Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 30 (2004) 686–702
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets

Chung-Ping Lin[a] and Bryan N. Danforth[b,*]

[a] *Department of Biological Sciences, College of Arts and Science, Tucker Hall, University of Missouri, Columbia, MO 65211, USA*
[b] *Department of Entomology, Comstock Hall, Cornell University, Ithaca, NY 14853-0901, USA*

Received 21 February 2003; revised 3 June 2003

Abstract

We analyzed 12 combined mitochondrial and nuclear gene datasets in seven orders of insects using both equal weights parsimony (to evaluate phylogenetic utility) and Bayesian methods (to investigate substitution patterns). For the Bayesian analyses we used relatively complex models (e.g., general time reversible models with rate variation) that allowed us to quantitatively compare relative rates among genes and codon positions, patterns of rate variation among genes, and substitution patterns within genes. Our analyses indicate that nuclear and mitochondrial genes differ in a number of important ways, some of which are correlated with phylogenetic utility. First and most obviously, nuclear genes generally evolve more slowly than mitochondrial genes (except in one case), making them better markers for deep divergences. Second, nuclear genes showed universally high values of CI and (generally) contribute more to overall tree resolution than mitochondrial genes (as measured by partitioned Bremer support). Third, nuclear genes show more homogeneous patterns of among-site rate variation (higher values of α than mitochondrial genes). Finally, nuclear genes show more symmetrical transformation rate matrices than mitochondrial genes. The combination of low values of α and highly asymmetrical transformation rate matrices may explain the overall poor performance of mitochondrial genes when compared to nuclear genes in the same analysis. Our analyses indicate that some parameters are highly correlated. For example, A/T bias was positively and significantly associated with relative rate and CI was positively and significantly associated with α (the shape of the gamma distribution). These results provide important insights into the substitution patterns that might characterized high quality genes for phylogenetic analysis: high values of α, unbiased base composition, and symmetrical transformation rate matrices. We argue that insect molecular systematists should increasingly focus on nuclear rather than mitochondrial gene datasets because nuclear genes do not suffer from the same substitutional biases that characterize mitochondrial genes.

Keywords: Insect phylogeny; Molecular evolution; Bayesian analysis

1. Introduction

It has been nine years since the literature on insect mitochondrial and nuclear genes was reviewed (Brower and DeSalle, 1994; Simon et al., 1994), and three years since the publication of Caterino et al.'s (2000) review of the state of insect molecular systematics. In that short time the field of insect molecular systematics has undergone some striking changes, dealing both with new methodologies (e.g., Bayesian methods Huelsenbeck et al., 2001, 2002), and with a recent flood of datasets based on combined nuclear and mitochondrial genes (see below). The existence of numerous combined nuclear + mitochondrial gene datasets provides an opportunity to examine across a broad array of insect groups both the utility of these two types of data and the patterns of nucleotide substitution that characterize nuclear and mitochondrial genes. We provide a brief review below of differences between mitochondrial and nuclear genes, before introducing our approach to analyzing the quality and characteristics of these two types of data in analyses of insect phylogeny.

Mitochondrial genes have been for many years the most commonly used source of data for studies of insect molecular phylogeny and phylogeography (Avise, 1987,

* Corresponding author. Fax: 1-607-255-0939.
E-mail address: bnd1@cornell.edu (B.N. Danforth).

1055-7903/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S1055-7903(03)00241-0

1994, 2000; Caterino et al., 2000; Harrison, 1989; Simmons and Weller, 2001; Simon et al., 1994). Mitochondrial genes are viewed as advantageous for phylogenetic analysis for several reasons. First, mitochondrial genes are generally easier to amplify than nuclear genes and conserved mitochondrial primers are widely available (see Simon et al., 1994). Second, mitochondrial genes lack non-coding regions (i.e., introns) that are common in single-copy nuclear genes. Third, mitochondrial genes are clonally inherited (through the maternal lineage) and non-recombining, making recombination, paralogy, and heterozygosity (heteroplasmy in mitochondrial genes) less of a problem for phylogenetic analysis. However, note that nuclear copies of mitochondrial genes may exist, creating problems for the analysis of mitochondrial gene sequences (Sunnucks and Hales, 1996; Zhang and Hewitt, 1996). Fourth, mitochondrial genes are generally thought to evolve at higher rates than nuclear protein-coding genes. In insects, mitochondrial genes are estimated to evolve 2–9 times faster than nuclear protein-coding genes (DeSalle et al., 1987; Monteiro and Pierce, 2001; Moriyama and Powell, 1997). For studies of closely related taxa that have diverged relatively recently, this is advantageous.

Mitochondrial genes have some clear disadvantages as well. Since all mitochondrial genes are linked on the same chromosome one could argue that they do not provide an independent estimate of phylogeny in the same way that unlinked single-copy, nuclear genes do (Harrison, 1989). Furthermore, the higher rate of substitution can be disadvantageous when one is trying to resolve divergences of more than 5–10 million years. Most importantly for phylogenetic analysis, mitochondrial genes have attributes that tend to lead to high levels of homoplasy when analyzed by standard phylogenetic methods, such as an extreme A/T bias in third positions (Frati et al., 1997; Mooers and Holmes, 2000).

Since the mid-1990's, thanks to work by Jerome Regier and colleagues as well as others (Brower and DeSalle, 1994; Friedlander et al., 1992, 1994), insect molecular systematists now have available protein coding nuclear genes that hold great promise for resolving deep (e.g., Cretaceous and older) divergences in insects. Such genes include EF-1α (Buckley et al., 2002; Caterino et al., 2001; Cho et al., 1995; Clark et al., 2000; Cognato and Vogler, 2001; Danforth, 2002; Danforth and Ji, 1998; Kjer et al., 2001; Mitchell et al., 1997; Reed and Sperling, 1999; Regier et al., 2000; Sipes and Wolf, 2001), PEPCK (Friedlander et al., 1996; Leys et al., 2002; Sota and Vogler, 2001; Wiegmann et al., 2000), DDC (Fang et al., 1997, 2000; Friedlander et al., 1998, 2000; Tatarenkov et al., 1999), *wingless* (Brower, 2000; Brower and DeSalle, 1998; Brower and Egan, 1997; Campbell et al., 2000; Morris et al., 2001), *white* (Baker et al., 2001), opsin (Ascher et al., 2001; Cameron and Mardulyn, 2001; Danforth et al., 2003; Hsu et al., 2001; Mardulyn and Cameron, 1999; Kawakita et al., 2003), *hunchback* (Baker and DeSalle, 1997), *period* (Regier et al., 1998), and others (see Brower and DeSalle, 1994; Caterino et al., 2000 for complete lists of nuclear protein coding genes used in insects). Nuclear genes have several desirable attributes relative to mitochondrial genes. First, nuclear genes generally have less biased base composition (but see Tarrío et al., 2001). Second, nuclear genes (generally) evolve more slowly than mitochondrial genes, and third, nuclear genes include both slowly evolving regions (exons) and more rapidly evolving regions (introns) (Brower and DeSalle, 1994; Friedlander et al., 1992, 1994). However, nuclear genes are often more difficult to work with than mitochondrial genes because they occur in lower copy number (and are therefore sometimes more difficult to amplify via PCR) and often involve two or more paralogous loci that may cause problems in phylogenetic analysis. *Wingless*, for example, occurs in at least five copies in insects and extreme caution needs to be taken when analyzing *wingless* sequences (Schubert et al., 2000).

When mitochondrial genes have been used in combination with nuclear genes it has generally been observed that the nuclear genes have greater resolving power (especially at deeper taxonomic levels), show lower levels of homoplasy (as measured by consistency index; CI), and provide greater bootstrap (Felsenstein, 1985) and Bremer (Bremer, 1988) support than mitochondrial genes (Baker et al., 2001; Brady, 2002; Danforth et al., 2003; Leys et al., 2000, 2002; Lin et al., submitted; Morris et al., 2002; Reed and Sperling, 1999). In a recent example (Baker et al., 2001), a comparison of mitochondrial (12S, 16S, and COII) and nuclear (*white*, *wingless*, EF-1α) genes showed striking differences between these data types: nuclear genes outperformed mitochondrial genes in most measures of phylogenetic utility, including tree resolution, consistency index, data decisiveness, and Bremer support (Table 3 in Baker et al., 2001). Others have noted the distinction between mitochondrial and nuclear genes and have commented on the generally better performance of the nuclear genes (e.g., Caterino et al., 2000; but see Monteiro and Pierce, 2001 for an alternative view).

It is now common practice among insect molecular systematists to combine one or more mitochondrial with one or more nuclear genes because the two types of data are unlinked and evolving under different evolutionary constraints. The existence of numerous combined insect mitochondrial and nuclear gene datasets provides an excellent opportunity to examine, in general, how the substitution patterns of mitochondrial and nuclear genes differ. By analyzing the two types of datasets in a combined analysis we can ask important questions, such as how do nuclear and mitochondrial genes compare in terms of phylogenetic utility? How do the details of the substitution process differ in mitochondrial and nuclear

genes? How does rate variation among sites within mitochondrial and nuclear genes compare? What attributes of the substitution process are correlated with dataset quality? And, finally, are there differences in mitochondrial and nuclear gene substitution patterns that could explain the (generally) better performance of nuclear genes when combined with mitochondrial genes?

In order to make comparisons among genes and gene regions we used a Bayesian framework. Bayesian methods are increasingly being used in evolutionary biology and systematics for inferring phylogeny, evaluating phylogenetic uncertainty (Huelsenbeck and Rannala, 1997; Huelsenbeck et al., 2000b; Lutzoni et al., 2001; Nielson, 2002), analyzing patterns of cospeciation (Huelsenbeck et al., 2000a), estimating ancestral states (Huelsenbeck and Bollback, 2001; Lutzoni et al., 2001), and estimating divergence times (Thorne et al., 1998; Kishino et al., 2001; Thorne and Kishino, 2002). Bayesian methods also provide an ideal framework for investigating and characterizing substitution patterns in molecular datasets (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 phylogeny can be effectively ignored (treated as a "nuisance parameter"; Huelsenbeck et al., 2001) so that estimates of substitution parameters are not dependent on any *particular* tree topology. This is advantageous because it means that estimates of parameter values incorporate uncertainty in tree topology that exists in most molecular phylogenetic studies. One can estimate parameter values using maximum likelihood (ML), but in that case a particular tree (which could be wrong) would have had to be specified for each analysis. While Bayesian (and ML) methods are being adopted by molecular systematists for tree reconstruction, systematists rarely examine in detail what the Bayesian (or ML) parameter estimates can tell us about the substitution patterns in general.

We used a Bayesian approach in order to compare the substitution patterns that characterize mitochondrial vs. nuclear genes. By comparing substitution patterns of the genes in an explicit, model-based way we hoped to detect general patterns that would explain why mitochondrial genes generally perform poorly in comparison to nuclear genes. By using a combination of the general time reversible model with some sites treated as invariant and the remaining sites assumed to follow a gamma distribution (GTR + I + G) and the general time reversible model with site-specific rates (GTR + SSR) we were able to explicitly compare rates of substitution among positions, as well as rate of transformations among bases within positions (Swofford et al., 1996). Understanding how substitution patterns differ between nuclear and mitochondrial genes would also provide some predictive power to those seeking to identify *new* promising genes for insect phylogenetic analysis. Our results corroborate earlier observations about how mitochondrial and nuclear gene substitution patterns differ, but also indicate some important (but overlooked) differences that characterize nuclear vs. mitochondrial genes.

2. Materials and methods

We obtained 12 combined mitochondrial and nuclear protein coding gene datasets from sources listed in Table 1. We selected studies in which the mitochondrial and nuclear gene datasets were (ideally, see below) >500 bp (in order to be able to infer the substitution patterns more accurately) and we sought datasets that utilized novel or previously unexplored nuclear or mitochondrial datasets. Our examples span both the Hemimetabola and the Holometabola and include seven orders: Hemiptera (true bugs), Thysanoptera (thrips), Phthiraptera (lice), Hymenoptera (wasps, ants, and bees), Coleoptera (beetles), Lepidoptera (moths and butterflies), and Diptera (flies). The datasets also span a range of divergence times from closely related and recently diverged taxa (e.g., *Uroleucon*, estimated to be <5 my old; Clark et al., 2000) to more ancient and more divergent taxa (e.g., Papilionidae and Membracinae, each estimated to be >50 my old; Caterino et al., 2001; Lin et al., submitted). Datasets were obtained either directly from the authors, or were downloaded from the Systematic Biology web site (http://www.systematicbiology.org/) or from TreeBase (http://www.treebase.org). We limited our comparisons to protein-coding genes because we did not want our results confounded by ambiguities resulting from alignment problems, which are common in ribosomal gene datasets (Wheeler et al., 2001; Whiting et al., 1997). We could have analyzed additional datasets. However, we excluded some datasets that were based on taxa already well represented in our sample of datasets (e.g., bees: Cameron and Mardulyn, 2001; Sipes and Wolf, 2001; Kawakita et al., 2003). We also avoided datasets in which there were substantial amounts (i.e., >50%) of missing data (e.g., Farrell et al., 2001; Moulton, 2000), and datasets in which there were significant levels of incongruence among genes (e.g., Sota and Vogler, 2001).

Datasets varied from 15 taxa to over 100 taxa (Table 1) and individual gene regions varied in size from 348 to over 1500 bp. Maximum likelihood (and presumably Bayesian) parameter estimates are known to be sensitive to taxon sampling (Sullivan et al., 1999; Yang and Yoder, 1999). Sullivan et al. (1999), based on

Table 1
Overview of the datasets

Datasets	Order	Family	No. taxa	Mitochondrial gene(s)	Nuclear gene (s)	Reference
Lice (*Columbicola*)	Phthiraptera	Philopteridae	15	COI (384 bp)	EF-1a (348 bp)	Johnson et al. (2003)
Aphids (*Uroleucon*)	Hemiptera	Aphididae	15	COI (799 bp) COII (596 bp) ND1 (559 bp)	EF-1a exons (877 bp) EF-1a intron (241 bp)	Clark et al. (2000)
Treehoppers (Membracinae)	Hemiptera	Membracidae	112	COI (1236 bp) COII (517 bp)	Wingless (373 bp)	Lin et al. (submitted).
Gall-inducing thrips	Thysanoptera	Phlaethripidae	24	COI (550 bp)	EF-1a exons (422 bp) EF-1a intron (100 bp) Wingless (445 bp)	Morris et al. (2001)
Bark beetles (*Ips*)	Coleoptera	Scolytidae	44	COI (769 bp)	EF-1a exons (684 bp) EF-1a introns (83 bp)	Cognato and Vogler (2001)
Stalk-eyed flies	Diptera	Diopsidae	35	COII (436 bp)	EF-1a (1031 bp) Wingless (619 bp) white (486 bp)	Baker et al. (2001)
Nymphalid butterflies	Lepidoptera	Nymphalidae	23	COI (310 bp) COII (669 bp)	Wingless (378 bp)	Brower and DeSalle (1998)
Nymphalid butterflies (*Bicyclus*)	Lepidoptera	Nymphalidae	60	COI (945 bp) COII (969 bp)	EF-1a (890 bp)	Monteiro and Pierce (2001)
Swallowtail butterflies (Papilionidae)	Lepidoptera	Papilionidae	37	COI (1530 bp) COII (684 bp)	EF-1a (995 bp)	Caterino et al. (2001)
Swallowtail butterflies (*Papilio*)	Lepidoptera	Papilionidae	25	COI (1532 bp) COII (687 bp)	EF-1a (1010 bp)	Reed and Sperling (1999)
Halictid bees	Hymenoptera	Halictidae	53	COI (1239 bp)	EF-1a exons (801 bp) EF-1a introns (448 bp) Opsin exons (489 bp) Opsin introns (169 bp)	Danforth et al. (2003)
Carpenter bees (*Xylocopa*)	Hymenoptera	Apidae	27	COI (600 bp)	PEPCK exons (478 bp) PEPCK introns (573 bp)	Leys et al. (2002)

real and simulated datasets, suggested that at least 20 taxa were needed for accurate estimation of rate parameters. Only two datasets fell below this threshold, and, as pointed out by Sullivan et al. (1999), the number of taxa needed to accurately estimate model parameters will vary from dataset to dataset. We see no reason why the datasets we have chose should not provide good estimates of substitution parameters.

For the mitochondrial datasets there was a preponderance of COI and/or COII datasets. Only one dataset (Clark et al., 2000) included ND1. For the nuclear genes our datasets include mostly EF-1α data (9 of 12 datasets), but we included studies based on *wingless* (4 of 12 studies), opsin (1 of 12), PEPCK (1 of 12), and *white* (1 of 12) (Table 1). In several cases (5 of 12) the nuclear gene datasets included intron sequences.

We initially performed an equal weights parsimony analysis on the combined mitochondrial and nuclear datasets. These trees were checked against results reported in the papers cited in Table 1 in order to make sure that our results matched the published trees. Using Paup* 4.0 b10 (Swofford, 2002) we calculated the base proportions for each dataset and data partition within datasets (e.g., nt1, nt2, nt3, and introns). We also used Paup* 4.0 to calculate the consistency index (CI), the number of parsimony informative sites, and the number of equally parsimonious trees for each gene. For parsimony analyses we performed 100 random sequence additions and TBR branch swapping.

In order to assess the relative contribution of each gene to the overall results, we calculated partitioned Bremer support (PBS; Baker and DeSalle, 1997; Bremer, 1988) 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. In this paper, we use CI and PBS/min steps as two possible measures of dataset quality. We are aware that other measures of dataset quality exist (e.g., number of resolved nodes, Yang, 1998), however, the combination of CI and PBS/min steps provides a useful quantitative measure of homoplasy and support, respectively.

For the Bayesian analyses we used MrBayes v. 3.0 (Huelsenbeck and Ronquist, 2001; http://www.morphbank.ebc.uu.se/mrbayes3/). We analyzed the datasets using several different models. First, we analyzed the combined dataset using a GTR + SSR model with rate catagories corresponding to gene. Second, we used a GTR + SSR model with character partitions corresponding to first (nt1), second (nt2), third (nt3) positions, and introns, within genes. These relative rate estimates provide a quantitative way of comparing the

rates of substitution among genes and among data partitions within genes. We compared the log likelihood of the trees obtained with data partitioned by gene and by codon × gene. Using the likelihood ratio test (Huelsenbeck and Crandall, 1997) we evaluated whether additional partitioning of the data into codon positions provided a significant improvement in log likelihood using a χ^2 distribution. Finally, we used a GTR + I + G model for analysis of each dataset individually. From the GTR + I + G analysis we obtained the instantaneous rate matrix (Q matrix), the shape parameter of the gamma distribution (α), and the proportion of sites estimated to be invariant (pi) for each gene and for each data partition within gene (Swofford et al., 1996). This gave us a method for comparing the relative symmetry of the Q matrix, as well as heterogeneity in rates of substitution among sites (α, pi). We also examined the correlations among parameter estimates.

Analyses consisted of running four simultaneous chains for 1×10^6 generations. Trees were sampled at intervals of 50 generations for a total of 20,000 trees. Stability of the process was achieved when likelihood values approached equilibrium, as determined by plotting the ln likelihood scores against generation time (Fig. 1a). We discarded the "burn-in" region (trees and parameter estimates obtained before equilibrium; in general 1×10^5 generations, or 2000 trees) and calculated the mean, variance, and 95% credibility intervals of the parameter estimates using MrBayes. Trees were represented as 50% majority rule consensus trees using Paup* (Fig. 1b).

3. Results

3.1. Comparison among genes in CI and partitioned Bremer support

Table 2 summarizes the parsimony results we obtained for each of the 12 datasets. In all 12 studies, analyses of the total combined dataset provide a strongly supported phylogeny for the group of species included in the study. Combined analyses generally yielded fewer than 10 trees and each study showed reasonably strong bootstrap support. Analysis of individual genes indicated that, for all datasets, the mitochondrial genes had lower CI than the nuclear genes, indicating that mitochondrial genes show consistently higher levels of homoplasy than nuclear genes. For 8 of 12 datasets the nuclear genes had the highest values of partitioned Bremer support (standardized by minimum number of steps; Table 2). In some cases (treehoppers, stalk-eyed flies, swallowtail butterflies, halictid and carpenter bees) the nuclear genes provided considerably more support than the mitochondrial genes. Based on the aphid dataset (Clark et al., 2000), ND1 performs far better than either COI or COII, and is comparable to EF-1α in providing support in the parsimony analysis.

For six studies analyzed (treehoppers, stalk-eyed flies, *Papilio*, halictid bees, and carpenter bees) the authors commented that the mitochondrial genes were of less phylogenetic utility than the nuclear genes. Authors of other studies did not comment on the phylogenetic utility of the different genes. Only Monteiro and Pierce (2001) commented that the mitochondrial and nuclear genes were of equal utility. This is supported as well by our analysis of partitioned Bremer support (Table 2).

3.2. Base composition

For all studies analyzed the mitochondrial genes showed greater base compositional bias than nuclear genes. The treehopper dataset (Lin et al., submitted) shows one of the most extreme mitochondrial base compositional biases with 75.8% A/T in the mitochondrial COI + COII region and only 39.5% A/T in the *wingless* fragment analyzed. These patterns are even more extreme when one looks at the individual codon positions primarily because of third positions. In the halictid bee dataset, for example, A/T bias in COI nt3 was 90.7%. Among the nuclear genes analyzed, most showed more or less even base composition. *Wingless* is exceptional in showing a high G/C bias in third position sites both in bees (data not shown) and in ants (Brady, 2002). For most datasets there was no significant heterogeneity among taxa in base composition (Table 2) and no clear patterns among genes.

3.3. Relative rates among data partitions

Table 3 shows the results of the GTR + SSR analyses with the datasets partitioned by gene and by codon position within gene. For all 12 datasets there was a significant improvement in log likelihood when we partitioned the datasets by codon position as well as by gene, indicating that there is substantial rate heterogeneity among sites within genes.

Comparisons of substitution rates among genes revealed that mitochondrial genes generally show higher rates of substitution than protein-coding regions of the nuclear genes (Table 4). The mitochondrial genes evolved as much as 6-fold faster in some datasets (e.g., gall-inducing thrips, halictid bees). In some Lepidopteran datasets (e.g., nymphalid and swallowtail butterflies) the mitochondrial and nuclear genes evolved at roughly the same rate. In only one case (nymphalid butterflies; Brower and DeSalle, 1998) did the nuclear gene (*wingless*) evolve faster than the mitochondrial genes (COI and COII). Interestingly, this is also a dataset in which the mitochondrial gene performed far better than the nuclear gene in terms of partitioned

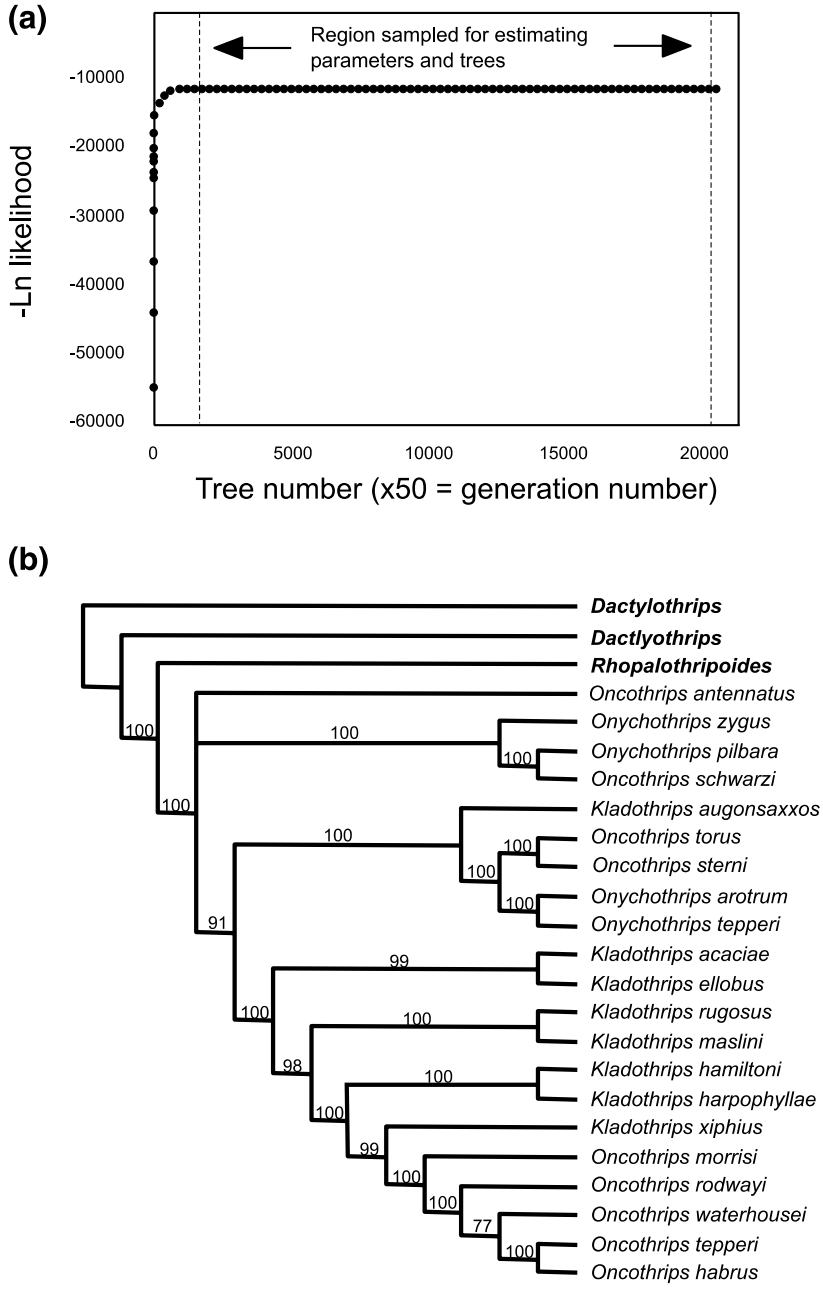

Fig. 1. (a) Relationship between likelihood score and the number of generations in the thrips dataset (Morris et al., 2001). In a typical Bayesian analysis, parameter estimates at the beginning of the run are poor (for example, all sites are initially assumed to evolve at a constant rate), and improvements in parameter estimates lead quickly to improved likelihood scores. After the "burn-in" the likelihood scores reach a plateau and the parameter estimates stabilize (region indicated by dashed lines). Mean and variance of the parameter estimates are calculated based on trees obtained after the "burn-in." (b) Trees obtained from the Bayesian analysis are represented as 50% majority rule consensus trees. Bayesian posterior probabilities are shown above the nodes for the thrips dataset (Morris et al., 2001). Posterior probabilities represent the proportion of the time each node was recovered during the stable part of the analysis. Outgroups are indicated in bold.

Bremer support (Table 2). As expected, nuclear introns evolve faster than the coding regions of the same genes (by as much as 5-fold; gall-inducing thrips).

A better understanding of rate variation is obtained when one looks at rates among codon positions within genes (see Figs. 2 and 3). In virtually all cases, nuclear gene third positions evolved at much lower rates than mitochondrial third positions. The one exception was the comparison of the nuclear genes *white*, *wingless*, and EF-1α with mitochondrial COII in stalk-eyed flies (Fig. 2f). In this case, nuclear *white* was not significantly different in third position rate than COII. Among the most striking differences in rate occurred in lice and in halictid bees (Fig. 3e), in which mitochondrial genes

Table 2
Summary of parsimony results

Data sets	Data partitions	A + T (%)	Base comp. hetero.	PI sites	CI[a]	PBS/min steps	MP trees	Author's comment	Reference
Lice	COI	62.4	$p = 0.997$	154	0.403	0.13	6	too fast	Johnson et al. (2003)
(*Columbicola*)	EF1a	51.4	$p = 0.999$	45	0.608	0.18	3	good	
Aphids	COI	77.2	$p = 1.00$	96	0.464	−0.01	293	nc[b]	Clark et al. (2000)
(*Uroleucon*)	COII	79.8	$p = 1.00$	74	0.482	−0.06	58	nc	
	ND1	83.7	$p = 1.00$	76	0.476	0.22	125	nc	
	EF1a	59.7	$p = 1.00$	89	0.612	0.16	1	nc	
Treehoppers	COI	70.2	$\mathbf{p < 0.001}$	726	0.113	0.59	4	bad	Lin et al. (submitted).
	COII	75.8	$\mathbf{p < 0.001}$	371	0.130	−0.04	72	bad	
	Wingless	39.5	$p = 1.00$	170	0.194	1.36	>1000	good	
Gall-inducing thrips	COI	72.6	$p = 0.999$	189	0.412	0.14	6	nc	Morris et al. (2001)
	EF1a	54.2	$p = 1.00$	56	0.595	0.25	216	nc	
	Wingless	45.6	$p = 1.00$	53	0.660	0.13	12	nc	
Bark beetles (*Ips*)	COI	67.4	$p = 0.999$	311	0.230	0.30	2	nc	Cognato and Vogler
	EF1a	57.9	$p = 1.00$	187	0.650	0.18	16	nc	(2001)
Stalk-eyed flies	COII	72.6	$p = 1.00$	164	0.292	0.18	8	bad	Baker et al. (2001)
	EF1a	53.6	$p = 0.999$	224	0.401	0.45	24	good	
	Wingless	58.8	$p = 0.734$	257	0.530	0.70	16	good	
	white	58.9	$p = 1.00$	186	0.426	0.51	6	good	
Nymphalid	COI and COII	76.3	$p = 0.999$	288	0.508	0.17	5	nc	Brower and DeSalle
butterflies	Wingless	44.9	$p = 0.999$	133	0.536	0.01	49	good	(1998)
Nymphalid	COI	69.0	$p = 1.00$	318	0.241	0.41	19	good	Monteiro and Pierce
butterflies	COII	76.0	$p = 1.00$	288	0.267	0.47	8	good	(2001)
(*Bicyclus*)	EF1a	49.0	$p = 1.00$	169	0.352	0.37	8050	good	
Swallowtail	COI and COII	74.0	$p = 0.986$	632	0.357	0.09	2	nc	Caterino et al.
butterflies	Ef1a	48.4	$\mathbf{p < 0.002}$	242	0.472	0.60	5	nc	(2001)
Swallowtail	COI and COII	73.5	$p = 0.651$	551	0.433	0.15	3	bad	Reed and Sperling
butterflies	EF1a	47.1	$p = 1.00$	160	0.576	0.34	40	good	(1999)
(*Papilio*)									
Halictid bees	COI	74	$p = 0.999$	453	0.201	0.24	1	bad	Danforth et al.
	EF1a	54.6	$p = 1.00$	274	0.47	0.37	53	good	(2003)
	Opsin	51.7	$p = 1.00$	127	0.505	0.50	>1000	ok	
Carpenter bees	COI	77.9	$p = 0.999$	169	0.405	−0.02	4	bad	Leys et al. (2002)
(*Xylocopa*)	PEPCK	60	$p = 0.694$	244	0.639	0.22	3	good	

[a] Excluding uninformative sites.
[b] No comment on dataset quality.

Table 3
LR tests of SSR models

Datasets	SSR by gene	SSR by gene + codon	DF	LR	*p* Value	Reference
Lice (*Columbicola*)	−4733.174	−4126.264	4	1213.819	<0.01	Johnson et al. (2003)
Aphids (*Uroleucon*)	−11402.831	−10946.480	8	912.701	<0.01	Clark et al. (2000)
Treehoppers	−102258.232	−95613.159	6	13290.156	<0.01	Lin et al. (submitted).
Gall-inducing thrips	−8213.593	−7655.666	6	1115.856	<0.01	Morris et al. (2001)
Bark beetles (*Ips*)	−17718.554	−15613.406	4	4210.297	<0.01	Cognato and Vogler (2001)
Stalk-eyed flies	−21288.052	−19366.906	8	3842.293	<0.01	Baker et al. (2001)
Nymphalid butterflies	−10562.678	−9885.644	6	1354.068	<0.01	Brower and DeSalle (1998)
Nymphalid butterflies (*Bicyclus*)	−27744.027	−25006.930	6	5474.195	<0.01	Monteiro and Pierce (2001)
Swallowtail butterflies	−38820.386	−35091.195	6	7458.383	<0.01	Caterino et al. (2001)
Swallowtail butterflies (*Papilio*)	−19260.036	−17512.347	6	3495.375	<0.01	Reed and Sperling (1999)
Halictid bees	−31178.490	−28676.587	6	5003.804	<0.01	Danforth et al. (2003)
Carpenter bees (*Xylocopa*)	−11183.598	−10731.714	4	903.768	<0.01	Leys et al. (2002)

Table 4
Summary of model parameters by dataset

Datasets	Data partitions	Total sites	A + T (%)	CI[a]	PBS/min steps	Pi	Alpha	Tree length	Relative rate	Reference
Lice	COI	384	62.4	0.403	0.13	0.398	0.331	26.58	1.63	Johnson et al. (2003)
(*Columbicola*)	EF1a	348	51.4	0.608	0.18	0.423	0.818	0.697	0.305	
Aphids	COI	799	77.2	0.464	−0.01	0.625	2.296	4.242	0.921	Clark et al. (2000)
(*Uroleucon*)	COII	596	79.8	0.482	−0.06	0.6	0.828	2.796	1.014	
	ND1	559	83.7	0.476	0.22	0.478	0.728	4.574	1.245	
	EF1a (exon)	877	55	0.471	0.02	0.454	1.612	0.288	0.558	
	EF1a (intron)	241	81.9	0.724	0.35	0.111	4.371	1.371	2.27	
Treehoppers	COI	1236	70.2	0.113	0.59	0.216	0.343	37.86	1.077	Lin et al. (submitted)
	COII	517	75.8	0.130	−0.04	0.027	0.301	40.53	1.139	
	Wingless	373	39.5	0.194	1.36	0.272	0.571	17.06	0.555	
Gall-inducing	COI	550	72.6	0.412	0.14	0.429	0.436	10.33	1.764	Morris et al. (2001)
thrips	EF1a (exon)	422	51.4	0.585	0.30	0.54	0.973	0.374	0.294	
	EF1a (intron)	100	68.9	0.561	0.19	0.106	2.399	2.06	1.527	
	Wingless	445	45.6	0.660	0.13	0.273	1.572	7.22	0.596	
Bark beetles (*Ips*)	COI	769	67.4	0.213	0.3	0.417	0.273	50.65	1.436	Cognato and Vogler
	EF1a (exon)	684	55.7	0.446	0.1	0.493	0.63	14.98	0.346	(2001)
	EF1a (intron)	83	73.7	0.580	0.32	0.199	4.233	14.49	1.362	
Stalk-eyed flies	COII	436	72.6	0.292	0.18	0.408	0.236	100.34	1.504	Baker et al. (2001)
	EF1a	1031	53.6	0.401	0.45	0.635	1.881	1.075	0.593	
	Wingless	619	58.8	0.530	0.70	0.244	1.067	1.913	1.126	
	white	486	58.9	0.426	0.51	0.54	2.281	2.556	1.251	
Nymphalid	COI	310	77.5	0.396	0.15	0.191	0.393	6.414	1.059	Brower and DeSalle
butterflies	COII	669	76	0.379	0.18	0.338	0.333	10.766	0.915	(1998)
	Wingless	378	44.9	0.445	0.01	0.188	0.618	2.804	1.101	
Nymphalid	COI	1256	69.0	0.241	0.41	0.496	0.652	7.141	1.27	Monteiro and Pierce
butterflies	COII	720	76.0	0.267	0.47	0.417	0.358	18	1.14	(2001)
(*Bicyclus*)	EF1a	941	49.0	0.352	0.37	0.594	0.776	1.12	0.542	
Swallowtail	COI	1530	72.3	0.237	0.09	0.238	0.227	8.118	1.017	Caterino et al. (2001)
butterflies	COII	684	77.4	0.256	0.07	0.467	0.448	20.68	1.062	
	Ef1a	995	48.4	0.268	0.60	0.609	1.096	2.662	0.931	
Swallowtail	COI	1532	71.9	0.370	0.11	0.534	0.605	2.57	1.181	Reed and Sperling
butterflies	COII	687	77.3	0.369	0.24	0.51	0.638	3.911	1.155	(1999)
(*Papilio*)	EF1a	1010	47.1	0.485	0.34	0.613	1.236	0.614	0.62	
Halictid bees	COI	1239	74	0.201	0.24	0.454	0.459	15.47	1.956	Danforth et al. (2003)
	EF1a (exon)	801	51.1	0.435	0.49	0.604	0.852	0.529	0.311	
	EF1a (intron)	448	63.4	0.526	0.27	0.314	2.42	1.296	0.732	
	Opsin (exon)	489	50.9	0.507	0.39	0.487	1.182	0.639	0.398	
	Opsin (intron)	169	56.2	0.609	0.58	0.087	3.93	0.985	0.652	
Carpenter bees	COI	600	77.9	0.327	−0.02	0.367	0.287	22.9	1.236	Leys et al. (2002)
(*Xylocopa*)	PEPCK (exon)	478	51.6	0.483	−0.01	0.344	0.936	0.908	0.562	
	PEPCK (intron)	573	68.8	0.576	0.35	0.073	4.86	1.69	1.118	

[a] Excluding uninformative sites.

evolve at up to 8 times faster than nuclear third positions. Introns evolve at approximately the same rate (aphids, halictid bees, and carpenter bees) or slightly faster (bark beetles) than nuclear third position sites.

There were no consistent differences in rate when comparing among mitochondrial genes: COI, COII, and ND1 seem to evolve at virtually the same rate when combined in the same analysis. In two cases, COII showed slightly (but significantly) lower rates of third-position substitution than COI (*Papilio* and *Bicyclus*), suggesting that COII might be a slightly better choice for recovering deeper divergences.

An examination of the 95% credibility intervals of the rates among sites indicates that, in general, the site-specific rates models show little variance around the overall rate estimated for each codon position (Figs. 2 and 3). In other words, while there is obviously rate variation within codon positions, the site-specific rates models explain much of the overall variance in rates among sites.

3.4. Transformation rate matrices (Q matrix)

One of the most obvious patterns to emerge from these comparisons is that the instantaneous rate matrix for mitochondrial genes is highly asymmetrical relative to that for nuclear genes (Figs. 4 and 5). For example, for the halictid bees (Danforth et al., 2003; Fig. 4) the

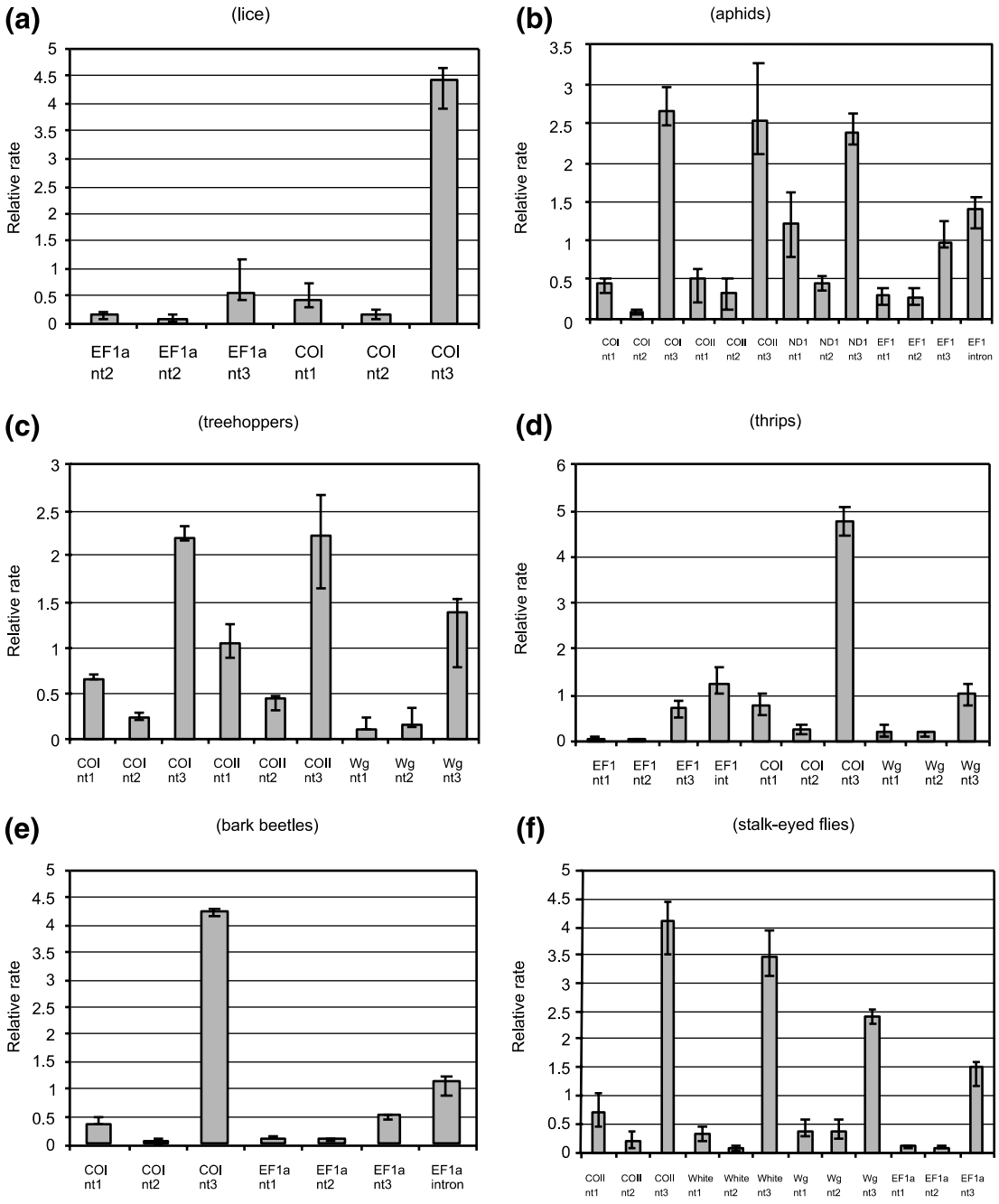

Fig. 2. Relative rates among data partitions based on GTR + SSR model with sites partitioned by gene and by codon position. Error bars indicate 95% credibility intervals: (a) Johnson et al. (2003), (b) Clark et al. (2000), (c) Lin et al. (submitted), (d) Morris et al. (2001), (e) Cognato and Vogler (2001), and (f) Baker et al. (2001).

mitochondrial gene (COI) shows a strikingly high rate of TC transitions relative to any other transformation. TC transitions occur 28 times faster than the next most frequent transformation (AT transversions), and 30 times faster than GA transitions (Fig. 4; note that the scale bars vary among graphs). In contrast, for coding or non-coding regions of the nuclear genes the instantaneous rate matrix is more symmetrical and also less skewed towards one type of change over another (Fig. 4). There is obviously an overall higher rate of GA and TC transitions, but this is only 2–10 times higher than the overall transversion rate. Furthermore, the rates of transitions are, overall, very close (with at most a 2-fold higher rate for one transition vs. the other) and the rates of transversions are, overall, very close (Fig. 4; the halictid dataset).

These patterns are evident in virtually all comparisons of the nuclear and mitochondrial genes we have made, but

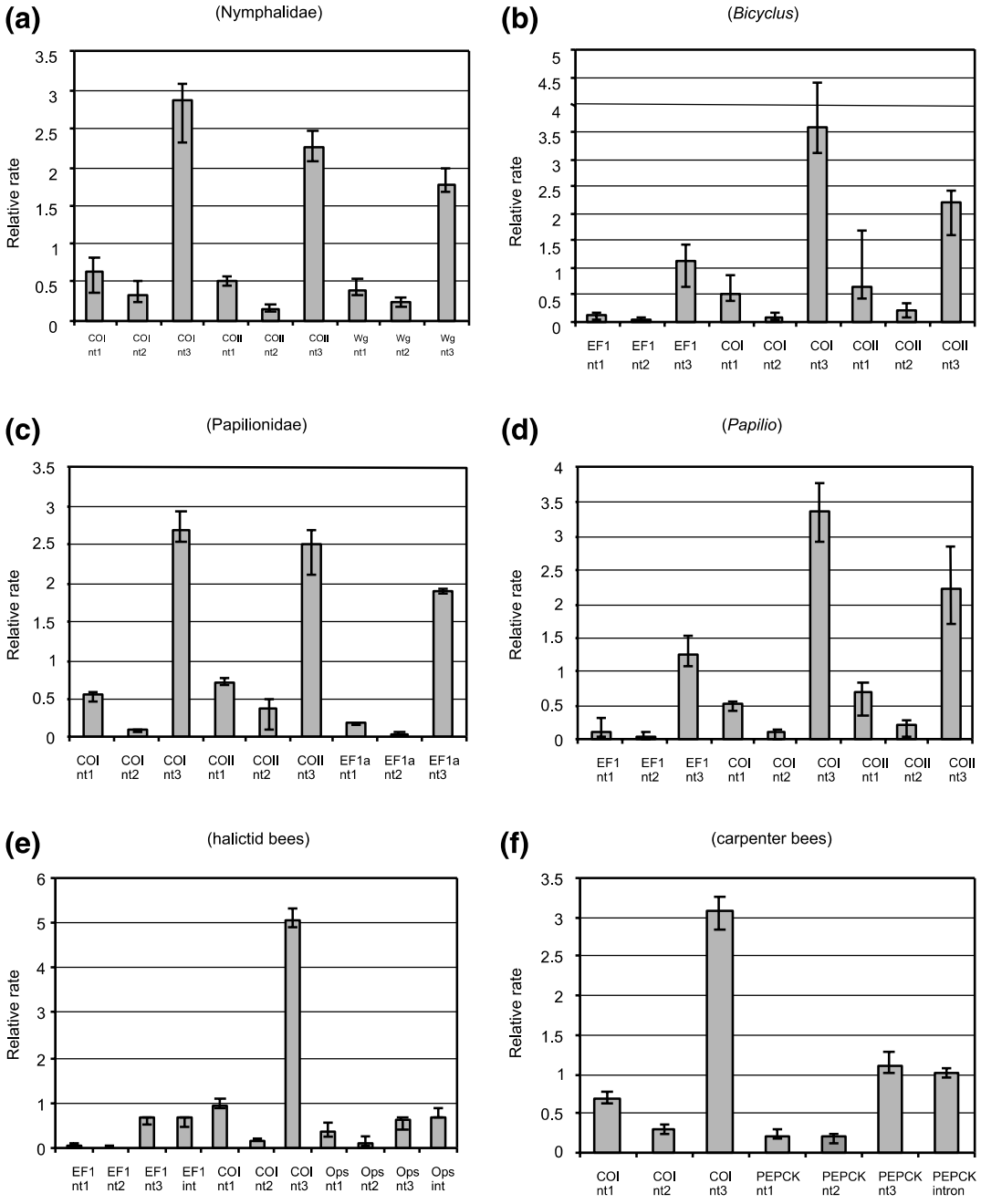

Fig. 3. Relative rates among data partitions based on GTR + SSR model with sites partitioned by gene and by codon position. Error bars indicate 95% credibility intervals: (a) Brower and DeSalle (1998), (b) Monteiro and Pierce (2001), (c) Caterino et al. (2001), (d) Reed and Sperling (1999), (e) Danforth et al. (2003), and (f) Leys et al. (2002).

are most striking in halictid bees (Fig. 4), all the lepidopteran datasets (Fig. 4; including Reed and Sperling, 1999), aphids (Clark et al., 2000), thrips (Fig. 5; Morris et al., 2001), and stalk-eyed flies (Fig. 5; Baker et al., 2001). In general, the skew in the mitochondrial transformation rate matrix is due to an excess of TC transitions, but for the aphid ND1 data set the skewed rate matrix is due to an excess of GA transitions. Nuclear gene introns show the least skewed transformation rate matrices (e.g., the halictid bee dataset; Fig. 4). The consequence of the highly skewed transformation rate matrix in mitochondrial genes is extraordinarily high levels of homoplasy that are not easily corrected for by a simple transition/transversion weighting scheme. Furthermore, highly skewed transformation rate matrices reduce the number of actual states that can occur at a nucleotide site from four (A, C, G, T) down to two (A, T), thus increasing homoplasy. Calculating a single transition/transversion ratio (as is common in molecular systematic studies) can obscure these patterns entirely.

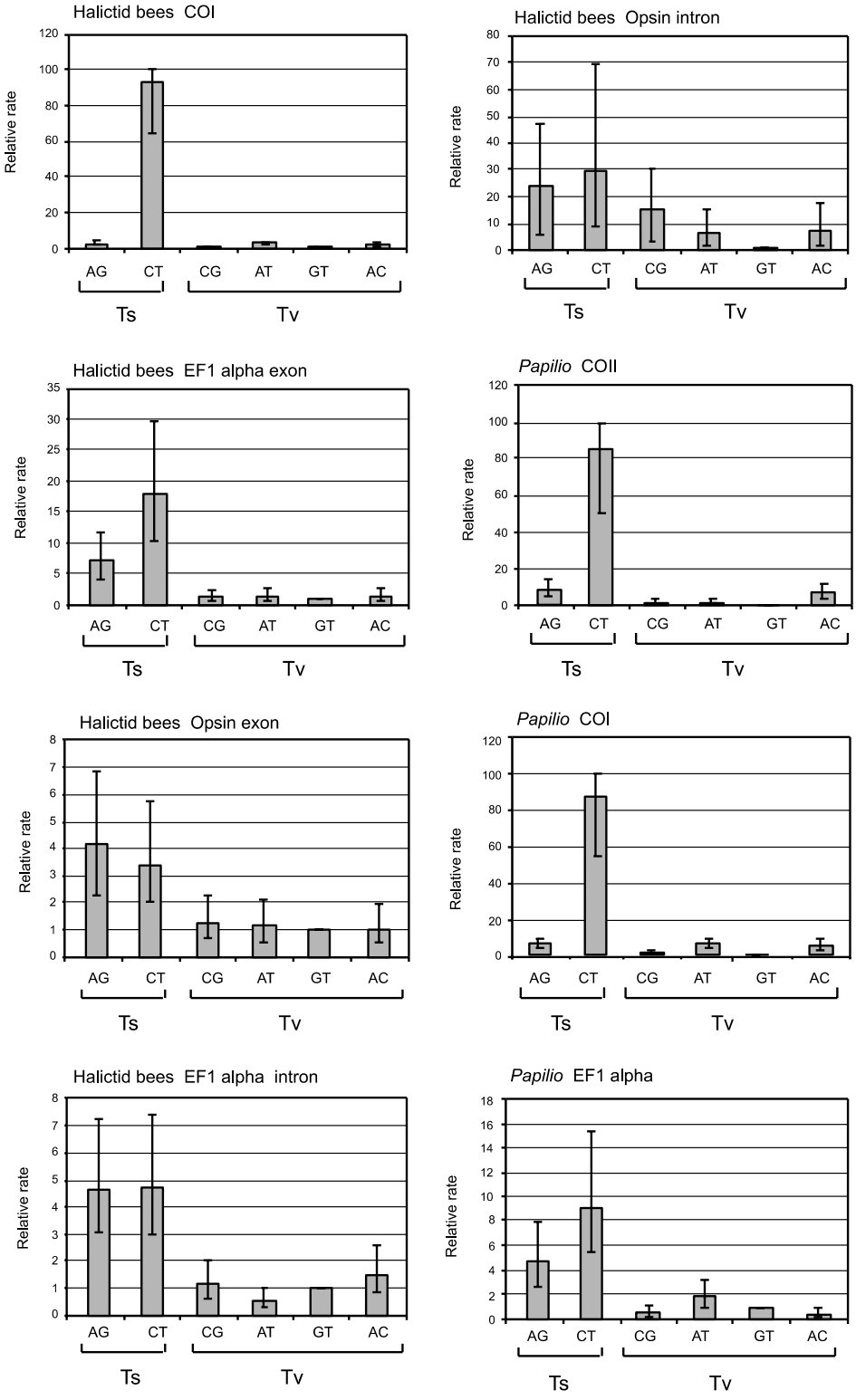

Fig. 4. Transformation rate matrices expressed graphically for different partitions of the data using the GTR + I + G model for halictid bees (Danforth et al., 2003) and *Papilio* butterflies (Reed and Sperling, 1999). Ts, transition; Tv, transversion.

The highly skewed rate matrices that characterize mitochondrial genes are evident in previous publications, but were generally not commented on by the authors. Monteiro and Pierce (2001) and Johnson and Whiting (2002), for example, show data on the instantaneous rate matrix for *Bicyclus* butterflies (Table 4) and lice in the suborder Ischnocera (Fig. 3, caption), respectively, which show these patterns, but did not note

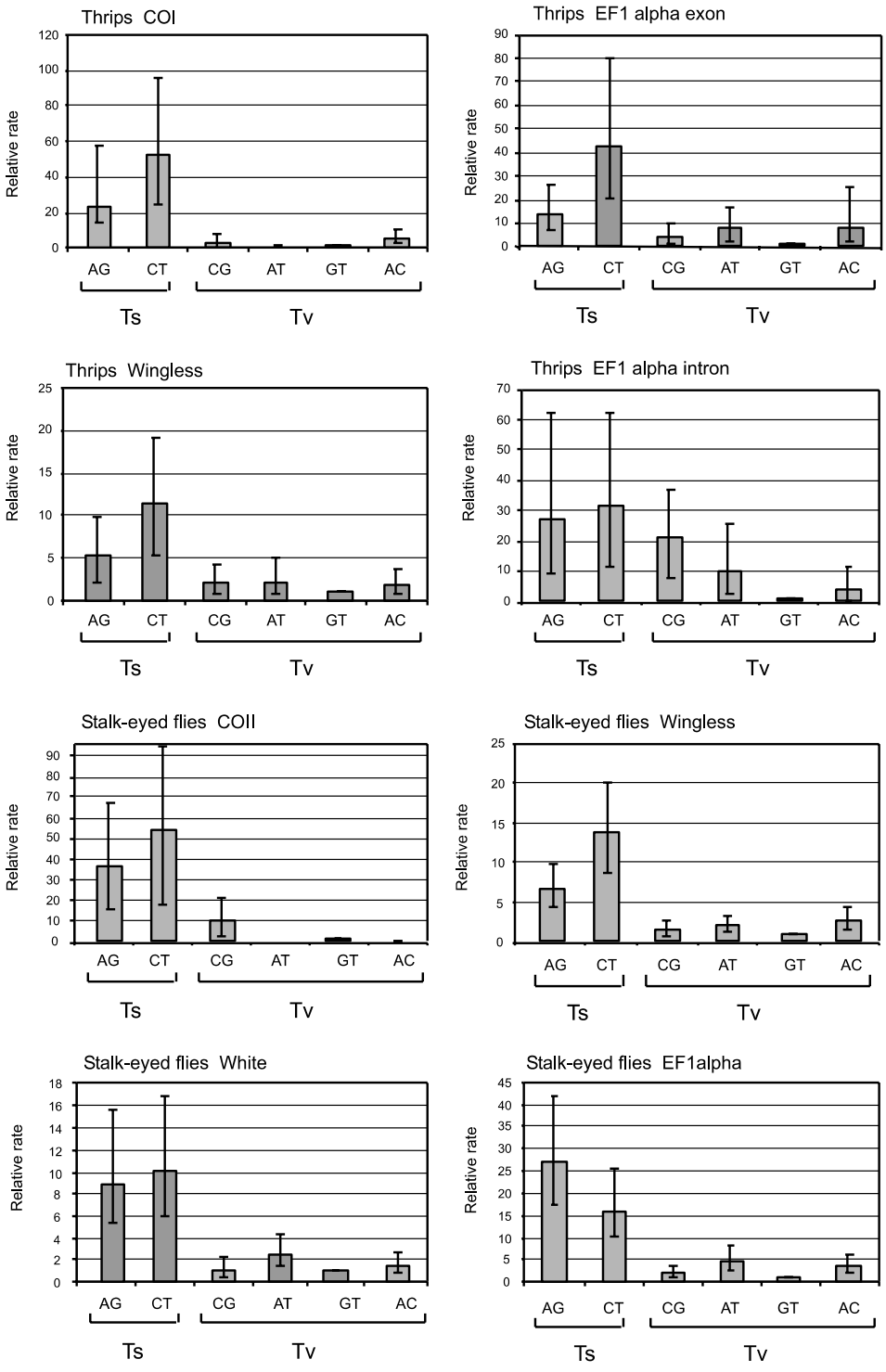

Fig. 5. Transformation rate matrices expressed graphically for different partitions of the data using the GTR + I + G model for gall-inducing thrips (Morris et al., 2001) and stalk-eyed flies (Baker et al., 2001). Ts, transition; Tv, transversion.

the striking differences between the nuclear and mitochondrial genes.

3.5. Shape of the gamma distribution (α)

Alpha (α), the shape of the gamma distribution describing among site rate variation, shows consistent differences among mitochondrial and nuclear genes. 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

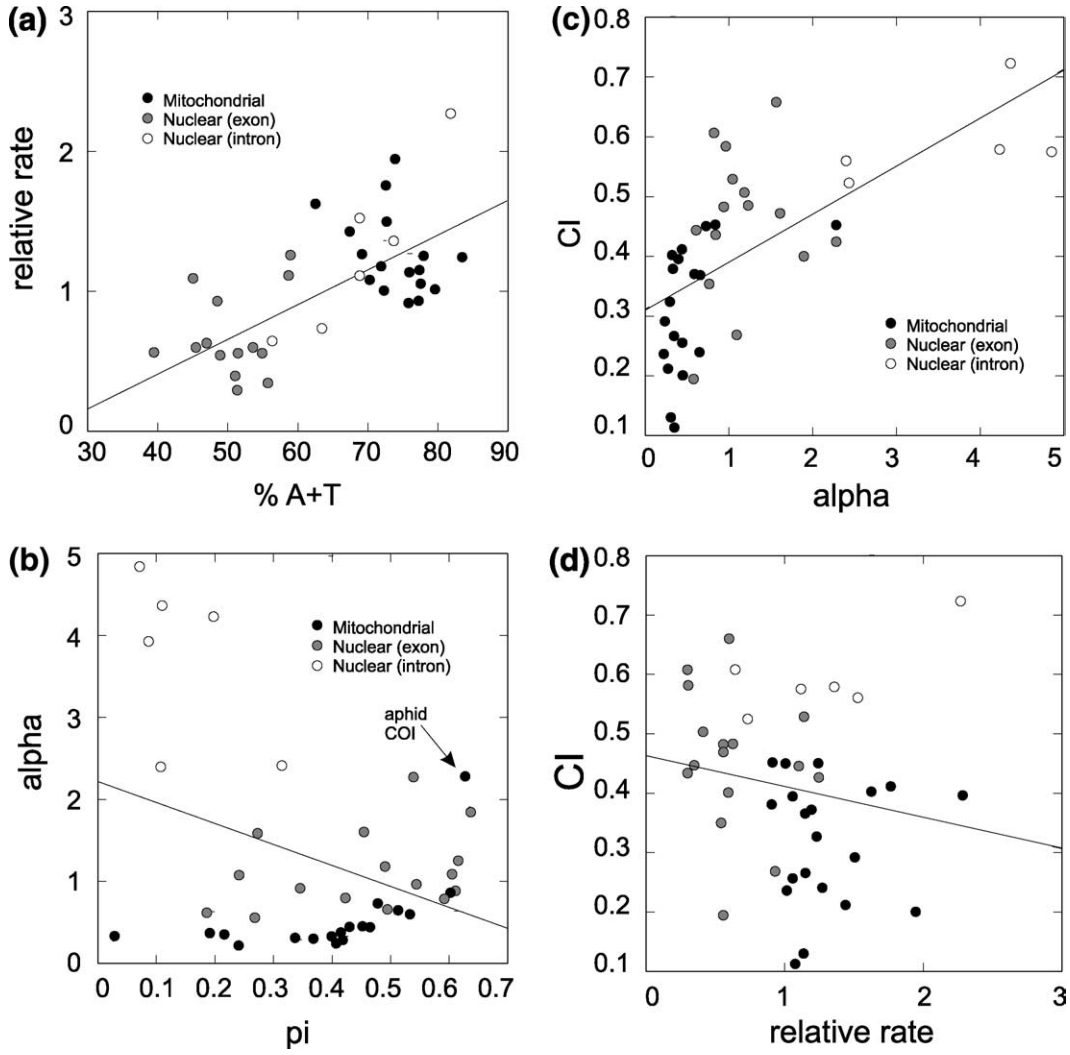

Fig. 6. (a) Relationship between rate of substitution and A/T bias. (b) Relationship between α (the shape of the gamma distribution) and pi (the proportion of invariant sites). (c) Relationship between CI (consistency index) and α. (d) Relationship between CI and relative rate. All regressions are significant except the relationship between CI and rate (see text).

or gene regions with a more even distribution of rates among sites.

For all studies analyzed, α is higher in coding regions of nuclear genes than it is in coding regions of the mitochondrial genes, indicating that nuclear genes show less rate heterogeneity among sites than mitochondrial genes. As might be expected, α for the non-coding regions was the highest for all datasets analyzed, presumably because all sites within introns are evolving at roughly the same rate. This may help explain why introns are such useful datasets in many phylogenetic studies involving nuclear genes (Danforth et al., 1999; Kawakita et al., 2003; Leys et al., 2002).

3.6. Exons vs. introns

Exons and introns of nuclear, protein-coding genes differed consistently in a number of ways. In all cases exons showed less skewed base compositional bias, lower overall rates of substitution, more heterogeneous patterns of among-site rate variation (as measured by α), more skewed transformation rate matrices, and (in 5 of 6 comparisons) lower values of CI. All of these differences are intuitive and expected given that exons are coding regions under selective constraints related to protein evolution. Nevertheless, the high values of α, the less skewed transformation rate matrices, and the (generally) higher values of CI that characterize introns indicate that introns, when they can be aligned, are capable of providing useful phylogenetic data with relatively low levels of homoplasy. The utility of introns has been emphasized recently by Kawakita et al. (2003).

3.7. Correlations among parameters

There was a significant positive correlation between base composition and relative rate for the 41 comparisons made in Table 4 ($r^2 = 0.667$; $p < 0.001$; Fig. 6a).

Gene regions with more A/T-biased base composition show, on average, higher rates of substitution. This pattern is largely explained by the difference in A/T-bias between nuclear and mitochondrial genes (Fig. 6a). Nuclear introns show a range of A/T bias as well as rates (Fig. 6a). This same pattern was observed by Jermiin and Crozier (1994).

Alpha (α), the shape parameter of the gamma distribution, showed a significant negative correlation with pi, the proportion of sites that are invariant ($r^2 = -0.365$; $p < 0.019$; Fig. 6b) when all partitions were analyzed. However, within coding regions of the mitochondrial and nuclear genes, there is evidently a positive correlation (Fig. 6b). This is likely a consequence of the fact that as more sites are allocated to the invariant sites category, the remaining sites will tend to show less rate heterogeneity.

CI, the consistency index, was positively correlated with alpha ($r^2 = 0.665$; $p < 0.0001$; Fig. 6c), suggesting that data partitions with lower values of among-site rate variation show less homoplasy. The aphid COI dataset was a significant outlier in comparison to other mitochondrial datasets in terms of alpha (Fig. 6c), indicating that not all mitochondrial genes show highly heterogeneous patterns of rate variation. However, one should be cautious about interpreting values of CI across datasets, since CI has been shown to be correlated with the number of taxa (Sanderson and Donoghue, 1989). CI showed a slight negative association with relative rate (Fig. 6d), but the correlation was not significant ($r^2 = -0.164$; $p < 0.305$). Partitioned Bremer support, another measure of dataset quality, showed no significant correlation with either relative rate ($r^2 = -0.142$; $p < 0.375$) or CI ($r^2 = -0.126$; $p < 0.434$).

4. Discussion

Our results indicate that nuclear genes have a slight advantage over mitochondrial genes in equal weights parsimony analysis. Nuclear genes had universally higher values of CI as compared to mitochondrial genes, and generally (8 of 12 comparisons) provided more in the way of partitioned Bremer support than the mitochondrial genes. These results corroborate the view among insect molecular systematists that mitochondrial genes show higher levels of homoplasy and are often of less utility, certainly at higher levels, than nuclear genes. However, our study also indicates that not all mitochondrial genes are the same. ND1, for example, contributed more in terms of partitioned Bremer support than EF-1α and far more than COI/COII in the aphid dataset (Clark et al., 2000). This pattern was also found by Baker and DeSalle (1997) in a study of Hawaiian drosophilids: ND1 performed far better than COI/COII in a combined analysis of nuclear and mitochondrial genes. Furthermore, in some datasets mitochondrial genes contributed more to Bremer support than the nuclear genes (e.g., aphids, bark beetles, and nymphalid butterflies; Table 2).

It is clear from our comparisons that there is much more heterogeneity in among-site rate variation (as indicated by the lower values of α) in mitochondrial genes than in nuclear genes. This, in part, accounts for the poor performance of mitochondrial genes relative to nuclear genes. With a few sites evolving at a very high rate, those sites will tend to saturate more quickly, leading to higher levels of homoplasy in mitochondrial datasets. That high values of α correspond with higher quality data is supported also by the positive correlation between α and CI (Fig. 6c). Yang (1998) came to a similar conclusion based on analyses of simulated datasets: high values of α (i.e., little among-site rate variation) yielded better performance than low values of α (based on the proportion of correct nodes recovered). The fact that simulation studies (Yang, 1998) and empirical studies (this study) come to the same conclusion suggests that α is an important predictor of dataset quality that few molecular systematists examine in detail. While for most of the data sets mitochondrial genes evolved faster than nuclear genes when compared on a codon-by-codon basis, the observation that mitochondrial genes evolve faster than nuclear genes is not universally true. We show above one example in which two mitochondrial genes (COI and COII) evolve more slowly than the nuclear gene (*wingless*; Brower and DeSalle, 1998).

One of the most interesting patterns to emerge from this analysis is that mitochondrial genes universally show highly asymmetrical patterns of among base substitution rates. In other words, the instantaneous rate matrix (Q) shows that mitochondrial genes (relative to nuclear genes) have a highly skewed distribution of transformation rates and these transformation rates do not necessary coincide with a simple transition/transversion bias. This may explain why the high levels of homoplasy in mitochondrial genes are so refractory to simple methods of a priori or a posterior weighting by codon position, or simple transition:transversion weighting. None of these methods can adequately "correct for" the biased transformation rate matrix. Only Cunningham's 6-parameter weighting method (Cunningham, 1997) would come close to accounting for the highly skewed rate matrix in mitochondrial genes.

The combination of low values of α and highly skewed transformation rate matrices may together explain the (overall) poor performance of mitochondrial genes relative to nuclear genes. Both properties of the nucleotide substitution process should lead to high levels of homoplasy, because both properties tend to limit the number of variable and/or alternative character states available. While a posteriori or a priori weighting (in

parsimony) or complex models that account for these biased substitution patterns (in maximum likelihood and Bayesian methods) may partially alleviate these problems, they cannot turn low-quality data into high-quality data. In our experience, choice of genes has a far greater impact on the phylogenetic results than choice of analytical method.

The Bayesian framework we have adopted in this study provides important insights into nucleotide substitution patterns and how they relate to phylogenetic utility of genes. Our methods could also be used to identify promising or detrimental attributes of genes prior to actually collecting a complete dataset (i.e., in the earliest stages of data evaluation). The properties that seem to characterize desirable genes and gene regions include more even base composition, higher values of α (i.e., less rate heterogeneity among sites; also see Yang, 1998), and a less skewed transformation rate matrix. It remains to be seen if these criteria can be put to practical use in choosing among genes or gene regions in the earliest stages of data collection. However, our results indicate that insect molecular systematists would be better off focusing their efforts on nuclear rather than mitochondrial genes (except in the case of very closely related taxa). Insect molecular systematists should also choose their datasets carefully, rather than relying on complex weighting schemes and highly parameterized models to correct for biased and/or skewed substitution patterns after the fact.

Acknowledgments

We are grateful to the authors who provided us with their mitochondrial and nuclear gene datasets. The following people commented on early drafts of this paper: Sean Brady, Karl Magnacca, John Ascher, Sarah Solomon, Curtis Ewing, Eduardo Almeida, Richard Harrison, Jeff Doyle, and Sedonia Sipes. Ignacio Cognato and one anonymous reviewer provided excellent reviews of the paper. This project was supported by National Science Foundation Research Grants in Systematic Biology (DEB-9815236 and DEB-0211701) to BND and a Doctoral Dissertation Improvement Grant to BND and C-PL (DEB-0104893).

References

Ascher, J.A., Danforth, B.N., Ji, S., 2001. Phylogenetic utility of the major opsin in bees (Hymenoptera: Apoidea): a reassessment. Mol. Phylogenet. Evol. 19, 76–93.

Avise, J.C., 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. 18, 489–522.

Avise, J.C., 1994. Molecular Markers, Natural History, and Evolution. Chapman & Hall, New York. 511pp.

Avise, J.C., 2000. Phylogeography: The History and Formation of Species. Harvard Univesity Press, Cambridge. 384pp.

Baker, R.H., DeSalle, R., 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst. Biol. 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). Syst. Biol. 50, 87–105.

Brady, S.G., 2002. Phylogenetics of army ants (Hymenoptera: Formicidae) based on morphological and molecular data. Ph.D. Dissertation, University of California, Davis.

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. Proc. R. Soc. B 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. Ann. Entomol. Soc. Am. 87 (6), 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 Mol. Biol. 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. Proc. R. Soc. Lond. B 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. Syst. Biol. 51 (1), 4–18.

Cameron, S.A., Mardulyn, P., 2001. Multiple molecular datasets suggest independent origins of highly eusocial behavior in bees (Hymenoptera: Apinae). Syst. Biol. 50 (2), 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). Mol. Biol. Evol. 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. Annu. Rev. Entomol. 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). Syst. Biol. 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. Mol. Biol. Evol. 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). Syst. Biol. 50 (6), 758–780.

Cunningham, C.W., 1997. Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. Syst. Biol. 46 (3), 464–478.

Danforth, B.N., 2002. Evolution of sociality in a primitively eusocial lineage of bees. Proc. Natl. Acad. Sci. USA 99 (1), 286–290.

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. Mol. Biol. Evol. 15 (3), 225–235.

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). Syst. Biol. 52 (1), 23–36.

Danforth, B.N., Sauquet, H., Packer, L., 1999. Phylogeny of the bee genus *Halictus* (Hymenoptera: Halictidae) based on parsimony and likelihood analyses of nuclear EF-1 sequence data. Mol. Phylogenet. Evol. 13 (3), 605–618.

DeSalle, R., Freeman, T., Prager, E.M., Wilson, A.C., 1987. Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*. J. Mol. Evol. 26, 157–164.

Fang, Q.Q., Cho, S., 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). Syst. Biol. 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). Mol. Phylogenet. Evol. 15, 473–486.

Farrell, B.D., Sequeira, A.S., O'Meara, B.C., Normark, B.B., Chung, J.H., Jordal, B.H., 2001. The evolution of agriculture in beetles (Curculionidae: Scolytinae and Platypodinae). Evolution 55 (10), 2011–2027.

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

Frati, F., Simon, C., Sullivan, J., Swofford, D.L., 1997. Evolution of mitochondrial cytochrome oxidase II gene in Collembola. J. Mol. Evol. 44, 145–158.

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). Mol. Phylogenet. Evol. 9, 131–140.

Friedlander, T.P., Regier, J.C., Mitter, C., 1992. Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. Syst. Biol. 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. Syst. Biol. 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). Mol. Biol. Evol. 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. Zool. J. Linn. Soc. 130, 213–234.

Harrison, R.G., 1989. Mitochondrial DNA as a genetic marker in population and evolutionary biology. Trends Ecol. Evol. 4, 6–11.

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). Biol. J. Linn. Soc. 72, 435–449.

Huelsenbeck, J.P., Bollback, J.P., 2001. Empirical and hierarchical Bayesian estimation of ancestral states. Syst. Biol. 50 (3), 351–366.

Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28, 437–466.

Huelsenbeck, J.P., Rannala, B., 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276, 227–232.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics (Oxford) 17 (8), 754–755.

Huelsenbeck, J.P., Rannala, B., Larget, B., 2000a. A Bayesian framework for the analysis of cospeciation. Evolution 54 (2), 352–364.

Huelsenbeck, J.P., Rannala, B., Masly, J.P., 2000b. Accommodating phylogenetic uncertainty in evolutionary studies. Science (Washington DC) 288 (5475), 2349–2350.

Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51 (5), 673–688.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Reverend Bayes meets Darwin: Bayesian inference in phylogeny and its impact on evolutionary biology. Science 294, 2310–2314.

Jermiin, L.S., Crozier, R.H., 1994. The cytochrome *b* region in the mitochondrial DNA of the ant *Tetraponera rufoniger*: sequence divergence in Hymenoptera may be associated with nucleotide content. J. Mol. Evol. 38, 282–294.

Johnson, K.P., Whiting, M.F., 2002. Multiple genes and the monophyly of Ischnocera (Insecta: Phthiraptera). Mol. Phylogenet. Evol. 22, 101–110.

Johnson, K.P., Crickshank, R.H., Adams, R.J., Smith, V.S., Page, R.D.M., Clayton, D.H., 2003. Dramatically elevated rate of mitochondrial substitution in lice (Insecta: Phthiraptera). Mol. Phylogenet. Evol. 26, 231–242.

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*). Mol. Biol. Evol. 20 (1), 87–92.

Kishino, H., Thorne, J.L., Bruno, W.J., 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol. Biol. Evol. 18, 352–361.

Kjer, K.M., Blahnik, R.J., Holzenthal, R.W., 2001. Phylogeny of Trichoptera (Caddisflies): characterization of signal and noice within multiple datasets. Syst. Biol. 50 (6), 781–816.

Leys, R., Cooper, S.J.B., Schwarz, M.P., 2000. Molecular phylogeny of the large carpenter bees, genus *Xylocopa* (Hymenoptera: Apidae), based on mitochondrial DNA Sequences. Mol. Phylogenet. Evol. 17 (3), 407–418.

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

Lin, C.-P., Danforth, B.N., Wood, T.K., submitted. Molecular phylogeny of the Membracinae: combined mitochondrial and nuclear DNA sequences resolve tribal and generic relationships. Syst. Biol.

Lutzoni, F., Pagel, M., Reeb, V., 2001. Major fungal lineages are derived from lichen symbiotic ancestors. Nature 411, 937–940.

Mardulyn, P., Cameron, S.A., 1999. The major opsin in bees (Insecta: Hymenoptera): a promising nuclear gene for higher level phylogenetics. Mol. Phylogenet. Evol. 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. Mol. Biol. Evol. 14, 381–390.

Monteiro, A., Pierce, N.E., 2001. Phylogeny of *Bicyclus* (Lepidoptera: Nymphalidae) inferred from COI, COII, and EF-1α gene sequences. Mol. Phylogenet. Evol. 18, 264–281.

Moriyama, E.N., Powell, J.R., 1997. Synonymous substitution rates in *Drosophila* mitochondrial versus nuclear genes. J. Mol. Evol. 45, 378–391.

Morris, D.C., Schwarz, M.P., Cooper, S.J.B., Mound, L.A., 2002. Phylogenetics of Australian *Acacia* thrips: the evolution of behaviour and ecology. Mol. Phylogenet. Evol. 25, 278–292.

Morris, D.C., Schwarz, M.P., Crespi, B.J., Cooper, S.J.B., 2001. Phylogenetics of gall-inducing thrips on Australian *Acacia*. Biol. J. Linn. Soc. 74 (1), 73–86.

Mooers, A.Ø., Holmes, E.C., 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15 (9), 356–365.

Moulton, J.K., 2000. Molecular sequence data resolves basal divergences within Simuliidae (Diptera). Syst. Entomol. 25, 95–113.

Nielson, R., 2002. Mapping mutations on phylogenies. Syst. Biol. 51 (5), 729–739.

Reed, R.D., Sperling, F.A., 1999. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus *Papilio*. Mol. Biol. Evol. 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. Mol. Biol. Evol. 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 Syst. Evol. 31, 179–186.

Sanderson, M.J., Donoghue, M.J., 1989. Patterns of variation in levels of homoplasy. Evolution 43, 1781–1795.

Schubert, M., Holland, L.Z., Holland, N.D., Jacobs, D.K., 2000. A phylogenetic tree of the *Wnt* genes based on all available full-length sequences, including five from the cephalochordate *Amphioxus*. Mol. Biol. Evol. 17, 1896–1903.

Simmons, R.B., Weller, S.J., 2001. Utility and evolution of cytochrome oxidase *b* in insects. Mol. Phylogenet. Evol. 20, 196–210.

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. Ann. Entomol. Soc. Am. 87, 651–701.

Sipes, S.D., Wolf, P.G., 2001. Phylogenetic relationships within *Diadasia*, a group of specialist bees. Mol. Phylogenet. Evol. 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*. Syst. Biol. 50, 39–59.

Sunnucks, P., Hales, H.F., 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I–II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). Mol. Biol. Evol. 13, 510–524.

Sullivan, J., Swofford, D.L., Naylor, G.J.P., 1999. The effects of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16, 1347–1356.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (and other methods), v. 4.0 β10. Sinauer Associates, Sunderland, MA.

Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.J., 1996. Phylogenentic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics, second ed. Sinauer, Sunderland, MA, pp. 407–514.

Tarrío, R., Rodríguez-Trelles, F., Ayala, F.J., 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. 18, 1464–1473.

Tatarenkov, A., Kwiatovski, J., Skarecky, D., Barrio, E., Ayala, F.J., 1999. On the evolution of Dopa decarboxylase (DDC) and *Drosophila* systematics. J. Mol. Evol. 48, 445–462.

Thorne, J.L., Kishino, H., 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51 (5), 689–702.

Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15, 1647–1657.

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. Syst. Biol. 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. Mol. Phylogenet. Evol. 15, 242–259.

Yang, Z., 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47 (1), 125–133.

Yang, Z.H., Yoder, A.D., 1999. Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. 48, 274–283.

Zhang, D.-X., Hewitt, G.M., 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol. Evol. 11, 247–251.