632	na	0.328	
	total	0.45	0.000	2177	915	0.242	0.413	na		
basal Lepidoptera	PEPCK	0.47	< 0.0001	623	312	0.398	0.204	0.109	0.230	Wiegmann *et al.*, 2000
(*n* = 25 taxa)	18S	0.51	0.933	1369	308	0.669	0.739	0.493	0.195	
	total	0.50	0.067	1992	620	0.469	0.403	0.278		
Corbiculate bees	LW opsin	0.55	1.000	502	151	0.505	0.560	0.116	1.443	Cameron & Mardulyn, 2001
(*n* = 20 taxa)	EF-1α	0.58	1.000	825	156	0.557	0.614	0.012	0.173	
	28S	0.41	0.905	748	168	0.559	0.511	0.085	1.297	
	total	0.51	< 0.0001	2075	475	0.529	0.544	0.081		
LT bees	argK	0.47	0.418	834	308	0.268	0.500	0.333	0.254	Danforth *et al.*, in prep.
(*n* = 58 taxa)	LW opsin	0.53	1.000	492	219	0.261	0.512	0.818	0.323	
	EF-1α	0.54	0.004	753	256	0.260	0.475	0.717	0.208	
	28S	0.40	1.000	694	230	0.412	0.619	0.266	0.403	
	total	0.48	< 0.0001	2773	1013	0.279	0.497	0.517		
Gall wasps	LW opsin	0.57	1.000	481	142	0.459	0.455	– 0.052	0.279	Nylander *et al.*, 2004
(*n* = 32 taxa)	EF-1α	0.60	1.000	367	87	0.508	0.652	0.474	0.235	
	28S	0.46	0.867	1154	198	0.423	0.446	0.211	0.180	
	total	0.51	0.655	2002	427	0.430	0.522	0.182		

approach is preferred because alignments differed in quality across the data sets and the Bayesian methods (described below) treat gaps as missing data. Trees were checked against results reported in the original papers (cited in Table 1) to ensure that our results matched the published trees. Using PAUP* 4.0 b10 (Swofford, 2002) we calculated the base proportions for each data set and data partition within data sets. We used PAUP* 4.0 also to calculate the consistency index (CI), the number of parsimony informative sites, and the number of equally parsimonious trees for each gene. For tree searches we performed 100 random sequence additions and TBR branch swapping.

Data decisiveness

We calculated data decisiveness (DD) for the separate and combined data sets as outlined in Goloboff (1991), using the following calculated values: the minimum number of steps possible for each data set (M), the mean length of 10 000 random tree topologies for each data set (S*), and the shortest tree possible with each data set (S). Data decisiveness, a measure of the degree to which the data provide a strong phylogenetic signal, is calculated as:

$$\mathrm{DD} = (\mathrm{S}^* - \mathrm{S})/(\mathrm{S}^* - \mathrm{M})$$

More decisive data sets allow the observer to more confidently choose some cladograms over others (Kitching *et al.*, 1998, p. 119).

Partitioned Bremer support

To assess the relative contribution of each gene to the overall results, we calculated partitioned Bremer support (PBS; Bremer, 1988; Baker & DeSalle, 1997) using TreeRot v.2 (Sorenson, 1999). We standardized the partitioned Bremer support by dividing the total Bremer support of each gene by the minimum number of steps for that gene (Baker *et al.*, 2001). This measure (PBS/min. steps) provides a quantitative measure of each gene's overall contribution to tree resolution (Baker *et al.*, 2001). We were unable to calculate partitioned Bremer support for the caddisfly data set using TreeRot, presumably because of the large size of the data set ($n = 101$ taxa).

Incongruence length difference test

We analysed incongruence among the ribosomal and protein-coding data sets within the same analysis using the incongruence length difference (ILD) test (Farris *et al.*, 1995) implemented in PAUP*.

Bayesian analyses

For the Bayesian analyses we used MrBayes v. 3.0 (Huelsenbeck & Ronquist, 2001; http://morphbank.ebc.uu.se/mrbayes3/) to analyse the data sets using two different models. First, we analysed the combined data set using a GTR + SSR model with rate catagories corresponding to a rate catagory for each ribosomal gene and separate rates for each of the codon positions within the protein-coding genes. This allowed us to compare rates among partitions and among genes. Second, we used a GTR + G model with a separate instantaneous rate matrix (Q matrix) and gamma distribution fit for each gene (using the 'unlink' command; see Appendix 3). From the GTR + G analysis we obtained the instantaneous rate matrix (Q matrix) and the shape parameter of the gamma distribution (α) for each gene (Swofford *et al.*, 1996). This allowed us to compare the relative symmetry of the Q matrix, as well as the heterogeneity in rates of substitution among sites (α). Appendix 3 provides the MrBayes blocks used for these two analyses. By analysing all data sets using the same standard set of models we could compare parameter values among data sets in a way not possible if we had applied different models to each data set. Our goal was not to reconstruct trees, but to understand nucleotide substitution patterns. We also examined the correlations among parameter estimates and parsimony results.

Analyses consisted of running four simultaneous chains for 1×10^6 generations. Trees were sampled at intervals of fifty generations for a total of 20 000 trees. We plotted the likelihood values against generation time to identify the region at which the likelihood values reached a stable plateau. We discarded the 'burn-in' region (in general 1×10^5 generations, or 2000 trees) and calculated the mean, variance, and 95% credibility intervals of the parameter estimates using MrBayes.

Results

Parsimony analyses

Data sets (whether protein-coding or ribosomal) differed little in the parsimony parameters we measured (Table 1, Figs 1–4), with no clear indication that either protein-coding genes or ribosomal genes consistently provided the more robust signal. Ribosomal genes exhibited slightly lower levels of homoplasy (Fig. 1), presumably because of their overall slower rate of nucleotide substitution (see below) or the inclusion of highly variable but unalignable regions in our analysis. Ribosomal genes also showed slightly higher levels of data decisiveness (Fig. 2) with eight data sets favouring the ribosomal genes and four data sets favouring the protein-coding genes (Table 1). There were no obvious differences among genes in terms of partitioned Bremer support (PBS; Fig. 3). Base composition among the ribosomal genes was more consistently G/C biased than the protein-coding genes (Fig. 4). Overall, our parsimony results provide little to support the view that one type of data (ribosomal or protein-coding) consistently outperforms the other, although ribosomal genes did seem to

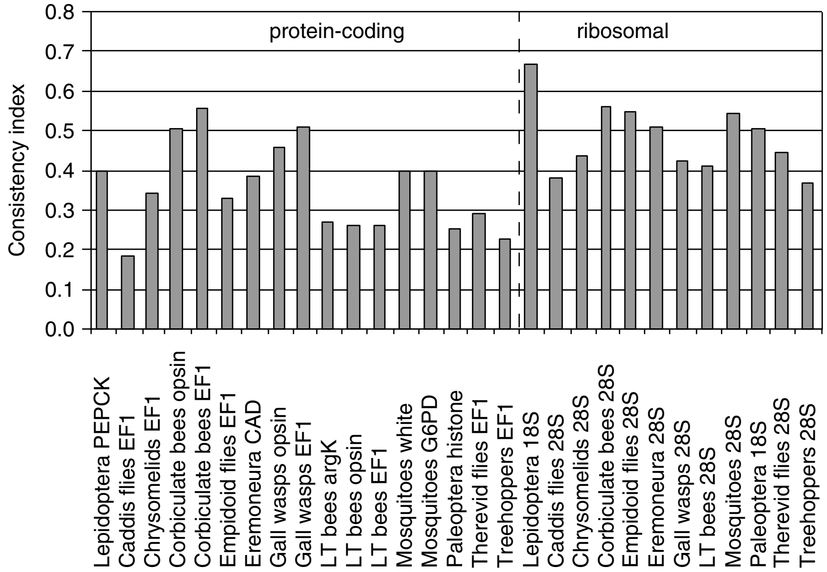

Fig. 1. Distribution of consistency index values for the genes analysed. Data from Table 1.

perform slightly better as measured by consistency index and data decisiveness.

In all but two studies [Yang *et al.*, 2000 ($P = 0.960$) and Cameron & Mardulyn, 2001 ($P = 0.350$)] there was significant incongruence between ribosomal and protein coding genes within the same analysis ($P < 0.01$ for all the remaining studies). This suggests that the ILD test may be overly sensitive to incongruence because no authors commented on significant *topological* incongruence among their ribosomal and protein-coding gene data sets. This test has been criticized by several authors (Barker & Lutzoni, 2002; Darlu & Lecointre, 2002; Dowton & Austin, 2002). Our survey of twelve data sets suggests that it is an overly sensitive measure of data set incongruence and most authors combine their data sets even when the ILD test suggests significant incongruence.

Bayesian analyses

Bayesian analyses using SSR models consistently showed that the ribosomal genes evolve at roughly the rate of protein-coding first or second position sites (Figs 5, 6). For example, in the fig wasp data set, 28S evolved at the same rate as opsin first positions, but faster than opsin second positions (Fig. 6f). Third position sites were universally the fastest sites. In some cases the ribosomal genes evolved considerably faster than first or second position sites, e.g. the chrysomelid beetles (Fig. 5c), mosquitoes

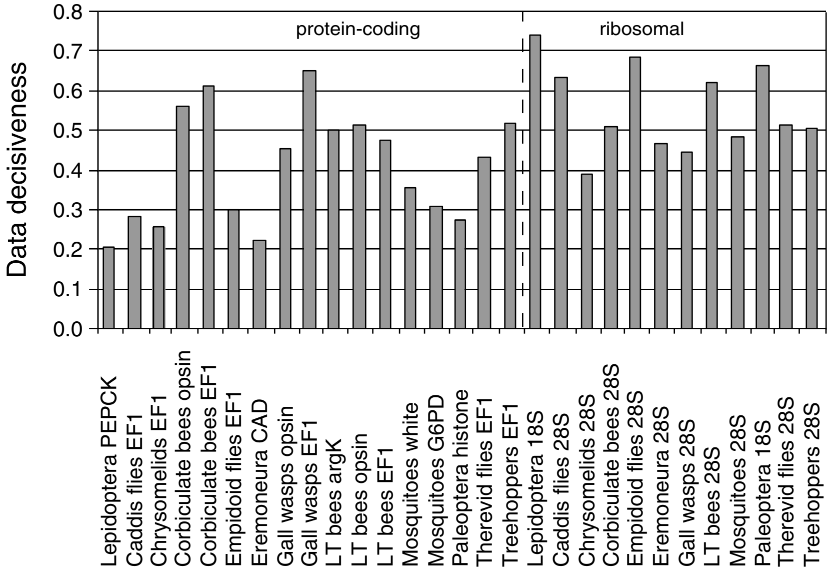

Fig. 2. Distribution of data decisiveness (DD) values for the genes analysed. Data from Table 1.

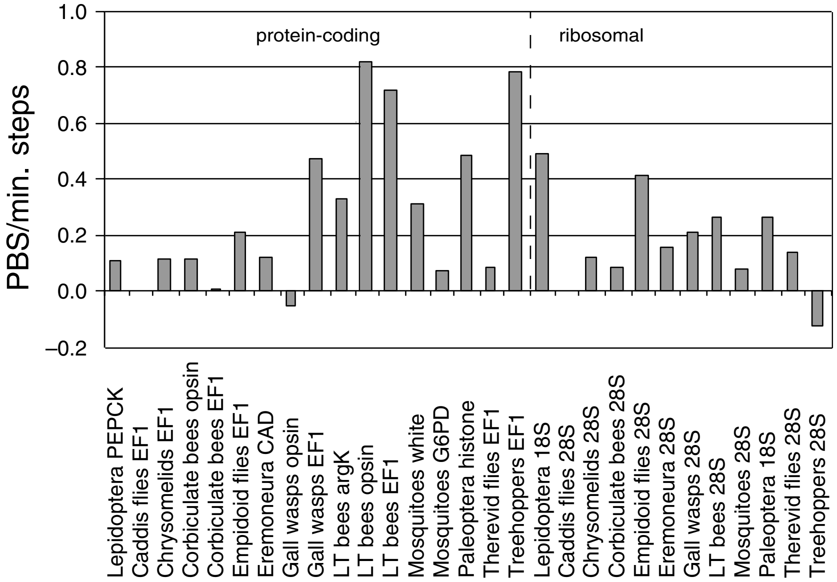

Fig. 3. Distribution of partitioned Bremer support (PBS) for the genes analysed. Data from Table 1.

(Fig. 6a) and corbiculate bees (Fig. 6d). In some cases (e.g. Palaeoptera and Chrysomelidae), the nuclear genes (histone and EF-1α) showed almost no first or second position variation. For this reason, neither histone nor EF-1α is likely to be a very good data set for ordinal level studies of insects. The differences in consistency index and data decisiveness described above may be attributed in large part to the lack of the rapidly evolving (and therefore more homoplasious) third position sites in the ribosomal genes (although there are regions of the ribosomal genes that evolve more rapidly).

Alpha (α), the shape of the gamma distribution describing among site rate variation, showed no consistent differences among ribosomal gene data sets and protein-coding nuclear gene data sets (Table 1). Lower values of α correspond to gene regions with greater rate heterogeneity among sites (e.g. a more *uneven* distribution of rates among sites). For example, low values of α correspond to genes with a few sites that change at a very high rate, and many sites that change at a very slow rate. Higher values of α correspond to genes or gene regions with a more even distribution of rates among sites. Alpha (α) is a correlate of data set quality in simulation studies (Yang, 1998). In terms of substitution parameters such as alpha, protein-coding and ribosomal genes were largely comparable with no evidence that the two data partitions behave substantially differently. Examination of the six-parameter rate matrices indicate that for both types of genes there is (not surprisingly) an excess of transitions vs. transversions, but neither gene showed the highly skewed patterns evident in mitochondrial genes (Lin & Danforth, 2004). Three data sets showed far greater alpha values than all the rest; however,

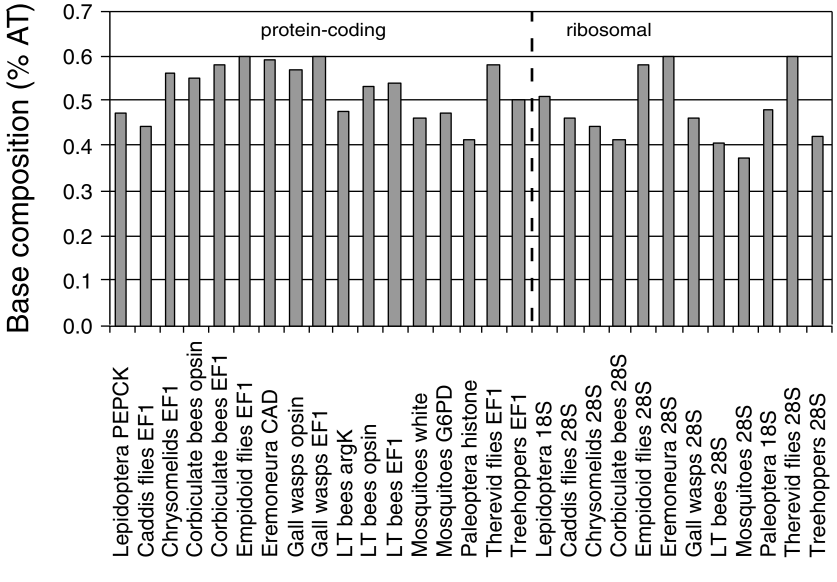

Fig. 4. Distribution of base compositional bias for the genes analysed. Data from Table 1.

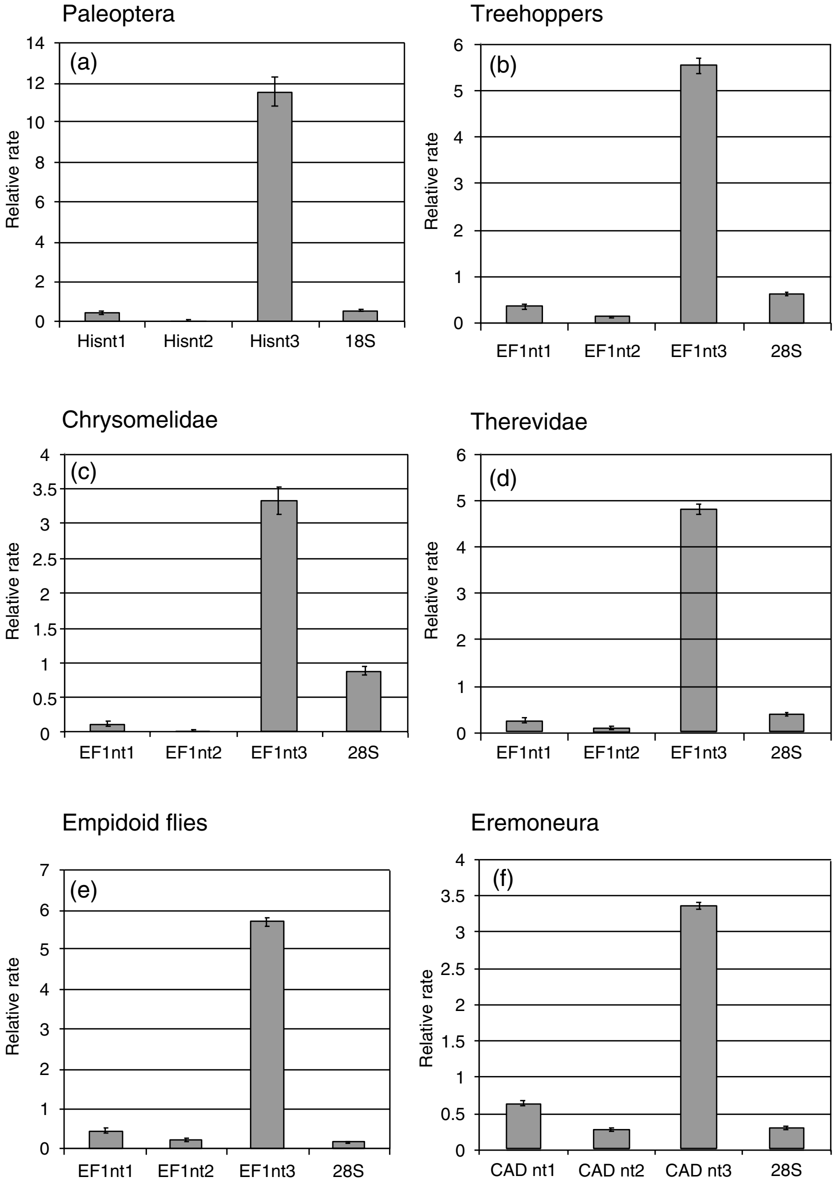

Fig. 5. Relative rates among codon positions and ribosomal genes for each data set. (a) Palaeoptera (Ogden & Whiting, 2003), (b) Treehoppers (Cryan *et al.*, 2000), (c) Chrysomelid beetles (Kim *et al.*, 2003), (d) Therevid flies (Yang *et al.*, 2000), (e) Empidoid flies (Collins & Wiegmann, 2002), (f) Eremoneura (Moulton & Wiegmann, 2004).

these were not consistently ribosomal or protein-coding data sets: 28S corbiculate bees, opsin corbiculate bees and 28S mosquitoes (Fig. 7a).

Bayesian parameter estimates indicate that the two partitions of the nuclear genome largely are comparable in substitution parameters. This is markedly different from the previous comparison of mitochondrial and nuclear protein-coding genes in which a suite of substitution parameters differed consistently between the nuclear and mitochondrial genomes (Lin & Danforth, 2004).

Correlations among parameters

The relationships among consistency index (CI), data decisiveness (DD), partitioned Bremer support (PBS) and α (the shape of the gamma distribution) are shown in Fig. 7. Only the relationship between CI and DD was significant ($r^2 = 0.526$; $P = 0.002$; Fig. 7b). Goloboff (1991) predicted that CI and DD would not necessarily be correlated, which is not supported by our analysis of these twelve empirical data sets. All other relationships were non-significant (Fig. 7). Contrary to the previous study (Lin & Danforth, 2004) there was no clear relationship between α and our measures of data set quality (Figs 7a, d).

Discussion

With the publication of four complete insect genomes (*Drosophila melanogaster*, Adams *et al.*, 2000, *Anopheles gambiae*, Holt *et al.*, 2002, *Apis mellifera*, http://www.hgsc.bcm.tmc.edu/projects/honeybee/ and *Bombyx*

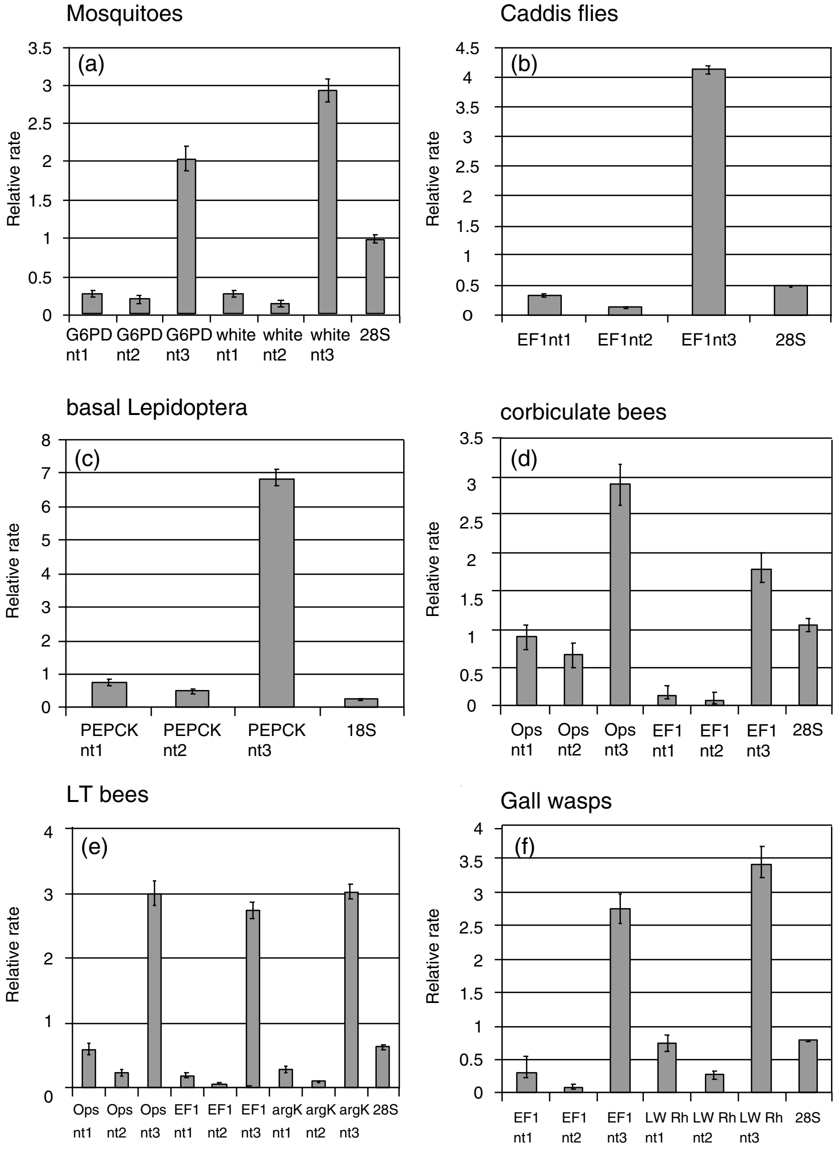

Fig. 6. Relative rates among codon positions and ribosomal genes for each data set. (a) Mosquitoes (Krzywinski *et al.*, 2001a,b), (b) Caddis flies (Kjer *et al.*, 2001), (c) basal Lepidoptera (Wiegmann *et al.*, 2000), (d) corbiculate bees (Cameron & Mardulyn, 2001), (e) long-tongued bees (Danforth *et al.* in prep.), (f) gall wasps (Nylander *et al.*, 2004).

mori, Xia *et al.*, 2004) in the past four years, insect molecular systematists now have a virtually limitless source of genes to choose from. But how do we make informed choices as to which genes will be useful phylogenetic markers? In the absence of a priori knowledge of what genes provide the most useful phylogenetic information, we may be unable to predict which genes will be the best for resolving deep phylogenies. Combined data sets with multiple genes give us the ability to evaluate both the relative phylogenetic utility of genes and the details of the nucleotide substitution patterns that characterize these genes. Critical evaluation of data set quality is an important first step in choosing genes for any phylogenetic study. Furthermore, Bayesian parameter estimates may give us important information on data set performance. We know of no previous study that evaluates the relative phylogenetic utility of nuclear ribosomal vs. nuclear protein-coding genes, perhaps because relatively few studies have combined the two into the same analysis.

Based on our analysis of twelve combined data sets consisting of at least one protein-coding gene and at least one ribosomal gene, there are no consistent differences in data set quality or substitution patterns as measured by the parameters we analysed. It is worth considering whether this conclusion is sufficiently justified by the analyses that we performed. First, it is possible that our methods were too crude to detect subtle differences in data set performance. A previous study using similar methods (Lin & Danforth, 2004) did detect consistent and striking differences between nuclear protein-coding genes and mitochondrial protein-coding genes in the same parameters we analysed here, so we suspect the parameters can reveal clear differences. Perhaps we could have partitioned the ribosomal gene regions more finely, for example into

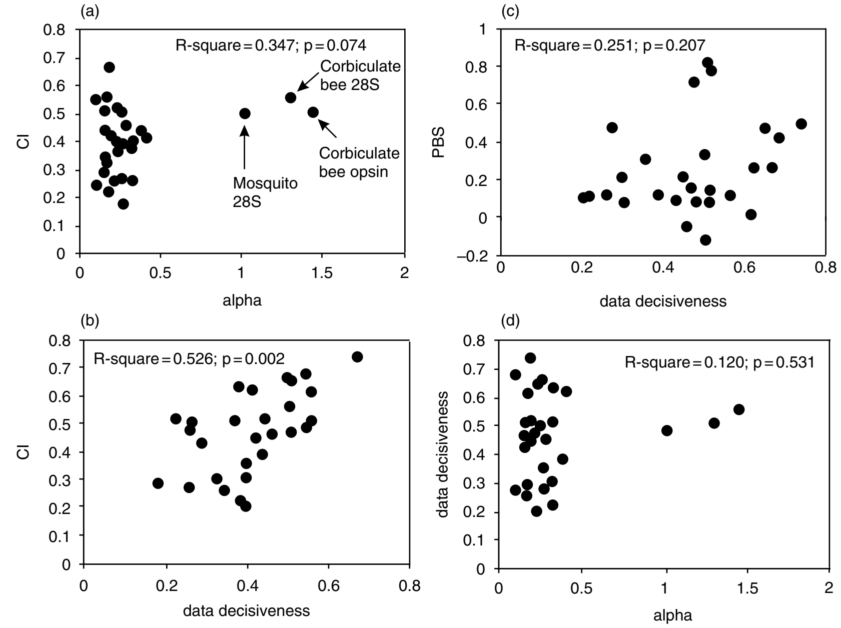

Fig. 7. (a) Relationship between consistency index and α (the shape of the gamma distribution). (b) Relationship between consistency index and data decisiveness. (c) Relationship between partitioned Bremer support and data decisiveness. (d) Relationship between data decisiveness and and α (the shape of the gamma distribution).

stems and loops or rapidly evolving and slowly evolving regions. Such an analysis would be informative, but partitioning the data in this way may be somewhat arbitrary, so we chose to treat the ribosomal genes as one substitution class. Finally, the studies we analysed span a wide range of taxonomic levels from those analysing phylogenetic relationships among orders (e.g. Ogden & Whiting, 2003) to generic and tribal-level relationships within a subfamily (e.g. Cameron & Mardulyn, 2001). Such differences in phylogenetic level may influence the perceived quality of the data and perhaps data set performance. Generally, higher-level relationships are more difficult to resolve than lower-level relationships. However, there were too few studies which combined nuclear protein-coding and nuclear ribosomal genes together to do more a detailed analysis at various phylogenetic levels. Furthermore, the studies we analysed all could be considered 'higher-level' studies of insects in that all analysed relationships above the generic and tribal level.

Given that nuclear ribosomal and protein-coding genes perform very similarly in combined phylogenetic analysis, what recommendations can we make as to the choice between nuclear protein-coding genes and nuclear ribosomal genes? The main advantage of the protein-coding genes appears to be in their ease of alignment and in the number of possible candidate loci that could be analysed. In addition, nuclear protein-coding genes exhibit variable rates of substitution such that some, rapidly evolving, genes may be good choices for recent divergences when analysed as nucleotide sequences (e.g. *period*, Regier *et al.*, 1998), whereas other more slowly evolving genes may be good choices for deep divergences when analysed as amino acid sequences (e.g. RNA polymerase II, Shultz & Regier, 2000). The ability to analyse protein-coding gene variation either as nucleotide sequences or as amino acid sequences greatly expands the range of taxonomic levels over which they can be applied (but see Simmons *et al.*, 2002).

Although nuclear protein-coding genes provide excellent data sets, they are often more difficult than ribosomal genes to amplify and sequence. Problems may arise for several reasons, including lack of highly conserved regions for primer design, long introns, and multiple paralogous copies. Ribosomal genes, in contrast, generally are easy to amplify and sequence and there is an expanding data set on these genes available in GenBank and EMBL. We would therefore recommend that researchers attempt to combine ribosomal genes with protein-coding genes whenever possible, as was done in the twelve studies that we analysed here. Such combined analyses may take advantage of the features of both types of data. For systematists working on higher-level studies the ribosomal genes may be amplified in all the taxa, including poorly preserved specimens and even fossil taxa. While the corresponding protein-coding gene data may not be available for all taxa, the results of the protein-coding genes may help to identify obvious errors in the ribosomal gene data sets, such as contaminant sequences and chimeric sequences (Kjer, 2004) and obvious errors in the alignment of the ribosomal gene sequences. Recent studies using multiple nuclear protein-coding genes plus nuclear ribosomal genes in combination with morphology indicate the power of this combined gene approach (Ward & Downie, 2005).

Both ribosomal and protein coding genes have additional advantages for higher-level studies that have been largely overlooked so far: the presence of highly conserved macromutational changes than can be coded for phylogenetic analysis. Such macromutational changes in the ribosomal genes include discrete and sometimes striking changes in the length of stems and/or loops. Coding these characters is possible using recently described methods (Billoud *et al.*, 2000; Lutzoni *et al*., 2000). Nuclear protein-coding genes possess additional (macromutational) information in the form of intron presence/absence (Moulton & Wiegmann, 2004) that can be coded for phylogenetic analysis. Introns have been shown to provide useful phylogenetic characters in a recent study of bees (Brady & Danforth, 2004).

Our major conclusion is that ribosomal and protein-coding nuclear genes differ little in both phylogenetic utility (assuming that one is willing to accept the ribosomal gene alignments, which are sometimes problematic) and in nucleotide substitution patterns. However, we view the protein-coding genes as a better alternative to ribosomal genes because they are not hindered by the uncertainty associated with alignment and because they occur in a practically limitless variety of genes characterized by variable rates of nucleotide and amino acid substitution. Nevertheless, nuclear ribosomal genes will continue to be a commonly used phylogenetic data sets because they are easy to amplify across a broad range of insect taxa and because there is already an enormous amount of data available for diverse insect orders.

Acknowledgements

We are grateful to the authors who generously and often very promptly provided us with their combined data sets. Eduardo Almeida, Sophie Cardinal, Robert Snyder and four reviewers provided very helpful comments on this paper. Peter Cranston provided numerous editorial and methodological comments on the paper. This project was supported by a National Science Foundation Research Grants in Systematic Biology (DEB-0211701) to B.N.D. and a Doctoral Dissertation Improvement Grant to B.N.D. and C.-P.L. (DEB-0104893).

References

Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Ascher, J.A., Danforth, B.N. & Ji, S. (2001) Phylogenetic utility of the major opsin in bees (Hymenoptera: Apoidea): a reassessment. *Molecular Phylogenetics and Evolution*, **19**, 76–93.

Baker, R.H. & Desalle, R. (1997) Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Systematic Biology*, **46**, 654–673.

Baker, R.H., Wilkinson, G.S. & Desalle, R. (2001) Phylogenetic utility of different types of data used to infer evolutionary relationships among stalk-eyed flies (Diopsidae). *Systematic Biology*, **50**, 87–105.

Barker, F.K. & Lutzoni, F.M. (2002) The utility of the incongruence length difference test. *Systematic Biology*, **51**, 625–637.

Belshaw, R. & Quicke, D.L.J. (2002) Robustness of ancestral character state estimates: evolution of life history strategy in ichneumonoid parasitoids. *Systematic Biology*, **51**, 450–477.

Billoud, B., Guerrucci, M.-A., Masselot, M. & Deutsch, J.S. (2000) Cirripede phylogeny using a novel approach: molecular morphometrics. *Molecular Biology and Evolution*, **17**, 1435–1445.

Brady, S.G. (2003) Evolution of the army ant syndrome: the origin and long-term evolutionary stasis of a complex of behavioral and reproductive adaptations. *Proceedings of the National Academy of Sciences (USA)*, **100**, 6575–6579.

Brady, S.G. & Danforth, B.N. (2004) Recent intron gain in elongation factor-1α (EF-1α) of colletid bees (Hymenoptera: Colletidae). *Molecular Biology and Evolution*, **21**, 691–696.

Bremer, K. (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution*, **42**, 795–803.

Brower, A.V.Z. (2000) Phylogenetic relationships among the Nymphalidae (Lepidoptera) inferred from partial sequences of the *wingless* gene. *Proceedings of the Royal Society of London B (Biological Science)*, **267**, 1201–1211.

Brower, A.V.Z. & DeSalle, R. (1994) Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. *Annals of the Entomological Society of America*, **87**, 702–716.

Brower, A.V.Z. & DeSalle, R. (1998) Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: the utility of wingless as a source of characters for phylogenetic inference. *Insect Molecular Biology*, **7**, 73–82.

Brower, A.V.Z. & Egan, M.G. (1997) Cladistic analysis of *Heliconius* butterflies and relatives (Nymphalidae: Heliconiiti): a revised phylogenetic position for Eueides based on sequences from mtDNA and a nuclear gene. *Proceedings of the Royal Society of London B (Biological Science)*, **264**, 969–977.

Buckley, T.R., Arensburger, P., Simon, C. & Chambers, G.K. (2002) Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. *Systematic Biology*, **51**, 4–18.

Cameron, S.A. & Mardulyn, P. (2001) Multiple molecular data sets suggest independent origins of highly eusocial behavior in bees (Hymenoptera: Apinae). *Systematic Biology*, **50**, 192–214.

Campbell, D.L., Brower, A.V.Z. & Pierce, N.E. (2000) Molecular evolution of the wingless gene and its implications for the phylogenetic placement of the butterfly family Riodinidae (Lepidoptera: Papilionoidea). *Molecular Biology and Evolution*, **17**, 684–696.

Caterino, M.S., Cho, S. & Sperling, F.A.H. (2000) The current state of insect molecular systematics: a thriving Tower of Babel. *Annual Review of Entomology*, **45**, 1–54.

Caterino, M.S., Reed, R.D., Kuo, M.M. & Sperling, F.A.H. (2001) A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera: Papilionidae). *Systematic Biology*, **50**, 106–127.

Cho, S., Mitchell, A., Regier, J.C., Mitter, C., Poole, R.W., Friedlander, T.P. & Zhao, S. (1995) A highly conserved nuclear gene for low level phylogenetics: Elongation factor-1 alpha recovers morphology-based tree for heliothine moths. *Molecular Biology and Evolution*, **12**, 650–656.

Clark, M.A., Moran, N.A., Baumann, P. & Wernegreen, J.J. (2000) Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. *Evolution*, **54**, 517–525.

Cognato, A.I. & Vogler, A.P. (2001) Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). *Systematic Biology*, **50**, 758–780.

Collins, K.P. & Wiegmann, B.M. (2002) Phylogenetic relationships and placement of the Empidoidea (Diptera: Brachycera) based on 28S rDNA and EF-1alpha sequences. *Insect Systematics and Evolution*, **33**, 421–444.

Cryan, J.R., Wiegmann, B.M., Deitz, L.L. & Dietrich, C.H. (2000) Phylogeny of the treehoppers (Insecta: Hemiptera:

Membracidae): evidence from two nuclear genes. *Molecular Phylogenetics and Evolution*, **17**, 317–334.

Danforth, B.N. (2002) Evolution of sociality in a primitively eusocial lineage of bees. *Proceedings of the National Academy of Sciences (USA)*, **99**, 286–290.

Danforth, B.N., Conway, L. & Ji, S. (2003) Phylogeny of eusocial *Lasioglossum* reveals multiple losses of eusociality within a primitively eusocial clade of bees (Hymenoptera: Halictidae). *Systematic Biology*, **52**, 23–36.

Danforth, B.N. & Ji, S. (1998) Elongation factor-1α occurs as two copies in bees: Implications for phylogenetic analysis of EF-1α sequences in insects. *Molecular Biology and Evolution*, **15**, 225–235.

Darlu, P. & Lecointre, G. (2002) When does the incongruence length difference test fail? *Molecular Biology and Evolution*, **19**, 432–437.

Dietrich, C.H., Rakitov, R.A., Holmes, J.L. & Black, W.C. (2001) Phylogeny of the major lineages of Membracoidea (Insecta: Hemiptera: Cicadomorpha) based on 28S rDNA sequences. *Molecular Phylogenetics and Evolution*, **18**, 293–305.

Dowton, M. & Austin, A.D. (2002) Increased congruence does not necessarily indicate increased phylogenetic accuracy – the behavior of the incongruence length difference test in mixed-model analyses. *Systematic Biology*, **51**, 19–31.

Fang, Q.Q., S.Cho, Regier, J.C. Mitter, C. Mathews, M. Poole, R.W. Friedlander, T.P. & Zhao & S. (1997) A new nuclear gene for insect phylogenetics: DOPA decarboxylase is informative of relationships within Heliothinae (Lepidoptera: Noctuidae). *Systematic Biology*, **46**, 269–283.

Fang, Q.Q. Mitchell, A. Regier, J.C. Mitter, C. Friedlander, T.P. & Poole, R.W. (2000) Phylogenetic utility of the nuclear gene dopa decarboxylase in noctuoid moths (Insecta: Lepidoptera: Noctuoidea). *Molecular Phylogenetics and Evolution*, **15**, 473–486.

Farris, J.S. Källersjö, M. Kluge, A.G. & Bult & M. (1995) Testing significance of incongruence. *Cladistics*, **10**, 315–319.

Friedlander, T.P. Horst, K.R. Regier, J.C. Mitter, C. Peigler, R.S. & Fang, Q.Q. (1998) Two nuclear genes yield concordant relationships within attacini (Lepidoptera: Saturniidae). *Molecular Phylogenetics and Evolution*, **9**, 131–140.

Friedlander, T.P. Regier, J.C. & Mitter, C. (1992) Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. *Systematic Biology*, **41**, 483–490.

Friedlander, T.P. Regier, J.C. & Mitter, C. (1994) Phylogenetic information content of five nuclear gene sequences in animals: Initial assessment of character sets from concordance and divergence studies. *Systematic Biology*, **43**, 511–525.

Friedlander, T.P. Regier, J.C. Mitter, C. & Wagner, D.L. (1996) A nuclear gene for higher level phylogenetics: Phosphoenolpyruvate carboxykinase tracks Mesozoic-age divergences within Lepidoptera (Insecta). *Molecular Biology and Evolution*, **13**, 594–604.

Friedlander, T.P. Regier, J.C. Mitter, C. Wagner, D.L. & Fang, Q.Q. (2000) Evolution of heteroneuran Lepidoptera (Insecta) and the utility of dopa decarboxylase for Cretaceous-age phylogenetics. *Zoological Journal of the Linnean Society*, **130**, 213–234.

Giribet, G. Edgecomb, G.D. & Wheeler, W.C. (2001) Arthropod phylogeny based on eight molecular loci and morphology. *Nature*, **413**, 157–161.

Goloboff, P.A. (1991) Homoplasy and the choice among cladograms. *Cladistics*, **7**, 215–232.

Hickson, R.E. Simon, C. & Perrey, S.W. (2000) The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. *Molecular Biology and Evolution*, **17**, 530–539.

Hillis, D.M. & Dixon, M.T. (1991) Ribsomal DNA: molecular evolution and phylogenetic inference. *Quarterly Review of Biology*, **66**, 411–453.

Holt, *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.

Hovmöller, R., Pape, T. & Källersjö, M. (2002) Palaeoptera problem: Basal pterygote phylogeny inferred from 18S and 28S rDNA sequences. *Cladistics*, **18**, 313–323.

Hsu, R., Briscoe, A.D., Chang, B.S.W. & Pierce, N.E. (2001) Molecular evolution of a long wavelength-sensitive opsin in mimetic *Heliconius* butterflies (Lepidoptera: Nymphalidae). *Biological Journal of the Linnean Society*, **72**, 435–449.

Huelsenbeck, J.P., Larget, B., Miller, R.E. & Ronquist, F. (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, **51**, 673–688.

Huelsenbeck, J.P. & Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford)*, **17**, 754–755.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R. & Bollback, J.P. (2001) Bayesian inference in phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Kawakita, A., Sota, T., Ascher, J.S., Ito, M., Tanaka, H. & Kato, M. (2003) Evolution and phylogenetic utility of alignment gaps within intron sequences of three nuclear genes in bumble bees (*Bombus*). *Molecular Biology and Evolution*, **20**, 87–92.

Kawakita, A., Sota, T., Ito, M., Ascher, J.S., Tanaka, H., Kato, M. & Roubik, D.W. (2004) Phylogeny, historical biogeography, and character evolution in bumble bees (*Bombus*: Apidae) based on simultaneous analysis of three nuclear gene sequences. *Molecular Phylogenetics and Evolution*, **31**, 799–804.

Kim, S.J., Kjer, K.M. & Duckett, C.N. (2003) Comparison between molecular and morphological-based phylogeny of galerucine/alticine leaf beetles (Coleoptera: Chrysomelidae). *Insect Systematics and Evolution*, **34**, 53–64.

Kitching, I.J., Forey, P.L., Humphries, C.J. & Williams, D.M. (1998) *Cladistics*, 2nd edn. The Systematics Association Publication no. 11. Oxford University Press, Oxford.

Kjer, K.M. (2004) Aligned 18S and insect phylogeny. *Systematic Biology*, **53**, 506–514.

Kjer, K.M., Blahnik, R.J. & Holzenthal, R.W. (2001) Phylogeny of Trichoptera (Caddisflies): characterization of signal and noise within multiple data sets. *Systematic Biology*, **50**, 781–816.

Krzywinski, J., Wilkerson, R.C. & Besansky, N.J. (2001a) Evolution of mitochondrial and ribosomal gene sequences in Anophelinae (Diptera: Culicidae): implications for phylogeny reconstruction. *Molecular Phylogenetics and Evolution*, **18**, 479–487.

Krzywinski, J., Wilkerson, R.C. & Besanksy, N.J. (2001b) Toward understanding Anophelinae (Diptera, Culicidae) phylogeny: Insights from nuclear single copy genes and the weight of evidence. *Systematic Biology*, **50**, 540–556.

Leys, R., Cooper, S.J.B. & Schwarz, M.P. (2002) Molecular phylogeny and historical biogeography of the large carpenter bees, genus *Xylocopa* (Hymenoptera: Apidae). *Biological Journal of the Linnean Society*, **77**, 249–266.

Lin, C.P. & Danforth, B.N. (2004) How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined data sets. *Molecular Phylogenetics and Evolution*, **30**, 686–702.

Lutzoni, F., Wagner, P., Reeb, V. & Zoller, S. (2000) Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Systematic Biology*, **49**, 628–651.

Mardulyn, P. & Cameron, S.A. (1999) The major opsin in bees (Insecta: Hymenoptera): a promising nuclear gene for higher level phylogenetics. *Molecular Phylogenetics and Evolution*, **12**, 168–176.

Mitchell, A., Cho, S., Regier, J.C., Mitter, C., Poole, R.W. & Mathews, M. (1997) Phylogenetic utility of Elongation factor-1-alpha in noctuoidea (Insecta: Lepidoptera): The limits of synonymous substitution. *Molecular Biology and Evolution*, **14**, 381–390.

Morris, D.C., Schwarz, M.P., Crespi, B.J. & Cooper, S.J.B. (2001) Phylogenetics of gall-inducing thrips on Australian *Acacia*. *Biological Journal of the Linnean Society*, **74**, 73–86.

Moulton, J.K. & Wiegmann, B.M. (2004) Evolution and phylogenetic utility of CAD (rudimentary) among Mesozoic-aged Eremoneuran Diptera (Insecta). *Molecular Phylogenetics and Evolution*, **31**, 363–378.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P. & Nieves-Aldrey, J.L. (2004) Bayesian phylogenetic analysis of combined data. *Systematic Biology*, **53**, 47–67.

Ogden, T.H. & Whiting, M.F. (2003) The problem with 'the Paleoptera problem:' sense and sensitivity. *Cladistics*, **19**, 432–442.

Ortiz-Rivas, B., Moya, A. & Martìnez-Torres, D. (2004) Molecular systematics of aphids (Homoptera: Aphididae): New insights from the long-wavelength opsin gene. *Molecular Phylogenetics and Evolution*, **30**, 24–37.

Reed, R.D. & Sperling, F.A. (1999) Interaction of process partitions in phylogenetic analysis: An example from the swallowtail butterfly genus *Papilio*. *Molecular Biology and Evolution*, **16**, 286–297.

Regier, J.C., Fang, Q.Q., Mitter, C., Peigler, R.S., Friedlander, T.P. & Solis, M.-A. (1998) Evolution and phylogenetic utility of the *period* gene in Lepidoptera. *Molecular Biology and Evolution*, **15**, 1172–1182.

Regier, J.C., Mitter, C., Peigler, R.S. & Friedlander, T.P. (2000) Phylogenetic relationships in Lasiocampidae (Lepidoptera): Initial evidence from elongation factor-1alpha sequences. *Insect Systematics and Evolution*, **31**, 179–186.

Schulmeister, S., Wheeler, W.C. & Carpenter, J.M. (2002) Simultaneous analysis of the basal lineages of Hymenoptera (Insecta) using sensitivity analysis. *Cladistics*, **18**, 455–484.

Shultz, J.W. & Regier, J.C. (2000) Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. *Proceedings of the Royal Society of London B (Biology Science)*, **267**, 1011–1019.

Simmons, M.P., Ochoterena, H. & Freudenstein, J.V. (2002) Amino acid vs. nucleotide characters: challenging preconceived notions. *Molecular Phylogenetics and Evolution*, **24**, 78–90.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. (1994) Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America*, **87**, 651–701.

Sipes, S.D. & Wolf, P.G. (2001) Phylogenetic relationships within *Diadasia*, a group of specialist bees. *Molecular Phylogenetics and Evolution*, **19**, 144–156.

Sorenson, M.D. (1999) *Treerot*, Version 2. Boston University, Boston, MA.

Sota, T. & Vogler, A.P. (2001) Incongruence of mitochondrial and nuclear gene trees in the carabid beetles *Ohomopterus*. *Systematic Biology*, **50**, 39–59.

Sullivan, J., Swofford, D.L. & Naylor, G.J.P. (1999) The effects of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Molecular Biology and Evolution*, **16**, 1347–1356.

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

Swofford, D.L., Olsen, G.J., Waddell, P.J. & Hillis, D.M. (1996) Phylogenentic inference. *Molecular Systematics*, 2nd edn (ed. by D. M. Hillis, C. Moritz and B. K. Mable), pp. 407–514. Sinauer, Sunderland, MA.

Tatarenkov, A., Kwiatovski, 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.

Ward, P.S. & Downie, D.A. (2005) The ant subfamily Pseudomyrmecinae (Hymenoptera: Formicidae): phylogeny and evolution of the big-eyed arboreal ants. *Systematic Entomology*, **30**, 310–335.

Wheeler, W.C., Whiting, M., Wheeler, Q.D. & Carpenter, J.M. (2001) The phylogeny of the extant hexapod orders. *Cladistics*, **17**, 113–169.

Whiting, M.F., Carpenter, J.C., Wheeler, Q.D. & Wheeler, W.C. (1997) The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. *Systematic Biology*, **46**, 1–68.

Wiegmann, B.M., Mitter, C., Regier, J.C., Friedlander, T.P., Wagner, D.M. & Nielsen, E.S. (2000) Nuclear genes resolve Mesozoic-aged divergences in the insect order Lepidoptera. *Molecular Phylogenetics and Evolution*, **15**, 242–259.

Xia, Q. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940. http://silkworm.genomics.org.cn/index.jsp.

Yang, Z. (1998) On the best evolutionary rate for phylogenetic analysis. *Systematic Biology*, **47**, 125–133.

Yang, L., Wiegmann, B.M., Yeates, D.K. & Irwin, M.E. (2000) Higher-level phylogeny of the Therevidae (Diptera: Insecta) based on 28S ribosomal and elongation factor-1alpha gene sequences. *Molecular Phylogenetics and Evolution*, **15**, 440–451.

Yang, Z.H. & Yoder, A.D. (1999) Estimation of the transition/transversion rate bias and species sampling. *Journal of Molecular Evolution*, **48**, 274–283.

Accepted 1 March 2005
First published online 21 April 2005

Appendix 1. GenBank numbers corresponding to the data analysed as part of the Cameron & Mardulyn (2001) study. EF-1α was added to the original Cameron & Mardulyn combined data set of 28S and LW (green) rhodopsin. Additional (mitochondrial) genes included in their study were cytB and 16S.

Species	28S (D2)	EF-1α	LW rhodopsin
Outgroups			
Melissodes rustica	AF181604	missing	AF091731
Centris inermis	missing	missing	AF181577
Anthophora pacifica	AF181605	AY208280	AF181578
Habropoda depressa	AF181606	missing	AF181579
Xylocopa virginica	AF181607	AY208290	AF091730
Ingroups			
Apis mellifera	AF181590	AY208278	AF091732
Apis nigrocincta	AF181591	AY208279	AF091728
Apis dorsata	AF066902	AY208277	AF091733
Bombus pennsylvanicus	AF181593	AY208281	AY268388 & AF091727
Bombus avinoviellus	AF181592	missing	AY268394 & AF091719
Bombus terrestris	AF181594	AY208282	AF091722
Trigona hypogea	AF181595	missing	AF091724
Scaptotrigona depilis	AF181596	AY208288	AF091729
Tetragona dorsalis	AF181597	AY208289	AF091726
Lestrimelitta limao	AF181598	AY208287	AF091723
Melipona compressipes	AF181599	missing	missing
Eufriesea caerulescens	AF181600	AY208283	AF091725
Euglossa imperialis	AF181601	AY208284	AF091720
Exaerete frontalis	AF181602	AY208286	AF091718
Eulaema meriana	AF181603	AY208285	AF091721

Appendix 2. GenBank numbers corresponding to the data analysed as part of the Moulton & Wiegmann (2004) study.

Species	CAD	28S–B	28S–C
Outgroup[1]	AY280675	AF503026	AF503071
Opetia	AY280692	AF502992	AF503013
Paraplatypeza	AY280693	AF502993	AF503014
Rhingia	AY280697	AF502998	AF503019
Drosophila	AAAB01008846	M21017	M21017
Musca	AY280689	AF503004	AF503025
Acarteroptera	AY280672	AF503032	AF503077
Atelestus	AY280700/01	AF502984	AF503005
Meghyperus	AY280688	AF502985	AF503006
Clinocera	AY280677	AF503038	AF503083
Dolichopus	AY280678	AF502989	AF503010
Empis	AY280681	AF503042	AF503087
Rhamphomyia	AY280696	AF503048	AF503093
Schistostoma	AY280698	AF503066	AF503110
Anthalia	AY280674	AF503056	AF503101
Leptopeza	AY280686	AF503059	AF503104
Platypalpus	AY280695	AF503063	AF503108
Lonchotoptera	AY280687	AF502991	AF503012

[1] Outgroup sequences were obtained from two different genera of Bombyliidae. For CAD we used *Bombylius major* and for the two fragments of 28S we used *Lordotus* sp.

Appendix 3

MrBayes blocks used to calculate relative rates and gamma shape parameters for the Wiegmann *et al.* (2000) data set on basal Lepidoptera.

(a) GTR + SSR model

```
begin mrbayes;
[data partition]
      set autoclose = yes;
      charset pepnt1 = 1355– 1991\3;
      charset pepnt2 = 1356– 1992\3;
      charset pepnt3 = 1357– 1990\3;
      charset 18 s = 1–1354 1978–92;
      partition codongene = 4:pepnt1,pepnt2,pepnt3,18 s;
      set partition = codongene;
[model = GTR + SSR with partition − specific substitution rates of GTR, character state frequencies]
      lset nst = 6;
      unlink revmat = (all);
      unlink statefreq = (all);
      prset applyto = (all) ratepr = variable;
[mcmc run]
      mcmc ngen = 1000000 printfreq = 50
      samplefreq = 50 savebrlens = yes;
end;
```

(b) GTR + G model

```
begin mrbayes;
[data partition]
      set autoclose = yes;
      charset pep = 1355– 1977;
      charset 18 s = 1–1354 1978–92;
      partition gene = 2:pep,18 s;
      set partition = gene;
[model = GTR + G with partition − specific gamma, substitution rates of GTR and character state frequencies]
      lset nst = 6 rates = gamma;
      unlink shape = (all);
      unlink revmat = (all);
      unlink statefreq = (all);
[mcmc run]
      mcmc ngen = 1000000 printfreq = 50
      samplefreq = 50 savebrlens = yes;
end;
```