Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 32 (2004) 110–122
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Molecular phylogenetics of the allodapine bee genus *Braunsapis*: A–T bias and heterogeneous substitution parameters

Michael P. Schwarz,[a,*] Simon M. Tierney,[a] Steven J. B. Cooper,[b] and Nicholas J. Bull[a]

[a] *Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia*
[b] *Evolutionary Biology Unit, South Australian Museum, Adelaide, SA 5000, Australia*

Received 27 May 2003; revised 14 November 2003
Available online 22 January 2004

Abstract

Extreme AT bias in Hymenopteran mitochondrial genes have created difficulties for molecular phylogenetic analyses, especially for older divergences where multiple substitutions can erode signal. Heterogeneity in the evolutionary rates of different codon positions and different genes also appears to have been a major problem in resolving ancient divergences in allodapine bees. Here we examine the phylogeny of relatively recent divergences in the allodapine bee genus *Braunsapis*. We examined heterogeneity in nucleotide substitution parameters for one nuclear gene and codon positions in two mitochondrial genes, exploring various phylogenetic analyses for recovering relationships among species from Africa, Madagascar, southern Asia, and Australia. We explored maximum parsimony, maximum likelihood, Log determinant and Bayesian analyses. Broad topological features of best fit trees tended to be similar for equivalent data sets (e.g., total, or with 3rd mt positions excluded), regardless of the analytic method used (e.g., maximum likelihood or Bayesian). Analyses that used the total data set without modelling partitions separately gave unlikely results, indicating that the Malagasy species was most closely related to Australian species. However, analyses that excluded 3rd mitochondrial positions, or modelled partitions separately, suggested that the Malagasy species falls within the African clade. The unlikely topologies apparently result from long branch attraction, and this problem is ameliorated where modelling allows more realistic estimates of base composition and evolutionary rates for 3rd mitochondrial positions. However, we found that even when codon positions are modelled separately, estimated evolutionary rates for 3rd mitochondrial positions are likely to underestimate true rates. Long branch attraction and multiple substitutions are likely to be much more difficult to circumvent in analyses that explore older, generic-level, divergences in allodapine bees where overwriting is expected to be much more extreme. Our results indicate an African origin for *Braunsapis*, followed by a single, very early, dispersal event into Asia and then by a later dispersal event into Australia. The Malagasy species is derived from within the African clade.
© 2003 Elsevier Inc. All rights reserved.

Keywords: Allodapini; Bee evolution; Phylogenetics; Long branch attraction; Rate heterogeneity; AT bias

1. Introduction

Allodapine bees are uniquely suited to phylogenetic-based studies of social evolution because of the wide range in both sociality and life history patterns (Schwarz et al., 1998). Recent studies have begun analysing the phylogenetic relationships of allodapine bees based on DNA sequence data and these studies have provided some important insights into allodapine evolution that were not evident from earlier, morphology-based studies. In particular, they have shown that the most basal allodapine clade is the tropical African genus *Macrogalea*, and that the split between the Australian exoneurine group (*Exoneura*, *Inquilina*, *Exoneurella*, and *Brevineura*) and the non-*Macrogalea* African fauna constitutes an ancient divergence in the allodapines (Schwarz et al., 2003). *Macrogalea* has a tropical African and Malagasy distribution, the exoneurines are restricted to Australia, and with the exception of the rare and poorly known palaearctic genus *Exoneuridia*, the remaining genera are largely restricted to sub-Saharan Africa. The one exception to this continentally defined distribution is the genus *Braunsapis*. Phylogenetically,

* Corresponding author. Fax: +61-8-8201-3015.
E-mail address: michael.schwarz@flinders.edu.au (M.P. Schwarz).

1055-7903/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2003.11.017

Braunsapis is a relatively distal allodapine genus (Bull et al., 2003) but has a very wide distribution, ranging from Madagascar, throughout sub-Saharan Africa, across southern and south eastern Asia, the Indonesian archipelago, New Guinea and down into tropical, subtropical, arid, and semi-arid regions of Australia. It is the only allodapine genus to span the tropical and southern regions of the Old World.

Despite the advances that allodapine molecular phylogenetic studies have produced, relationships among some allodapine clades have either remained unresolved or are not well supported (Bull et al., 2003; Schwarz et al., 2003). A substantial part of this problem appears to derive from the extreme AT richness in hymenopteran 3rd mt codon positions, which appears to lead to saturation in transitions and degradation of useful signal for maximum parsimony (MP) analyses. Although maximum likelihood (ML) methods can account for multiple substitutions at fast evolving sites (Swofford et al., 1996), there remain problems in applying ML models. Different partitions may show quite different patterns of nucleotide substitution, but current software, with the exception of some recent Bayesian analytical approaches (Huelsenbeck and Ronquist, 2001; MRBAYES version 3b4, Simon and Larget, 2000) does not allow simultaneity in tree searching and the fitting of separate models to multiple partitions.

In this report we explore a variety of molecular phylogenetic analyses for investigating the phylogeny of *Braunsapis* using DNA sequences from one nuclear and two mitochondrial gene regions. We use species from Australia, southern Asia, Madagascar, and Africa.

2. Materials and methods

2.1. Taxa used

A recent molecular phylogenetic study (Schwarz et al., 2003) provides strong support for the monophyly of *Allodape* + *Braunsapis*. However, one *Braunsapis* species of particular interest, *B. trochanterata*, closely approaches *Allodape* in both adult and larval morphology (Michener, 1971, 1977). We therefore chose four species of *Allodapula* (see Table 1), which falls within the sister clade to *Braunsapis* + *Allodape* (Michener, 1977; Reyes, 1998), as the outgroup to examine the phylogenetic position of *B. trochanterata*. We included four species of *Allodape* and 16 species of *Braunsapis* in our study. *Braunsapis* taxa include one species from Madagascar, four from Australia, four from southern Asia, and five from Africa. Unfortunately, the taxonomy of *Braunsapis* is poorly developed and five species in our study could not be assigned to current nomens. Four species of *Braunsapis* have been described from Madagascar (Pauly et al., 2001) but our specimens cannot be firmly assigned to any of these and there are likely to be at least several species that are currently undescribed (Schwarz unpub.). We therefore refer to our Malagasy species as '*Braunsapis* Madagascar sp.' One of our Australian species (from the Pilbara in Western Australia), one species from Ko Chang Island in Thailand and another species from Thompson's Falls in Kenya also do not fit currently described species, and are referred to as '*Braunsapis* Pilbara sp.,' '*B.* KoChang sp.,' and '*B.* Kenya sp.,' respectively. Voucher specimens are deposited at the South Australian Museum. Species names and collecting localities are given in Table 1.

2.2. DNA extraction, amplification, and sequencing

Bees were collected from field-collected nests and transferred to 99% ethanol. For DNA extraction, bees were removed from ethanol and their thoraces were detached and pressed firmly between filter paper to remove excess ethanol. DNA was extracted from thoraces using the techniques described by Schwarz et al. (2003).

One nuclear and two mitochondrial gene regions were amplified and sequenced for this study. For all species we used bi-directional sequencing. For most species we sequenced two or more specimens, but for species where specimens were scarce we used multiple extractions from the one individual. The mitochondrial regions were from the protein coding genes cytochrome oxidase *b* (Cyt b; 429 bp sequenced) and cytochrome oxidase I (COI; 1279 bp), and the nuclear region was from the F1 copy of elongation factor-1α (EF-1α; 457 bp). The primers used for PCR amplification of the Cyt *b* region were designed by Y. C. Crozier (La Trobe University, Melbourne, Australia): cb1: 5′-TAT GTA CTA CCA TGA GGA CAA ATA TC-3′ and cb2: 5′-ATT ACA CCT CCT AAT TTA TTA GGA AT-3′. For the COI gene region, PCR amplification used two sets of primers. The first set utilised in previous allodapine studies (Bull et al., 2003; Schwarz et al., 2003) were designed by Lunt et al. (1996) and consists of 620bp at the 3′ end of *COI*: UEA7: 5′-TAC AGT TGG AAT AGA CGT TGA TAC-3′ and UEA10: 5′-TCC AAT GCA CTA ATC TGC CAT ATT A-3′. The second set: M414: 5′-CCT TTT ATA ATT GGA GGA TTT GG-3′ (LC01490; Folmer et al., 1994) and M399: 5′-TCA TCT AAA AAC TTT AAT TCC TG-3′ (designed from bee *COI* sequences by S. Cooper) allowed amplification of a contiguous 759 bp region of *COI* that is upstream and overlapping by 91 bp the region amplified using the Lunt et al. *COI* primers. The *COI* sequences from each region were combined, giving 1285 bp of total *COI* sequence, and were treated as a single entity in the analyses. Primers for the EF-1α F1 region (designed by Danforth and Ji, 1998) were: EF-1For2: 5′-AAG GAG GC[C/G] CAG GAG ATG GG-3′, EF-1Rev2: 5′-[T/C]TC [G/C]AC [T/C]TT CCA TCC GTA

Table 1
Locality origins of specimens used in this study

Genus	Species	Collecting locality
		Africa
Allodapula	*melanopus*	Wilderness, Western Cape, South Africa
Allodapula	*empeyi*	Cape St.Francis, Eastern Cape, South Africa
Allodapula	*variegata*	Wilderness, Western Cape, South Africa
Allodapula	*acutigera*	Kleinmond, Western Cape, South Africa
Allodape	*mucronata*	Kleinmond, Western Cape, South Africa
Allodape	*friesei*	Kleinmond, Western Cape, South Africa
Allodape	*punctuata*	Cape St.Francis, Eastern Cape, South Africa
Allodape	*exoloma*	Cape St.Francis, Eastern Cape, South Africa
Braunsapis	*trochanterata*	Matemwe, Zanzibar Is., Tanzania
Braunsapis	Kenya sp.	Thompson's Falls, Kenya
Braunsapis	*boussouyi*	Soutpansberg Ranges, Northern Province, South Africa
Braunsapis	*albipennis*	Soutpansberg Ranges, Northern Province, South Africa
Braunsapis	*vitrea*	Soutpansberg Range, Northern Province, South Africa
Braunsapis	*paradoxa*	Cape St. Francis, Eastern Cape, South Africa
		Madagascar
Braunsapis	Madagascar sp.	Antananarivo
		India
Braunsapis	*indica*	Nilgiri Hills, Tamil Nadu
Braunsapis	*pictarsis*	Nilgiri Hills, Tamil Nadu
		Thailand
Braunsapis	*mixta*	Kanchanaburi, Kanchanaburi Province
Braunsapis	*philippinensis*	Kanchanaburi, Kanchanaburi Province
Braunsapis	KoChang sp.	Ao Khlong Prao, Ko Chang Is.
		Australia
Braunsapis	Pilbara sp.	Pilbara, Western Australia
Braunsapis	*falcata*	Lake Gilles, South Australia
Braunsapis	*unicolor*	Lake Gilles, South Australia
Braunsapis	*protuberans*	Great Sandy National Park, Queensland

CC-3′. PCR amplification and sequencing procedures are described in Bull et al. (2003)

Forward and reverse sequences were compared for each gene fragment, and sequences were manually edited and aligned using SeqEd 1.03 (Applied Biosystems).

2.3. Base composition, saturation, and substitution rates

Like many Hymenoptera (Danforth et al., 2003; Jermiin and Crozier, 1994; Leys et al., 2000), allodapines show strong AT bias for mitochondrial genes, especially for 3rd codon positions. This can exacerbate saturation problems for older divergences, or where analyses cover a wide range in hierarchical levels (Whitfield and Cameron, 1998), and some recent allodapine studies (Bull et al., 2003; Reyes et al., 1999; Schwarz et al., 2003) recommended excluding 3rd codon mitochondrial transitions. Heterogeneity in base composition among taxa can also provide problems for phylogenetic analyses if they do not reflect shared ancestry, leading to spurious attraction between taxa with similar compositional bias (Lockhart et al., 1994). We measured base composition for codon positions separately for each gene fragment and examined whether base composition for each partition varied among taxa using χ^2 tests implemented in PAUP*b4.10 (Swofford, 2002).

One of the greatest advantages of ML methods over MP is that they incorporate specific models for evolutionary change, including the likelihood of multiple substitutions at sites. However, applying an inappropriate model to a data partition can lead to spurious results (Arbogast et al., 2003; Bruno and Halpern, 1999). This can be potentially overcome by partitioning data into sets, each of which is likely to conform to a specific model, and carry out separate analyses on each partition. In practice though, this often entails significant problems: (i) most available software for ML analyses does not allow simultaneity of effective tree searching with partitioned model fitting; (ii) criteria for combining data from different model/partitions are problematic at best (Barker and Lutzoni, 2002; Cunningham, 1997; Dowton and Austin, 2002); and (iii) for many data sets, partitioning leads to a loss of power because of the reduced information content for partitions. Many studies have combined data to overcome problems (i) and (iii), and have either relied on combinability criteria that now appear to be problematic (Cunningham, 1997), or else have hoped that

heterogeneity in evolutionary dynamics will not lead to major problems.

Nucleotide composition data (see Results below) suggested differences between 1st, 2nd, and 3rd codon positions, with 3rd codons showing extreme A–T bias, but equivalence in composition between the *Cyt b* and *COI* gene fragments. Although we found differences in bias among the EF-1α codon positions, the number of parsimony informative changes overall was very small ($N = 51$). We therefore defined four partitions for our data—1st, 2nd, and 3rd positions for the mt genes combined, and a single partition for EF-1α. We used Bayesian analysis (MrBayes 3.0b4) to estimate base frequencies, transition matrices, proportion of invariant sites and Γ-shapes for each partition by unlinking estimates of these for each partition. Analyses were begun with prior probabilities randomly drawn from the default distributions of MrBayes 3.0b4. We used default priors because these are flat distributions and generally uninformative, a property we considered desirable given we did not wish to introduce specific beliefs about substitutional dynamics, since these dynamics were the subject of investigation. Analyses were run for 2 million generations, sampling every 500th generation, and log likelihoods were examined to determine when stationarity was reached. Three runs were used to check that alternate prior probabilities did not lead to substantially different parameters at stationarity. For all runs, stationarity was reached after about 10^5 generations and parameter estimates were based on the last 1000 trees (i.e., last half million generations).

2.4. Maximum parsimony analyses

Maximum parsimony analyses were carried out using PAUP*b4.10. All heuristic searches were carried out with 50 random addition sequences of taxa, keeping best trees only, holding 5 trees at each step, and with TBR swapping on all trees. The same searching protocol was used for bootstrap analyses, with support based on 500 bootstrap pseudoreplicates. The possible effects of saturation at 3rd mitochondrial positions was explored by using three separate analyses: (i) all codon positions and all substitutions equally weighted; (ii) transitions at 3rd mt codon positions weighted zero; and (iii) 3rd mt codon substitutions excluded entirely.

2.5. Log determinant analyses

In cases where taxa differ significantly in base pair composition, both MP and ML analyses may converge on spurious topologies if similar biases among some taxa do not reflect shared ancestry. Log determinant (log Det) analyses are relatively robust to variation in base composition (Lockhart et al., 1994) and provide a means for exploring whether topological features result from compositional biases. We therefore performed log Det analyses to examine whether inclusion of 3rd mt positions (where we found significant heterogeneity in base composition among taxa, see below) may have resulted in misleading topologies. We removed a portion of sites estimated to be invariant, based on a preliminary GTR + I ML analysis.

2.6. Maximum likelihood analyses

We explored two different approaches for ML analyses. First, we carried out a single model ML analysis applied to all sites combined (referred to as the 'total data set'). A single, randomly chosen, best tree from the MP analysis (using all sites with equal weighting) was used as a starting tree for a heuristic search with TBR swapping and estimation of the rate matrix, proportion of invariant sites and Γ-shape, assuming four rate categories. The resulting parameters were then used in a bootstrap analysis with 500 pseudoreplicates. Second, we omitted 3rd mt positions from the data set because this partition differed strongly from others in compositional bias and evolutionary rate, and estimated all parameters for a single model using the preceding procedure, again followed by bootstrap analysis with 500 pseudoreplicates.

2.7. Bayesian analyses

Bayesian analyses provide some substantial advantages over ML analyses with heuristic searches, particularly with respect to computational burden for large data sets (Huelsenbeck and Ronquist, 2001). Another major advantage is the ability of recent software (MrBayes version 3.04b, Huelsenbeck and Ronquist, 2001) to allow different models to be applied to separate partitions, while simultaneously searching for optimal trees. However, the recency of Bayesian analytic approaches to phylogenetic inference means that such analyses should be treated cautiously (Huelsenbeck et al., 2002) especially with respect to choosing priors and the interpretation of posterior probabilities (e.g., Cummings et al., 2003; Suzuki et al., 2002; Wilcox et al., 2002). Over-parameterization needs to be avoided if subsequent models are based on too few data (Huelsenbeck and Ronquist, 2001), and the choice of priors needs to be realistic, with assessment of whether outcomes are robust with respect to chosen priors. We grouped our data into the four partitions described above, because of expected differences in evolutionary rates, but similarity in nucleotide composition, among mitochondrial genes for each mt codon position, and because our nuclear gene fragment differed strongly in composition from the mt genes. Likelihood ratio tests have previously shown that the GTR + I + G model is the most suitable for analyses of *COI* and EF-1α data

from allodapine bees (Bull et al., 2003). We, therefore, used a GTR + I + G model for each partition, and partition rates, substitution matrices, nucleotide compositions, proportion of invariant sites, Γ-shapes were estimated separately for each partition by unlinking each of these estimates across partitions. We ran analyses following the methods described in *Base composition, saturation and substitution rates* above, for the same duration with three multiple runs, each with three cold and one hot chain, to assess robustness of default priors.

2.8. Concordance among partition models

A potentially major problem in fitting models of evolutionary change to fast evolving partitions is that once homoplasies reach a certain, saturated, level, further change may not be empirically discernible. Detecting such saturation is important because once it is reached, estimated distances may no longer reflect the degree of sequence divergence and true branch length (in the sense of actual substitutional events) will become longer than either measured or predicted branch length. Traditionally, saturation for transitions or transversions for various partitions has been explored by plotting uncorrected 'p' distances (which do not take multiple substitutions into account) against reversible model distances (e.g., Tamura-Nei) which can incorporate overwriting. However, implementation of distances like the Tamura-Nei index do not allow estimated Γ-shapes and proportions of invariant sites to be taken into account. For our analyses we reasoned that while models fitted separately to each partition may be expected to have different dynamics, we might nevertheless expect that each model should provide some measure of overall change since divergence from any node. In an ideal situation we might expect that separate estimates of divergence between any species-pairs, calculated for each partition, would be linearly related. However, if the degree of homoplasic change for a particular partition is underestimated, distances based on that partition should show asymptotic behaviour compared to distances where homoplasic change is better accounted for. In particular, where saturation occurs for a fast evolving partition, we might expect that distances for highly diverged taxa would not have Euclidean properties, and instead one might expect successively greater underestimation of distances as true divergence rates increase (Arbogast et al., 2003). This should lead to asymptotic relationships between distance measures based on partitions with very different evolutionary rates.

3. Results

3.1. Base composition bias and saturation

Chi-square tests for differences in base composition among taxa (Table 2) indicated highly significant variation for 3rd codon positions of both *COI* and *Cyt b* ($P < 0.00001$ for both gene regions) but no differences for all other gene/position partitions ($P > 0.9$ for all seven tests). Both mitochondrial gene fragments showed strong AT bias, especially for 3rd positions, but no apparent bias was observed for EF-1α (Table 2). Plots of uncorrected 'p' distances against TN distances (not shown here) suggested saturation problems for 3rd mt transitions but not for other partitions where scattergrams indicated largely linear plots for both transitions and transversions. This is very similar to findings by Reyes et al. (1999) and Schwarz et al. (2003).

Bayesian estimates for GTR substitution matrices, proportions of invariant sites, and Γ-shapes are given in Table 3. These data indicate very different evolutionary dynamics for each of the partitions, not surprising given the differences in nucleotide compositions and likely constraints for silent and non-silent substitutions for mitochondrial sites, and expected rate differences for nuclear and mitochondrial genes. In particular, Γ-shapes varied enormously among the partitions and the proportion of invariant sites for 3rd mt positions

Table 2
Base frequencies for codon positions of each gene and χ^2 tests of bias among taxa

Codon position	A	C	G	T	% A–T bias	χ^2	*P*
COI							
1st	0.3377	0.1195	0.1854	0.3575	69.52	19.01	1.0
2nd	0.2119	0.2028	0.1349	0.4503	66.22	6.93	1.0
3rd	0.4484	0.0243	0.0174	0.5100	95.84	273.40	<0.0001
Cyt *b*							
1st	0.3310	0.1336	0.1522	0.3832	71.42	11.71	1.0
2nd	0.2588	0.2067	0.0876	0.4470	70.58	4.73	1.0
3rd	0.4355	0.0341	0.0172	0.5132	94.87	171.49	<0.0001
EF-1α							
1st	0.3159	0.1411	0.3620	0.1811	49.70	1.39	1.0
2nd	0.3226	0.2427	0.1448	0.2899	61.25	0.23	1.0
3rd	0.1249	0.2827	0.3879	0.2045	32.94	20.72	1.0

Table 3
Substitution rate matrices, proportion of invariant sites, and γ-shapes (α), estimated using Bayesian analyses for 1nt, 2nt, and 3nt mitochondrial partitions and for EF-1α

Codon position	C	G	T	*p*(inv)	α
1st position mt					
A	2.38645	8.19672	6.37129	0.74901	17.75119
C		1.15956	27.00412		
G			1		
2nd position mt					
A	1.67814	6.55482	2.39422	0.84411	9.10456
C		0.87403	1.72418		
G			1		
3rd position mt					
A	0.08921	68.00393	0.54653	0.01289	0.32401
C		0.83733	46.40634		
G			1		
EF-1α					
A	0.21353	3.73207	0.51704	0.65991	12.37643
C		0.15040	5.14581		
G			1		
Total data set					
A	1.26503	14.64909	14.62201	0.50140	0.61559
C		1.49928	29.09948		
G			1		

Estimates are based on means of 1000 saved trees representing generations $1.5–2 \times 10^6$ in the MCMC analysis.

was extremely low (1.3%). Base substitutions also varied enormously among partitions, with extremely high $A \leftrightarrow G$ rates for 3rd mt positions and very high $C \leftrightarrow T$ rates for 1st and 3rd mt positions. Along with the differences among partitions in base composition, these figures strongly suggest that models applied to a single combined data set are unlikely to reflect actual substitutional dynamics.

3.2. Maximum parsimony

The unweighted MP analysis with all sites included and all substitutions equally weighted gave two equally most parsimonious trees which differed only in the position of *Braunsapis falcata* within the Australian group. These trees (not shown here) are represented as a strict consensus tree in Fig. 1A where bootstrap values are indicated for nodes with >50% support. Support for the monophyly of both *Allodape* and *Braunsapis* was very strong. *B. trochanterata* was recovered as the most basal of the *Braunsapis* species, though support for this was not high (59%). Support for the remaining African species forming a monophyletic group was moderately high (88%). Two topological features were unexpected. First, the Madagascan species, *B.* Madagascar sp., was placed as a sister species to the Western Australian species, *B.* Pilbara sp., though with low support (68%). Another anomalous result was that the undescribed species from Thailand, *B.* KoChang sp., was recovered as a sister species to the Australian species *B. unicolor*, with moderate support (77%). Support for monophyly of the clade containing the Australian group, the Thai species and *B.* Madagascar sp. was moderately high (89%).

The analysis where 3rd mt transitions were weighted zero gave two equally most parsimonious trees. A strict consensus of these trees differed from Fig. 1A only in that *B. falcata* was placed as the sister taxon to *B.* Madagascar sp. + *B.* Pilbara sp., and *Allodape punctuata*, *A. exoloma*, and *A. friesei* collapsed into a polytomy. Bootstrap support values (where >50%) were not substantially different from the combined data analysis (Fig. 1A), except that there was <50% support for a sister-group relationship between *B.* KoChang sp. and *B. unicolor* and support for a sister relationship between *B. indica* and *B. pictarsis* dropped to 66%.

The analysis where 3rd mt positions were excluded gave six equally most parsimonious trees differing in bifurcations among some *Allodape* species and in bifurcation order among some Asian and Australian species. These trees are summarised in Fig. 1B as a 50% majority rule tree with bootstrap support indicated for nodes with >50% bootstrap support. The major difference in topology between this analysis and the two previous analyses is that *B.* Madagascar sp. was now grouped with the African species (excepting *B. trochanterata*), which was still recovered as the most basal *Braunsapis* species, and the Thai species, *B.* KoChang sp., was placed within the Asian group, which now formed a sister clade to a monophyletic Australian

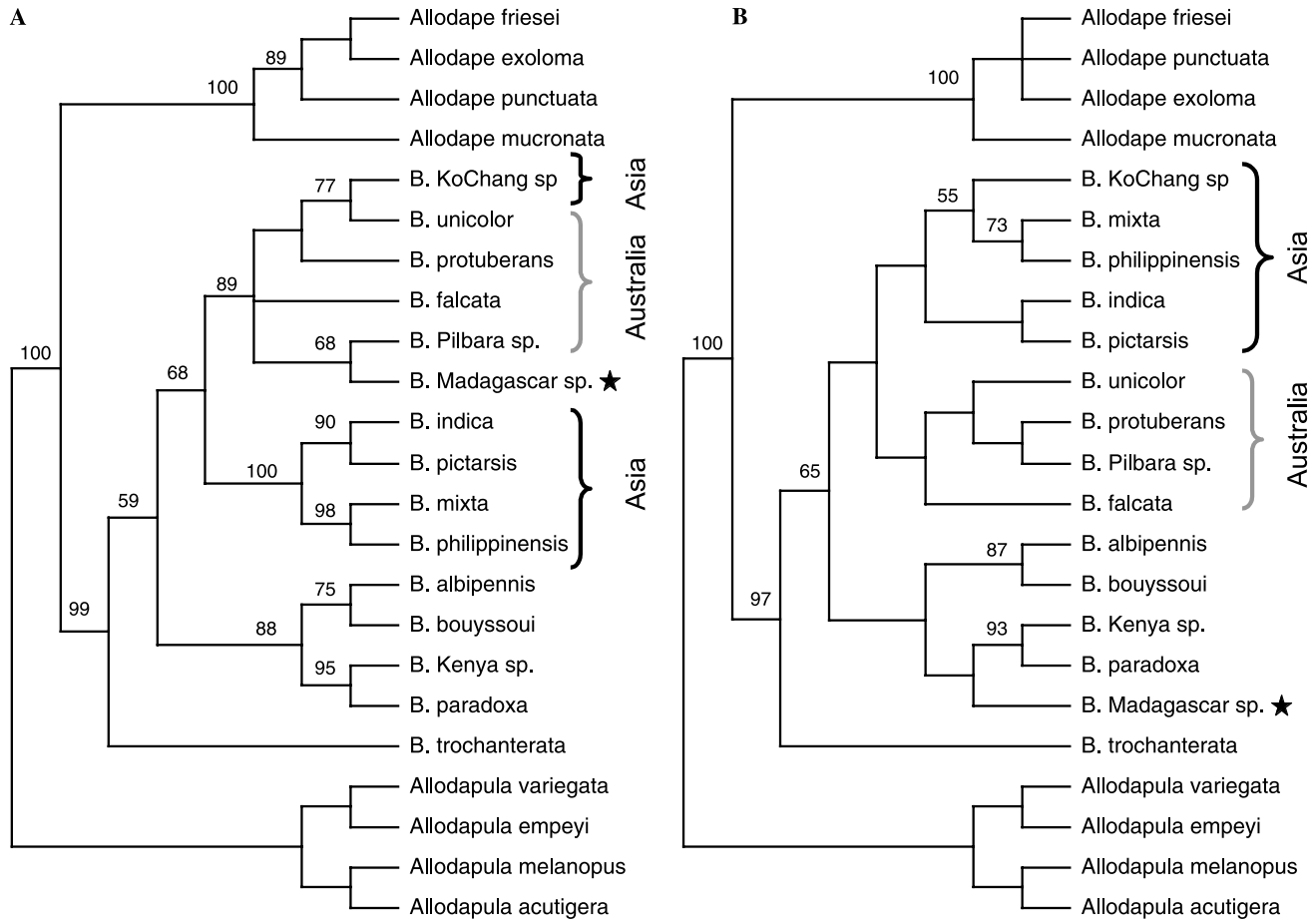

Fig. 1. (A) Strict consensus of two most parsimonious trees based on the total data set with equal weighting of all sites. Bootstrap values are based on 500 pseudoreplicates. Asian species are indicated by black brackets, Australian species by grey brackets, and the Madagascan species by a star. All other species are African. (B) Fifty percentage majority rule consensus of 6 equally most parsimonious trees based on a data set with 3rd mitochondrial codon positions excluded. Bootstrap values are based on 500 pseudoreplicates.

clade. However, most nodes in the analyses received very low support (<50%).

3.3. Log determinant analyses

The estimated proportion of invariable sites for the combined data analysis was 0.6468 and the heuristic search returned a single best tree that was virtually identical to the strict consensus of the combined data MP analysis (Fig. 1A), except that *B. falcata* was placed as sister group to *B. protuberans* and *B. trochanterata* formed a sister clade to the remaining African species, rather than being the most basal member of *Braunsapis*. The estimated proportion of invariable sites for the analysis with 3rd mitochondrial positions excluded was 0.7036. This tree was identical to the MP analysis where 3rd mitochondrial positions were excluded (Fig. 1B), except that the bifurcation order among the Australian species was changed to *B.* Pilbara sp. + (*B. falcata* + (*B. unicolor* + *B. protuberans*)). Bootstrap support for bifurcations among these taxa in the MP analysis was <50%.

The broad congruence between MP and log Det analyses for the two data sets firstly suggests that topologies from the MP analyses are unlikely to be the result of base composition bias among taxa, and secondly, that whatever factors lead to the placement of *B.* Madagascar sp. in either the Australian or African clades, and *B.* KoChang sp. in the Australian or Asian clades, are due to the inclusion/exclusion of 3rd mt positions which have a similar effect for both analytic methods.

3.4. Maximum likelihood

Estimated substitution matrices, proportions of invariant sites and Γ-shape parameters for the two ML analyses (total data set and the data set excluding 3rd mt positions) are given in Table 4. The best fit trees for each analysis are summarised as phylograms in Figs. 2A and B, respectively, where bootstrap values >50% are also indicated. Both analyses recovered *B. trochanterata* as a sister clade to the remaining African

Table 4
Substitution rate matrices, proportion of invariant sites, and γ-shapes, estimated under GTR + I + Γ models for (i) the total data set, and (ii) all genes combined but excluding mitochondrial 3rd codon positions

Codon position	C	G	T	*p*(inv)	γ-shape
Total data set					
A	1.18741	13.15698	13.25910	0.50762	0.60235
C.		1.25759	24.97010		
G			1		
Excluding mitochondrial nt3 sites					
A	1.10018	5.85012	3.80034	0.64482	0.48540
C		0.55059	8.67459		
G			1		

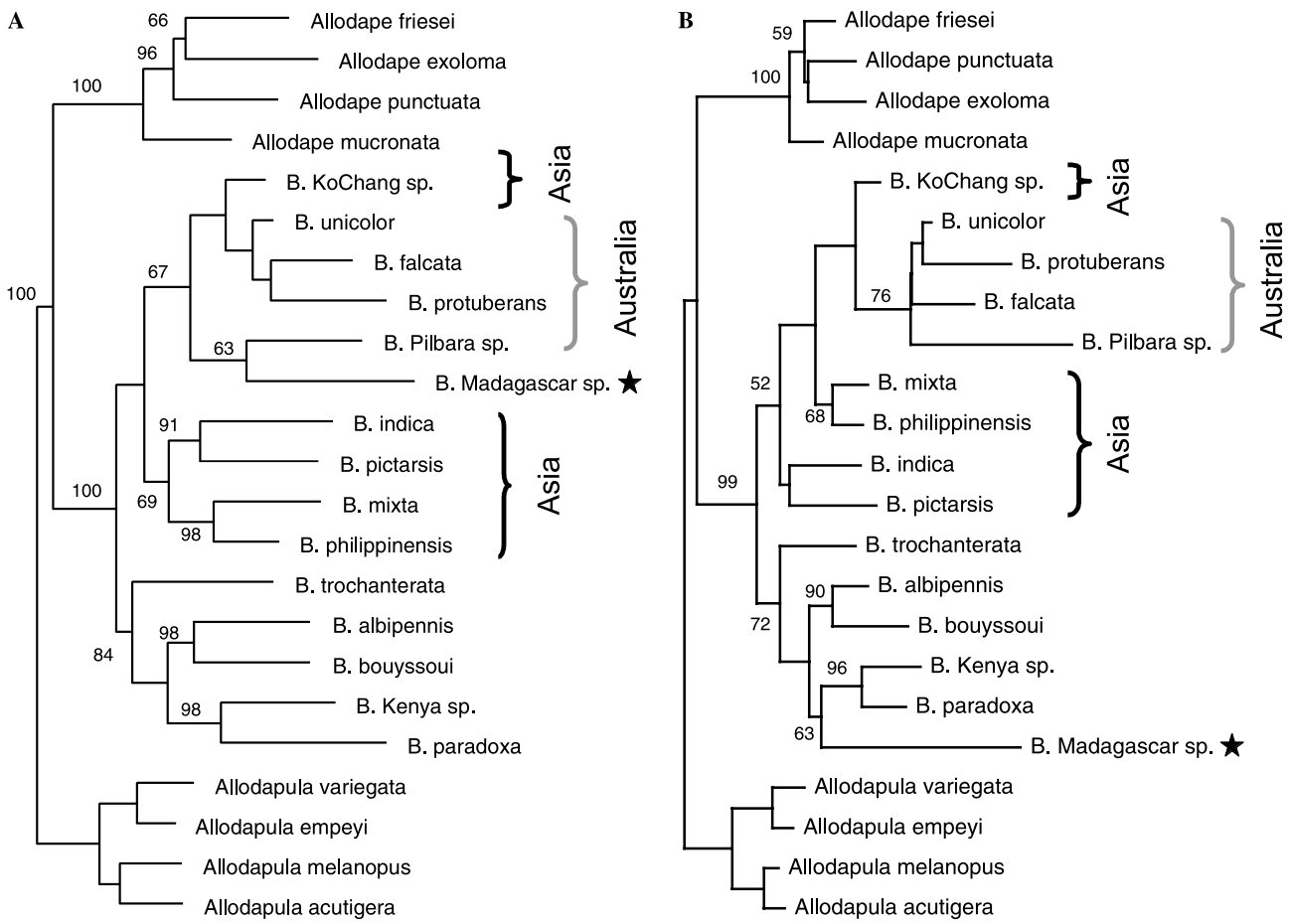

Fig. 2. (A) Best fit maximum likelihood tree based on a single GTR + I + Γ model applied to the total data set. Bootstrap values >50% are based on 200 pseudoreplicates. Asian species are indicated by black brackets, Australian species by grey brackets, and the Madagascan species by a star. All other species are African. (B) Best fit maximum likelihood tree base on a single GTR + I + Γ model applied to the data set excluding 3rd mitochondrial codon positions. Bootstrap values >50% are based on 200 pseudoreplicates.

taxa, rather than as sister group to the remaining *Braunsapis* as a whole. With this exception, the ML analyses were broadly congruent with the two MP analyses: the total data set placed *B.* KoChang sp. as sister species to *B. unicolor* + (*B. falcata* + *B. protuberans*) and *B.* Madagascar sp. as sister group to *B.* Pilbara sp., whereas the data set excluding 3rd mt positions led to *B.* Madagascar sp. being grouped with the African species and *B.* KoChang sp. being sister clade to a monophyletic Australian group. Bootstrap support tended to be higher for most nodes than under the corresponding MP analyses

3.5. Bayesian analyses

The three Bayesian runs all gave identical topologies with highly similar estimates for the different substitution parameters. Substitutional parameters from the first

run are given in Table 3. A phylogram summarising results from the last 0.5 million generations of the same run, with posterior probabilities and branch lengths representing mean values over the 1000 sampled trees from these generations (Fig. 3A). These analyses all placed *B.* Madagascar sp. within the African clade (excepting *B. trochanterata*) and monophyly of this clade was supported by a high posterior probability (100%). *B. trochanterata* formed a basal branch to this clade with a posterior probability of 90%, and the remaining *Braunsapis* species formed a monophyletic group with a high posterior probability (98%). The Australian species also formed a highly supported (99%) monophyletic group, with the Thai *B.* KoChang sp., falling into a highly supported (100%) sister group relationship to the Australian clade.

The recovery of *B.* Madagascar sp. as a member of the African clade in the partitioned Bayesian analysis (rather than as a member of the Australian group, as indicated by the combined data ML analysis), could be due to either partitioning per se or differences in the algorithms used by ML and Bayesian analyses. To assess these possibilities we carried out another Bayesian analysis, this time combining all partitions and fitting a single model to the combined data, thereby mirroring the procedure used in the combined ML analysis. We employed the same priors and run procedures as for the partitioned analyses and again carried out three separate runs. These analyses all produced congruent phylograms (Fig. 3B) which placed *B.* Madagascar sp. A as a sister group to *B.* Pilbara sp. with high support (100%) and with high support for monophyly of an Australian + *B.* Madagascar sp. + *B.* KoChang sp.clade (100%). This suggests that the unlikely relationship between the Australian species and *B.* Madagascar sp. was due to applying a single model to the combined partitions, rather than any algorithm specific to either PAUP or MrBayes.

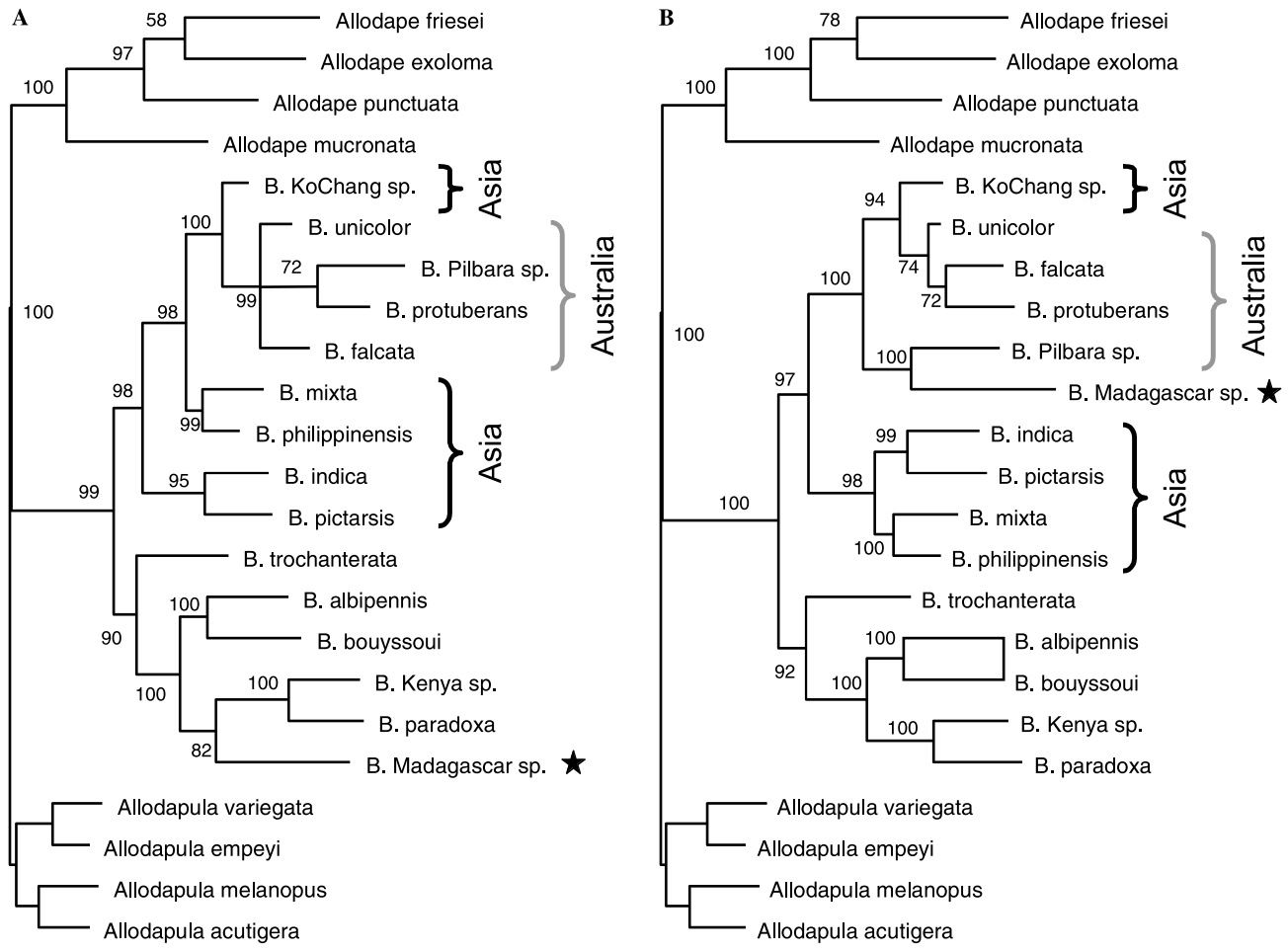

Fig. 3. (A) Consensus phylogram from a Bayesian analysis with four separately modelled partitions comprising 1st, 2nd, and 3rd codon mitochondrial positions and the combined EF-1α sites. Posterior probabilities and mean branch lengths are derived from 1000 trees taken from generations 3.5–4 million, sampling every 500th generation. Asian species are indicated by black brackets, Australian species by grey brackets, and the Madagascan species by a star. All other species are African. (B) Consensus phylogram from a Bayesian analysis with a single model applied to the combined data set. Posterior probabilities and mean branch lengths are derived from 1000 trees taken from generations 3.5–4 million, sampling every 500th generation.

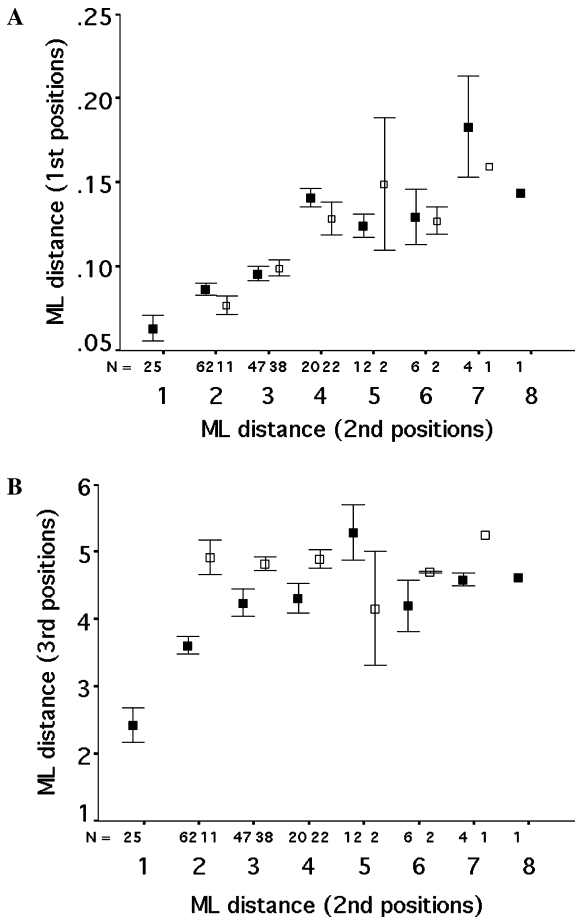

Fig. 4. (A) Plots of mean distances (±95% confidence limits) based on maximum likelihood GTR + I + G distances calculated from a model fitted to 1st mitochondrial positions, versus distances based on a separate ML model fitted to 2nd mt codon positions. Hollow squares indicate mean distances between ingroup and outgroup taxa, and filled squares represent mean distances between ingroup taxa. (B) Plots of mean distances (±95% confidence limits) based on maximum likelihood GTR + I + G distances calculated from a model fitted to 3rd mitochondrial positions, versus distances based on the same ML model fitted to 2nd mt codon positions as for (A). Hollow squares indicate mean distances between ingroup and outgroup taxa, and filled squares represent mean distances between ingroup taxa.

It is likely that the partitioned Bayesian analyses gave a more credible topology than combined analyses because they involved fitting a more reasonable model to the fast evolving 3rd mitochondrial positions (Table 2). Nevertheless it is possible that the 3rd position model still underestimated evolutionary rate. We examined plots of 1st and 3rd mt position ML distances (see Section 2) against 2nd mt position estimates for evidence of asymptotic behaviour. Because scattergrams contained so many pairwise data points it was difficult to discern patterns. We therefore combined 2nd position distance estimates into eight categories (0–0.01, 0.011–0.02, etc.) and then calculated mean (±95% C.I.) values for both 1st and 3rd position distance estimates. These means were calculated separately for distances between ingroup and outgroup taxa (where underestimation is more likely because of greater divergence times) and among ingroup and outgroup taxa alone (where underestimation is less likely to be a problem because divergences will be more recent). The resulting graphs (Figs. 4A and B) suggested an approximately linear relationship between 1st and 2nd distance estimates for both within and between ingroups and outgroups, but strong asymptotic behaviour for 3rd position distances. In fact distances between ingroup and outgroup species pairs showed a flat relationship, with a mean value that was similar in magnitude to the asymptotic value reached by distances among ingroup and outgroup species pairs. This result suggests that for greater divergences between taxa pairs, GTR models fitted to 3rd codon positions are likely to increasingly underestimate the actual amount of evolutionary change.

4. Discussion

4.1. Comparison of analytic approaches

MP, logDet, and ML analyses based on the combined data all suggested two topological features that raise biogeographical problems. First, they suggested that the undescribed Thai species, *B.* KoChang sp., fell within the Australian clade and, second, they suggested that the Malagasy *B.* Madagascar sp., was a sister species to the Western Australian *B.* Pilbara sp.. It is not difficult to posit the former topological feature as reasonable: the existence of myriad islands that could form 'stepping stones' between Australia and the Asian mainland allows the possibility of northwards migration from a member of a relatively derived Australian group to Thailand. However, the sister-group relationship between *B.* Pilbara sp. and *B.* Madagascar sp. is highly problematic. There are currently very few islands between Madagascar and Australia and these are separated by very extensive ocean expanses.

Reyes et al. (1999) recommended exclusion of 3rd codon position mitochondrial transitions for allodapines under MP analysis because of saturation problems. Interestingly, our MP analysis where this was followed gave very similar results to the combined data set analysis, suggesting that weighting these substitutions zero did not avoid homoplasy-related problems. This finding suggests that transversion substitutions also are homoplasious, although the transversion substitution rate is much lower than the transition rate (Table 3). Analyses where 3rd mt positions were excluded gave quite different topologies and instead placed *B.* Madagascar sp. within the African group, or else lead to it being placed in a basal polytomy under a 50% majority

rule criterion. Interestingly, these analyses also recovered the Thai species as being either a member of a monophyletic Asian clade or as the sister species to a monophyletic Australian clade, within a monophyletic Asian + Australian clade. The fact that log Det analyses for combined and restricted data sets gave similar patterns suggests that this switch in topology between analyses based on the combined and restricted data sets is not due to base composition bias at 3rd mt positions.

One major advantage of ML models over MP analyses is that the former are able to incorporate the effects of multiple substitutions, ameliorating the misleading effect of homoplasies expected at fast evolving sites. However, our ML models which included the fast evolving 3rd mt positions, where it was thought that Γ-shaped rate distributions might account for rate heterogeneity, still lead to unlikely topologies, where the Malagasy species was included within the Australian clade and the Australian clade included an Asian species. These a priori unlikely results were not found for the Bayesian analyses where the four different partitions were allowed to have quite different substitutional dynamics. Instead, the Malagasy species was firmly grouped with the African species, and the Australian taxa were recovered as a well supported monophyletic group. However, the Bayesian analyses where a single model was fitted to the combined data placed *B.* Madagascar sp. within the Australian clade, strongly suggesting that a lack of partitioning, rather than Bayesian analysis per se, lead to this unlikely result.

Although the partitioned Bayesian analyses did not lead to unlikely topologies, plots of ML distance estimates for the three mitochondrial positions suggested that for greater divergences, the third position model increasingly underestimated the extent of likely divergence. This suggests that for analyses covering deeper divergences, or when including clades with very high evolutionary rates, long branch attraction (LBA) will still be a problem. In our case, the extreme compositional bias for 3rd mt positions likely exacerbated this problem, and it is difficult to see how this could be overcome.

Our analyses therefore suggest the following important results, and these may help in the development of analytic strategies for future studies where partitions exhibit quite different evolutionary patterns. First, MP and ML analyses where 3rd mitochondrial positions were excluded resulted in more feasible topologies than where 3rd mt positions were included. Second, estimates of branch lengths based on the reduced data set (excluding 3rd mt sites) indicated likely long branches for *B.* Madagascar sp. and *B.* Pilbara sp., suggesting that the unfeasible topological features from the combined data set analyses were due to LBA arising from 3rd mt positions, where substitution rates are very high and with an extreme AT bias. Bayesian analyses, where substitutional dynamics were allowed to vary across partitions, produced feasible results, without the unlikely topological features of other analyses. Bayesian analyses for a combined data, single model approach gave similar results to the combined data ML analysis, suggesting that partitioning per se helped reduce likely LBA effects. However, plots of ML genetic distances based on the partitioned Bayesian analyses (Fig. 4) suggested that distances based on separately modelled 3rd mt positions still underestimated likely divergence levels. We did not explore how such underestimation may be overcome and this could form a productive question for future studies.

Given that MP and ML analyses where 3rd mt positions were excluded gave broadly similar topologies to our Bayesian analyses where they were included, the question arises as to whether analyses should merely exclude problematic data, rather than attempt more complex model fitting. In cases where sequences are very large it may be easier to simply exclude problematic partitions, but for situations like our study this would result in the loss of useful data, especially for recent divergences where rate underestimation is unlikely to be important and where the number of informative sites is low. As a first step, our results suggest that the likelihood of LBA effects due to saturation should be assessed with distance measures based on separate partition models and that simple but commonly employed tests, such as using plots of uncorrected 'p' versus TN distances, should be treated very cautiously. Where LBA is likely to arise, credibility of deeper nodes should be examined by excluding problematic partitions for MP and ML analyses. Use of separate models for different partitions may overcome such problems, though our results suggest the potential for problems where models may still underestimate true levels of divergence.

4.2. Historical biogeography and the evolution of Braunsapis

All analyses indicated that *Allodape* and *Braunsapis* were both monophyletic and bootstrap support for monophyly was extremely high. On the basis of morphological data, (Michener, 1971, 1975a,b, 1977) argued that *B. trochanterata* closely approaches *Allodape* in a variety of characters, especially larval traits, but believed that a number of synapomorphies suggested its inclusion in *Braunsapis*. All of our well supported results strongly suggest that *B. trochanterata* is the most basal member of the African *Braunsapis* clade, but some analyses placed the divergence between the African and the Asian + Australian below this species clade. Taken as a whole, our results do not refute this possibility, and molecular support for the position of *B. trochanterata* is ambivalent.

Allodape is most abundant and diverse in temperate regions of Africa but extends into tropical Africa, *Braunsapis* is predominantly tropical and subtropical, and *B. trochanterata* is restricted to tropical east Africa. This strongly suggests that the divergence between *Allodape* and *Braunsapis* occurred in tropical Africa. An early divergence between the African and Australian + Asian groups is strongly suggested by most analyses, excepting those problematic results suggesting an unlikely position for *B.* Madagascar sp. Monophyly of the Asian + Australian species was indicated by all analyses, suggesting that these species evolved from a single common ancestor species that dispersed from Africa to southern Asia, relatively soon after the origin of the genus, followed by subsequent radiation within southern Asia and then dispersal into Australia. Monophyly of the Australian *Braunsapis* was supported by the partitioned Bayesian analyses, and all other analyses where 3rd mt codon positions were excluded, suggesting a single dispersal event into Australia, at least for the species included in our study, which come from the far eastern, far western, and semi-arid southern-central regions of the continent. Interestingly, radiation within Australia has involved colonization of tropical, sub-tropical, arid, and semi-arid regions, but no species are recorded from the southern temperate region where the exoneurine genera *Exoneura* and *Brevineura* are most abundant and diverse. For the Australian *Braunsapis* clade, internal branch lengths for the ML restricted data set analysis and the Bayesian analyses are all very short, suggesting very rapid diversification following colonization of Australia. Similarly, rapid diversification following colonization of Australia has also been shown for the exoneurine allodapines (*Exoneura, Exoneurella, Brevineura* and *Inquilina*), which diverged from the African taxa well before the origin of *Braunsapis* (Bull et al., 2003).

The distribution of *Braunsapis* in Australia, covering tropical, subtropical, arid, and semi arid regions, but excluding cool temperate regions, suggests that the spread of this genus may have been limited more strongly by low maximum winter temperatures, rather than by low minimum temperatures or rainfall. This could be related to the genus-wide pattern of year-round brood rearing (Aenmey, 2002; Maeta et al., 1992; Michener, 1962, 1971; Pauly et al., 2001; Zammit and Schwarz in review) whereby larvae are present throughout the year. In contrast, members of the exoneurine group display marked seasonality in brood rearing, are largely restricted to temperate regions, and overlap with *Braunsapis* only in sub tropical and southern semi-arid regions. Tierney et al. (2000) found marked phylogenetic conservatism in key life history parameters for the exoneurine genera and it seems very likely that such conservatism is more widespread among allodapine genera. It is possible that one or more traits of *Braunsapis*, such as year-round brood rearing or intolerance of low winter-maximum temperatures, also exhibit phylogenetic inertia, limiting opportunities for radiation into temperate regions. Studies on the phylogenetic distribution and life history traits of *Braunsapis* species occurring at the southern marginal latitudes may throw light on this possibility.

Acknowledgments

We are indebted to Prof. S. Hanrahan for vital logistic support for fieldwork in Africa. We thank Kathy Saint for valuable help with DNA techniques and Tracy Aenmey, Meg Schwarz, Zeta Steen, and John Zammit for help with field work, often under arduous conditions. The work here was supported by Australian Research Council Grant A19942062 to M. Schwarz and R. Crewe.

References

Aenmey, T.K., 2002. Life cycle and social organisation of Southern African bees from the genus Braunsapis (Hymenoptera:Apidae). Masters thesis, Faculty of Science, Witwatersrand University, Johannesburg, South Africa.

Arbogast, B.S., Edwards, S.V., Wakely, J., Beerli, P., Slowinski, J.B., 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu. Rev. Ecol. Syst. 33, 707–740.

Barker, F.K., Lutzoni, F.M., 2002. The utility of the incongruence length difference test. Syst. Biol. 51, 625–637.

Bruno, W.J., Halpern, A.L., 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16, 564–566.

Bull, N.J., Schwarz, M.P., Cooper, S.J.B., 2003. Molecular phylogeny and evolution of the Australian allodapine bees. Mol. Phylogenet. Evol..

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K., 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52, 477–487.

Cunningham, C.W., 1997. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. 14, 733–740.

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., Shuqing, J., 2003. Phylogeny of eusocial Lasioglossum reveals multiple losses of eusociality within a primitively eusocial clade of bees. Syst. Biol. 52, 23–36.

Dowton, M., Austin, A.D., 2002. Increased congruence does not necessarily indicate increased phylogenetic accuracy—the behaviour of the incongruence length difference test in mixed-model analyses. Syst. Biol. 51, 19–31.

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Mol. Marine Biol. Biotechnol. 3, 294–299.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.

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

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.

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

Lockhart, P.J., Steel, M.A., Hendy, M.D., Penny, D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11, 605–612.

Lunt, D.H., Zhang, D.X., Szymura, J.M., Hewitt, G.M., 1996. The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies. Insect Mol. Biol. 5, 153–165.

Maeta, Y., Sakagami, S.F., Michener, C.D., 1992. Laboratory studies on the behavior and colony structure of *Braunsapis hewitii*, a xylocopine bee from Taiwan. Univ. Kansas Sci. Bull. 54, 298–333.

Michener, C.D., 1962. Biological observations on the primitively social bees of the genus *Allodapula* in the Australian region (Hymenoptera, Xylocopinae). Insectes Sociaux 9, 355–373.

Michener, C.D., 1971. Biologies of African allodapine bees (Hymenoptera, Xylocopinae). Bull. Am. Mus. Nat. Hist. 145, 219–302.

Michener, C.D, 1975a. Larvae of African allodapine bees. 2. *Braunsapis* and *Nasutapis*. J. Entomol. Soc. S. Afr. 38, 223–242.

Michener, C.D., 1975b. A taxonomic study of African allodapine bees. Bull. Am. Mus. Nat. Hist. 155, 67–240.

Michener, C.D., 1977. Discordant evolution and classification of allodapine bees. Syst. Zool. 26, 32–56.

Pauly, A., Brooks, R.W., Nilsson, L.A., Persenko, Y.A., Eardley, C.D., Terzo, M., Griswold, T., Schwarz, M., Patiny, S., Munzinger, J., Barbier, Y., 2001. Hymenoptera Apoidea de Madagascar. Ann. Zool. Wet. 128, 1–390.

Reyes, S.G., 1998. A cladistic analysis of the bee tribe Allodapini (Hymenoptera: Apidae: Xylocopinae). Philipp. Ent. 12, 55–84.

Reyes, S.G., Cooper, S.J.B., Schwarz, M.P., 1999. Species phylogeny of the bee genus *Exoneurella* Michener (Hymenoptera: Apidae: Allodapini): evidence from molecular and morphological data sets. Systematics 92, 20–29.

Schwarz, M.P., Bull, N.J., Hogendoorn, K., 1998. Evolution of sociality in the allodapine bees: a review of sex allocation, ecology and evolution. Insect Soc. 45, 349–368.

Schwarz, M.P., Bull, N.J., Cooper, S.J.B., 2003. Molecular phylogenetics of allodapine bees, with implications for the evolution of sociality and progressive rearing. Syst. Biol. 52, 1–14.

Simon, D., Larget, B., 2000. Bayesian analysis in molecular biology and evolution (BAMBE), version 2.03 beta. Department of Mathematics and Computer Science, Duquesne University.

Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99, 16138–16143.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.0b10. Sinauer, Sunderland, MA.

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

Tierney, S.M., Cronin, A.L., Loussert, N., Schwarz, M.P., 2000. The biology of *Brevineura froggatti* and phylogenetic conservatism in Australian allodapine bees. Insectes Sociaux 47, 96–97.

Whitfield, J.B., Cameron, S.A., 1998. Hierarchical analysis of variation in the mitochondrial 16s rRNA gene among Hymenoptera. Mol. Biol. Evol. 15, 1728–1743.

Wilcox, T.P., Zwickl, D.J., Heath, T.A., Hillis, D.M., 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol. Phyl. Evol. 25, 361–371.